2019 Fifth International Conference on Data Science and Engineering (ICDSE-2019)

Table of Contents

Real Time Arrhythmia Monitoring with Machine Learning Classification and IoT

Devadharshini M S[1], Heena Firdaus A S[2], Sree Ranjani R[3] and Devarajan N[4]
[1,2,4] *Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, India.*
[3]*Indian Institute of Technology Madras, Chennai, India.*
[1]*devadharshini2307@gmail.com* ,[2]*asheena789@gmail.com*, [3]*sreeranjanirajendran@gmail.com*

***Abstract*— Among the applications of Internet of Things (IoT) smart health care and management is a significant one. The wireless technologies and wearable sensors enable effective monitoring of the patients. This voluminous data from the wearable sensors can be processed, analyzed and classified using the machine learning algorithms. Many techniques are emerging for the biomedical data mining. The proposed method can be used for arrhythmia patient monitoring and classification. The arrhythmia patient can be continuously monitored using the wearable sensors and mobile application. This method benefits the patient as they have the freedom to be mobile and monitored in their usual environment. The data from electrocardiogram (ECG) sensors are categorized into various groups using machine learning algorithm. In this paper, a method for monitoring, visualizing the arrhythmia of patients and classification of the data for the hospitals in order to extend their patient care using the mobile application is proposed.**

Keywords— Arrhythmia, Machine Learning, Health Monitoring, IoT, Wearable Technology, Classification.

I. INTRODUCTION

Arrhythmia is the condition in which there is an irregular beating of heart and this is categorized under the pathologies of heart [1]. The condition of arrhythmia patients has to be continuously monitored as the information about the heartbeat is necessary to diagnose the type of arrhythmia and to follow appropriate medical procedures. Certain types of arrhythmia do not require medical attention but it has to be monitored as it supplements signs of prevailing heart diseases. Usually the patients are diagnosed using the ECG monitors or Holter monitors that is available in the clinics. This is where the problem arises, for arrhythmia diagnosing ECG of patients has to be continuously monitored for minimum of 24-48 hours. Moreover, in case of the developing countries the clinics and medical experts are insufficient. The patient care and diagnosis can be improved by extending the patient monitoring in casual environment of patient. This method enhances the quality of diagnosis as the condition of patient is monitored in normal house environment rather than a clinical setup. Physiological monitoring of patients has to be accurate, easy to use and available at nominal cost [2]. Ambulatory electrocardiography can be used for the remote patient monitoring as per the guidelines published by the American College of Cardiology (ACC) and the American Heart Association (AHA) [3]. Therefore, to monitor the patients in their usual environment the wearable sensors can be used. The interest over the wearable sensors are rising in recent years. In order to achieve long term recording of physiological information of patients and manage those data, the researchers have considered the application of wearable sensor technology [4-6].

A standardized Body Area Network (BAN) framework is being used for incorporation of the wireless sensors, which is aimed at improving the free and casual mobility of the patients in a daily situation while being monitored by a wireless wearable system [7]. This concept is advantageous to existing system as it does not need any technical skills to operate. The overall idea presented in this paper includes arrhythmia patient monitoring system, arrhythmia classification using machine learning and the data visualization. The existing frameworks either consists of the monitoring methods [16] or the classification algorithms as mentioned in [8]. Mostly a complete system for arrhythmia that can be equipped in real time in hospitals is not proposed. Thus, in this paper we propose a complete system for arrhythmia patient monitoring and categorization using IoT and Machine learning algorithm.

II. METHODOLOGY

The ultimate aim of our concept is to provide a system for arrhythmia patient monitoring and arrhythmia classification to improve the diagnosis. This system can be used in the hospitals for extending the health care to patients. In this method patients having irregular heartbeats are assigned with unique number with which the mobile application can be accessed. The application has different views for doctors and patients with restricted access. Our proposed concept monitors the patients continuously using the wearable sensors and transmits the data to the android/iOS application which can be viewed by both doctors and patients. The patient who is suspected to have arrhythmia is continuously monitored for the physiological parameters like heart rate and temperature using the wearable sensors. In case of detection of irregular heart beat the android/iOS application notifies the patient to wear the wireless ECG sensor. The application classifies the state of patients into *16* groups of arrhythmias as in [17] automatically using machine learning algorithms. The application enables two different views for patients and doctor. The patients can only view their corresponding medical records while doctors can view records of all patients. The mobile application is designed with enhanced security and privacy. This architecture consists of five major components: data sensing, data transmission, cloud storage, classification and visualization as shown in Fig.1.

Data sensing method involves monitoring patients which is performed using the wearable sensors for monitoring ECG, pulse rate and temperature. The location of patients is also tracked using GPS (Global Positioning System). in order to provide medical services in case of emergencies. All the sensors are typically aggregated to a microcontroller which reads the data from sensors. The data sensing consists of wearable sensors and GPS.

Data transmission component consists of a wi-fi module which acquires the data from microcontroller and transmits it to the mobile application via the cloud. This transmission must be done in a secured way in order to achieve data privacy.

Cloud Storage includes the major component of the system-storage. The system is designed in a way such that the

biomedical information of patients is stored for a long time. This long-time storage also assists the health professionals in diagnosing arrhythmia.

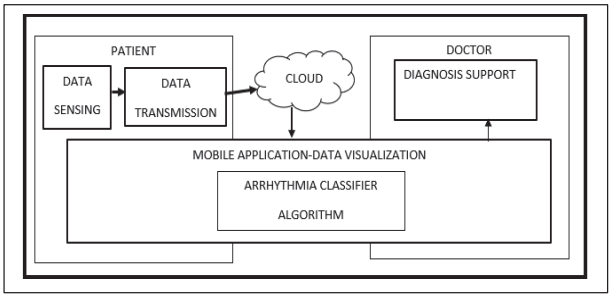


Fig.1. Overall architecture of arrhythmia patient monitoring and classification.

A. Data Sensing

The data sensing that is being performed using the wearable sensors with the design qualities that does not hinder the patient's mobility, light weight and easy to operate. The wearable devices are aggregated to form a Wireless Personal Area Network (WPAN). The wearable devices acquire the physiological data- heart rate and temperature. The data from ECG is acquired when the application intimates irregular heart beat detection. The location of patients is also tracked in order to provide services during the medical emergencies. The heart rate, temperature and ECG of the patients are acquired continuously from the BAN and the location is tracked using GPS.

The values from the sensors are acquired by the microcontroller. This physical phenomenon of sensing data in the distributed fashion is stated as energy efficient sensing mechanism [9]. Schemes like energy efficient sensing mechanism is implemented using IoT based sensing architecture.

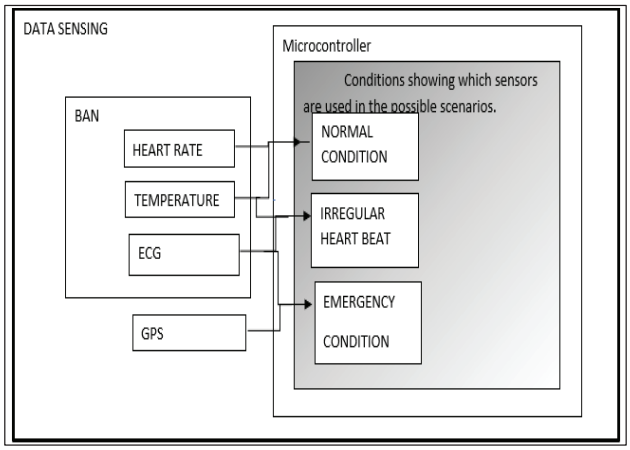


Fig.2. Data sensing using the architecture of IoT for implementing energy efficient sensing mechanism and to overcome the wearable issues.

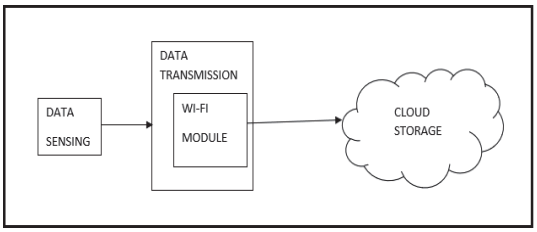


Fig.3. Data flow from sensors to cloud storage

The decision-making process is made and intimated using the mobile application. In real time, the patient is assumed to be in normal condition. In this state, only heart rate and temperature are monitored by the microcontroller that can be programmed to compare heart rate value periodically. If the value varies from the threshold range then the signal is sent to the application which intimates the patient to wear ECG monitor, and the microcontroller gets the data from ECG sensor. Moreover, this method overcomes major issues like reliability, privacy issues, user interface as mentioned in wearable sensors [10]. This architecture is shown in Fig.2.

B. Data Transmission and Cloud Storage

The components in this system are used for transmitting the data to the doctors and patients with the help of mobile application. The recordings of the physical parameters of the patient are transmitted via the cloud. The microcontroller transmits the data to cloud using the wi-fi module. The recordings of the patient are transmitted to the cloud for long term storage as shown in the Fig.3. This method improves scalability and extends the benefits like data accessibility on demand both from patients and doctors. The cloud storage has to maintain data security and privacy. While storing electronic medical data of an individual, much importance should be given to data privacy. This is implemented by restricting the data access by users. That is, the patients can access their corresponding records only. And only authorized doctor can access the records via the mobile application. Our concept incorporates secured cloud storage methods as discussed in [11-13].

C. Classification and Visualization

The data from the cloud is accessed by the mobile application which makes decisions on the sensors and classifies the arrhythmia as shown in Fig.4. The mobile application processes the data from the sensors and makes the decision regarding the usage of sensors and intimates the patients about it.

The classification of arrhythmia is done using the resampling method and random forest classifier with selected features [8]. The arrhythmia classification method involves training the classifier with the data set from clinical records and prior ECG sensor data. After the training process the classifier is provided with the current sensor data which classifies arrhythmia into *16* categories and has *96%* accuracy [8]. Visualization is necessary as it would be easy for both doctors and patients to access the voluminous data and analyses by the proposed system.

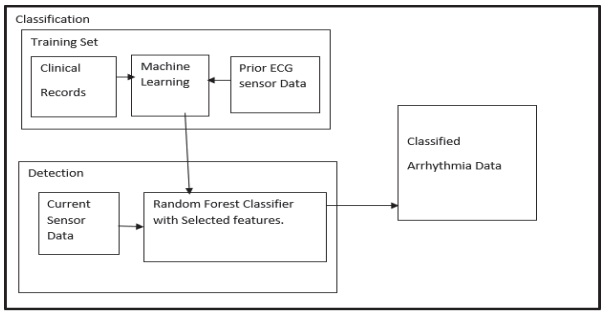


Fig.4. depicts the data flow in classification component.

The data visualization is considered as the independent and important research area [14]. The visualization of data is achieved using the interactive mobile application. The categorized data are presented using different colors. We are equipping this method as color distance and color categories enhances the identification and understanding of the differences in data [15].

III. Results and Discussion

This system is proposed for real time implementation in the hospitals specifically for arrhythmia patients. Fig.5. shows the ECG sensor reading in a normal patient. The system is mainly controlled over the mobile application as shown in Fig.6.

Fig.5. Example for reading from ECG sensor in a normal patient.

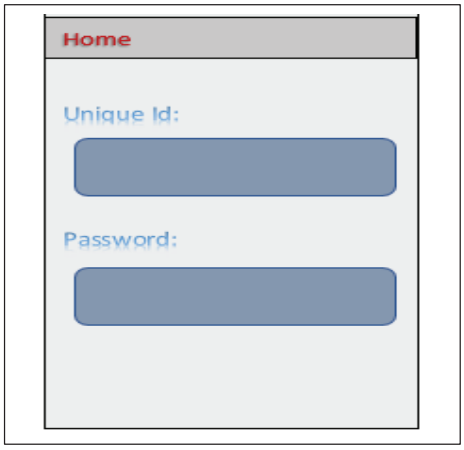


Fig.6. Home Screen of Mobile Application

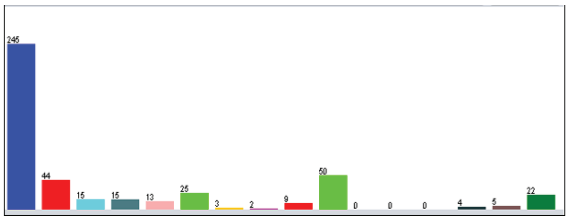


Fig.7.Graph representing different classes of arrhythmia as in the original data set.

This is how the home screen of mobile application will look like where the patient has to enter the unique number and password in order to use the services. The classifier is trained using the original data set that is available in the UCI repository. The two images are samples depicting the visuals of the proposed system.

The random ensemble classifier mentioned in [8] is incorporated in the mobile application as it is suitable for solving the particular problem. This is implemented using the open source software TensorFlow. The required libraries can be imported and develop the application which will be able to classify the arrhythmia. The sample data set is taken from the UCI repository [17] and the figure 7 shows the different classes of arrhythmia before resampling of the data. This database consists of 452 samples and 274 features [17].

The sample data set is resampled using the methods mentioned in [8]. Data resampling is necessary to achieve accuracy over small set of samples. The volume of patients varies but the accuracy of application has to be maintained so the feature selection is done. There are several algorithms for the arrythmia classification, but for the proposed system this algorithm is preferred to maintain the accuracy and reliability for varying volume of data. The heart rate of the patient is continuously monitored and we can analyze the type of arrhythmia with respect to different heart rates. Fig.8. shows arrhythmia classes in the graph plotted with respect to heart rate. The sample data is processed to get the required details. The different colors in the graph represent the arrhythmia classes as shown in Fig.9. The categories of arrhythmia as in [17] is shown in the Fig.10.

Table I. 16 Classes of Arrhythmia [17]

Class Code	Class Name
1	Normal
2	Ischemic changes (Coronary Artery Disease)
3	Old Anterior Myocardial Infarction
4	Old Inferior Myocardial Infarction
5	Sinus tachycardy
6	Sinus bradycardy
7	Ventricular Premature Contraction (PVC)
8	Supraventricular Premature Contraction
9	Left bundle branch block
10	Right bundle branch block
11	1. degree Atrio Ventricular block
12	2. degree AV block
13	3. degree AV block
14	Left Ventricule hypertrophy
15	Atrial Fibrillation or Flutter
16	Others

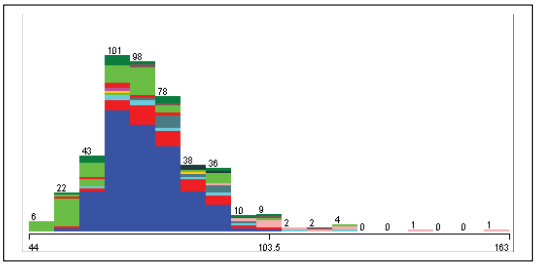


Fig.8. Arrhythmia classes plotted with respect to heart rate

Fig.9. 16 classes of arrhythmia color code used in graph.

IV. CONCLUSION

It is very important to detect and diagnose the arrhythmia appropriately in order to prevent the loss of human life. In this paper, a proper system for arrhythmia patient monitoring is proposed to enhance hospital services. The working of each module of the system is discussed. IoT and machine learning equipped system offers observations and recordings for a longer-period of time. The classified data and visualizations improve the diagnosis process. Thus, the proposed real time monitoring system for arrhythmia patients will reduce the risk of human life loss with the wearable sensors and machine learning technique. In future some other parameters like blood pressure can be monitored in order to enhance the early detection and accurate diagnosis of arrhythmia patients.

REFERENCES

[1] Krasteva, V.; Jekova, I. (2007): QRS Template Matching for Recognition of Ventricular Ectopic Beats. Ann Biomed Eng 35 (12), 2065-76.

[2] Silva, I.; Moody, G. B.; Celi, L. (2011): Improving the Quality of ECGs Collected Using Mobile Phones: The PhysioNet/Computing in Cardiology Challenge 2011. Computing in Cardiology 2011 38 (1),273-6.

[3] M. H. Crawford, "ACC/AHA Guidelines for ambulatory electrocardiography.," Journal of the American College of Cardiology, vol. 34, pp. 912-48, 1999.

[4] A. Pantelopoulos and N. Bourbakis, “A survey on wearable sensor-based systems for health monitoring and prognosis,” IEEE Trans. Sys., Man, and Cybernetics, Part C: Applic. and Reviews, vol. 40, no. 1, pp. 1–12, Jan 2010.

[5] D. Son, J. Lee, S. Qiao, R. Ghaffari, J. Kim, J. E. Lee, C. Song, S. J. Kim, D. J. Lee, S. W. Jun, S. Yang, M. Park, J. Shin, K. Do, M. Lee, K. Kang, C. S. Hwang, N. Lu, T. Hyeon, and D.-H. Kim, “Multifunctional wearable devices for diagnosis and therapy of movement disorders,”Nature Nanotechnology, pp. 1–8, 2014.

[6] A. Page, O. Kocabas, T. Soyata, M. Aktas, and J.-P. Couderc, “Cloud-Based Privacy-Preserving Remote ECG Monitoring and Surveillance,” Annals of Noninvasive Electrocardiology (ANEC), 2014. [Online]. Available: http://dx.doi.org/10.1111/anec.12204

[7] I. Sachpazidis, A. Stassinakis, D. Memos, S. Fragou, S. Nachamoulis, A. Vamvatsikos, A. Stavropoulou, M. Fonseca, R. Magalhães, B. Valente, A. D'Aquila, M. Fruscione, J. Ferreira, and C. Aguiar, "@HOME ein neues Eu-projekt zum Tele Home Care," Biomed Tech (Berl), vol. 47, pp. 970-2, 2002.

[8] Saumendra Kumar Mohapatra, Mihir Narayan Mohanty, "Analysis of Resampling Method for Arrhythmia Classification Using Random Forest Classifier with Selected Features", Data Science and Business Analytics (ICDSBA) 2018 2nd International Conference on, pp. 495-499, 2018.

[9] T. Torfs, V. Leonov, C. Van Hoof, and B. Gyselinckx, “Body-heat powered autonomous pulse oximeter,” in 5th IEEE Conf. on Sensors, Oct 2006, pp. 427–430.

[10] T. Martin, E. Jovanov, D. Raskovic, "Issues in wearable computing for medical monitoring applications: a case study of a wearable ecg monitoring device", Wearable Computers 2000. The Fourth International Symposium on, pp. 43-49, 2000.

[11] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, “Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption,” IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 1, pp. 131–143, Jan. 2013.

[12] S. Ruj, M. Stojmenovic, and A. Nayak, “Privacy preserving access control with authentication for securing data in clouds,” in Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (Ccgrid 2012), ser. CCGRID 12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 556–563.

[13] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, “Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption,” IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 1, pp. 131–143, 2013.

[14] E. R. Tufte and P. Graves-Morris, The visual display of quantitative information. Graphics press Cheshire, CT, 1983, vol. 2.

[15] C. Healey, “Choosing effective colours for data visualization,” in Proceedings of the 7th conference on Visualization. IEEE, 1996, pp. 263–270.

[16] Rune Fensli, Einar Gunnarson, Torstein Gundersen, "A wearable ECG-recording system for continuous arrhythmia.

[17] Gunvenir, H. A., Acar, B., Demiroz, G., & Cekin, A. (1997). Supervised machine learning algorithm for arrhythmia analysis. Computers in cardiology, 433-436.

A Nonlinear Multivariate Approach to Identify Differentially Coexpressed Pathways

Mrityunjay Sarkar
Dept. of ECE
Durgapur Institute of Advanced Technology and Management
Durgapur, India
mrityu1488@gmail.com

Aurpan Majumder
Dept. of ECE
National Institute of Technology, Durgapur
Durgapur, India
aurpan.nitd@gmail.com

Sanjay Dhar Roy
Dept. of ECE
National Institute of Technology, Durgapur
Durgapur, India
sanjay.dharroy@ece.nitdgp.ac.in

Abstract— **Research on Gene Regulatory Networks (GRN) is primarily guided by differential co-expression analysis. Through this approach, we can observe a biologically significant difference in gene regulatory control under varied metabolic conditions. Although quite successful, the traditional analysis gets restricted either from the point of view of a single gene-specific topology where only pairwise co-expression plays a major role or in the type of gene coexpression. In the latter context, state of the art examples mostly considers linear correlative regulatory patterns in a network. In the current work, we overcome these limitations using an established gene regulatory framework which accounts for the complete co-expression structure. An overall improvement on the regulatory functions could be observed incorporating nonlinear gene to gene regulation. In this regard, the nonlinear differential orchestrated action of genes in different conditions present in certain unreported gene regulatory pathways of p-53 dataset have been found to be significant, where linear measures of differential regulations reflected meager importance.**

Keywords- Differential Coexpression; Mutual Information; Polynomial Regression; Spline Regression; Weight; HUB Gene.

I. INTRODUCTION

Cutting edge research methodologies explore through the differential structure of a GRN across conditions for better understanding of its varied functionalities under dissimilar stages. Differential gene expression analysis happened to be the sprout in understanding the differential behavior of a gene in a GRN across conditions. For many years these differentially expressed (DE) genes were believed to be the key biomarkers in any form of the disease [1]. However, simple DE analysis could not proceed in this long run of research due to certain shortfalls. As stated in [2] small changes in expression values for single genes are not detectable by a simple two-sample test (*t-test*) with multiple testing corrections. Again, traditional DE analysis always considers each gene separately and ignores the dynamic gene to gene interaction pattern in a network [3]. Changes in the coding region and the posttranslational modifications (e.g., phosphorylation, acylation, methylation, etc.) can modify the protein activity without any change in the gene expression level. However, this activity can alter the gene to gene interaction pattern [4] craving path for differential connectivity amongst genes of a GRN placed under unalike states.

In this connection, the problem of connectivity issue can be addressed in two ways. First utilizing Topological Overlap (TO) based measure [5], where the TO concept is implemented to select those genes having a disjoint set of neighbors across conditions. In our previous works [6,7] we have shown that based upon the weight metric it can again be subdivided into two segments, weighted TO and unweighted TO. The second approach to address the connectivity pattern is via differential co-expression (DC) based measure. The interrelation between DC and connectivity is reflected by almost all the methods while computing DC, namely DCe, DCp, WGCNA, LRC, and ASC [8]. The main drawback of these approaches is consideration of only pairwise co-expression, and the complete co-expression structures between genes are ignored. Again, these techniques consider linear co-regulation amongst genes by calculating the differential co-expression via correlation. But there always remains a fair chance of nonlinear association amongst genes present in a GRN [9,10]

DiffCoEx [11], and GSNCA [12] are two stalwart techniques which address our first problem by considering the net co-expression change of the networks. However, in DiffCoEx, genes are first clustered based on the change in coexpression, and the algorithm gets restricted to focus on individual gene clusters only**.** Accordingly, it is unable to interpret the contributory role of every gene related to changes in connectivity patterns present throughout the network. Although the GSNCA approach overcomes this limitation by unveiling the inherent DC structure utilizing the complete co-expression network but gets narrowed down with linear regulatory concepts only.

Our current work is dedicated to solving this problem, assuming a complete gene to gene differential co-expression network as framed in GSNCA [12], but this time incorporating some non-linearity amongst genes using

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

mutual information (mi), spline regression (spline_reg) and polynomial regression (poly_reg) as possible measures of co-expression.

We have demonstrated our algorithm and the various aspects of it using the NCI 60 cell lines (p53) dataset [13,14]. The varying gene importance in terms of weighted network connectivity across different conditions is illustrated via the union of a first and second minimal spanning tree [12] (henceforth will be represented as MST2) based graphical visualization**.** As described in [12] genes at the center of MST2 plots, also called hub genes, happen to be the highest weighted genes playing a pivotal role in regulating various significant biological pathways. Our analyzed results from nonlinear regulation context show that for almost all such pathways there is a significant change in the weight of the hub gene across conditions leading to the formation of separate hub genes having varied pathway regulatory functionalities in different states. This is followed by asserting the nonlinear GSNCA significance of all concerned regulatory pathways.

II. METHODOLOGY

The simulation set up considers two biological conditions having a different number of samples. As given in [12], here we assume *n1* and *n2* samples of p genes present across two different conditions. In this context, we initiate our problem by developing an adjacency matrix M of size $p \times p$ with matrix elements m_{ij}= the non-linear dependency between gene *i* and *j*. Thus the implementation of three non-linear measures gives us three adjacency matrices, namely *M_MI*, due to mutual information, *M_PR*, due to polynomial regression, and *M_SR*, due to spline regression. Unless otherwise mentioned, we maintain these notations throughout the paper. Following a similar line of thought as given in [12], we can say that for any fully connected co-expression network (across any condition) there will be *p* nodes and p(p-1)/2 edges, where the weight of an edge between node *i* and *j* will be equal to $1-|m_{ij}|$. The importance of a gene in a GRN can be inferred by assessing the impact of that gene on all other genes present in the network. This assessment is done by computing all possible cross-correlative measures (m_{ij}) between the concerned gene and all other genes present in the GRN [12]. Next, a gene-specific weight factor w_i is considered, which is proportional to i^{th} gene's cross-correlation values. Finally, a weight vector *w* is estimated for all genes, as shown below:

$$w_i = \sum_{j \neq i} w_j m_{ij} \quad \text{where, } 1 \leq i \leq p \tag{1}$$

Matrix form Eq. (1) is

$$(M - I)w = w \tag{2}$$

Accordingly, the GSGNCA score of a GRN is calculated as

$$w_{GSGNCA} = \sum_{i=1}^{p} \left| w_i^{(1)} - w_i^{(2)} \right| \tag{3}$$

Here $w_i^{(1)}$ and $w_i^{(2)}$ indicate the weight of the i^{th} gene under conditions 1 and 2, respectively.

In this work, we propose an extension in the type of regulation involved in the GSNCA architecture. The existing work has focused on linear regulation between network genes, whereas we have extended further considering the possibility of nonlinear differential regulations. In short, the initial gene co-expression measurement by correlation is replaced by various nonlinear approaches, like mutual information (mi), polynomial regression (poly_reg) and spline regression (spline_reg). The reason behind taking both polynomial and spline regression measures is to check how the network structure, along with hub gene changes if we incorporate more smoothness in terms of regulation. The entire algorithm is given in Fig. 1.

Fig. 1: Gene Specific Generalized Network Co-Expression Analysis (GSGNCA)

1. Calculation of Gene-Coexpression.
   - Set the mode of operation among three non-linear measures.
   - Compute the condition-specific intergene dependencies accordingly.
2. While (each and every gene pair is considered *i*=1,2,.....,p and *j*=1,2,.....p)
   begin
   begin
   - Evaluation of weight vector (for each pathway) using equation 1.

   end
   - Calculation of GSGNCA measure for each pathway using equation 3.

   end
3. Visualization of MST2 plots by the combination of two non-linear gene dependency matrices.
4. Significance calculation of each pathway by permutation test.

In this context, the initial step is, to begin with, some selected p53 signaling pathways. The selected pathways possess more than 10 and less than 500 genes which are fed into our algorithm. This restriction in selection is being maintained to reduce the time complexity of the analysis. The stepwise descriptions of the algorithm are given below:

- Step 1: Here, we calculate the gene to gene dependency in terms of cross-mi / spline-reg/ poly-reg in two different conditions. Following the user-defined functions of adjacency given in [15] which include *NonLinMI*, *NonLinSP,* and *NonLinPR* we compute the symmetric uncertainty based mutual information, spline regression and polynomial regression measures respectively amongst the genes for a specific p53 pathway. The user-defined functions given above invoke three R package functions, namely *mutualInfoAdjacency*, *sm.spline* (with an order of 3, for cubic spline) and *adjacency.polyReg* [16,17].
- Step 2: Here, we compute the weight vectors corresponding to the genes of a specific pathway placed under two different conditions following

equations 1 and 2. Based on these weight vectors, we can determine the W_{GSGNCA} score of a particular pathway via equation 3.

- Step 3: At this stage, we determine the MST2 plots of a particular pathway in each condition for all the three types of adjacency measures. As given in [12], from this visualization, we can understand the importance of the associated weights on a gene under two different conditions. In this connection, genes having an appreciable difference in weights are considered to be significant.
- Step 4: At this final step, we judge the statistical significance of the W_{GSGNCA} score obtained for every pathway using a random permutation of expression levels of the participating genes.

III. RESULTS AND DISCUSSION

We have tested our algorithm on a publicly available NCI 60 cell line (p53) dataset. p53 acts as a major tumor suppressor protein. This dataset contains 50 samples where 17 cell lines carry wild type (WT) TP53, and the remaining 33 cell lines carry mutated (MT) TP53 [13,14]. To have a complete analysis, we feed all the p53 gene names in DAVID [18,19] and obtain a set of KEGG [20] pathways with significant p-values ($<10^{-2}$) as well as a high number of participating genes (>30). This set is intersected with the parent set of selected pathways given in supplementary of [12]. In this operation, 18 pathways match which comprises SET I. There remain 57 unmatched (not reported earlier) pathways which are significant according to DAVID comprising SET II.

These gene pathways are fed to the algorithm to analyze GSGNCA. In this context, function *NonLinMI* is used to compute nonlinear differential co-expression using mutual information-based adjacency measure. Parallelly, *NonLinSP,* and *NonLinPR* are used to determine the nonlinear differential adjacency measures using spline regression and polynomial regression, respectively. As a comparative measure, we also explore the same using simple linear coexpression as conducted earlier [12]. The p-value significant GSGNCA scores using the three nonlinear and one linear method for certain pathways possessing appreciable difference in weight connectivity of the hub genes are listed in Tables1 and 2.

The crux portion of our analysis lies in obtaining the MST2 plots for each measure of adjacency corresponding to a network of genes present in a pathway under the two different conditions. The plots under the two different conditions help us to understand the differential pattern of connectivity amongst the intra network genes. We specially focus on the hub genes as they possess maximum weight in a network and also act as a potential pathway regulator. Let us assume that for a pathway, the hub genes are HUB1 and HUB2 in conditions 1 and 2, respectively. Therefore, to have a compact differential orchestrated action of the genes present in the pathway, HUB1 and HUB2 must be different. This is followed by the fact that the hub genes should possess different weight components in the two states of the network.

Table1: Significance analysis of GSGNCA scores of certain relevant pathways from SET I

Name of the Pathway	p-value			
	Mutual Information	Polynomial Regression	Spline Regression	Correlation
PEROXISOME	0.687	0.376	0.356	0.009
DRUG_METABOLISM_OTHER_ENZYMES	0.454	0.150	0.762	0.039
BLADDER_CANCER	0.122	0.338	0.483	0.033
MELANOMA	0	0.590	0.493	0.087
CHRONIC_MYELOID_LEUKEMIA	0.031	0.641	0.521	0.902

Table 2: Significance analysis of GSGNCA scores of certain relevant pathways from SET II

Name of the Pathway	p-value			
	Mutual Information	Polynomial Regression	Spline Regression	Correlation
LYSOSOME	0.062	0.036	0.214	0.450
LONG_TERM_POTENTIATION	0.609	0.228	0.048	0.112
B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.269	0.197	0	0.479
T_CELL_RECEPTOR_SIGNALING_PATHWAY	0.030	0.081	0.724	0.629
ENDOMETRIAL_CANCER	0.070	0.167	0.957	0.068
ALZHEIMERS_DISEASE	0.805	0.631	0.030	0.981

Table3: Weight difference and functionalities of HUB genes present in GSGNCA significant pathways

SET I				
Pathway (**Method**)	Hub Gene in wild type (WT) TP53		Hub Gene in mutated (MT) TP53	
	Name (**Weighted values across WT-TP53 and MT-TP53**)	Functionality	Name (**Weighted values across WT-TP53 and MT-TP53**)	Functionality
PEROXISOME (**Cor**)	PEX10 (**1.401-0.752**)	Mutation of PEX gene is responsible for Peroxisome biogenesis disorder (PBD) [21]	ACOX1 (**0.912-1.397**)	Responsible for generation of peroxisomalstraight-chain acyl-CoA oxidase enzyme [Genetic Home Reference, url: https://ghr.nlm.nih.gov/gene/ACOX1]
DRUG_METABOLISM_OTHER_ENZYMES (**Cor**)	CDA (**1.285-1.053**)	Role of this gene in multiple drug-Metabolism Enzymes are given in [22]	XDH (**1.068-1.242**)	Participation via controlling of mammalian Target of Rapamycin (mTOR) signaling [23]
BLADDER_CANCER (**Cor**)	EGFR (**1.375-1.281**)	Overexpression of EFGR plays a key role in Bladder Cancer [24]	CCND1 (**1.043-1.358**)	Acts as a potential target in a bid for the proliferation of Bladder Cancer [25]
MELANOMA (**Cor**)	PIK3CB (**1.393-1.111**)	Act as a suppressor of Melanoma [26,27]	FGFR1 (**0.945-1.336**)	Severely co-expressed along with FGF1 in melanoma [28]
CHRONIC_MEYOLID_LEUKEMIA (**MI**)	ABL1 (**1.293-1.139**)	All members of ABL family along with BCR-ABL participate in CML [29]	MDM2 (**0.871-1.179**)	Overexpression of MDM2 eliminates the tumor suppressor function of p53 genes, thus having an adverse effect [30]
SET II				
Pathway (**Method**)	Hub Gene in wild type TP53		Hub Gene in mutated TP53	
	Name (**Weighted values across WT-TP53 and MT-TP53**)	Functionality	Name (**Weighted values across WT-TP53 and MT-TP53**)	Functionality
LYSOSOME (**MI**)	ACP5 (**1.43-0.996**)	It is one of the lysosomal acid phosphateses. Deficiency of which results in relatively mild phenotypes [31]	LAPTM5 (**1.061-1.298**)	It is a Lysosomal-Associated gene usually associated with the spontaneous regression of neuroblastomas [32]
LYSOSOME (**Poly_Reg**)	CTNS (**1.642-1.315**)	Control lysosomal activity by the generation of cystinosin protein [Genetic Home Reference, URL: https://ghr.nlm.nih.gov/gene/CTNS].	LAPTM5 (**1.164-1.709**)	Same as above
LONG_TERM_POTENTIATION (**Spline_Reg**)	CAMK2A (**1.423-0.917**)	CaMKII family participates in LTP with the spatiotemporal dynamics of CaMKII activation in individual dendrite spines in LTP [33]	CAMK2G (**1.077-1.399**)	Same as CAMK2A
B_CELL_RECEPTOR_SIGNALING_PATHWAY (**Spline_Reg**)	CD22 (**1.377-0.644**)	It is exclusively expressed on B cells which regulates adhesion and B cell receptor (BCR) signaling as an inhibitory co-receptor of the BCR [34]	RASGRP3 (**1.028-1.487**)	It acts as mediators for a multitude of receptor-coupled mechanisms [35]
T_CELL_RECEPTOR_SIGNALING_PATHWAY (**Poly_Reg**)	CD3D (**1.611-1.688**)	Activate the T-Cell Response [36]	ZAP70 (**1.573-1.695**)	Activate T-Cell Response by antigen receptor [37]
ENDOMATERIAL_CANCER (**MI**)	PTEN (**1.294-1.034**)	Acts as a tumor suppressor which negatively regulates the PI3K-AKT signaling pathway postulated in the pathogenesis of endometrial carcinoma [38]	EGF (**0.901-1.315**)	EGF and its receptor (EGFR) play a crucial role in endometrial cancer cells by constituting a principal growth-promoting pathway [39]
ALZHEIMERS_DISEASE (**Spline_Reg**)	ADAM17 (**1.325-0.901**)	Plays a pivotal role in numerous human diseases like Alzheimer because of its diversified role in many cellular mechanisms [40]	COX5B (**1.006-1.5**)	Reduction of expression level triggers different neurodegenerative diseases like Alzheimer's [41]

In this connection, we especially check the MST2 plots of a pathway generated by the measure having the least p-value significance. The sole intention is to confirm that the hub genes show a significant differential pattern. A brief discussion on the biological importance of the hub genes on the chosen (linear/ nonlinear measure specific) significant KEGG pathways is given in Table 3.

In this regard, the comparisons highlighted through the two-tier assessment of the p53 pathways (significance value as well as HUB gene functioning) ascertains the reliability of linear correlative measure on the first set (SET I) of filtered pathways [12]. On the other hand, the potentiality of the second set (SET II) of filtered pathways, basically tumorigenesis and metabolism-related are unveiled mostly by different nonlinear regulatory measures.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed an extension to a previously reported multivariate statistical test which works on the complete DC structure of a GRN to identify the significant changes between two distinct biological conditions. The interesting part of this algorithm compared to others is that almost all other approaches deal with the pairwise measure of co-expression only, making them unable to detect in parallel all changes in the co-expression structure. The basic GSNCA algorithm [12] overcomes this limitation by considering all possible cross-correlations of all genes in a network. However, it stands limited to linear correlative measures only. In this article, we have tried to resolve this by considering some nonlinear approaches such as mutual information (mi), polynomial regression (poly_reg) and spline regression (spline_reg) as potential measures of co-expression.

We have tested our advanced prototype (GSGNCA) along with the conventional linear GSNCA algorithm on NCI 60 cell line (p53) dataset, which acts as a major tumor suppressor protein. Results obtained by applying various measures have been tabularized with supportive arguments.

To justify the pros and cons of the linear measure over nonlinear ones, we have focused equally on the pathway significance as well as the importance of hub genes across conditions on some selected significant pathways. The intra-controlling capability of a gene (especially the hub gene) in a GRN is guided by the connectivity factor and the weight of a gene is proportional to its cross-correlation with all other genes. Thus the hub gene possesses high weight indicating strong connectivity with all others in one condition and low weight in the other condition implying weak regulatory action. To validate our argument, we have listed both the weight differences of the HUB genes across conditions in Table 3. Accordingly, in two different conditions, we obtain in general two different hub genes with maximum connectivity. Hence, the altered HUB genes make the KEGG pathways act differentially across conditions. To assess the differential attribute of the hub genes, we have not only considered the biological evidence from the literature corresponding to the concerned pathways but the weighted difference of hubs across conditions as well.

In this line of research, up-gradation can be conducted on the differential network analysis in genomics (DINGO) algorithm which first divides the conditional dependencies into global and group-specific components and then jointly estimates the group-specific conditional dependencies [42].

REFERENCES

[1] N. Puthiyedth, C. Riveros, R. Berretta, and P. Moscato, "Identification of Differentially Expressed Genes through Integrated Study of Alzheimer's Disease Affected Brain Regions" ,*PLOS one*, 11(4)., 2016

[2] VK.Mootha et al., "PGC-1alpha-responsive genes involved in oxidative Phosphorylation are coordinately downregulated in human diabetes". *Nat. Genet*., 34, pp. 267–273, 2003

[3] F. Emmert-Streib , GV. Glazko, "Pathway analysis of expression data: deciphering functional building blocks of complex diseases".,*PLoSComput. Biol*., 7, e1002053, 2011.

[4] Ad. Fuente, "From 'differential expression' to 'differential networking' identification of dysfunctional regulatory networks in diseases", *Trends Genet* 26, pp. 326–333, 2010.

[5] M. Ray, W.X. Zhang, "Analysis of Alzheimer,s disease severity across brain regions by topological analysis of gene co-expression networks", *BMC Systems Biology*, vol. 4, October 2010.

[6] A. Majumder and M. Sarkar : "Exploring Different Stages of Alzheimer's Disease through Topological Analysis of Differentially Expressed Genetic Networks" : International Journal of Computer Theory and Engineering, Vol.6, No. 5, pp. 386-391, October 2014

[7] M. Sarkar, A. Majumder, "TOP: An Algorithm in Search of Biologically Enriched Differentially Connective Gene Networks". In: Proceedings of 5th Annual International Conference on Advances in *Biotechnology* (BIOTECH 2015), pp. 124-133. GSTF, Singapore, 2015.

[8] J. Yang, H. Yu, B-H. Liu,*et. al.* "DCGLv2.0: An R package for unveiling differential regulation from differential co expression" ,*PLoS ONE*, 2013, 8(11): e79729

[9] A. Majumder, M. Sarkar , "Paired Transcriptional Regulatory System for Differentially Expressed Genes", *Lecture Notes on Information Theory*, Vol.2, No. 3, pp. 266-272, 2014.

[10] X. Guo X, Ye Zhang, W. Hu, H. Tan, X. Wang, "Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation", *PLOS one*, Vol. 9, Issue. 2, 2014.

[11] B.M. Tesson, R. Breitling, R.T. Jansen, "DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules", *BMC Bioinformatics*, 11:497, 2010

[12] Y. Rahmtallah, F. Emmert-Streib, G. Glazko, "Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets", *Bioinformatics*, Vol. 30, No. 3, pp. 360-368, 2014.

[13] M. Olivier et al., "The IARC TP53 database: new online mutation analysis and recommendations to users" ,*Hum. Mutat*., 19, pp. 607-614, 2002

[14] A. Subramanian et al.. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles" ,*Proc. Natl. Acad. Sci*. USA, 102, PP. 15545–15550, 2005

[15] M. Sarkar M, A. Majumder, "Quantitative Trait Specific Differential Expression (qtDE)", *Procedia Computer Science***46** :706-718, 2015

[16] P. Langfelder, S. Horvath, "WGCNA: an R package for weighted correlation network analysis" ,*BMC Bioinformatics* 9:559, 2008

[17] J.O. Ramsay, N.E. Heckman, B.W. Silverman, "Penalized Regression with model-based penalties", *Behavior Research Methods, Instruments, & Computers*, Springer, Vol. 29, Issue 1, pp. 99-106, 1997.

[18] D.W. Huang, B.T. Sherman, R.A. Lempicki , "Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources", *Nature Protoc.***4(1)** :44-57, 2009

[19] D.W. Huang, B.T. Sherman, R.A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists", *Nucleic Acids Res.***37(1)**: 1-13, 2009.

[20] M. Kanehisa M, S. Goto, KEGG: "Kyoto encyclopaedia of genes and genomics". *Nucleic Acids Research.***28**: 27-30, 2000

[21] D.S. Warren, B.D. Wolfe ,S.J. Gould: "Phenotype-genotype relationships in PEX10-deficient Peroxiso me biogenesis disorder patients", *Human Mutat.***15(6)** :509-521, 2000

[22] S.N.Iyer, A.V. Tilak , M.S. Mukherjee , R.S. Singhal , "Genotype Frequencies of Drug- Metabolizing Enzymes Responsible for Purine andPyrimidine Antagonists in a Healthy Asian-Indian Population", *Biochemical Genetics*, Vol. 50, Issue. 9, pp. 684-693, 2012

[23] J.M. Rosenbluth, D.J. Mays, M.F. Pino, L.J. Tang , J.A. Pietenpol , "A Gene Signature-Based Approach Identifies mTOR as a Regulator of p73", *Molecular and Cellular Biology*, Vol. 28, No. 19, pp. 5951-5964, 2008

[24] L. Chang et. al., "Restoration of LRIG1 suppresses bladder cancer cell growth by directly targeting EGFR activity", *Journal of Experimental and Clinical Cancer Research*, 2013

[25] Z. Liang et. al., "MicroRNA-576-3p Inhibits Proliferation in Bladder Cancer Cells by Targeting Cyclin D1", *Molecules and Cells***38(2)**: 130-137, 2015

[26] H. Bonnevaux et. al., "Concomitant Inhibition of PI3Kß and BRAF or MEK in PTEN- Deficient / BRAF-Mutant Melanoma Treatment: Preclinical Assessment of SAR260301 Oral PI3Kß-Selective In hibitor", *Molecular Cancer Therapeutics*, Vol. 15, Issue 7, 2016

[27] Y. Nakanishi et. al., "Activating Mutations in PIK3CB Confer Resistance to PI3K Inhibition and Define a Novel Oncogenic Role for p110ß", *Cancer Research*, Vol. 76, Issue. 5, 2016

[28] L. Xerri . et. al., "Expression of FGF1 and FGFR1 in human melanoma tissues", *Melanoma Res*. **6(3)** : 223-230, 1996

[29] J.Y.J Wang, "The Capable ABL: What Is Its Biological Function?" ,*Molecular and Cellular Biology*, Vol. 34, No. 7, pp. 1188-1197, 2014

[30] M. Zhou, A.M. A.M. Yeager, S.D. Smith, H.W. Findley, "Overexpression of the MDM2 gene by childhood acute lymphoblastic leukemia cells expressing the wild-type p53 gene", *Blood*, Vol. 86, Issue. 5, pp. 1608-1614, 1995

[31] A. Suter et. al., "Overlapping functions of lysosomal acid phosphatase (LAP) and tartrateresistant acid phosphatase (Acp5) revealed by doubly deficient mice", *Development*, 128, pp. 4899-4910, 2001.

[32] J. Inoue et. al., "Lysosomal-Associated Protein Multispanning Transmembrane 5 Gene (LAPTM5) Is Associated with Spontaneous Regression of Neuroblastomas" , *PLoS One***4(9)**, 2009.

[33] J. Lisman , "The CaM kinase II hypothesis for the storage of synaptic memory (Review)", *Trends in Neurosciences*, Vol. 17, Issue 10, pp. 406-412, 1994

[34] N. Seiger et. al., "CD22 ligation inhibits downstream B cell receptor signaling an d Ca(2+) flux upon activation", *Arthritis Rheum.***65(3)**:770-779, 2013

[35] Z. Nagy et. al, "Function of RasGRP3 in the formation and progression of human breast cancer", *Molecular Cancer*, **13:96**, 2014.

[36] J. Liu J, S. Wu, M. Li, X. Wang, Y. Tang, "LncRNA expression profiles reveal the co-expression network in human colorectal carcinoma", *Int. J. Clin. Exp. Pathol.***9(12)** : 1885-1892, 2016.

[37] H. Wang et. al., "ZAP-70: An Essential Kinase in T-cell Signaling", *Cold Spring HarbPerspectBiol***2(5)**, 2010.

[38] B. Djordjevic et.al., "Clinical assessment of PTEN loss in endometrial carcinoma: immunehistochemistry outperforms gene sequencing", *Modern Pathology***25** : 699-708, 2012

[39] L. Albiter et.al., "EGFR isoforms and gene regulation in human endometrial cancer cells", *Molecular Cancer***9:166**, 2010.

[40] M. Gooz, "ADAM-17: The Enzyme That Does It All", *Crit. Rev. Biochem. Mol. Biol.***45(2)** :146-149, 2010.

[41] N. Safavizadeh, S.A. Rahmani, M. Zaefizadeh, "Investigation of cytocrom c oxidase gene subunits expression on the Multiple sclerosis" ,*Indian J. Hum. Genet*.,**19(1) :** 18-25, 2013

[42] M.J. Ha, V. Baladandayuthapani, K-A Do. , "DINGO: Differential Network Analysis in Genomics", *Bioinformatics* 31.21, pp.3413-3420, 2015

An Empirical Study to Detect the Collision Rate in Similarity Hashing Algorithm Using MD5

Tushaar Gangavarapu
Worldwide Deals, Automated Advertising
Amazon.com, Inc.
Bangalore, India
tushaargvsg45@gmail.com

Jaidhar C.D.
Department of Information Technology
National Institute of Technology Karnataka
Mangalore, India
jaidharcd@nitk.edu.in

***Abstract*—Similarity Hashing (SimHash) is a widely used locality-sensitive hashing algorithm employed in the detection of similarity, in large-scale data processing, including plagiarism detection and near-duplicate web document detection. Collision resistance is a crucial property of cryptographic hash algorithms that are used to verify the message integrity in internet security applications. A hash function is said to be collision-resistant if it is hard to find two different inputs that hash to the same output. In this paper, we present an empirical study to facilitate the detection of collision rate when SimHash is employed to check the integrity of the message. The analysis was performed using bit sequences with length varying from 2 to 32 and Message Digest 5 (MD5) as the internal hash function. Furthermore, to enable faster collision detection with more significant speedup and efficient space utilization, we parallelized the process using a distributed data-parallel approach with synchronous computation and optimum load balancing. Collision detection is desirable, owing to its applicability in digital signature systems, proof-of-work systems, and distributed content systems. Our empirical study revealed a collision rate of 0% to 0.048% in SimHash (with MD5) with the variation in the length of the bit sequence.**

***Index Terms*—Collision Rate, Collision Search, Integrity, MD5, SimHash**

I. Introduction

In today's world of open communication and computing, providing a way to check the integrity of the stored messages or transmitted messages through an unreliable medium is of vital importance [1]. The integrity of the message guarantees that the message is not tampered with, in the transit and is usually achieved by utilizing hash functions. It is quite evident from the pigeonhole principle that every hash function with fewer outputs than inputs would result in some of the inputs hashing to the same output, i.e., the collision of hashes is plausible with most hashing schemes [2], [3]. Collision resistance is a vital property of a cryptographic hash function, which ensures the *difficulty* of finding two distinct inputs that hash to the same output value. While collision resistance is desirable, it does not imply the non-existence of collisions.

Cryptographic hash functions are customarily designed to ensure collision resistance. The *birthday paradox* gives a definitive upper bound on the collision resistance, i.e., if an attacker computes $\sqrt{2^N}$ hash operations (for a hash digest of $N-$bit size) on random input, then it is likely that matching outputs exist [4]. Most hash functions including Message Digest 5 (MD5) [5] and Secure Hash Algorithm 1 (SHA-1) [6] that were estimated to be collision-resistant, were later broken [7], [8], [9]. The impact of collisions is essentially application-dependent, and determining the collision rate can help estimate the collision resistance of a hash function. The use of hash functions in the security of digital signature schemes, proof-of-work systems, distributed content systems, data integrity schemes, e-cash, group signature, and a multitude of other cryptographic protocols makes it almost mandatory to determine the collision rate of the underlying hash functions.

In 1994, MD4 [10] was broken by attacking the last two rounds [4], [11], [12]. MD5 was broken in 1998, using the modular differential attack, and the collisions can be generated in about 15 minutes to an hour [13], [8], which is estimated by exploiting the weakness in the internal structure of MD5. Other hash functions including MD4, RIPEMD [14], and HAVAL-128, [15] can also be broken using a modular differential attack [8], [7]. SHA-1 is not broken yet, but a collision was found with the complexity of less than 2^{69} hash operations [9], [4]. It can be noted that the existing literature does not provide any collision information or collision search strategies for collision detection in SimHash [16], which is a locality-sensitive hashing scheme.

For currently unbroken cryptographic hashing schemes, there do not exist any known internal structural weaknesses, thus implying that the collision rate detection is the only guaranteed way of proving their collision resistance. SimHash, developed by Moses Charikar, is widely used in detecting similarity in large-scale data processing applications. When SimHash is used to check the message integrity, the need for the detection of its collision rate becomes vital. In this paper, we present an efficient empirical analysis of the collision rates in SimHash algorithm through a distributed data-parallel dictionary-updation approach, with optimal load balancing and synchronous computation. This study employs MD5 as the internal hashing scheme, and the analysis is presented for bit sequences ranging from 2 to 32 bits in length. Furthermore, the execution time taken to measure the collision rate is also detailed, to give an overall estimate of the time taken to identify collisions in SimHash.

978-1-7281-2087-4/19/$31.00

The rest of the paper is structured as follows: Section II presents a brief overview of the SimHash algorithm. Section III reviews the existing literature and work previously carried out in this domain. The proposed methodology to compute the collision rates for SimHash is presented in great detail, in Section IV. Section V presents the obtained experimental results, followed by conclusions and discussion on future research possibilities in Section VI.

II. Background: Review of SimHash

While most hash algorithms including MD5, SHA-256, and HAVAL-128 hash different inputs (even with the slightest of the variations) to entirely different hash digests, SimHash hashes similar inputs (in terms of the Hamming distance) to similar (closer) hash digests. Consider the following example:

```
phrase1 = "magic is all within you"
phrase2 = "magic is all in you"
phrase1.MD2 = 923ce24b045b25ad82341c2a8ac65f65
phrase2.MD2 = c0d972488d0c98763ab1f596a63e35f3
hammingDistance(phrase1.MD2, phrase2.MD2) = 65

phrase1.SimHash = 2da266b7f30b82d9
phrase2.SimHash = 2da366b773a382fd
hammingDistance(phrase1.SimHash, phrase2.SimHash) = 7
```

The SimHash algorithm uses an internal hashing algorithm to hash shingles or $n-$grams obtained from a given phrase. Each hash digest corresponding to each $n-$gram is then utilized to arrive at the final similarity hash digest. Algorithm 1 details the entire procedure employed to obtain the SimHash digest for a given input phrase, the specified value of n in $n-$grams (shingle size), and the defined internal hashing scheme.

Algorithm 1: SimHash Algorithm

Input: input phrase, shingle size, hash algorithm
Output: SimHash digest

```
n−grams ← inputPhrase.shingles(shingleSize)
hashDigests ← []
for shingle ∈ n−grams do
    hashDigest ← binary(shingle.hashAlgorithm)
    hashDigests.append(hashDigest)
SimHashBits ← [0] * len(hashDigests[0])
for hashDigest ∈ hashDigests do
    for idx ← 0 to len(hashDigest) do
        if hashDigest[idx] = 1 then
            SimHashBits[idx] ← SimHashBits[idx] + 1
        else
            SimHashBits[idx] ← SimHashBits[idx] − 1
SimHashDigest ← string.empty
for idx ← 0 to len(SimHashBits) do
    if SimHashBits[idx] > 0 then
        SimHashDigest.append('1')
    else
        SimHashDigest.append('0')
return SimHashDigest
```

The hash digests obtained through SimHash for similar input phrases often have low Hamming distance and higher Jaccard similarity. This property of SimHash is extremely practical in near-duplicate detection [17], [18]. By using SimHash for near-duplicate detection, we can reduce the time complexity from O(N^2) for pair-wise comparison to O(N).

III. Related Work

In the past, many cryptographic algorithms including MD4, MD5, SHA-0, and SHA-1, were broken by exploiting the structural weaknesses of the underlying hashing schemes [8], [9], [7]. Most of the studies concerning SimHash in the existing literature aim at evaluating the applicability of this locality-sensitive algorithm to near-duplicate detection in data processing applications, including plagiarism checking [19] and email spam detection [20].

Sood and Loguinov [21] proposed a significantly faster and a greater space-efficient approach to detect similar document pairs in large-scale data collections. Their bit-flipping algorithm resulted in certain performance overhead. Fu *et al*. [22] presented a document-based query searchable encryption scheme over encrypted cloud document, based on similarity hashing and trie based indexing. Jiang and Sun [23] proposed a semi-supervised SimHash algorithm to search high-dimensional data. Their algorithm learned the optimal feature weights from prior knowledge, to relocate the data, ensuring that similar data inputs have similar hash digests. Ho *et al*. [20] employed the SimHash algorithm with a parallel processing framework and meet-in-the-middle attack, to detect spam emails.

The existing research only presents the applications of the SimHash algorithm without any reference to its collision rate (and thus, the collision resistance). Hence, we conclude that there exist no state-of-the-art studies concerning the determination of the collision rates (resistance) for the SimHash algorithm.

IV. Methodology

Collision detection aims at finding two distinct inputs (here bit sequences) hashing to the same digest. Firstly, 2^n distinct bit sequences of length (n) varying from 2 to 32 are generated. Then, the SimHashes for the generated bit sequences are computed using the procedure in Algorithm 1, with an internal hashing scheme as MD5 and the shingle size of two. All the hash digests are stored in a hash map (dictionary) to ensure a constant lookup complexity (O(1)). In the hash map, we store the obtained hash digest as the key and the count of its occurrence as the corresponding value.

The computation for lower-order sequences (up to 16 bits in length) is manageable and does not require any parallel considerations. However, for higher-order bit sequences, the computational complexity of collision detection is very high, especially in terms of the time taken. Thus, the need to parallelize the entire process of collision detection becomes more relevant when dealing with higher-order bit sequences. In this study, we employ a distributed data-parallel approach using OpenMP [24], [25], MPI [26], and multiprocessing (in Python), to reduce the time complexity of the SimHash collision detection process efficiently.

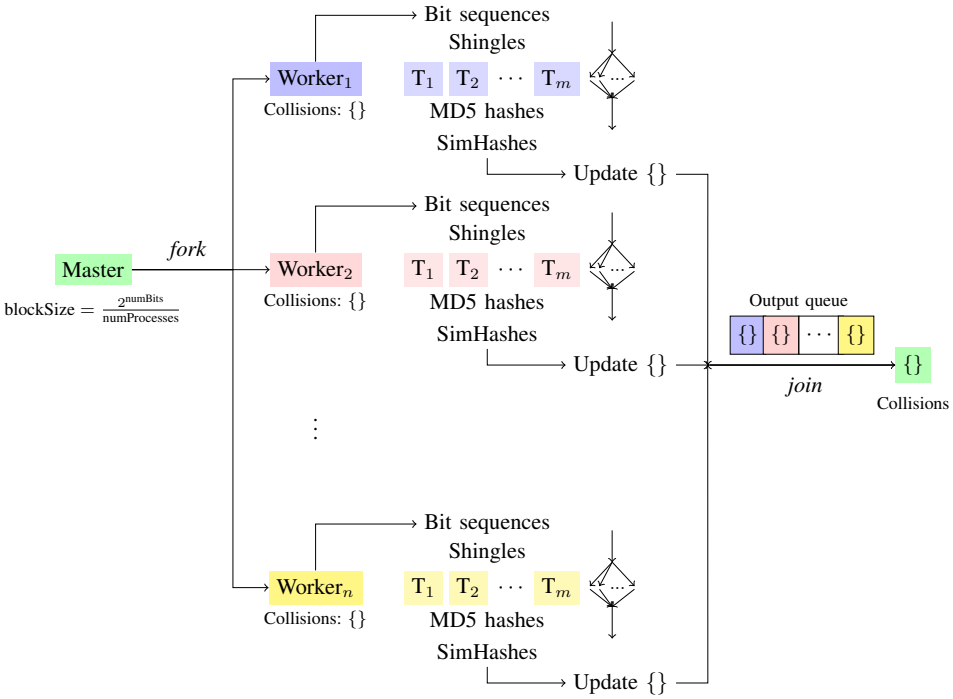


Fig. 1. Distributed data-parallel approach with optimal load balancing and synchronization to detect the collisions in SimHash (with MD5).

TABLE I
COLLISION RATE IN SIMHASH USING MD5 AS THE INTERNAL HASHING ALGORITHM.

#Bits	#Collisions	Collision rate (%)	#Processes	Time (s)
2	0	0	1	0.00007000
4	0	0	1	0.00032000
8	0	0	1	0.00613700
16	0	0	1	29.3268820
20	120	0.01144409180	4	464.700325
24	8,128	0.04844665527	4	6235.20520
28	41,523	0.01546852291	16	104092.872
32	1,438,275	0.03348744940	64	225431.219

The entire distributed data-parallel workflow employed in the detection of SimHash (with internal MD5 hashing scheme) collisions is depicted in Fig. 1. In our distributed data-parallel approach, the master process computes the block size as $\frac{2^{\text{numBits}}}{\text{numProcesses}}$. The master process (denoted as *Master*, in Fig. 1) then divides the task into several worker processes (denoted by *Worker*$_i$, $i \in [1, n]$, in Fig. 1), which then compute the workload corresponding to the predetermined block size. Different worker processes are then run on multiple machines with identical computing power. Each worker process maintains a collision dictionary into which it updates the SimHashes and their counts. The workload per worker process involves the generation bit sequences, $n-$grams (shingles), and MD5 hashes, along with SimHash computations for all the shingles. Each process spawns several threads (denoted by T_i, $i \in [1, m]$, in Fig. 1) to distribute the computation of MD5 hashes and SimHashes, thus ensuring further parallelization. Synchronization among the threads during the updation of the collision dictionary is ensured through locks. Once a worker process completes its workload, it enqueues its corresponding collision dictionary into a process output queue maintained by the master process. All the collision dictionaries from the process output queue are then merged by adding the values (counts) for same keys (SimHash digests) across various dictionaries.

Furthermore, we recorded the execution times, to measure the overall time taken in the identification of the collision rate (and thus, the collision resistance) for a given bit sequence length. Execution time for every bit sequence length (2 to 32) is collected eight times to overrule the bias caused due to any other system processes that are not under the control of the experimenter. Moreover, with every run of the experiment, the order of experimentation for a specific bit length was shuffled to ensure an unbiased measurement of the time taken. The individual measurements were then averaged to obtain the overall time taken to identify collisions accurately.

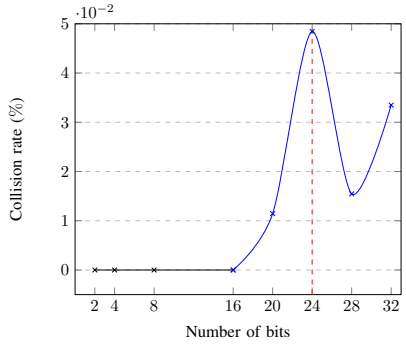


Fig. 2. A graph depicting the variation in the collision rate with the increasing number of bits.

V. Experimental Results

All the results presented in this study are obtained using multiple nearly identical machines with an i5 7200U at $4\times$ 3.1 GHz processor, an 8 GB DDR3 at 1333 MHz memory, and $10/100/1000$ Gigabit LAN network.

The collision rates are computed as $\frac{\text{numcollisions}}{2^{\text{numBits}}}$. The variation in the collision rate (%) plotted against the variation in the number of bits is presented in Fig. 2. It can be observed that the collision rates for lower-order bit sequences (2 to 16 bits) is 0%. However, a maximum collision rate of 0.048% can be observed for $24-$bit sequences (marked with a red dotted line in Fig. 2). Table I tabulates the experimental results obtained for bit sequences with length varying from 2 to 32 bits.

It is evident from Fig. 2 that the collision rate increases for higher-order collisions (with a maximum value at 24 bits). It can also be observed from Table I that the time taken in the determination of the collision rate increases exponentially with the increase in the number of bits. Approximately a duration of a day and four hours for $28-$bit sequences, and two days and 14 hours for $32-$bit sequences was required to determine their respective collision rates. Distributed data parallelization with synchronization and optimal load balancing resulted in a greater speedup and more efficient storage utilization than the sequential counterparts.

VI. Conclusions

Evaluating the collision resistance of a cryptographic hashing algorithm plays a pivotal role in applications requiring integrity, such as digital signature schemes, e-cash, and proof-of-work systems. SimHash is a widely used locality-sensitive algorithm used in many large-scale data processing applications. In this paper, we presented a distributed data-parallel framework with synchronization and optimal load balancing, to detect the collision rates of the SimHash algorithm with a more significant speedup and efficient storage utilization. We presented our analysis using bit sequences with length varying from 2 to 32 bits. It was observed that the time taken to detect the collisions increases exponentially with the increase in the number of bits. As a part of the future work, we aim at analyzing the bit patterns of the SimHash digests in great detail, to try and exploit any internal structural weaknesses.

References

[1] H. Krawczyk, M. Bellare, and R. Canetti, "Hmac: Keyed-hashing for message authentication," Tech. Rep., 1997.

[2] S. Goldwasser and M. Bellare, "Lecture notes on cryptography," *Summer course "Cryptography and computer security" at MIT*, vol. 1999, p. 1999, 1996.

[3] J. Floyd, "What do Hash Collisions Really Mean?" Jul 2008, [Online; accessed 21. Dec. 2018] URL: https://permabit.wordpress.com/2008/07/18/what-do-hash-collisions-really-mean.

[4] R. Pass, "Lecture 21: Collision-Resistant Hash Functions and General Digital Signature Scheme," *Course on Cryptography at Cornell University*, Nov 2009, uRL: https://www.cs.cornell.edu/courses/cs6830/2009fa/scribes/lecture21.pdf.

[5] R. Rivest, "The md5 message-digest algorithm," Tech. Rep., 1992.

[6] J. H. Burrows, "Secure hash standard," DEPARTMENT OF COMMERCE WASHINGTON DC, Tech. Rep., 1995.

[7] X. Wang, D. Feng, X. Lai, and H. Yu, "Collisions for hash functions md4, md5, haval-128 and ripemd." *IACR Cryptology ePrint Archive*, vol. 2004, p. 199, 2004.

[8] X. Wang and H. Yu, "How to break md5 and other hash functions," in *Annual international conference on the theory and applications of cryptographic techniques*. Springer, 2005, pp. 19–35.

[9] X. Wang, Y. L. Yin, and H. Yu, "Finding collisions in the full sha-1," in *Annual international cryptology conference*. Springer, 2005, pp. 17–36.

[10] R. Rivest, "The md4 message-digest algorithm," Tech. Rep., 1992.

[11] B. Den Boer and A. Bosselaers, "An attack on the last two rounds of md4," in *Annual International Cryptology Conference*. Springer, 1991, pp. 194–203.

[12] H. Dobbertin, "Cryptanalysis of md4," in *International Workshop on Fast Software Encryption*. Springer, 1996, pp. 53–69.

[13] J. Katz, A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 1996.

[14] H. Dobbertin, "Ripemd with two-round compress function is not collision-free," *Journal of Cryptology*, vol. 10, no. 1, pp. 51–69, 1997.

[15] J. Seberry, "Haval a one-way hashing algorithm with variable length of output 1 yuliang zheng josef pieprzyk," 1993.

[16] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 380–388.

[17] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting near-duplicates for web crawling," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 141–150.

[18] S. Buyrukbilen and S. Bakiras, "Secure similar document detection with simhash," in *Workshop on Secure Data Management*. Springer, 2013, pp. 61–75.

[19] C. Sadowski and G. Levin, "Simhash: Hash-based similarity detection," 2007.

[20] P.-T. Ho, H.-S. Kim, and S.-R. Kim, "Application of sim-hash algorithm and big data analysis in spam email detection system," in *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*. ACM, 2014, pp. 242–246.

[21] S. Sood and D. Loguinov, "Probabilistic near-duplicate detection using simhash," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1117–1126.

[22] Z.-J. Fu, J.-G. Shu, J. Wang, Y.-L. Liu, and S.-Y. Lee, "Privacy-preserving smart similarity search based on simhash over encrypted data in cloud computing," *ŁŁ*, vol. 16, no. 3, pp. 453–460, 2015.

[23] Q. Jiang and M. Sun, "Semi-supervised simhash for efficient document similarity search," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 93–101.

[24] B. Chanduka, T. Gangavarapu, and C. D. Jaidhar, "A single program multiple data algorithm for feature selection," in *Intelligent Systems Design and Applications*. Cham: Springer International Publishing, 2020, pp. 662–672.

[25] T. Gangavarapu, H. Pal, P. Prakash, S. Hegde, and V. Geetha, "Parallel openmp and cuda implementations of the n-body problem," in *Computational Science and Its Applications – ICCSA 2019*. Cham: Springer International Publishing, 2019, pp. 193–208.

[26] M. Snir, S. Otto, S. Huss-Lederman, J. Dongarra, and D. Walker, *MPI–the Complete Reference: The MPI core*. MIT press, 1998, vol. 1.

Jeffries-Matusita distance as a tool for feature selection

Rikta Sen
Faculty of Software and Information Science
Iwate Prefectural University
Iwate, Japan
rikta.sen06@gmail.com

Saptarsi Goswami
A.K.Choudhury School of IT
Calcutta University
Kolkata, India
line 5: saptarsi007@gmail.com

Basabi Chakraborty
Faculty of Software and Information Science
Iwate Prefectural University
Iwate, Japan
basabi@ iwate-pu.ac.jp

***Abstract*— Feature selection is one of the most important preprocessing steps in Machine Learning. This can be broadly divided into search based methods and ranking based methods. The ranking based methods are very popular because they need much lesser computational power. There can be many different ways to rank the features. One of the ways to measure effectiveness of a feature is by evaluating its ability to separate the classes involved. These interclass Separability based measures can be directly used as a feature ranking tool for binary classification problems. Bhattacharya Distance which is the most popular among them has been used majorly in a recursive setup to select good quality feature subsets. Jeffries-Matusita (JM) distance improves Bhattacharya distance by normalizing it between 0 and 2. In this paper, we have ranked the features based on JM distance. The results are comparable with mutual information, Relief and Chi Squared based measures as per experiments conducted over 24 public datasets but in much lesser time. JM distance also provide some intuition about the dataset prior to any feature selection or machine learning algorithm. A comparison has been done on classification accuracy and JM scores of these datasets, which can provide a good intuition on how good a dataset is for classification and point out the need of or lack of further feature collection.**

Keywords— Bhattacharya Distance, Chi Squared, Feature selection, Jeffries-Matusita (JM) distance, Mutual information, Relief

I. INTRODUCTION

All machine learning or statistical models aims at estimating the function' f 'which maps the input or independent variables to the output or the dependent variable. The two major reasons for such estimation are [1] namely prediction and inference. For prediction' f ' can be treated as a black box whereas for inference we have to understand more about importance of individual features, interactions of the features and final interpretability of the model. Feature selection which can be described as a process of selecting features which are relevant and are not redundant for a particular task, can help in both these objectives particularly in inferring the model. From last few years the number of features that are captured and can be stored has increased exponentially because of the surge in Internet of Things (IOTs), Sensors, Social Media Applications etc. contributing as source and then cloud computing making the storage of the features easier. As noted by Tang et al., the growth in UCI dataset repository has been exponential [2]. However, as noted by Shu et al. [3] in their recent work, not all these features that are getting captured and recorded are helping the models rather the irrelevant and redundant features have a tendency to confuse the learning algorithm and thereby affecting the accuracy negatively.

Feature Selection can be broadly classified into filter based and wrapper based approaches [4],[24]. Filter based methods are model agnostic and primarily depends on the characteristics of the feature subsets whereas the wrapper methods evaluates the feature subset based on its performance with a fixed machine learning model. Filter based approaches are more generic and computationally less expensive. A filter approach can work as a search based or as a ranking based method. As the state space to be searched in case of feature selection is combinatorial in nature, search methods need significant computational power. The ranking based method produces a rank of features and uses different techniques to select few among them. Though the ranking based methods ignore the feature interaction, as noted by Prati, these methods are simple, scalable and yields good performance from empirical studies [5]. Shu et al., as a part of their empirical study over eight publicly available bio medical data observed that the univariate methods to be more stable as compared to the multivariate methods. It may also be noted that for quite a few of the domains, feature ranking has been observed to be used. Few such examples are in Alzheimer's disease classification from structural MRI [6], Software Quality [7], textual data [8], gene Expression Data [9] etc. As suggested by Goswami et al. [10] a measure like multi-variety score (MVS) can be used to classify the datasets into strong correlated, weak correlated, weak independent and strong independent categories. Subsequently ranking based feature selection methods can be applied on strong independent and weak independent datasets.

The measures that are used for ranking of the features generally are either statistics based or are information theory based. Some such important measures which are popular and hence widely followed are Information Gain, Symmetrical Uncertainty, Relief, One- R, Chi – Squared etc. [11]. Most of these measures focus on the predictive power of the features with respect to the class. Another way of measuring the effectiveness of the features may be by their ability of separating the classes in a dataset. Bhattacharya Distance proposed by Anil Kr. Bhattacharya in 1930's measures divergence between two probability distributions. For binary classification problems the class labels divide each of the features into two distributions and hence measures like Bhattacharya Distance can be used to measure the separability between these feature class distributions. Class separability scores like Bhattacharya Distance can also be used for feature ranking especially for binary classification problems. Jeffries-Matusita distance is an improvement over Bhattacharya Distance, which standardized the distance between 0 to 2 for a*n* easy comparison across datasets. As per literature study, though Bhattacharya Distance has been

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

used for feature ranking [12], there does not seem to be any research effort where JM distance is used for feature ranking.

In this paper, Jeffries-Matusita distance has been used as a feature ranking measure for binary classification problem over 24 publicly available datasets. The results have been compared with three popular ranking based measures including Information Gain, Relief and Chi Squared. An analysis of the JM value with the classification accuracy has been done, to understand if such analysis reveals any more information over and above the selection of the features.

Rest of the paper is organized as following, in Section II, related works on feature ranking have been discussed. In Section III, some general measures for evaluating relevance of a feature is discussed. In Section IV, details of the experimental setup are furnished. In Section V, the results have been presented with a critical analysis of the same. Section VI, contains the conclusion.

II. Related Work

As mentioned in Section I, ranking based feature selection is being applied in many application domains. Winker et al. has applied feature ranking mechanism to identify type 1 diabetes susceptibility genes [13]. Hulse et al. did an extensive study on bioinformatics dataset using ranking based on Chi-Squared, Information Gain and Relief and a metric proposed by the authors [14]. One of the earlier works of feature selection using Bhattacharya Distance was done as early as 1996. Here, a recursive algorithm to select optimal feature subset was proposed for L-class problem under normal multi-distribution assumption [15]. The formulation has strong theoretical foundation however may turn computationally complex. C. C. Reyes et al. have extended the concept of Bhattacharya Distance to Bhattacharya Space for a problem of texture segmentation [16]. In 2006, Xuan et al. proposed a Bhattacharya Distance based recursive feature selection algorithm which used PCA based pre-processing and attempts to find $m \times n$ transformation matrix, which converts n dimensional original feature space to m dimensional reduced feature set while minimizing the class error probability[17]. However, the process needs significant computational power and works under the strong assumption of equal prior probability. In a paper in 2015 [12], the authors used Bhattacharya Distance for feature ranking directly on 8 biomedical datasets which has two classes. JM distance which is an improvement over Bhattacharya distance have been used in feature extraction [18] and also in some search based methods of feature selection [19]. However as per our study of literature there seem to be very less work where JM Distance is used for feature ranking.

III. Feature Ranking Measures

In this section, some of the important feature ranking measures have been discussed. Some of the most important feature ranking measures for classification are as follows:

A. Mutual Information

Mutual Information is an information theoretic measure which expresses the dependency of one variable on another variable. If mutual information is used between a feature and the class, then this gives one basis to measure relevance of a feature. Mutual information can be calculated as follows:

$$MI(Class, A) = H(Class) + H(A) - H(Class, A) \quad (1)$$

H indicates entropy, entropy of a random variable is calculated as:

$$H(A) = -\sum_a P_a(A) \log p_a(A) \quad (2)$$

Mutual Information is one of the most used measures in feature selection. Over the years, there have been several improvements over mutual information. Normalized mutual information maps the value of mutual information between 0 and 1. The work produced by Estévez et al. [20] is an important reference of feature selection using Mutual Information.

There are other information theoretic measures like information gain and symmetrical uncertainty which are minor variation of the same concepts. In this paper, for the experimental study, information gain has been used.

B. Chi Squared (χ^2)

The idea behind Chi Squared is first to assume that the feature and the class are independent of each other. Then these expected values are compared with the actual values. χ^2 is given by the following equation

$$\chi^2 = \frac{\sum_{k=1}^{c} n_k (\mu_k^j - \mu_j)^2}{\sigma_j^2} \quad (3)$$

Where k denotes the class and j denotes the feature.

C. Bhattacharya Distance

Bhattacharya Distance is used to measure similarity between two probability distributions. For a binary classification problem each feature can give two probability distributions based on the class values. If these two distributions are considered to be normal then Bhattacharya distance is calculated as following:

$$D_B(p,q) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2}\right) \quad (4)$$

where p and q respectively denotes the two probability distribution created by a feature.

D. Jeffries-Matusita distance

This is a distance measure, which improves Bhattacharya Distance by scaling it between 0 and 2.

$$J_M(p,q) = \sqrt{2(1 - e^{-D_B(p,q)})} \quad (5)$$

JM distance has been seen to be particularly useful for remote sensing.

E. Relief

Relief is calculated iteratively based on random instances from the dataset. A weight corresponding to the feature is calculated based on nearHit (closed instance from the same class) and nearMiss (closest distance from different class). w_i is initialized to 0 and then in each step it is updated as the following:

$$w_i = w_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \quad (6)$$

Finally, an average of w_i is taken over the iterations.

For a more broad based understanding of different types of measures like distance measures, information measures, dependency measures and consistency measures the work by Huan Liu [21] can be referenced.

IV. MATERIAL AND METHODS

In this section, different details of the empirical study that we conducted are furnished which can be used to reproduce the results. Datasets are taken from the public UCI data repository [22].

Dataset Description: Dataset description is shown in Table I.

TABLE I. DESCRIPTION OF DATASETS

Datasets	No. of features	No. of classes	No. of Instances	Description
Sonar	61	2	208	This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network [1]. The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.
Ion (Ionosphere Data Set)	34	2	351	This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere.
Bupa (Liver Disorders Data Set)	7	2	345	The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the dataset constitutes the record of a single male individual.
Heart	14	2	270	This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient.
Biodeg (QSAR biodegradation Data Set)	41	2	1055	The QSAR biodegradation dataset was built in the Milano Chemometrics and QSAR Research Group (UniversitÃ degliStudi Milano â€“ Bicocca, Milano, Italy).The data have been used to develop QSAR (Quantitative Structure Activity Relationships) models for the study of the relationships between chemical structure and biodegradation of molecules.
Apndcts (Appendicitis Data set)	8	2	106	The data represents 7 medical measures taken over 106 patients on which the class label represents if the patient has appendicitis (class label 1) or not (class label 0).
Mcg (MAGIC Gamma Telescope Data Set)	11	2	19020	The data are MC generated (see below) to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique.
Twonorm	21	2	7400	This is an implementation of Leo Breiman's twonorm example. It is a 20 dimensional, 2 class classification example. Each class is drawn from a multivariate normal distribution with unit variance.
Best cancer (Breast Cancer Wisconsin (Diagnostic) Data Set)	32	2	569	Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
Diabetes (Pima Indians Diabetes Database)	9	2	768	The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
Prostate Cancer	10	2	100	This dataset contains 8 active variables that are radius, texture, perimeter, area, smoothness, compactness, symmetry and fractal dimension.
Lung Cancer	7	2	59	This data set contains 4 active variables including age, smokes, areaQ, and alcohol, and task is to predict lung cancer.
Cryotherapy	7	2	90	This dataset contains information about wart treatment results of 90 patients using cryotherapy.
Fertility diagnosis	10	2	100	100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria. Sperm concentrations are related to socio-demographic data, environmental factors, health status, and life habits.
Indian Liver Patient dataset (ILPD Dataset)	10	2	583	This data set contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos.
Banknote authentication	5	2	1372	Data were extracted from images that were taken from genuine and forged banknote-like specimens.
Faults (Steel Plates Faults Data Set)	27	2	1941	A dataset of steel plates faults, classified into 7 different types. The goal was to train machine learning for automatic pattern recognition.
kc2 (KC2 Software defect prediction)	22	2	522	One of the NASA Metrics Data Program defect data sets. Data from software for science data processing. Data comes from McCabe and Halstead features extractors of source code. These features were defined in the 70s in an attempt to objectively characterize code features that are associated with software quality.

Phoneme	5	2	5404	The aim of this dataset is to distinguish between nasal (class 0) and oral sounds (class 1). The dataset originates from the European ESPRIT 5516 project: ROARS. The aim of this project was the development and the implementation of a real time analytical system for French and Spanish speech recognition.
pc1 (PC1 Software defect prediction dataset)	23	2	1109	One of the NASA Metrics Data Program defect data sets. Data from flight software for earth orbiting satellite. Data comes from McCabe and Halstead features extractors of source code.
Climate model	21	2	540	This dataset contains records of simulation crashes encountered during climate model uncertainty quantification (UQ) ensembles. The goal is to use classification to predict simulation outcomes (fail or succeed) from input parameter values, and to use sensitivity analysis and feature selection to determine the causes of simulation crashes.
SPECTF	45	2	349	The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.
Satellite	37	2	5100	The satellite dataset comprises of features extracted from satellite observations. In particular, each image was taken under four different light wavelength, two in visible light (green and red) and two infrared images. The task of the original dataset is to classify the image into the soil category of the observed region.
Japanese Vowels	12	2	640	This dataset records 640 time series of 12 LPC cepstrum coefficients taken from nine male speakers.

- 'R' has been as the computational environment [23].
- NaiveBayes have been used as the machine learning classifier.
- 'R' Packages SpatialEco and FSelector have been used for the calculation of JM Distance, Information Gain (IG), Chi-Square (CS) and Relief.
- Default parameters have been used for computation of relief.
- 80% of each the datasets has been taken as training and 20% as testing. This has been repeated 10 times with different seeds and the average value has been reported.
- The classification accuracies have been computed for 10%, 25%, 50% and 75% for each of the methods and the maximum of them have been reported.

V. RESULTS AND ANALYSIS

The classification accuracies of the four methods JM Distance, Information Gain, Chi Squared and Relief for feature selection algorithms are compared in subsection A. In subsection B, these methods are compared in terms of execution time. Finally, in section C, an investigation has been performed with the top 10% JM values of the datasets with the classification accuracy, to understand if these values give any general idea of the separability of the classes in the dataset.

A. Comparison on classification accuracy

In Table II, the maximum classification accuracy of these methods is reported.

It can be observed from the Table II that

- In terms of maximum wins, JM distance is 3rd followed Chi Square and Information Gain.

TABLE II. CLASSIFICATION ACCURACY COMPARISON

Dataset	JM	IG	Relief	CS
Sonar	0.6929	0.6929	**0.731**	0.6952
Ion	**0.9394**	0.893	0.893	0.8746
Bupa	0.5652	**0.5739**	0.5768	**0.5739**
Heart	0.8519	**0.8556**	0.8444	**0.8556**
Biodeg	0.718	**0.7393**	0.7318	**0.7393**
Apndcts	**0.8773**	**0.8773**	**0.8773**	**0.8773**
Mgc	0.7401	**0.7547**	0.7296	**0.7547**
Twonorm	**0.977**	**0.977**	**0.977**	**0.977**
Brest cancer	0.943	**0.9465**	0.943	**0.9465**
Diabetes	**0.776**	0.7747	0.7675	0.7747
Prostate Cancer	0.83	**0.85**	0.835	**0.85**
Lung Cancer	**0.9833**	**0.9833**	**0.9833**	**0.9833**
Cryotherapy	**0.8889**	**0.8889**	**0.8889**	**0.8889**
Fertility_Diagnosis	0.87	**0.885**	**0.885**	**0.885**
Indian LiverPatientsdataset(ILPD)	**0.5754**	0.5595	0.5629	0.5586
Banknote authentication	**0.8796**	0.8749	0.8749	**0.8796**
Faults	0.5185	0.526	**0.6573**	0.5735
kc2	**0.8476**	0.8429	0.8381	0.8438
Phoneme	0.7684	0.7654	0.7594	**0.7705**
pc1	**0.9135**	0.9027	0.9077	0.9041
Climate Model	**0.8954**	**0.8954**	**0.8954**	**0.8954**
SPECTF	0.7471	0.75	0.74	**0.7571**
Satellite	0.9841	**0.9861**	0.9856	**0.9861**

JapaneseVowels	**0.9599**	0.9588	0.9568	0.9584

A more detailed comparison is enclosed in Fig. 1 below:

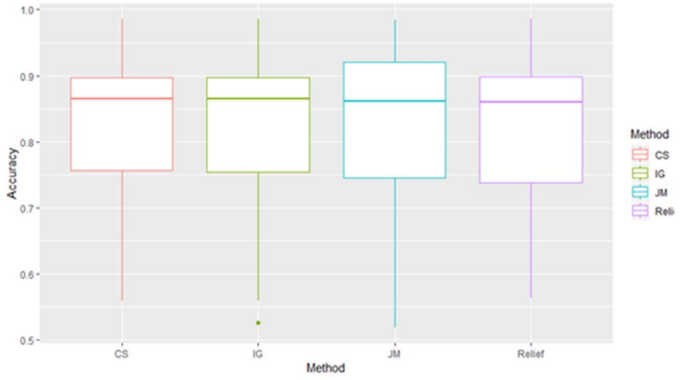


Fig. 1. Classification Accuracy of all the methods.

From Fig. 1, it can be concluded that

- All the methods are quite comparable, though JM distance does not win in terms of number of datasets, the range of classification accuracies produced by JM is very much comparable with all other methods.

B. Comparison on execution time

In this section, analysis of the comparison on the basis of execution time has been enclosed. This is enclosed in the below Table III.

TABLE III. COMPARISON OF EXECUTION TIME FOR ALL DATASETS

Data set	JM	IG	Relief	CS
Sonar	0.1813	0.2402	24.26	0.2052
Ion	0.1239	0.1782	15.34	0.1548
Bupa	0.02858	0.04127	3.355	0.04341
Heart	0.04915	0.08396	4.776	0.06185
Biodeg	0.3029	0.405	52.18	0.3405
Apnd	0.02617	0.03484	1.418	0.03281
Mgc	1.168	1.68	314.68	1.621
Twonorm	0.7747	1.17	212.52	1.107
Brest cancer	0.1382	0.1989	23.19	0.1878
Diabets	0.05705	0.08227	9.174	0.07453
Prostate Cancer	0.024608	0.04075	1.464	0.03183
Lung Cancer	0.014483	0.02413	0.5226	0.01766
Cryotherapy	0.01957	0.03044	1.0064	0.02745
Fertility_Diagnosis	0.02696	0.03891	1.53	0.03238
ILPD	0.05823	0.08041	10.98	0.07577
Banknote authentication	0.05848	0.08939	9.526	0.08543
Faults	0.3262	0.4407	66.18	0.4637
kc2	0.09688	0.1419	13.62	0.1299
Phoneme	0.243	0.3236	41.3	0.3334
pc1	0.1891	0.2307	28.52	0.2219
ClimateModel	0.09906	0.1383	12.48	0.12572
SPECTF	0.1711	0.2152	21.79	0.1994
Satellite	0.9317	1.296	237.9	1.28
JapaneseVowels	0.7298	1.0048	158.91	0.9991

From Table III, it can be observed that

- Relief is the most expensive of all the four.
- IG, JM and CS are comparable.

The comparison of these three methods has been shown in Fig. 2.

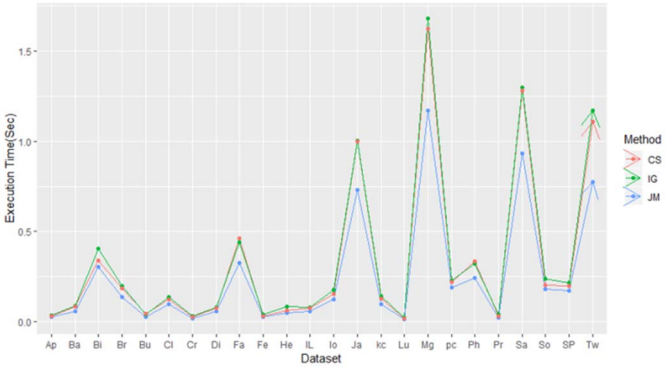


Fig. 2. Comparison of JM, CS and IG on the basis of execution time.

It can be observed that JM which is the blue line consistently takes the lowest time as compared to other datasets.

C. Comparison of classification accuracy and top JM values of the datasets

An invsetigation on the classification accuracies achieved on the datasets has been done with the average of top 10% JM values. The comparison has been demonstrated in the Fig. 3 below. The datasets having a value greater than one have been marked as ;High', the ones having a value greater than 0.5 and less than 1 have been marked as medium and ones having less than 0.5 has been marked as low as far as class reparability is concerned.

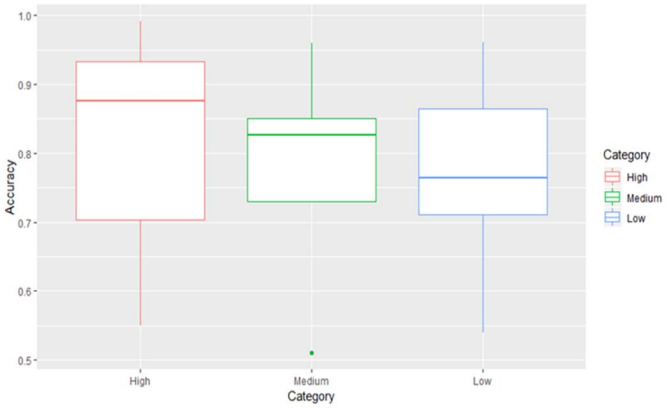


Fig. 3. Comparison of top JM values across datasets.

As observed from Fig. 3

- The datasets marked as high have displayed higher classification accuracy on average as compared to Medium.
- Similarly, datasets marked as medium showed higher classification accuracy on average as compared to Low.
- The maximum values for 'high' marked datasets are also much more than the 'Medium' class.

- Similarly, the minimum values of the 'low' marked datasets are much less than the 'Medium' marked datasets.

VI. CONCLUSION

Feature selection is one of the most critical methods of machine learning. As search based methods become prohibitively expensive for larger dimensions, feature ranking is considered as a viable tool for many application domains. Most of the works on feature ranking have focused on information theoretic measures; however, class separability measures are relatively unexplored may be with the exception of Bhattacharya distance. In this paper, JM distance have been used as a tool for ranking based feature selection. The proposed method has been compared with standard feature ranking measures like information gain, chi-squared, relief etc. over 24 publicly available datasets. The results in classification accuracy is quite comparable with other methods, however JM distance takes much lesser time as compared to all other methods. A comparison has been done on average of top 10% JM value of the features with the classification accuracy. The analysis clearly reveals some relation among them and can be extended to much further and in depth study of the same.

VII. REFERENCES

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. Springer, 2013.

[2] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review BT - Data classification: algorithms and applications," 2014.

[3] W. Shu, W. Qian, and Y. Xie, "Incremental approaches for feature selection from dynamic data with the variation of multiple objects," *Knowledge-Based Syst.*, vol. 163, pp. 320–331, 2019.

[4] S. Goswami, A. K. Das, A. Chakrabarti, and B. Chakraborty, "A feature cluster taxonomy based feature selection technique," *Expert Syst. Appl.*, vol. 79, pp. 76–89, 2017.

[5] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *Proceedings of the International Joint Conference on Neural Networks*, 2012, pp. 1–8.

[6] I. Beheshti and H. Demirel, "Feature-ranking-based Alzheimer's disease classification from structural MRI," *Magn. Reson. Imaging*, vol. 34, no. 3, pp. 252–263, 2016.

[7] T. M. Khoshgoftaar, K. Gao, and A. Napoliano, "An Empirical Study of Feature Ranking Techniques for Software Quality Prediction," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 22, no. 02, pp. 161–183, 2012.

[8] A. Rehman, K. Javed, H. A. Babri, and M. Saeed, "Expert Systems with Applications Relative discrimination criterion – A novel feature ranking method for text data," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3670–3681, 2015.

[9] Shaohong Zhang, Hau-San Wong, Ying Shen, and Dongqing Xie, "A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 4, pp. 1257–1263, 2012.

[10] S. Goswami, A. Chakrabarti, and B. Chakraborty, "A proposal for recommendation of feature selection algorithm based on data set characteristics," *J. Univers. Comput. Sci.*, vol. 22, no. 6, pp. 760–781, 2016.

[11] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.*, vol. 21, no. 1, pp. 119–135, 2011.

[12] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Comput. Biol. Med.*, vol. 66, pp. 1–10, 2015.

[13] C. Winkler *et al.*, "Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes," *Diabetologia*, vol. 57, no. 12, pp. 2521–2529, 2014.

[14] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "A comparative evaluation of feature ranking methods for high dimensional bioinformatics data," *Proc. 2011 IEEE Int. Conf. Inf. Reuse Integr. IRI 2011*, pp. 315–320, 2011.

[15] X. Guorong, C. Peiqi, and W. Minhui, "Bhattacharyya distance feature selection," *Proc. - Int. Conf. Pattern Recognit.*, vol. 2, pp. 195–199, 1996.

[16] C. C. Reyes-Aldasoro and A. Bhalerao, "The Bhattacharyya space for feature selection and its application to texture segmentation," *Pattern Recognit.*, vol. 39, no. 5, pp. 812–826, 2006.

[17] G. Xuan, X. Zhu, P. Chai, Z. Zhang, Y. Q. Shi, and D. Fu, "Feature selection based on the Bhattacharyya distance," in *Proceedings - International Conference on Pattern Recognition*, 2006, vol. 3, pp. 1232–1235.

[18] M. R. P. Homem, N. D. A. Mascarenhas, and P. E. Cruvinel, "The linear attenuation coefficients as features of multiple energy CT image classification," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 452, no. 1–2, pp. 351–360, 2000.

[19] S. B. Serpico, M. D'Inca, F. Melgani, and G. Moser, "Comparison of feature reduction techniques for classification of hyperspectral remote sensing data," *Image Signal Process. Remote Sens. VIII*, vol. 4885, pp. 347–358, 2003.

[20] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[21] L. Huan and Y. Lei, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, 2005.

[22] C. D. a. G. Dua, "UCI Machine Learning Repository," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[23] R Development Core Team, "R: A Language and Environment for Statistical Computing," Viena, Austria, 2018.

[24] Goswami S, Chakrabarti A. Feature selection: A practitioner view. International Journal of Information Technology and Computer Science (IJITCS). 2014 Oct;6(11):66.

Change Analysis of Indian Metropolitan Cities through a Spatiotemporal Ontology

Saritha S
Department of Computer Science
Cochin University of Science and Technology
Kochi, Kerala, India
sarithas@cusat.ac.in

G Santhosh Kumar
Department of Computer Science
Cochin University of Science and Technology
Kochi, Kerala, India
san@cusat.ac.in

***Abstract*— Spatial regions are being continuously monitored for their ever changing pattern with respect to time. A constant watch on the patterns of a region will help in understanding the changes that is happening around and enable the decision makers to act accordingly for betterment of the society. It is essential to monitor the cities and their associated growing regions to suffice for the growth of a nation. This is the underlying motivation of this study. In this work, a spatiotemporal ontology for change analysis is applied to understand the growing pattern of Indian Metropolitan cities on a semantic level. Change analysis of the cities are studied in terms of landscape metrics incorporated in the ontology. A conclusive study of growth pattern of the case studies is presented in this paper.**

Keywords—change analysis, ontology, spatiotemporal, cities

I. INTRODUCTION

Change analysis [1] of a spatial region is of immense importance, as it plays a crucial role for town-planners and decision makers to understand the developing/developed land of the region. Such an understanding will help in the proper planning of utilities and facilities to mankind for better social life. This takes us to one step ahead of *change detection* [2], which is termed as *change analysis*. The process is usually achieved with the input from remote sensing images and other supporting data. In this work, a spatiotemporal ontology [3] is used to find the change patterns of a spatial region at different temporal resolutions. Ontologies are formal mechanisms of defining the properties and relationships of objects or entities that exist in the domain of interest [4]. They help to generalize information and minimize complexity, and the model is used to solve problems in the same domain. The objective of this work is to perform the change analysis of Indian metropolitan cities using the spatiotemporal ontology proposed by us [3]. Specific study areas are chosen for the study. After applying the spatiotemporal ontology, a detailed analysis of the changes detected in the temporal resolution is sought. The detailed analysis aids to find out how the landscape has grown over the years. It will also serve as an insight to deduce the general characteristics of the Indian metropolitan cities.

A first attempt to quantify the changes that has happened in a landscape is seen in [5] using remote sensing images and landscape metrics. An urban sprawl index is also proposed in the same work. A massive study in the same concept is performed in [6] which takes into consideration around 77 urban regions worldwide. A similar approach combining census approach is taken in [7] to study the diversity associated with the cities in the world. The analysis of study in all these works for the different parameters vary with respect to the case studies chosen. Analogous to spatiotemporal ontology, a spatiotemporal reasoning approach is proposed in [8] using a 4D-fluent approach. Generally change detection are done in two categories namely, pixel – based change detection and object-based change detection. The common techniques to depict changes are contingency tables and maps of changed regions [9, 10, 11] in pixel-based techniques.

The paper is organized as follows. Section 2 points to the details of study area chosen. The next section, Section 3 describes the methods used in this study and also briefs about the spatiotemporal ontology. Section 4 is a detailed analysis of the change patterns observed in the case studies. The work is concluded in section 5.

II. STUDY AREA

India is a land of diversities. The diversification of the country is attributed to many factors like the unique linguistic states, culture difference between northern and southern states, landscape patterns of the states and population of the state. The diversification has evolved over the years. Of course, along these years the landscape of the region also has undergone drastic change. To study how the landscape has changed, two metropolitan cities, which has immensely contributed to the growth of the nation, namely, Mumbai and Bangalore are taken as case studies.

Mumbai (formerly known as Bombay) is a metropolitan city situated on the west coast of India. It is one of the most populous cities of India. Over the past decade, Mumbai has grown as the financial and commercial centre of India, thus supporting the nation's economy. Bangalore (officially known as Bengaluru), a metropolitan city of Indian sub-continent and is accepted as a twin-town/sister city of many cities world-wide, including San Francisco, USA. It is the third most populous city of India and is located in southern India. The city is regarded as the "Silicon Valley of India" and is the leading IT exporter of the country. Bangalore city also accounts for the nation's development in terms of IT sector. The study areas are marked in Fig 1.

III. METHODOLOGY

The method of study adopted in this paper can be broadly done in two phases. In the first phase, remote sensing images are to be classified using the appropriate classifier model. Obviously, before classification, image pre-processing techniques have to be applied to rectify the radiometric errors. Common classifier models like Naïve Bayes', Neural Networks, Support Vector Machines, Support Tensor Machines are experimented with to find the appropriate model. In the second phase, classified images of different temporal resolutions are taken into the spatiotemporal ontology which is built to quantify the change patterns evolved in a region. The spatiotemporal ontology has been built with traditional features associated with remote sensing images and is supported by landscape metrics [8].

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

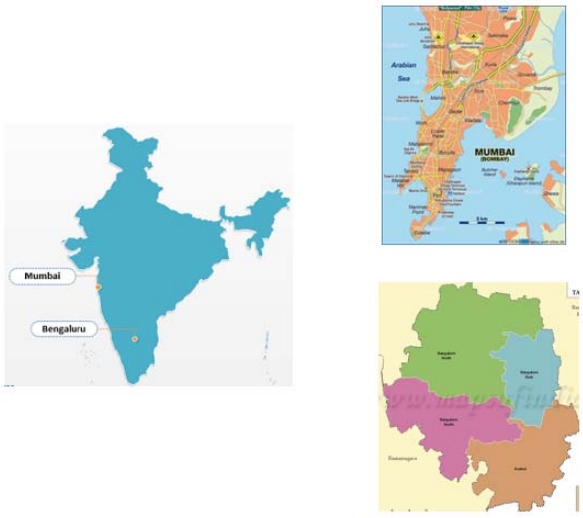


Fig 1. Study Area – Mumbai and Bangalore
(Map Courtesy – http:// www.mapsofindia.com, https://www.reddit.com/r/bangalore, https://www.reddit.com/r/mumbai)

A. Classification

The remote sensing images acquired from the Landsat 6, 7 and 8 satellite systems are chosen for Mumbai and Bangalore city. The details of the satellite systems are given in [12]. To classify the images, a new set of features proposed by us in [13], namely, intra-spectral and inter-spectral features are used. The classification is done with the help of Support Tensor Machines [14]. The remote sensing images at different temporal resolutions are acquired and are classified accordingly. The classified images are the input to the next section.

B. Evaluative Study of Landscape Metrics

Generally, the change analysis is quantified through the "class area" difference that has occurred in each land use or land cover. The way to quantify the changes in structure in terms of shape, morphology, texture and position is less attempted. In order to quantify those changes, the support of landscape metrics is sought with. Due to the huge number of landscape metrics defined in the literature, there is always a confusion of what are the relevant ones. It is also necessary to identify the landscape metrics which help in understanding the change patterns that has occurred in the region.

Class Area (CA) and Landscape Area (LA) are the important landscape metrics and are also needed to find others. There are numerous patches which may belong to the same class, hence class area is relevant to understand the constitution of the landscape. So to gather an information on the number of patches that are present in a landscape, Mean Patch Size is also taken into consideration. Variation of the mean patch size will help to understand the growth/retardation of a particular class.

Different shape metrics are mean perimeter-area ratio, mean shape index and Area Weighted Mean Fractal Dimension (AWMFD). Literature points to the fact that the first kind of shape metric varies with respect to the size of the patches and does not provide information on the shape of the structure. Experiments from [15] show that mean shape index is not behaving uniformly on the same kind of changes that has happened in different spatial regions and hence cannot be assessed properly. Area weighted mean fractal dimension is a function of area and perimeter. It is computed over each patch and is average on each class in the landscape. Its value is dependent on the area of the patch and can be sought as a reasonable measure for finding shape complexity. An analogous measure to this is Contrast Weighted Edge Density (CWED) which standardizes edge to a per unit area. Due to this standardization scheme, a comparison of different landscape regions are possible in terms of CWED. To understand the shape better, to measure elongation or compactness of a region, Contiguity Index (CI) is also needed.

Aggregation metrics helps to assess a landscape with respect to aggregation or clunking of different patch types. The two metrics chosen for study are Contagion Index (CONTG) and Interspersion and Juxtaposition Index(IJI). Contagion Index is derived based on cell adjacencies, whereas Interspersion and Juxtaposition index is based on patch adjacency. In the latter index, the adjacency of each patch is evaluated with respect to all other patch types. From the studies of [15] it is clear that contagion index measures both interspersion and dispersion, and is trying to measure one value, keeping other constant and is thus not supportive for finding the same. Hence Interspersion and Juxtaposition index indicates only interspersion and is used for change analysis.

Diversity metrics indicate the richness and evenness of a landscape. These two factors are the structural indicators of a landscape. Richness indicates diversification and evenness indicates uniformity of classes in the landscape. Shannon's Diversity Index (SDI) is the most prominent diversity indicator. The absolute value of the same does not indicate anything, where the relative comparison at different temporal resolutions is a good indicator of how diversity changed or evolved. An evenness index called Shannon's Evenness Index (SEI) is also taken into consideration that aids in isolating the evenness component of diversity.

An extensive work in this regard by the same authors can be found in [15]. The detailed formulae of land scape metrics and the associated terms are given in [15].

C. Spatiotemporal Ontology

An initial attempt to study changes occurred in Indian cities is done in [16]. In [16], a quantitative index is computed for regions which indicate the growth of the region. To bridge the gap between conventional nominal values and semantic understanding of changes that has occurred in a region, a spatiotemporal ontology [3] is proposed by the same authors. The ontology helps to conceptualize the metrics affecting the changes in a region. The ontology models the different types of features associated with images and also the relevant landscape metrics, which aid in understanding the change patterns of a region. Hence with the aid of the ontology, the change analysis of a region can be performed on a semantic level. The semantic levels possible with the ontology proposed by us are the shape complexity of a region in terms of class labels, structural growth of a region and the level of interspersion between different classes in a region. Thus the ontology aids in performing the change analysis study in a more semantic manner within different time frames of our choice.

The spatial information of the region chosen for understanding change patterns has to be modeled in classes,

object properties, data properties and individuals of the ontology. The entire ontology is built under *SaptiotemporalEnity*. It is composed of subclasses like *SpatialEntity*, *TemporalEntity*, *ChangeModel*. The spatial information is demonstrated in *SpatialEntity*. The main sub concepts/classes of the spatial domain are *Regions*, *Labels* and *BoundingBox*. The *Regions* depict the spatial areas for the land under study. *Regions* can come under different *Labels*. The *Labels* class is further divided into appropriate labels like *Builtup*, *Baregrounds*, *Vegetation*, and *Water*. The regions boundaries are marked through *BoundingBox*. The **BoundingBox** identify the four corners of the region. The *Region* class has the most important data property as *hasLabel*, which indicates the label on the region. The class is also associated with the property called *RegionFeatures*, which is further split into *ColorFeatures*, *TextureFeatures*, *ShapeFeatures*, *Indices* and *Metrics*. *Indices* and *Metrics* are two new features introduced exclusively from the perspective of remote sensing images. Thus the features are a combination of conventional features along with landscape metrics and defined indices. The appropriate indices are chosen from [15]. Thus the data property Metrics can be written as

(hasClassArea, hasAWFractalDimension, hasCWEdgeDensity, hasContiguityIndex, hasIJIIndex, hasSDIIndex, hasSEIIndex, hasSplittingIndex).

The class *TemporalEntity* stores the temporal information. The *TimeOfInterest* is the subclass which is expanded as *TimeOfInterest (day, month, year)*. This entity denotes the acquisition time of the satellite image. The *Change Model* from the *SpatialEntity* and *TemporalEntity* is mapped to *SpatiotemporalEntity*. It is associated with low-level features such as *morphology, shape, position and texture*.

The ontology is also incorporated with SWRL rules and associated axioms as well as temporal entity to model the changes that has evolved. Details of SWRL rules and inferred axioms can be sought in [3]. The ontology is developed in Ontology Web Language –DL (OWL-DL) using Protégé. The ontology is hosted on

https://ontohub.org/repositories/spatiotemporal_ontology

Thus all the above mentioned landscape metrics will aid us to assess a landscape pattern and helps to understand the morphology, texture, shape and position of the region under consideration.

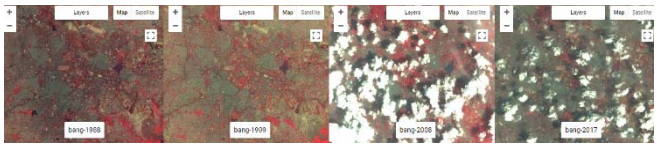
Metrics associated with morphology, texture, shape and position will help us to understand the pattern of the landscape at different temporal resolutions and provide a brief understanding of the same. For example, for understanding the shape complexity in two aspects namely (a) regular/irregular and (b) strip\planar, indices like Area Weighted Mean fractal dimension and Contiguity Index can be sought for. On closer inspection of the values of Area Weighted Mean fractal dimension, shapes with simple perimeter approach a value of 1 and a value of 2 when the shapes become more complex. Due to irregular land use, this parameter serves as one of the best for measuring raggedness of urban cities. The value of contiguity index helps to measure the elongation and compactness of a region. Likewise, the Interspersion and Juxtaposition index aids in understanding the texture composition of a region in terms of its different patches/classes.

IV. Results and Discussions

As mentioned in Section 2, the study area chosen are the two metropolitan cities of India (a) Mumbai and (b) Bangalore. Sample raw images of Mumbai city in 1988, 1999, 2008 and 2017 are given in Fig 2a. The images of Bangalore city on the same temporal domain is given in Fig 2b.

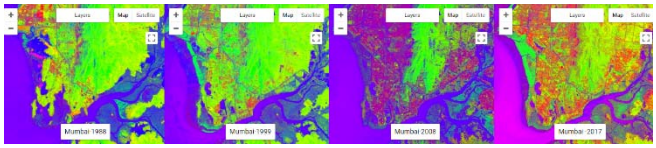
The first step in this study is to apply classification model to classify the remote sensing images. The second step is to run the spatiotemporal ontology on the classified images to observe the changes. The quantitative values resulting from the landscape metrics will aid in judging the landscape pattern evolution. Only the results of second phase are analyzed in detail, as the first phase is not the contribution in this work. Fig 3a and Fig 3b represents the classified images of Mumbai and Bangalore city respectively for the different temporal resolution. It is to be noted that the features used for classification using Support Tensor Machines are intra-spectral and inter-spectral features, which are the contribution of same authors.

(a)

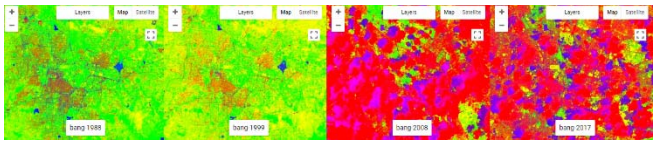


(b)

Fig 2. (a) Mumbai (b) Bangalore



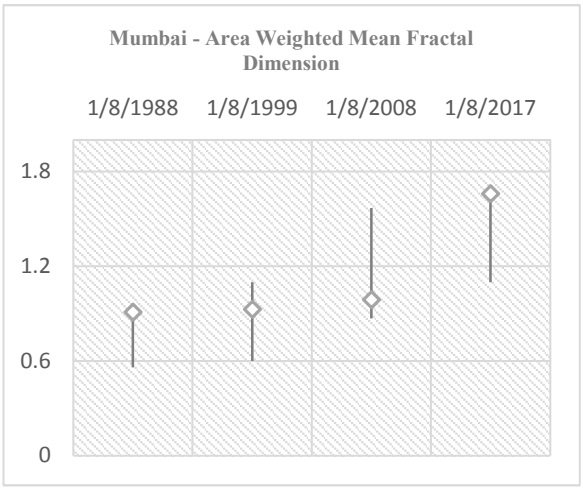
(a)



(b)

Fig 3. Classified Images of (a) Mumbai (b) Bangalore

The landscape metric, Area Weighted Mean Fractal Dimension, is an indicator of the shape complexity associated with the region. The range of values for this metric is from 1 to 2. When the values approaches 1, the shapes have simple perimeter and when the values approaches 2, the shapes become more complex. Fig 4 presents the shape complexity associated with Mumbai and Bangalore at different temporal resolutions. From the observations in the figure, each vertical line represents the highest and lowest value associated with that region. The marker in the vertical line is an indicator of weighted average value of the fractal dimension.



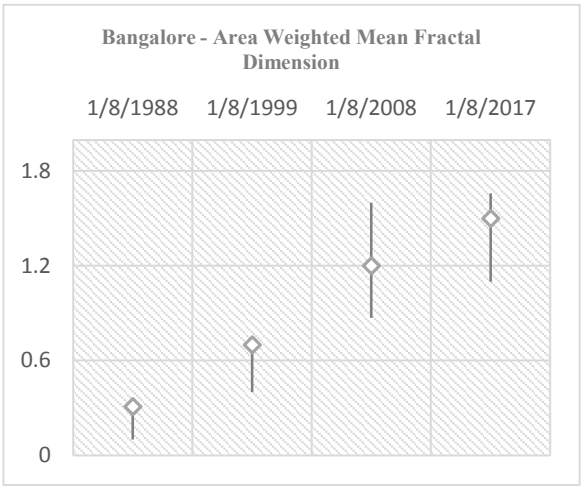
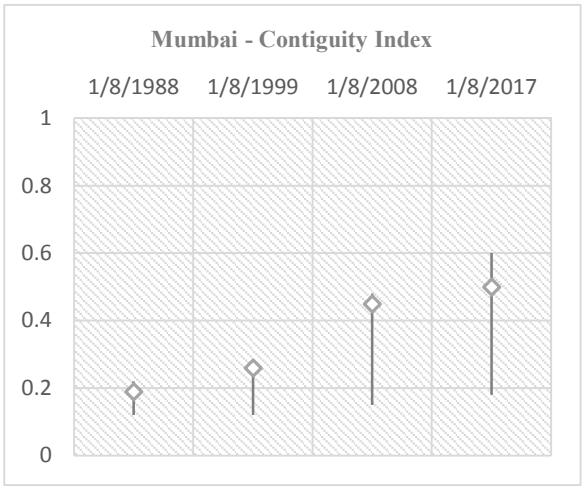


Fig 4 Area Weighted Fractal Dimension depicting shape complexity

The graph of Mumbai points to the following facts - (a) a drastic change in the shape complexity of the city is seen between 1999 -2008 , (b) the period between 1988 -1999 does not exhibit a radical change, only a minor variation in the fractal dimension values are noted , (c) the average value of 1999 – 2008 lies in the lower range of vertical line, thus indicating the presence of a large number of patches with simple shapes and (d) the change in the period 2008- 2017 has happened in almost all patches as the average value has risen to 1.8.

Similarly, an examination of Bangalore city points to the following facts - (a) the change happened between the years 1988- 1999 are appreciable as the average values in the two time ranges differ by 0.5, (b) almost all patches of Bangalore city has undergone changes in the period 1999 -2008, thus indicating a mid-value average in the vertical line.

The morphological pattern associated with the regions of the landscape is given by the landscape metric called contiguity index. It measures the elongation or compactness of a region thus naming it as either strip or planar. The range of values are from 0 to 1. As the regions elongate, the value approaches zero and as it become more compact the value approaches 1. Fig 5 depicts the study of contiguity index for the two metropolitan cities. The two plots almost behave similarly, with the range of values in the same time slots is almost same, but with different averages. This leads to the conclusion that the way the structure of regions in two metropolitan cities has evolved under the considered temporal resolutions is almost same.



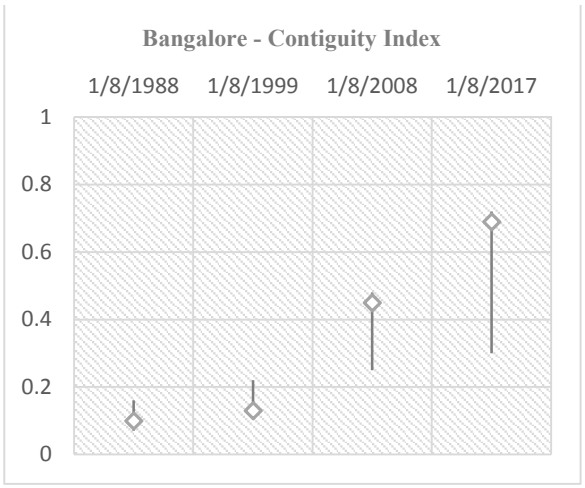


Fig 5. Contiguity Index depicting morphological feature

To understand the texture composition of the region and also to understand the patch dispersion as adjacent or disjoint, the landscape metric Interspersion and Juxtaposition index comes into play. This is a relative index which indicates the level of interspersion as a percentage of maximum possible value for a given number of patch types. So the value is dependent on the number of patch types. The range of values is from 0 to 100. Higher value indicates a composition of smooth and adjacent same patch types and as the value goes down the texture become rough and patches will be disjoint. Fig 6 shows the Interspersion and Juxtaposition Index values for Mumbai and Bangalore cities. On analysis of the indices, it is seen that for Mumbai city, considering the entire temporal resolution, in 2017, there are equal number of rough and smooth as indicated by the average value. It has to be assumed that the urban structure contributing to this factor is equally spread in the region along with non-urban lands. Looking into Bangalore city, the average value for the given number of patches is low in the first two temporal domains, and high for the second two temporal resolutions. This leads to the

conclusion that the urban growth which has happened in 2008 -2017 in Bangalore has been reflected in almost all the patches.

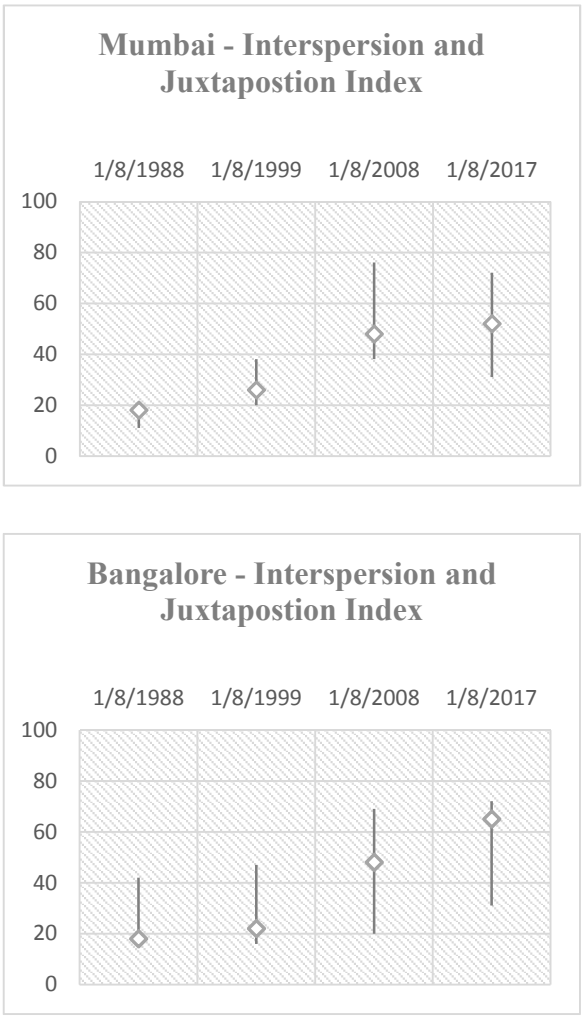


Fig 6. Intersppersion and Juxtapostion Index depicting texture analysis

The results are validated with [17] and [18], which are authorized results from Government of India. The analysis of this work is done Google Earth Engine Python API.

The interpretation of the study can be summarized as- (a) Mumbai and Bangalore city has grown tremendously in the period 1999 -2008 as evident from the landscape metrics given by the ontology. (b) most of the regions in Bangalore city has moved to urban land, whereas in Mumbai, the development as urban area in certain regions are more acute than others. (c) the shape patterns has grown more complex in the cities during the interval 2008 -2017, indicating urban growth at the border lands and (d) the morphological structure of the regions in both the cities show synonymous behavior towards an elongated structure.

V. Conclusion

This work presents a case study of two Indian metropolitan cities, Bangalore and Mumbai, to observe the spatiotemporal change patterns that has evolved in the temporal domain. The spatiotemporal ontology proposed with the support of features, indices and landscape metric and aided by SWRL rules and axioms is used to quantify the change patterns. The change patterns were quantified in terms of landscape metrics and is semantically mapped as well. Thus the study on a temporal range has helped to infer spatiotemporal semantics of the metropolitan cities of India, namely, Mumbai and Bangalore. The ontology can also be extended to support sudden and dynamic events occurring in a region. The case studies can be further expanded to study all metropolitan cities and will also help to identify cities whose growth are much in par with metropolitan ones.

References

[1] Radke, Richard J., et al. "Image change detection algorithms: a systematic survey." IEEE transactions on image processing14.3 (2005): 294-307.

[2] Bovolo, Francesca, and Lorenzo Bruzzone. "The time variable in data fusion: A change detection perspective." IEEE Geoscience and Remote Sensing Magazine 3.3 (2015): 8-26.

[3] Saritha, S., and G. Santhosh Kumar. "Spatiotemporal Ontology for Understanding Semantics in Change Patterns of Remote Sensing Images." In *International Conference on Innovative Computing and Communications*, pp. 307-313. Springer, Singapore, 2019.

[4] Bechhofer, Sean. "OWL: Web ontology language." Encyclopedia of database systems. Springer US, 2009. 2008-2009.

[5] Yuan, Fei, et al. "Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing." Remote sensing of Environment 98.2-3 (2005): 317-328.

[6] Huang, Jingnan, Xi X. Lu, and Jefferey M. Sellers. "A global comparative analysis of urban form: Applying spatial metrics and remote sensing." Landscape and urban planning 82.4 (2007): 184-197.

[7] Schneider, Annemarie, and Curtis E. Woodcock. "Compact, dispersed, fragmented, extensive? A comparison of urban growth in twenty-five global cities using remotely sensed data, pattern metrics and census information." Urban Studies 45.3 (2008): 659-692.

[8] McGarigal, Kevin, and Barbara J. Marks. "FRAGSTATS: spatial pattern analysis program for quantifying landscape structure." Gen. Tech. Rep. PNW-GTR-351. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station. 122 p 351 (1995).

[9] Tewkesbury, Andrew P., Alexis J. Comber, Nicholas J. Tate, Alistair Lamb, and Peter F. Fisher. "A critical synthesis of remotely sensed optical image change detection techniques." Remote Sensing of Environment 160 (2015): 1-14.

[10] Alphan, Hakan, Hakan Doygun, and Yüksel I. Unlukaplan. "Post-classification comparison of land cover using multitemporal Landsat and ASTER imagery: the case of Kahramanmaraş, Turkey." Environmental monitoring and assessment 151, no. 1 (2009): 327-336.

[11] Olofsson, Pontus, Giles M. Foody, Stephen V. Stehman, and Curtis E. Woodcock. "Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation." Remote Sensing of Environment 129 (2013): 122-131.

[12] https://landsat.usgs.gov/ - Accessed on December 2017.

[13] Saritha, S., and G. Santhosh Kumar. "Inter-Spectral and Intra-Spectral Features for Effective Classification of Remotely Sensed Images." Procedia Computer Science 115 (2017): 549-555.

[14] Hao, Zhifeng, et al. "A linear support higher-order tensor machine for classification." IEEE Transactions on Image Processing 22.7 (2013): 2911-2920.

[15] Saritha, S., and G. Santhosh Kumar. "Analysis of the smart growth of kochi city through landscape metrics." IEEE Region 10 Symposium (TENSYMP), 2017. IEEE, 2017.

[16] Saritha, S., and G. Santhosh Kumar. "Change detection in urban landscapes: a tensor factorization approach." *Spatial Information Research*: 1-14.

[17] https://www.rvo.nl/sites/default/files/Smart%20Cities%20India.pdf. Retrieved February 26, 2018.

[18] http://smartcities.gov.in/content/innerpage/no-of-smart-cities-in-each-state.php. Retrieved August 18, 2017.

Q-value Learning Automata (QvLA)-RACH Access Scheme for Cellular M2M Communications

Nasir A. Shinkafi
Communication Research Group, Department of Electrical Engineering Bayero University, Kano
P.M.B. 2011, Kano, Nigeria
nasir.shinkafi@umyu.edu.ng

Lawal Muhammad Bello
Communication Research Group, Department of Electrical Engineering Bayero University, Kano
P.M.B. 2011, Kano, Nigeria
lmbello.ele@buk.edu.ng

Dahiru Sani Shu'aibu
Communication Research Group, Department of Electrical Engineering Bayero University, Kano
P.M.B. 2011, Kano, Nigeria
dsshuaibu.ele@buk.edu.ng

Ibrahim Saidu
Department of Information and Communication Technology, Usmanu Danfodio University, Sokoto, P.M.B
2346, Sokoto, Nigeria
ibrahim.saidu@udusok.edu.ng

***Abstract*— This paper introduces a QvLA-RACH access scheme to reduce resource wastage by enhancing slot utilization in cellular M2M communications. The proposed QvLA-RACH scheme employs a Q-value update technique to reduce idle slots by controlling the probability of collision. the scheme also uses a cooperative Q-Learning strategy to update the Q-value. The performance of the proposed scheme is evaluated compared to the existing scheme using extensive simulation. The results show that the proposed QvLA-RACH scheme achieves better performance in terms of throughput and access delay by 49.7% and 18%, respectively.**

Keywords— Machine-to-Machine, Access Barring, LTE Network, RACH congestion, ALOHA protocol, Q-learning, Q-value, Quality of Service, Learning Automata.

I. INTRODUCTION

With the rapid evolution of telecommunication technologies and the proliferation of portable devices such as laptops, smartphones, tablets, PDAs, etc. have led to the continuous growth of service coverage and capacity as well as emergence of various new applications [1]. Notable amongst the new area of applications is Machine to Machine (M2M) communications. M2M communications involve fully automated interactions between users without human intervention. There are a wide range of M2M applications including e-Health monitoring, smart homes, smart cities, smart metering etc [2]. To effectively deploy this technology ubiquitously, cellular networks are considered as a suitable infrastructure for the deployment of M2M communications [3]. However, cellular networks are mainly designed for Human-to-Human (H2H) communications, which requires specific basic service requirements such as high data rates; and is largely characterized by a higher proportion of downlink (DL) traffic compared to the M2M communication which has mainly small data payloads and mostly uplink traffic (UL) [3-5]. Furthermore, M2M is characterised by a massive number of users which is projected to be higher than the number of H2H users, as forecasted in [6]. These features of M2M have been identified as the reason for the congestion problem on the Random Access Channel (RACH) of cellular networks [5, 7]. The RACH congestion or overload challenge is caused by the simultaneous and pervasive access attempts by the M2M users when accessing the cellular networks, resulting in resource wastage and poor Quality of Service (QoS) [2, 5, 7, 8]. Therefore, efficient RACH access techniques are essential to overcome these challenges and guarantee the realization of cellular M2M communication by minimising the resource wastage and improve QoS performance.

Various RACH access schemes have been proposed to address these problems [2, 3, 5, 7-11]. Some of the solutions proposed by the Third Generation Partnership Project (3GPP) include Physical Random Access Channel (PRACH) resource separation, Extended Access Barring (EAB), slotted access and dynamic resource allocation schemes [7]. Recent literature includes the Reinforcement Learning (RL) techniques such as Q-learning (QL) and Learning Automata (LA) based RACH access schemes [8, 11]. Amongst these access schemes, some focus on minimising collisions to improve resource utilisation [3, 9], others such as QL-RACH use RL to enhance throughput and minimise access delay using an intelligent slot assignment mechanism [8], and Frame-Based QL-RACH (FB-QL-RACH) provides a separate frame for H2H back-off [10]. In [11], a LA-QL-RACH uses LA to improve the RACH scheme by categorising M2M into three (3) QoS classes: High (H), Medium (M) and Low (L). The categorisation improves the RACH throughput performance by minimizing the level of interaction and collision of M2M with H2H users. However, the scheme induces a Q-learning punishment regulating technique using the steady state LA feedback which produces suboptimal Q-values that increase the chances of producing idle slots.

In this paper, a Q-value Learning Automata QL-RACH (QvLA-RACH) scheme is proposed. The scheme introduces a Q-value update integration technique to minimize resource wastage by reducing idle slots and controlling the probability of collision. Also, the technique employs a cooperative Q-Learning strategy to update the Q-value adopted from [12]. The strategy improves the accuracy and speed of the LA. The performance of the proposed QvLA-RACH is evaluated compared to the existing scheme via simulation. The results illustrate that the proposed scheme achieves superior performance compared to the LA-QL-RACH scheme in terms of an increase in throughput and decrease in access delay.

The rest of the paper is organized as follows; Section II presents an overview of the LA-QL-RACH scheme and provides the details of the proposed QvLA-RACH scheme. Section III presents the performance evaluation and the paper is concluded in section IV.

II. LA-QL-RACH AND QVLA-RACH SCHEME

A. LA-QL-RACH Scheme

The LA-QL-RACH employs a LA technique adopted from [11] to classify M2M users into x QoS classes: High (H), Medium (M), and Low (L) where $x \in \{H, M, L\}$. In each class x, the scheme monitors the status of the preambles and whenever any preamble experiences collision it returns a probability $\rho_x^{coll}(t)$ per LA cycle. At the end of each cycle, eNode B (eNB) monitors the value of ρ_x^{coll} for QoS class x and generates LA feedback, $r_x(t)$ by comparing it with the expected value of v as in (1). Full details of the LA-based RACH technique is provided in section 4 of [11].

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

$$r_x(t) = \begin{cases} 0 & if\ \rho_x^{coll}(t) < v \\ 1 & if\ \rho_x^{coll}(t) \geq v \end{cases} \quad (1)$$

and;

$$v = 1 - 2e^{-1} \quad (2)$$

where *v* is the expected value of the probability of collision for a given priority class which refers to the convergence of a preamble in the collided state compared with the maximum throughput achieved through the s-Aloha of $2e^{-1}$[11]

When the scheme is stable;

$$\rho_x^{coll}(t) = v \quad (3)$$

and RACH allocation probability ($\varphi_x(t)$) is updated in a reinforcement manner as in (4).

$$\varphi_x(t+1) = \begin{cases} \varphi_x(t) + L & if\ r_x(\mathrm{t}) = 1 \\ \varphi_x(t) - L & if\ r_x(\mathrm{t}) = 0 \end{cases} \quad (4)$$

where L is the LA update variable.

At this stage, the penalty factor (*R*) in QL-RACH as presented in (5) by [8];

$$Q' = (1 - \gamma)Q + \gamma R \quad (5)$$

where Q' is the new Q-value and γ is the learning rate;

is controlled by the outcome of the LA feedback to produce the new Q-value according to (6);

$$R = \begin{cases} +1 & if\ r_x(\mathrm{t}) = 0 \\ -1 & if\ r_x(\mathrm{t}) = 1 \end{cases} \quad (6)$$

The scheme provides a significant improvement in the RACH throughput and repetitive collisions experienced in the dedicated M2M slots are avoided. However, when the probability of collision increases due to increase in the demand for resources from a given class, unit feedback, $r_x(t)$ with value of 1, is produced. The unit feedback as in (6) leads to penalising the affected M2M slot by lowering its Q-value by one. However, at steady state, when every M2M has acquired a slot, the slots that had not suffered collision will have higher Q-values than those that had experienced collision. Since subsequent resource allocation requests are directed to slots with higher Q-values, those with lower Q-values continue to remain idle and hence unutilised, leading to resource wastage. Therefore, at this stage, the unit feedback is ineffective in regulating the *R*, which consequently produces sub-optimal Q-values and hence increases the chances of producing idle slots.

B. Proposed QvLA-RACH Scheme

To address the challenges identified in LA-QL-RACH, a QvLA-RACH scheme is proposed to enhance slot utilization and eliminate resource wastage. The scheme employs a Q-value update integration technique to address the problem of suboptimal Q-value by reducing the chances of producing idle slots. The technique is triggered according to the LA feedback such that the RACH resource allocation probability ($\varphi_x(t)$) increases with probability ρ_x^{coll} and decreases with probability $1 - \rho_x^{coll}(t)$ to produce a corresponding Q-value in each case. The new Q-value (Q'), which is collision dependent is updated according to [12] as shown in (7);

$$Q' = \sum_{x\in\{H,M,L\}} \zeta_x \times Q \quad (7)$$

where $\zeta_{x\in\{H,M,L\}}$ is a weighting parameter for a given priority class that is generated whenever collision occurs and is derived as presented in (8);

$$\zeta_{x\in\{H,M,L\}} = \frac{\rho_x^{coll}(t)}{\sum_{i=1}^{N} \rho_x^{coll}(t)} \quad (8)$$

where N is the total number of M2M users.

The updated strategy in (7) produces a new Q-value to replace (5) for the QL-RACH scheme. The proposed QvLA-RACH scheme is implemented using the following algorithm 1.0;

Algorithm 1.0 QvLA-RACH algorithm implementation on LA-QL-RACH for collided M2M users during RACH Contest; M2M – Machine to Machine; RACH – Random Access Channel; LA – Learning Automata; QL – Q-Learning; QvLA – Q-value Learning Automata

1: **for** every M2M user employing LA-QL-RACH to contend RACH resources **do**
2: Calculate probability of collision ($\mathrm{P}_x^{coll}(t)$) and compare it with the expected value v
3: Generate LA feedback ($\mathrm{r}_x(\mathrm{t})$) value of 0 or 1 according Step 2
4: Record Q from (5)
5: **end for**
6: Calculate $\varphi_x(t)$ according to (4)
7: **If** probability of collision is less than the expected value **then**
8: Decrement $\varphi_x(t)$
9: **else if** probability of collision is higher or equal to the expected value, **then**
10: Increment $\varphi_x(t)$
11: **end else if**
12: Calculate $\zeta_{x\in\{H,M,L\}}$ according (8)
13: Calculate Q' according (7)
14: Go to Step 2
15: **end if**

III. PERFORMANCE EVALUATION

A. Simulation Scenario

Simulation has been used to assess the performance of the proposed QvLA-RACH scheme compared to the LA-QL-RACH using Matlab Simulator. As presented in Table 1, one RA slot is assumed to occur per unit LA cycle and fifty preambles are reserved in each RA slot for use by the M2M users. The users are also assumed to be within an eNB coverage area in a single cell of an LTE network. The table presents details of the parameters used in this simulation based on LTE standard.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
PRACH configuration index	12
RA-slot period	1ms, 1 cycle
1 RA-slot	50 preambles
Preamble format duration	1ms
Back-off period / AC-Barring Time	28ms
Number of allowed retransmissions	7
RACH allocation probability ($\varphi_h, \varphi_m, \varphi_l$)	0.5, 0.3, 0.2
Learning Rate	0.01

As shown in the table, a PRACH configuration index of 12 has been selected to decide the PRACH preamble timing and PRACH preamble type as well as determines when the user should transmit RACH. The index also indicates which frame and sub-frame the M2M user is allowed to transmit a PRACH preamble. For example, the index of 12 means the user is allowed to transmit RACH at subframe number 0,2,4,6,8 of any System Frame Number. Each frame (10ms) consists of 10 sub-fames of 1ms each. Each sub-frame has 2 slots of 0.5ms. Also, ac-Barring time of 28ms and learning rate of 0.01 have been considered as the Back-off period for retransmission and learning convergence respectively. The values of the RACH allocation probability (φ_x) have been selected as the ratio of the pre-allocated preambles per class x for use by all the M2M users. Additionally, the method used by [8] has been adopted in presenting our results. The method used the s-ALOHA throughput capacity (e^{-1}) in Erlangs (E) at both the upper and lower limits as a threshold for traffic prediction. An Erlang is a unit of traffic density in a telecommunications system such that one Erlang is the equivalent of one continuous call in a specific channel over a fraction of time that the channel was in use either for offered traffic or useful throughput. An upper limit of 0.3 E is selected as it is closer to the critical point whereas a lower limit of 0.1 E is chosen as it is away from the critical point. These limits are assumed to be the average peak hour load which H2H could generate during its interaction with M2M and are used as a measure of RACH stability.

B. Simulation Results and Discussion

In this section, we evaluate the performance of the proposed QvLA-RACH scheme against the existing LA-QL-RACH scheme in terms of RACH throughput and access delay for the three QoS classes. Fig. 1 illustrates a comparison of the throughput performances of the three (3) QoS classes (H, M, L) for the proposed QvLA-RACH and LA-QL-RACH schemes when the total generated traffic is above the s-ALOHA capacity. As shown in the figure, above the s-ALOHA capacity signifies that the overall total load (generated traffic) is a combination of the fixed H2H traffic load (0.3 E) with a variable but additional M2M traffic load.

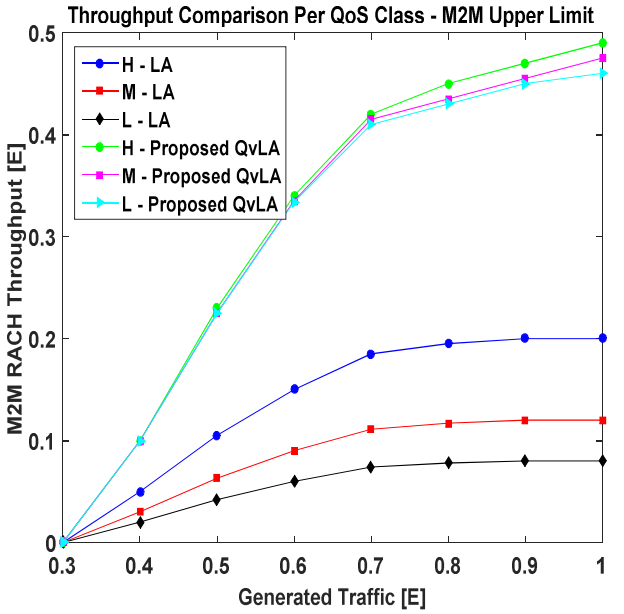


Fig. 1. Throughput Comparison per priority class - M2M Upper Limit (H2H=0.3E)

The Fig. 1 depicts the behaviors of the schemes showing impact of the Q-value update integration technique on resource allocation. It also shows that the performance of each QoS class with the Q-value update is higher compared to the existing scheme. For instance, when the generated traffic is 1.0 E, the throughput for the H QoS class for the LA-QL-RACH is 20%, it is 49% for the same class for the QvLA-RACH. The 29% improvement is due to the reduction of idle slots by the Q-value update integration technique. Furthermore, this shows that in the QvLA-RACH scheme all available resources within a given LA cycle are utilised and that each QoS class takes resources according to its demand.

Fig. 2 shows a comparison of the throughput performance of the three (3) QoS classes for the QvLA-RACH and LA-QL-RACH scheme when the generated traffic is below the s-ALOHA capacity. Below the s-ALOHA capacity, the H2H traffic is fixed at a lower limit of 0.1 E. At 1.0 E, when the throughput for the H QoS class for the LA-QL-RACH scheme is 36%, it is 85.7% for the same QoS class belonging to the QvLA-RACH scheme. The improvement of 49.7% between the QvLA-RACH and LA-QL-RACH at 1.0 E is realised for the same reasons observed in Fig. 1. However, a proportionate distribution of resources to the three QoS classes is observed in the two schemes showing there is a negligible impact on the distribution pattern, rather an enhancement in the slot utilisation has been recorded.

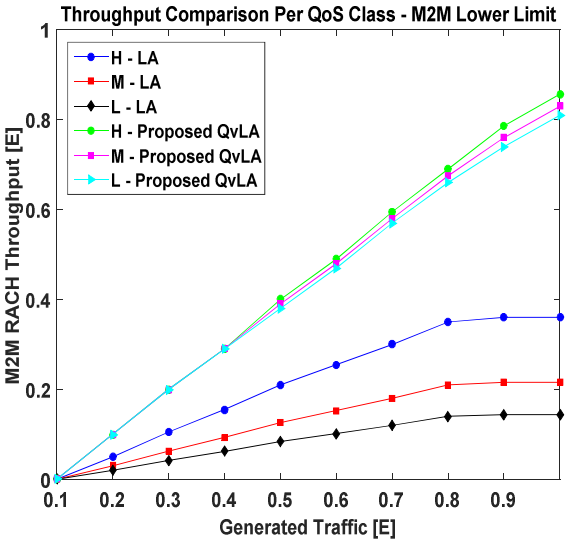


Fig. 2. Throughput Comparison per priority class - M2M Lower Limit (H2H=0.1E)

The overall improvement in the throughput performance of the QvLA-RACH scheme is due to the Q-value integration update technique applied on the LA-QL-RACH scheme to enhance slot utilization.

Fig. 3 presents the average access delay against the number of M2M users for the three (3) QoS classes within 200 LA cycles compared with the existing scheme. The figure shows a reduction in the average access delay of the QvLA-RACH scheme when compared with the LA-QL-RACH scheme. In the new scheme, all the M2M users from the three (3) QoS classes appear to experience lower average access delay with the increase in the number of access requests per class. For example, when the RACH access request is 15,000, the QvLA-RACH QoS classes have lower access delays of 18%, 7% and 8% for the H, M and L classes, respectively as compared to the LA-QL-RACH.

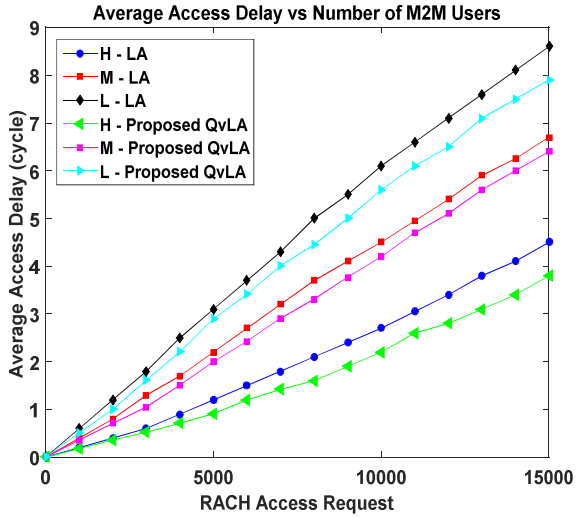


Fig. 3. The average access delay vs the number of M2M users for three QoS classes

Therefore, it can be shown that the average access delay for each priority class has reduced, since the percentage of resources allocated to that class depends on the demand of that class and the idle slots are reduced.

IV. Conclusion

In this paper, a QvLA-RACH scheme is proposed to provide efficient resource utilization in the previous scheme (LA-QL-RACH) by minimising the chances of producing idle slots. The scheme employs a Q-value update integration technique to reduce idle slots using LA feedback to produce optimal Q-values according to the probability of collision. Simulation is used to evaluate the performance of the proposed QvLA-RACH scheme against the compared scheme. The results show that the scheme improves the RACH throughput by 49.7% and reduces the access delay by 18%. Overall, the proposed QvLA-RACH scheme performs better than the existing LA-QL-RACH scheme in terms of throughput and access delay for the three QoS classes.

References

[1] Rapeepat Ratasuk, Athul Prasad, Zexian Li, Amitava Ghosh, and Mikko A. Uusitalo, "Recent Advancements in M2M Communications in 4G Networks and Evolution Towards 5G"*IEEE Intelligence in Next Generation Networks (ICIN)*, pp.52-57, 2015.

[2] S. K. Tan, M. Sooriyabandara, and Z. Fan, "M2M communications in the smart grid: Applications, standards, enabling technologies, and research challenges," *International Journal of Digital Multimedia Broadcasting*, 2011.

[3] Chang-Yeong Oh, Duckdong Hwang, and Tae-Jin Lee, "Joint Access Control and Resource Allocation for Concurrent and Massive Access of M2M Devices" *Wireless Communications, IEEE,* pp. 4182 - 4192, 2015.

[4] Cisco San Jose. Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019 technical report, CA, USA; 2015.

[5] Biral A, Centenaro M, Zanella A, Vangelista L, Zorzi M. "The challenges of M2M massive access in wireless cellular networks". *DigitCommun Netw.* 2015; 1(1):1-19.

[6] ETSI, (2011) "Standards on Machine to Machine Communications", Mobile world congress, Barcelona, Feb. 2011.

[7] Monowar Hasan, and Ekram Hossain, "Random Access for Machine-to-Machine Communication in LTE-Advanced Networks: Issues and Approaches" *IEEE Communications Magazine, vol. 51,* June, pp. 86 – 93, 2013.

[8] L. M. Bello, P. Mitchell, and D. Grace, "Application of Q-Learning for RACH Access to Support M2M Traffic over a Cellular Network," *in European Wireless Conference; Proceedings*, pp. 1-6, 2014.

[9] Ningbo Zhang, Guixia Kang, Jing Wang, Yanyan Guo, and Fabrice Labeau, "Resource Allocation in a New Random Access for M2M Communications" *Communications Letters, IEEE, vol. 19,* May, pp. 83-846, 2015.

[10] L. M. Bello, P. Mitchell, and D. Grace, "Frame based back-off for Q-learning RACH access in LTE networks," in *Telecommunication Networks and Applications Conference(ATNAC), 2014 Australasian,* pp. 176-181, 2014.

[11] Morvari F, Ghasemi A. "Priority-based adaptive access barring for M2M Communications in LTE networks using learning automata". *Int J Commun Syst.* 2017; e3325. https://doi.org/10.1002/dac.3325

[12] Mao Y, Yantao T, Xinyue Qi, "Cooperative Q-Learning Based on Learning Automata", in *Proceedings of the IEEE International Conference on Automation and Logistics Shenyang, China*, August 2009, pp. 1973 – 1978, 2009.

Application of Adaptive Neuro-Fuzzy Inference System for the prediction of Early Age Strength of High Performance Concrete

Deepak Kumar Sinha
Dept. of civil engineering
Dr. BR Ambedkar National Institute Of Technology
Jalandhar, India
deepaksinha.official@gmail.com

S. Rupali
Dept. of civil engineering
Dr. BR Ambedkar National Institute Of Technology
Jalandhar, India
satavalekarr@nitj.ac.in

Shailja Bawa
Dept. of civil engineering
Dr. BR Ambedkar National Institute Of Technology
Jalandhar, India
bawas@nitj.ac.in

***Abstract*— This paper deals with the design of a model based on Adaptive Neural Fuzzy Inference System (ANFIS) for the prediction of early age (3 days) compressive strength of concrete. The model is generated by a dataset having 8 parameters. these are converted into 7 inputs viz. Cement, Flyash, BFS, water, Superplasticizer, Coarse aggregates, and fine aggregates and 1 output i.e. 3 days compressive strength. The model was trained and tested using hybrid method of learning. The results produced Training and checking errors as 0.153MPa and 1.212MPa respectively making ANFIS very much appropriate for this purpose.**

Keywords*— *Artificial Intelligence, ANFIS, Fuzzy Logic, Concrete Mix Design, HPC, ANN, Fuzzy Sets.

I. INTRODUCTION

Concrete is broadly used construction material because of its ease of production and placement. It is a heterogeneous material, which is composited of aggregates and Binder. Basic ingredients of concrete are cement or lime as the binder, water, coarse aggregate and fine aggregates [1]. Concrete gets its strength by chemical reactions between cement and water. the reactions make the concrete stronger with time under normal conditions [2]. Thus, the compressive strength of concrete can be said to be a function of various raw materials of the mix and time. Early age strength (less than 3days) of concrete affects the long term strength and performance of concrete[3]. Therefore, an understanding of the same is essential for ensuring adequate durability and safety of the structure.

In present times, the use of High Performance Concrete (HPC)[4]has been quite popularized. It is a kind of concrete which possesses any of the following properties: high compressive strength, high elastic modulus, fire resistance, sulfate resistance, high early strength, self-compacting, high density, etc. To achieve such concrete various types of admixtures or additives are used in concrete[5]. These can be chemical or mineral admixtures. Concrete containing Flyash[6] and GGBS[7] improves the longtime strength of concrete [8]. It also helps in reducing carbon footprint[9] in construction by reducing the amount of cement used.

Concrete Mix design simply means proportioning the various components to achieve a desired strength and workability[10]. Even though concrete's production is easy, its mix design is a complicated process. It requires tedious calculations and experimentations based on empirical relationships and trial and error in the presence of expert supervision.

With the availability of new cementitious materials, such as flyash and Blast Furnace Slag [7], the number of components in a concrete mix has increased recently. Hence the number of properties to be adjusted has also increased making the empirical methods insufficient. We, therefore, need a system that is more natural as well as scientific. Artificial Intelligence[11] can play a crucial role here. An important step of mix design is predicting the compressive strength of the concrete.

There are various types of artificial intelligence approaches. Some of them are Artificial Neural Network[12], Fuzzy logic[13], Genetic Algorithm, ANFIS, etc.

Neural networks are similar to human neurons and they have a similar capability of learning using a dataset. they are good at linear and non-linear curve fitting or regression analysis, but don't say anything about the relationship between the variables in a reasonable way that we can understand. Fuzzy logic, on the other hand, resembles the human way of thinking and reasoning. It involves the creation of various rules involving membership functions that can be easily understood to predict the output for a given set of inputs. However, it does not have learning capabilities. ANFIS utilizes the reasoning capability of fuzzy logic and learning capability of ANN, hence it is the best tool available to predict random outcomes like concrete compressive strength.

These tools have recently been applied in various fields of civil engineering. Natraja et al. [14] presented the development of standard concrete mix modeling using fuzzy inference systems in 5 layers. They used backpropagation for training and achieved satisfying results as compared to conventional methods. The models produced an error of about 5%. Chopra et al.[15] found out the proper Neural Network architecture for concrete and compared it with Random Forest, and Decision Tree model concluding that NN outperforms others in this case of study. RMSE of DT, RF, and NN were respectively 3.9, 2.6, and 0.8 MPa respectively. Akhund et al. [16] applied a Fuzzy Expert System in various problems in civil engineering and found it

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

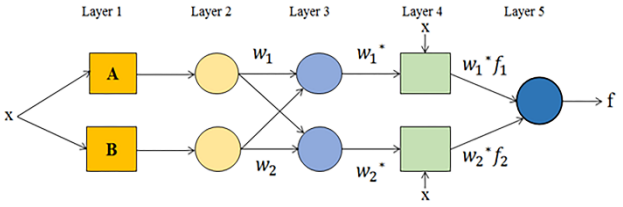


Fig. 1. Architecture of ANFIS

to perform suitably. Thus, Machine learning is found to be useful in civil engineering applications.

This study presents the application of ANFIS for the prediction of concrete compressive strength and checks for its suitability in this case. The aim of the study was to achieve an RMSE below 5mpa, which is the average error achieved using conventional mix design.

II. Architecture of ANFIS

Adaptive Neuro-Fuzzy Inference System (ANFIS)[17] is composed of five layers and is based on a feed-forward neural network (see Fig. 1). The layers are as follows: fuzzifying layer, ruleset layer, normalizing layer, de-fuzzifying layer, and Single Summation neuron.

The advantages of ANFIS can be listed as :

- It uses the ANN's learning ability to find a pattern
- It develops a Fuzzy Expert System based on the training in the form of a FIS file which helps in understanding the pattern between the data.

III. Datasets

Concrete compressive strength dataset by Yeh[18] is used in this study. Out of 1030 samples of mix design, 134 samples corresponding to 3 days (early age) compressive strength of concrete are selected. From this dataset, 3 datasets are produced 80% for training, 10% for checking/testing and 10% for validation. The datasets have 7 input variables viz. Cement, Blast Furnace Slag, FlyAsh, water, superplasticizer, Coarse Aggregates, and Fine Aggregates and 1 output variable i.e. compressive strength. Table I shows the range of data being used.

The inputs are now expressed as a ratio of 1st variables that is cement to reduce the computational load (See Fig. 2). This

S.No.	cement	BFS/cement	Flyash/cement	Water/cement	SP/cement	C.A/cement	F.A/cement	Comp. strength
1	139.60	1.50	0	1.37536	0	7.5	5.78008	8.06342
2	349.00	0.00	0	0.55014	0	3	2.31203	15.0491
3	198.60	0.67	0	0.96677	0	4.92648	4.15659	9.13142
4	310.00	0.00	0	0.61936	0	3.13225	2.74387	9.86640
5	374.00	0.51	0	0.45481	0.02700	2.47620	2.02326	34.3979
6	313.30	0.84	0	0.56016	0.02745	3.34152	1.95276	28.7994
7	425.00	0.25	0	0.36117	0.03882	2.00494	2.08729	33.3982
8	425.00	0.25	0	0.35623	0.04376	2.20235	1.89105	36.3009
9	375.00	0.25	0	0.3376	0.0624	2.27226	2.64693	28.9993
-	-	-	-	-	-	-	-	-
134	480	0	0	0.4	0	1.95	1.50208	24.3936

Fig. 2. Inputs converted to ratio of 1st variable (cement)

helps in increasing the accuracy of the model as well as reducing the workload of the computer.

IV. ANFIS Model Designing

The model was created using Neuro-Fuzzy Designer in MATLAB 2018. At first training dataset is loaded, then checking dataset is loaded by clicking load data (See Fig. 3). Now, the initial FIS is generated. There are two methods in MATLAB ANFIS toolbox for generating the initial FIS viz. *Grid-partition* and *subtractive-clustering*. With the growing number of input variables, grid partition FIS leads to a paralyzed calculation system. By using Subtractive Clustering, it is easy to build a ruleset model without much computational power. In subtractive clustering, there are 4 parameters (See Fig. 4).

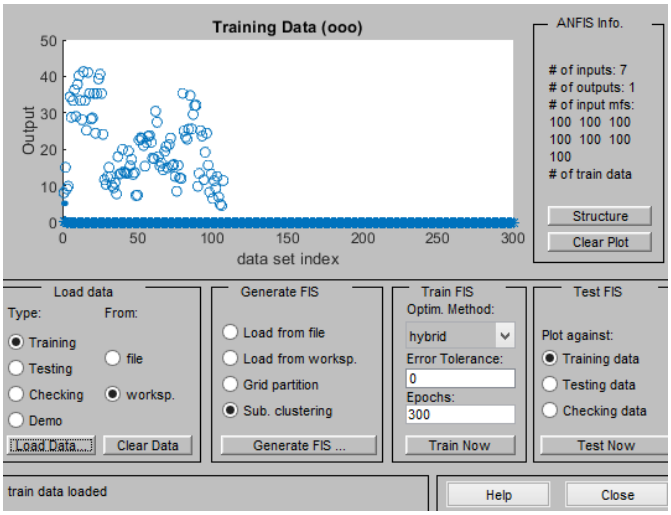


Fig. 3. ANFIS design process

TABLE I. Range of Variables Used In Dataset

	Cement (kg/m3)	Blast Furnace Slag (kg/m3)	Flyash (kg/m3)	Water (kg/m3)	Superplasticizer (kg/m3)	Coarse Aggregates (kg/m3)	Fine Aggregates (kg/m3)	3-Day Compressive Strength (MPa)
Min	102	0	0	121.75	0	822	605	4.56
Max	531	282.8	174.74	214.6	32.2	1134.5	992.6	41.3

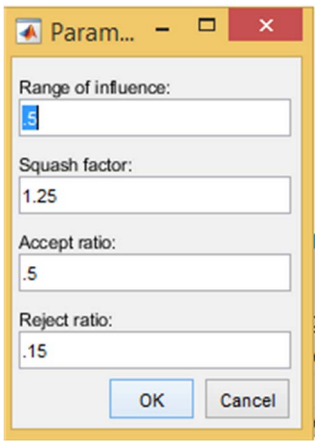


Fig. 4. Subtractive Clustering Parameters

There was not much information about these parameters. It is though known that Range of influence determines the radius of clusters, hence lower its value, more is the number of cluster points, hence more is the accuracy of the model. To find out the most optimum value for the parameters, we conducted Orthogonal Array Testing(OAT)[19]. Table II shows the results of OAT.

From the trials, it is concluded that by reducing the range of influence, accuracy increase. On this basis, the final model is as follow: Range of influence=0.02, squash factor=0.25,Accept ratio=0.5, and Reject ratio= 0.15.

ANFIS info:

Number of nodes: 1642

Number of linear parameters: 816

Number of nonlinear parameters: 1428

Total number of parameters: 2244

Number of training data pairs: 108

Number of checking data pairs: 14

Number of fuzzy rules: 102

The ANFIS structure contains 5 layers (See Fig. 5)

TABLE II. RESULTS OF ORTHOGONAL ARRAY TESTING

S.No.	Training Parameters				Avg. Testing Error
	Range of influence	*Squash factor*	*Accept ratio*	*Reject ratio*	
1.	0.3	1.25	0.5	0.15	1.1
2.	0.4	1.25	0.5	0.15	6.922
3.	0.7	1.25	0.5	0.15	14.561
6.	0.4	0.4	0.5	0.15	35.223
7.	0.4	0.8	0.5	0.15	66.792
8.	0.4	1.0	0.5	0.15	101.425
9.	0.4	1.2	0.2	0.15	94.233

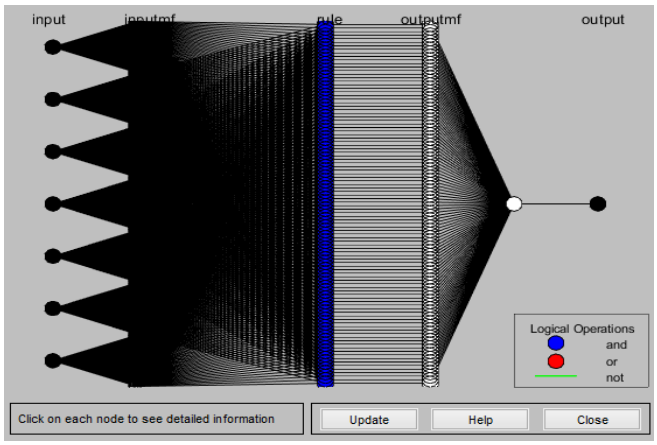


Fig. 5. Final ANFIS model structure

V. RESULTS AND DISCUSSION

The FIS file so created produce the following Root Mean Square Error (RMSE):

Minimal training RMSE = 0.152700 MPa

Minimal checking RMSE = 1.211921 MPa

Finally, we test the model with the validation dataset. The average testing error is **0.95887 MPa** (See Fig. 6).
Further, the relationships between various variables with compressive strength can be analyzed using the Surface Viewer option (See Fig. 7-10). Fig. 7 shows that with increasing cement content, the compressive strength increases up to a point and then decreases. Fig. 8 shows that with increasing water content beyond the ratio of 1, the compressive strength falls sharply, which is in tally with a theoretical understanding of concrete[20]. Fig. 9 shows how superplasticizer and water content are related. It can be seen that with a low amount of water and increased amount of superplasticizer, concrete strength is very high, it is also in agreement with theory[21]. Fig. 10 shows the relationship between coarse and fine aggregates. it is seen that maximum compressive strength is near the diagonal. It means when the

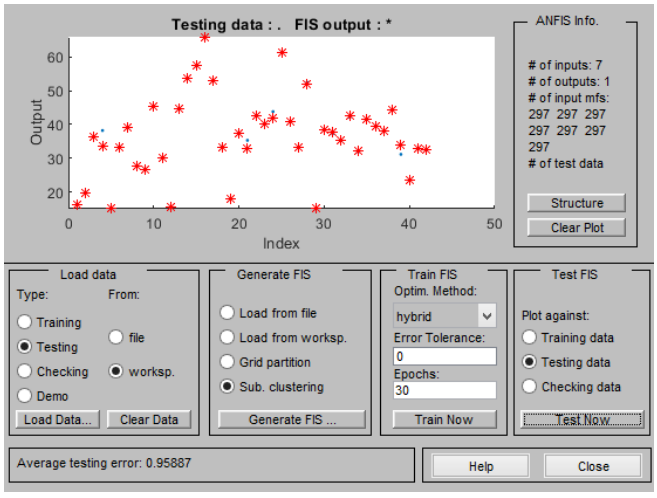


Fig. 6. Testing of Validation Set

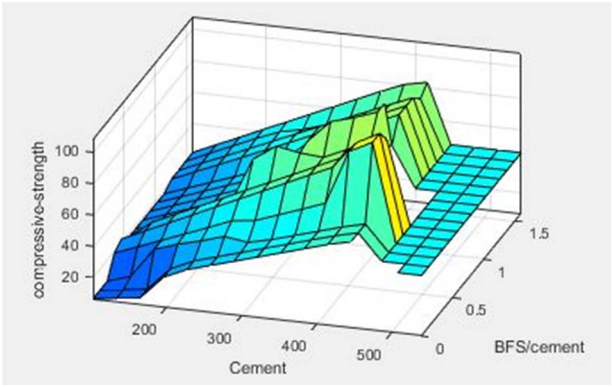


Fig. 7. Surface diagram of cement and BFS

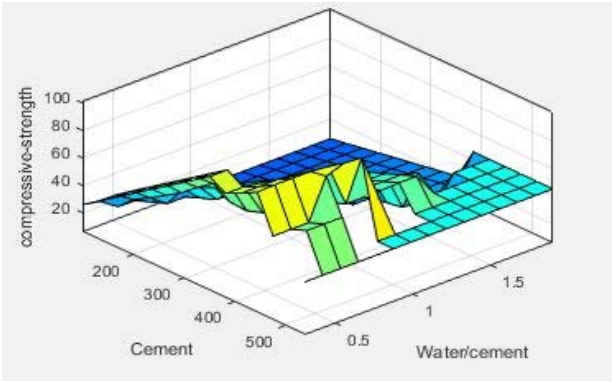


Fig. 8. Surface diagram of cement and water

coarse and fine aggregates are more or less equal, concrete has high compressive strength. It is also in agreement with previous research[22]

In the rule viewer window, the sliders can be used to change the input variables to get the corresponding output (See Fig. 11). This feature is quite user-friendly.

VI. CONCLUSION

In civil engineering, the mix design of concrete is very important. Knowing the early age compressive strength accurately can make further testing and calculations easier and minimize the errors in the mix design process.

In this study, the ANFIS based model was created using the Subtractive Clustering approach using the MATLAB program. The subtractive clustering parameters were obtained through trials. Finally, the developed model performs reasonably good for the prediction of early-age strength. The model produced an RMSE of 0.95887 MPa. This error is very less as compared to that of an ANN model [23], which produces an RMSE of 3.69 MPa.

It was found that MATLAB is very user-friendly yet advanced software. It provided the option to use both the command-based interface as well as a graphical interface.

The software gives an option to graphically assess the relationships of different components of concrete with each other based on the ANFIS model, giving the user an understanding of the basic behavior of the concrete mix. It can be used as a research tool for non-conventional mixes for which standardization does not exist.

The only drawback of using ANFIS is that it requires a large dataset for learning. Overall, ANFIS is one of the best realizable tools for predicting the early age compressive strength of concrete as presented by this study.

This study can be further extended to determine 28 days strength from the early age strength. Also, the extensive Mix Design process can be carried out by making several such models. This can help in reducing wastes and time involved in mix design, which in turn can help in reducing costs as well as environmental pollution.

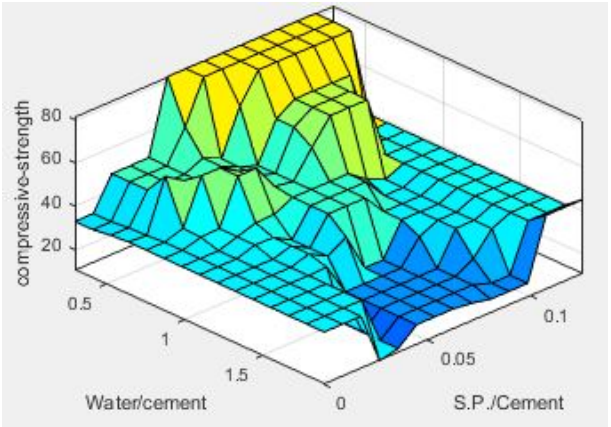


Fig. 9. Surface diagram of water and Superplasticizer

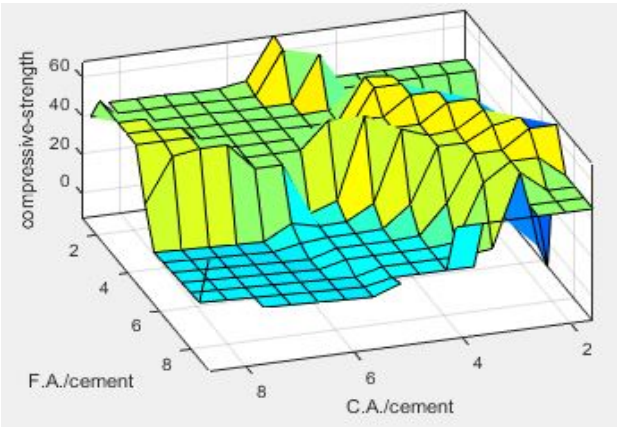


Fig. 10. Surface diagram of Fine aggregates and coarse aggregates

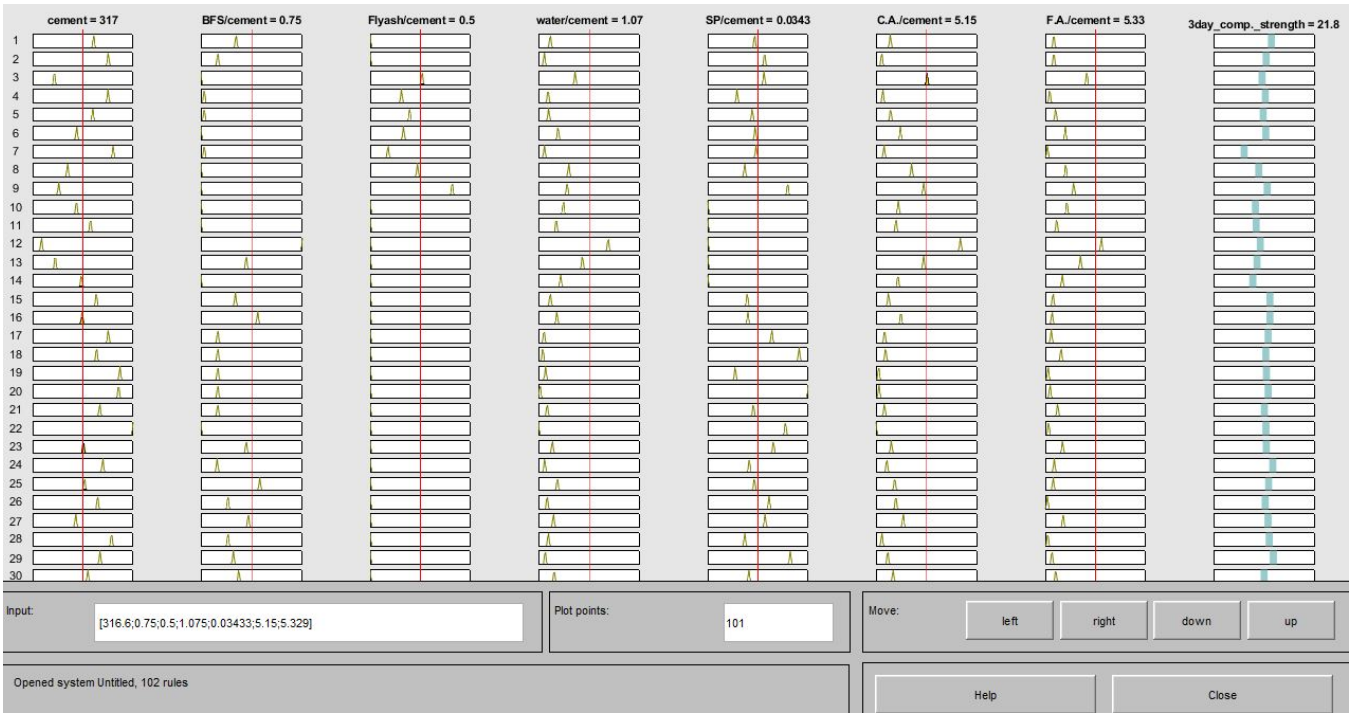


Fig. 11. Rule Viewer of the FIS model

REFERENCES

[1] H. Van Damme, "Concrete material science: Past, present, and future innovations," *Cem. Concr. Res.*, vol. 112, pp. 5–24, Oct. 2018.

[2] C. F. Ferraris, "Concrete mixing methods and concrete mixers: State of the art," *J. Res. Natl. Inst. Stand. Technol.*, vol. 106, no. 2, p. 391, Mar. 2001.

[3] M. Nehdi and A. M. Soliman, "Early-age properties of concrete: overview of fundamental concepts and state-of-the-art research," *Proc. Inst. Civ. Eng. - Constr. Mater.*, vol. 164, no. 2, pp. 57–77, Apr. 2011.

[4] A. Neville and P.-C. Aïtcin, "High performance concrete—An overview," *Mater. Struct.*, vol. 31, no. 2, pp. 111–117, Mar. 1998.

[5] International Union of Testing and Research Laboratories for Materials and Structures., *Materials and structures.*

[6] U. S. F. H. Administration, *Fly Ash*. 1999.

[7] U. S. F. H. Administration, *Ground Granulated Blast-Furnace Slag*.

[8] B. K. and H. B. A. Varun, "EFFECT OF ADDITION OF FLYASH AND GGBS ON CEMENT CONCRETE IN FRESH AND HARDENED STATE," *Int. J. Adv. Eng. Res. Dev.*, vol. 5, no. February, 2018.

[9] A. L. Radu, M. A. Scrieciu, and D. M. Caracota, "Carbon Footprint Analysis: Towards a Projects Evaluation Model for Promoting Sustainable Development," *Procedia Econ. Financ.*, vol. 6, pp. 353–363, Jan. 2013.

[10] K. Day, *Concrete mix design, quality control and specification E&FN Spon*. 1999.

[11] M. K. Alsedrah, "Artificial Intelligence Advanced Analysis and Design," *Res. Gate*, no. March, 2018.

[12] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharm. Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000.

[13] C. H. Chen, "Fuzzy logic and neural network handbook." 1996.

[14] M. C. Nataraja, M. A. Jayaram, and C. N. Ravikumar, "A Fuzzy-Neuro Model for Normal Concrete Mix Design.," *Eng. Lett.*, vol. 13, no. 2, pp. 98–107, 2006.

[15] P. Chopra, R. K. Sharma, M. Kumar, and T. Chopra, "Comparison of Machine Learning Techniques for the Prediction of Compressive Strength of Concrete," *Adv. Civ. Eng.*, vol. 2018, pp. 1–9, 2018.

[16] M. A. Akhund, "A Review on Expert System and its Applications in Civil Engineering.".

[17] J. R. Jang, "ANFIS : Adaptive-Network-Based Fuzzy Inference System," vol. 23, no. 3, 1993.

[18] I-Cheng Yeh, "UCI Machine Learning Repository: Concrete Compressive Strength Data Set." [Online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Data.xls. [Accessed: 30-Mar-2019].

[19] G. W. Delius, *Orthogonal Arrays (Taguchi Designs)*. University of York, 2004.

[20] S. B. Singh, P. Munjal, and N. Thammishetti, "Role of water/cement ratio on strength development of cement mortar," *J. Build. Eng.*, vol. 4, pp. 94–100, Dec. 2015.

[21] P.-C. Nkinamubanzi, S. Mantellato, and R. J. Flatt, "Superplasticizers in practice," *Sci. Technol. Concr. Admixtures*, pp. 353–377, Jan. 2016.

[22] C. S. Poon and C. S. Lam, "The effect of aggregate-to-cement ratio and types of aggregates on the properties of pre-cast concrete blocks," *Cem. Concr. Compos.*, vol. 30, no. 4, pp. 283–289, Apr. 2008.

[23] I.-C. Yeh, "Modeling Concrete Strength with Augment-Neuron Networks," *J. Mater. Civ. Eng.*, vol. 10, no. 4, pp. 263–268, 2002.

A Graph based Keyword Extraction from Twitter using Node and Edge Weight

Ritika
Computer Science and Engineering,
UIET, Panjab University,
Chandigarh, India
ritikagosain93@gmail.com

Mukesh Kumar
Computer Science and Engineering,
UIET, Panjab University,
Chandigarh, India
mukesh_rai9yahoo.com

Pretti Aggarwal
Computer Science and Engineering,
UIET, Panjab University,
Chandigarh, India
pree_agg@pu.ac.in

***Abstract*—With the advent of mobile and internet technology, social media has experienced a tremendous growth over the past few years. Social networking site such as Twitter provides the platform for content generation, information dissemination and communications. Summarizing the textual contents or extracting the influential words is a challenging task. Therefore an unsupervised graph based approach is proposed for automatic Keyword Extraction from the collection of Tweets using Node and Edge Weight (KETNEW). The node weight and edge weight depends on various parameters such as tweet frequency, position of node, clustering coefficient, co-occurrence frequency and shared neighbors. The important aspect of our approach is that it is a hybrid approach i.e. a graph based approach (structural approach) which depends upon statistical as well as linguistic features. Apart from this, it also depends on the position of the word in the text. These numerous features made it perform better than existing techniques.**

Keywords—twitter, graph, unsupervised, keyword

I. INTRODUCTION

Keyword Extraction is a task of automatically extracting influential terms/words of a document that best describes the subject of that document or the topic of discussion [1]-[4]. Keywords are defined as one or more entities in a document that represents the content of that document. It has an important role in Information Retrieval, index construction, text mining, text summarization, text classification, cluster analysis, topic detection, recommender system, search engine optimization and natural language processing.

In today's context, social media changes the way of communication between people such as business executives, individuals as well as public. Facebook, Twitter, LinkedIn, Instagram etc. are the examples of various communication channels used presently. Twitter is one of the highest-ranking social media websites where people communicate with each other using short messages called "tweets" which are limited to 280 characters only. It is an online platform where people share information such as breaking news, current affairs, articles, events, trolls etc. Many corporations and news agencies also use this micro-blog service to broadcast news articles and advertisements among the people.

The proposed method aims to build an unsupervised graph based keyword extraction system on a collection of tweets. Tweets are represented as a Graph-of-Words where words in tweets represent nodes in graph and edges are defined by the relationship between these words. Node weight and edge weight depends on various parameters such as tweet frequency, position of words in the tweet, clustering coefficient and co-occurrence frequency, shared neighbors respectively as described below in Section 3. Based on these weights, NI score is computed and e top ranked keywords are extracted from graph. The proposed technique KETNEW is compared with other techniques from literature such as TFIDF, TextRank and TKG with different centrality measures. A variant of KETNEW is also presented in which instead of using clustering coefficient, eigenvector centrality is used while calculating node weight. The proposed approach is tested on five topics of FSD dataset. The results are evaluated and compared with other techniques. It is observed that KETNEW and a variant of KETNEW are superior to all other techniques.

The remaining paper is structured as follows: In Section 2 we discuss literature review. Section 3 introduces the proposed methodology. Section 4 describes the performance of experiments to demonstrate the significance of the proposed approach. Section 5 contains a conclusion and outlines some future scope.

II. LITERATURE SURVEY

Keyword extraction techniques can be roughly classified as supervised or unsupervised [4]. Supervised approach uses documents with predefined labels based on which it builds its own model for keyword extraction. Unsupervised approach uses no training data. Unsupervised approach can be further classified as simple statistical, linguistics, graph-based and other approaches [4]. Simple statistical approaches are language and domain independent. The statistical information of words such as n-gram, term frequency, document frequency, co-occurrence, tfidf etc. can be used to find keywords in the document. Linguistic approaches are language dependent. These approaches analyze the content of the document on the basis of linguistic features such as syntax, semantics, morphology etc. Graph based approaches generate a graphical representation of the textual data and using centrality measures or some other parameters, it determines the importance of the nodes. This section provides the brief overview of literature work.

Bag-of-Words Model:

Bag-of-words model is used for keyword extraction in information retrieval. It is also called vector space model. In this model, a document is represented as a bag containing all the terms free from stopwords, punctuation marks, special characters and the word order. TFIDF score is calculated for each term which determines the importance of the term in the document. The score is calculated using Eq. (1)

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

$$TF(i) = \frac{f(i)}{n}$$

$$IDF(i) = \log\left(\frac{N}{df(i)}\right)$$

$$TF - IDF(i) = TF(i) * IDF(i) \quad (1)$$

Graph based Centrality Measures:

In this approach, the text is represented as a Graph-of-words where each term in the text represents a node and a pair of consecutive terms represents an edge in the graph. Using centrality measures, the most central nodes can be identified and are considered as keywords. Some related work is as follows: Abilhoa, W. D., & de Castro, L. N. presents an approach called as TKG (Twitter Keyword Graph) which is a graph-based approach for extracting keywords from tweets [3]. The importance of a node is calculated using different centrality measures such as closeness centrality etc. Abilhoa, W. D., & de Castro, L. N. also proposed an unsupervised graph based approach in which text is represented as a graph and three centrality measures: closeness, eccentricity and degree centrality are applied [4]. Eccentricity and closeness centrality are used to decide the rank of the nodes and degree centrality is used as a tie breaker. Eccentricity does not perform well in case of disconnected graph due to infinite path length.

Other Related Work:

Mihalcea, R., & Tarau, P. presented the TextRank which is an unsupervised graph based keyword extraction algorithm that considers edge weights while calculating the node score [5]. It is based on the Google's Page Rank algorithm [6]. Bellaachia, A., & Al-Dhelaan, M. proposed an approach called as NE Rank that takes into account node weights in addition to edge weights while calculating the node score. It basically uses node weights with text rank when computing importance of the node. Node weight depends on only statistical information i.e. TFIDF [7]. Biswas, S. K., Bordoloi, M., & Shreya, J. proposed an unsupervised graph based keyword extraction approach KECNW which uses NE Rank for calculating the node score where node score depends on various parameters such as centrality measure, strength, importance of neighboring nodes, position, term frequency. A graph is built in which each term represent a node and pair of consecutive terms represents edges. The approach does not depend on linguistic features of the text except for stopwords removal [8].

This paper proposes an unsupervised hybrid approach for keyword extraction i.e. an approach which is based on the graphical model, statistical information as well as linguistic features. It also depends on the position of the node in the text. The importance of a node is based on NI score which is calculated using both node weight and edge weight which further comprise of various influencing parameters such as statistical measure, centrality measure as well as positional parameter.

III. PROPOSED KETNEW MODEL

The proposed model KETNEW aims to build a keyword extraction system on a collection of tweets. The model considers tweets as a Graph-of-Words. NI (Node Importance) score is calculated for each word (node) in the graph based on which top-ranked keywords are extracted. The model is divided into three phases: Data Collection, Data Preprocessing and Keyword Extraction. Each phase is discussed as below.

3.1 Data Collection (Phase 1)

Data Collection is the process of gathering the data and storing it in some data file for the purpose of analysis. The data is publically available on the internet which is collected using the Twitter Streaming API and a python library known as tweepy.

3.2 Data Preprocessing (Phase 2)

The data collected in phase 1 is not in required form. It contains a lot of noise which needs to be filtered-out before further analysis. Preprocessing is further divided into three subphases.

3.2.1 Sub-Phase 1:

1. *Remove RT, @username*: RT is the reserved word which is used to re-post a tweet of some other user. @username is used to mention a particular user. RT, @username is removed as they do not provide any significant information about the subject of the tweets.
2. *Remove symbol hashtag:* Symbol '#' is prefixed with some entity to tag tweets related to that entity. Symbol # provides no meaningful information. So any # appearing in the tweet is removed.
3. *Remove URL:* The URL is usually used to navigate to a page for a detailed description of the tweet. URL doesn't provide any significant information and therefore it is removed.
4. *Remove Retweets:* Tweets may contain some retweets. Retweet is a tweet from a user which is re-posted by some other user. As a result, the similar text is shown up many times. This may skew our results, as one user's tweet will occur many times in our analysis. Hence, filter out these duplicate tweets before further processing as they provide no additional information.

3.3.3 Sub-Phase 3:

1. *Remove Unimportant/inferior terms*: Terms having frequency less than *ATF* (Average Term Frequency) are considered as least frequency terms. These are relatively less important terms. Remove all these terms from the tweets. *ATF* is calculated using Eq. (2):

$$ATF = \frac{\sum_{i=1}^{n} f(i)}{n} \quad (2)$$

where *f(i)* denotes the frequency of *i* i.e. the number of occurrences of term *i* in the tweets and *n* is number of unique terms in tweets.

2. *Omit additional white spaces:* Remove extra white spaces between the terms in the tweets.

3.3 Keyword Extraction (Phase 3)

After the preprocessing phase is complete, next phase is Keyword Extraction phase which is discussed as below.

3.3.1 Graph-of-Words

After the preprocessing phase is complete, next phase is Graph-of-Words representation. Let *G* be an undirected weighted graph having a collection of nodes and edges.

Each term in preprocessed tweets represents a node in the graph and consecutive terms in the tweets represent an edge between two nodes in the graph.

3.3.2 Calculating Node and Edge Weight

The parameters used in calculating node weight and edge weight are defined as follows:

1. *Tweet Frequency:* It is similar to the document frequency of a term. It is defined as number tweets containing this term. It is calculated using Eq. (3)

$$twf(i) = \sum_{t=1}^{N} bit \quad (3)$$

where *bit* value is 0/1. Looping through all the *N* tweets, if a term *i* occurs in the tweet *t*, set the bit value for corresponding tweet to 1, otherwise 0. *twf* of a term is the total number of set bits for that term.

2. *Position of a node:* Position of a term is also an important parameter while extracting keywords as suggested by Hotho, Nürnberger, and Paab. Position of the term also contributes to the node weight in the graph [8]. Positional weight is calculated using Eq. (4):

$$a)\ first(i) = \frac{\sum_{t=1}^{N} bit}{N}$$

$$b)\ last(i) = \frac{\sum_{t=1}^{N} bit}{N} \quad (4)$$

where *N* is the number of tweets. While calculating *i*, if term *i* occurs at first position in tweet *t*, then the corresponding bit is set to 1, otherwise 0. Similarly while calculating *last(i)*, if term *i* occurs at last position in tweet *t*, then the corresponding bit is set to 1, otherwise 0.

3. *Clustering Coefficient:* Clustering Coefficient *cc(i)* of a node *i* is defined as the fraction of possible interconnections between neighbors of a node [9]. It is calculated using Eq. (5).

$$cc(i) = \frac{2\, l_i}{d_i(d_i - 1)} \quad (5)$$

where d_i denotes the number of neighbors of node i i.e. degree of node i and l_i denotes number of links between neighbors of node i.

4. *Co-occurrence frequency:* It is defined as the frequency of occurrence of consecutive terms in tweets. It is calculated using Eq. (6).

$$CF(i,j) = \frac{f(i,j)}{f(i)+f(j)+f(i,j)} \quad (6)$$

where $f(i), f(j)$ is the frequency of occurrence of term i, j respectively and $f(i,j)$ is thefrequency of occurrence of pair of terms i, j in any order [10].

5. *Shared Neighbors:* It is defined as the number of common neighbors between nodes i, j in the graph. This value $SN(i,j)$ is normalized in range 0 to 1.

After calculating the above parameters, node weight and edge weight is calculated using Eq. (7) and Eq. (8) respectively and is normalized in range 0 to 1.

A. *Node Weight:* Node Weight plays a crucial role in extracting influential keywords and is calculated using Eq. (7)

$$node\_weight(i) = twf(i) + first(i) + last(i) + cc(i) \quad (7)$$

B. *Edge Weight:* Edge represents the relationship between the nodes in the graph. An edge is defined by the co-occurrence of terms in the tweets. Edge weight is calculated using Eq. (8).

$$edge\_weight(i,j) = CF(i,j) + SN(i,j) \quad (8)$$

where $CF(i,j)$ is co-occurrence frequency of node i, j and $SN(i,j)$ is the shared neighbors between two nodes i, j.

3.3.3 Keyword Extraction based on NI Score

After building Graph-of-Words and calculating node and edge weights, next step is to calculate NI Score for each node v_i in the graph. Keywords are extracted on the basis of NI (Node Importance) score. Psuedocode for computing NI score is given in Algorithm 1.

After calculating NI score, sort all the nodes on the basis of NI score and assign a rank to each node. In case, two nodes have the same score, a rank resolution parameter clustering coefficient is used to break the tie between two nodes. Pseudocode for this is given in Algorithm 2.

Algorithm 1: Calculating NI score:

Inputs:
Graph G: An undirected weighted Graph of words having $node\_weight(v_i)$,
d: *damping factor* $= 0.85$

Output:
$S(v_i)$: *Score of each node* v_i

for each *node* v_i *in the Graph G*
  $W(i) = node\_weight(v_i)$
  for each *node* v_j *adjacent to node* v_i *in G*
    $W(j) = node\_weight(v_j)$
    $W(j,i) = edge\_weight(v_j, v_i)$
    for each *node* v_k *adjacent to* v_j *in G*
      $W(j,k) = edge\_weight(v_j, v_k)$
      $sum[j] = \sum W(j,k)$
    end for
    $Q[j] = \sum \frac{W(j,i)}{sum[j]}\ W(j)$
  end for
  $T[i] = \sum Q[j]$
  $S(v_i) = (1-d).W(i) + d.W(i).T[i]$
end for
return $S(v_i)$

Algorithm 2: Node Ranking:

Let v_i and v_j be two nodes/terms in the sorted list:

1. if $S(v_i) > S(v_j)$
2. $\quad$ *then* $Rank(v_i) < Rank(v_j)$
3. *else if* $S(v_i) = S(v_j)$
4. $\quad$ *if* $CC(v_i) > CC(v_j)$
5. $\quad\quad$ *then* $Rank(v_i) < Rank(v_j)$

In the end, k top ranked nodes are selected as keywords, where $k \geq 1$ and is an integer.

KETNEW with a variation:

KETNEW variant is also introduced which uses eigenvector centrality instead of clustering coefficient while calculating node weight.

- *Eigenvector Centrality:* It measures the influence of the node in the network. Unlike degree centrality which only depends upon the number of nodes connected to it, eigenvector centrality $ec(i)$ also depends upon the importance of the nodes connected to it [9].

$$node\_weight(i) = twf(i) + first(i) + last(i) + cc(i) \quad (8)$$

where $ec(v_i)$ is the eigenvector centrality of vertex i.

Node weight is calculated using Eq. (8). Top keywords are extracted using NI score and eigenvector centrality.

IV. Evaluation and Comparison

The proposed system is validated on five topics of FSD (First Story Detection) dataset namely: (i) Ending of Nasa's Space Shuttle Program (ii) Plane carrying Russian hockey team crashes (iii) Fire in children's camp on Utoya island (Norway) (iv)Google's plan to buy Motorola Mobility (v) S&P downgrades US credit rating. Dataset is available online. Each topic has collection of tweets and relevant keywords corresponding to it.

To evaluate the performance of our system, retrieved keywords are matched against relevant keywords and quantitative analysis is performed in terms of precision, recall and F-measures which are calculated using Eq. (9), (10) and (11) respectively.

$$Precision = \frac{|Relevant \cap Retrieved|}{|Retrieved|} \quad (9)$$

$$Recall = \frac{|Relevant \cap Retrieved|}{|Relevant|} \quad (10)$$

$$F\ measure = \frac{2}{\frac{1}{Precison} + \frac{1}{Recall}} \quad (11)$$

$Precision$ is calculated as the number of retrieved keywords which are relevant to the total number of retrieved keywords. $Recall$ is calculated as the total number of retrieved keywords which are relevant to the total number of relevant keywords. $F\ Measure$ is calculated by taking weighted harmonic mean of $Precision$ and $Recall$.

The proposed technique is then compared with other existing techniques namely TFIDF, TextRank, TKG and its variants. The performance measure of the proposed system and different techniques are shown below in Table 1-5 when top 10 keywords are extracted from tweets.

TABLE 1 PERFORMANCE MEASURE ON TOPIC I

Model	*Precision*	*Recall*	*F-Measure*
Bag-of-Words (TF-IDF)	4/8	4/8	0.44444
Text Rank	5/10	5/8	0.55555
TKG with degree Centrality(DC)	5/10	5/8	0.55555
TKG with betweenness Centrality(DC)	4/10	4/8	0.44444
TKG with Eccentricity(EC)	4/10	4/8	0.44444
TKG with closeness Centrality(CC)	6/10	6/8	0.66666
TKG with Eigenvector Centrality(EC)	6/10	6/8	0.666666
Proposed KETNEW	7/10	7/8	0.77777
Proposed KETNEW with variation	7/10	7/8	0.77777

TABLE 2 PERFROMANCE MEASURE ON TOPIC II

Model	*Precision*	*Recall*	*F-Measure*
Bag-of-Words (TF-IDF)	0/10	0/8	----------
Text Rank	6/10	6/8	0.66666
TKG with degree Centrality(DC)	6/10	6/8	0.66666
TKG with betweenness Centrality(DC)	7/10	7/8	0.77777
TKG with Eccentricity(EC)	4/10	4/8	0.44444
TKG with closeness Centrality(CC)	6/10	6/8	0.66666
TKG with Eigenvector Centrality(EC)	6/10	6/8	0.666666
Proposed KETNEW	7/10	7/8	0.77777
Proposed KETNEW with variation	7/10	7/8	0.77777

TABLE 3 PERFORMANCE MEASURE ON TOPIC III

Model	*Precision*	*Recall*	*F-Measure*
Bag-of-Words (TF-IDF)	2/10	2/7	0.23529
Text Rank	3/10	3/7	0.35294
TKG with degree Centrality(DC)	3/10	3/7	0.35294
TKG with betweenness Centrality(DC)	4/10	4/7	0.40758
TKG with Eccentricity(EC)	4/10	4/7	0.40758
TKG with closeness Centrality(CC)	4/10	4/7	0.40758
TKG with Eigenvector Centrality(EC)	6/10	4/7	0.40758
Proposed KETNEW	5/10	5/7	0.58823
Proposed KETNEW with variation	4/10	4.7	0.47058

TABLE 4 PERFORMANCE MEASURE ON TOPIC IV

Model	*Precision*	*Recall*	*F-Measure*
Bag-of-Words (TF-IDF)	4/8	4/8	0.44444
Text Rank	5/10	5/8	0.55555
TKG with degree Centrality(DC)	5/10	5/8	0.55555
TKG with betweenness Centrality(DC)	4/10	4/8	0.44444
TKG with Eccentricity(EC)	4/10	4/8	0.44444
TKG with closeness Centrality(CC)	6/10	6/8	0.66666
TKG with Eigenvector Centrality(EC)	6/10	6/8	0.666666
Proposed KETNEW	7/10	7/8	0.77777
Proposed KETNEW with variation	4/10	4/5	0.53333

TABLE 5 PERFORMANCE MEASURE ON TOPIC V

Model	*Precision*	*Recall*	*F-Measure*
Bag-of-Words (TF-IDF)	4/8	4/8	0.44444
Text Rank	5/10	5/8	0.55555
TKG with degree Centrality(DC)	5/10	5/8	0.55555
TKG with betweenness Centrality(DC)	4/10	4/8	0.44444

TKG with Eccentricity(EC)	4/10	4/8	0.44444
TKG with closeness Centrality(CC)	6/10	6/8	0.66666
TKG with Eigenvector Centrality(EC)	6/10	6/8	0.666666
Proposed KETNEW	7/10	7/8	0.77777
Proposed KETNEW with variation	4/10	4.5	0.53333

Table 1-5 shows *Precision, Recall* and *F – Measure* values represent graphical representation of the quantitative analysis. These values signify the efficiency of system in extracting relevant keywords. Higher the values, better the system is.

From the results presented above in Table 1-5, it is observed that the proposed technique outperforms all other existing techniques in terms of quantitative measures: *Precision, Recall* and *F Measure*. KETNEW variant performs almost same as KETNEW. Centrality measure eccentricity does not perform well if the graph is not connected i.e. disconnected. All other techniques are superior to TFIDF. TKG with closeness centrality and eigenvector centrality performs better than TextRank and other centrality measures. Average Performance analysis of all the techniques is presented below in Fig.1.

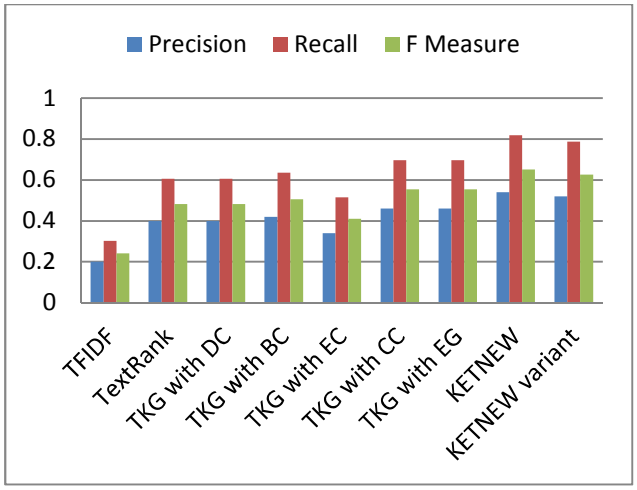


Fig. 1. Average Performance Analysis for different techniqes

V. Conclusion and Future Scope

This paper addressed an automatic keyword extraction approach KETNEW which make use of an undirected weighted graph in which node weight and edge weight depends on various influencing parameters. Majority of approaches have been discussed in literature survey but an important characteristic of KETNEW is that, besides depending on linguistic features, it does not require much knowledge of these features and therefore it is easily portable to other domains, language and text corpus. Results show that the keyword extraction not only depends upon frequency of occurrence but also on some other influencing parameters such as parameters such as tweet frequency, position of node, clustering coefficient, co-occurrence frequency and shared neighbors. The results show that KETNEW outperforms the existing techniques. A variation of KETNEW is also proposed that make use of eigenvector centrality instead of clustering coefficient and almost performs similar to KETNEW.

In future, the proposed work can be extended in the following directions: Other influencing parameters or centrality measures can be used for more good results. Second, build an approach for automatic text extraction from heterogeneous data such as text, videos, and images.

VI. References

[1] Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences, 39(1), 1-20.

[2] Palshikar, G. K. (2007, December). Keyword extraction from a single document using centrality measures. In International Conference on Pattern Recognition and Machine Intelligence (pp. 503-510). Springer, Berlin, Heidelberg.

[3] Abilhoa, W. D., & de Castro, L. N. (2014). TKG: A Graph-Based Approach to Extract Keywords from Tweets. In Distributed Computing and Artificial Intelligence, 11th International Conference (pp. 425-432). Springer, Cham.

[4] Abilhoa, W. D., & De Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation, 240, 308-325.

[5] Mihalcea, R., & Tarau, P. (2004). : Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

[6] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.

[7] Bellaachia, A., & Al-Dhelaan, M. (2012, December). Ne-rank: A novel graph-based keyphrase extraction in twitter. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on (Vol. 1, pp. 372-379). IEEE.

[8] Biswas, S. K., Bordoloi, M., & Shreya, J. (2018). A graph based keyword extraction model using collective node weight. *Expert Systems with Applications*, *97*, 51-59.

[9] Al-Taie, M. Z., & Kadry, S. (2017). Python for Graph and Network Analysis. Springer International Publishing.

[10]Sonawane, S. S., & Kulkarni, P. A. (2014). Graph based representation and analysis of text document: A survey of techniques. International Journal of Computer Applications, 96(19)

Smart Fruit Warehouse and Control System Using IoT

Anurag Shukla*, Gazal Jain†, Kavyansh Chaurasia‡ and Venkanna U., Member IEEE§
Department of Computer Science, DSPM-IIIT Naya Raipur, Atal Nagar, India
Email: *anurag17100@iiitnr.edu.in, †gazal17100@iiitnr.edu.in, ‡kavyansh17100@iiitnr.edu.in, §venkannau@iiitnr.edu.in

***Abstract*—We live in an era where technology and automation is impacting every domain possible, increasing efficiency, productivity, scalability and at the same time cutting down investments and efforts. IoT (Internet of Things) has given us the power of connecting everyday objects to internet and hence they can be controlled and monitored from anywhere in the world. However, traditional warehouses still remains an exception. Humans toiling their way through the day, controlling and monitoring all the environmental conditions of the warehouses, is the general Picture of warehouses around the globe today and that comes with a cost of human errors resulting in wastes of resources and decreasing efficiency. Therefore, we require an automated smart warehouse which not only provides multi-parameter monitoring and control but is able automatically vary the parameters according to the environment. Existing solutions fail to address this problem which are either impractical or offer partial solution. In this paper we present a smart warehousing system which through various sensors connected to a micro-controller performs not only multi-parameter monitoring by sending the data to the cloud but also provides automated control by processing the data in the server and sending results back to the micro-controller. Moreover, the system also provides with a website and an android application which present various graphs and give complete control and automation through a tap of a button.**

***Index Terms*—IoT, smart fruit warehouse system, warehouse, automation, micro-controller**

I. Introduction

Indian economy is an agrarian economy since the dawn of civilization on Indian subcontinent, now with over 70% of its rural population engaged in activities related to agriculture [1]. India is the second largest producer of Fruits and Vegetables in the world, accounting for 11% of all the worlds produce. Majority of this produce is damaged due to lack of proper warehousing mechanisms and is declared unfit for consumption. Environmental factors mainly temperature and humidity play the most important role in determining the produces quality. Hence, there is a crucial need of multi-parameter monitoring and control system to check the fruits quality. Traditional warehouses, unfortunately still relies on human practices like checking the thermometer and humidity meter manually, analyzing it and then taking the decision for regulation of parameters inside the warehouse. Humans are prone to errors; hence manual control of warehousing systems increases the probability of damage to the stored fruits/vegetables. Hence, an automated warehousing system is required which can not only monitor the environmental parameters of the warehouse but can analyze the data, take actions and can regulate the parameters based on the results.

978-1-7281-2087- 4/19/$31.00 © 2019 IEEE

However, a smart and automated warehousing system does involve some challenges and there are few important features that an automated warehousing system should have. The monitoring should be multi-parameter, completely automated, transparent and the data should be presented to user in a visually friendly manner. The data thus collected should be analyzed and the system should take actions based upon the results. Therefore, the system should provide an automated control based on the parameters value collected in real time. Solutions like [2] implements the monitoring by sensing the parameters and sending them to cloud but it doesnt analyze the data collected. While other solutions like [3] analyses the data but it is too simple to spit out major conclusions. Moreover, the analysis and computations on the data shouldnt be done on the remote micro-controller that collects the data because generally they dont have the specifications required to handle the large amount of data generated leave alone the computations. Hence, the system should be completely centralized i.e. the server should perform all the manipulations and computations and send the results back to the micro-controller. The centralized system will further add the advantage of monitoring and controlling the warehouse globally through the internet. Another solution, [4] implements automated monitoring mechanisms but the it runs the whole algorithm in the remote controller itself. However both the solutions [3] and [4] respectively does not allow the user to control the warehouse parameters from the internet. [3] on the other hand just let the user to turn off or on the fan and doesnt let the user to control the parameters to a specific value.

This paper presents an IoT based smart warehouse control and monitoring system . The major contributions of the paper are as follows:

1) The proposed system is fully centralized i.e. the warehouse can be controlled and monitored from anywhere around the globe.
2) It plots graphs and collect data for different parameters, analyses them and take decisions based on the results.
3) It sets the parameter values (Temperature, humidity and light) based upon the type of fruit stored and can detect any malfunctions whatsoever and notifies the user. The platform is hooked up with algorithms take takes care of the duration of storage and will notify the user on completion of the duration.

4) The solution provides both a web interface and an android app that gives a rich user interface to interact with the system.
5) The proposed system also saves the data for all sessions for any future references.

Henceforth, the paper is organized as follows: Section II explains various works on smart fruit warehouse system. The proposed prototype and working model is described in section III. Section IV describes experimental setup. The results are illustrated in section V. Finally, Section VI concludes the paper.

There have been many efforts to implement a fully automatic control and monitoring warehousing systems. This section mentions different implementations, ideas and their limitations.

Temperature and humidity monitoring system for storage rooms of industries [2] implements a monitoring system that not only sends the value of various parameters to the cloud but can also verify those values using a table. However, this system does not give an interface where the user can see the data. Also, there is no clear mention of how the type of fruit is decided for comparing the required parameters with the table. Moreover, this system also fails in providing a control system.

Implementation of IoT based Smart Warehouse Monitoring System [5] implements Multi parameter monitoring system which monitors the environment of the warehouse and sends the data over the cloud to ThingsSpeak [6]. Using MATLAB [7] supported by ThingsSpeak the data collected can be analyzed and visualized. While this method is successful in automating the multi-parameter monitoring of warehouse, it does not provide a system for controlling the environment parameters i.e. this system fails to provide the action mechanism that could be taken after the data analysis.

Design of fruits warehousing monitoring and control system based on Wi-Fi [4] implements both a monitoring and control system for warehouse. It is mainly divided into two parts the monitoring node and remote intelligent control terminal. The analysis of the data is done in the monitoring node itself. They proposed that to keep it online, the data can be sent to the cloud by using the control system. This system diverts from the centralized behaviour. Even though a server is used, the processing is done in the monitoring node. The server is just used as a way to store data on the cloud and all other analysis and processing is done remotely which in a way also questions the transparency of the system.

II. Proposed Solution

The proposed solution can be split into two modules the hardware module which sense the data and provides the input in a specific format and the software module which works on the input to provide automation and control.

Fig. 1 represents sensor node which consists of temperature, humidity and light sensors along with a micro-controller. Micro-controller receives respective values from each of the sensors and sends it to server. The server analyses the data

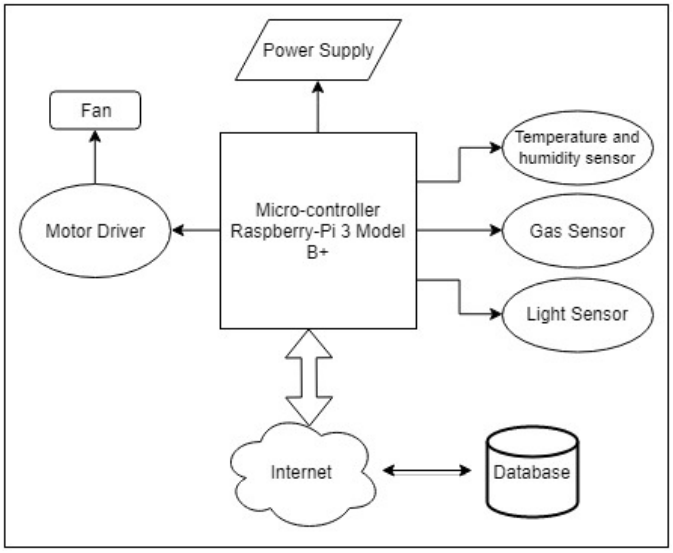


Fig. 1. Block Diagram of proposed solution

which includes determination of optimal parameters for storage, determination of any malfunction, ensuring of storage for optimal days and controlling the operation of the warehouse and produces an output which is sent back to the micro-controller. Moreover, the server stores the data and also produces plots which can be seen on the web page or the mobile application for better visualization of data. The micro-controller regulates the parameter controlling unit in accordance to the response of the server which also includes the change of parameters requested by the user. Rest of this section explains the two modules i.e. hardware module and software module in detail.

A. Hardware Module

1) Micro-Controller: Raspberry Pi [8] is a small computer board that has a 1.2 GHz quad-core processor which works in conjuncture with the 40 input/output and power Pins which can be connected to different types of sensors and actuators to sense Pis environment and also perform required calculations on the recorded data. It has its own Linux based operating system a.k.a Raspbian [9] in which its programmed to collect the data from the sensors and process it as per requirement. For GUI, it has Ethernet, Wi-Fi and HDMI support to connect it to a computer. Its a robust, low-cost minicomputer that has adequate processing capabilities needed for construction of our sensing nodes.

2) DHT11 Temperature and Humidity Sensor: It features a calibrated digital signal output with temperature and humidity sensing capabilities. It has an integrated high-performance 8-bit micro-controller. Each DHT11 sensors features extremely accurate calibration of humidity calibration chamber. The calibration coefficients are stored in the OTP program memory. The single-wire serial interface system is integrated to become quick and easy. Its small size, low power, signal transmission

distance up to 20 meters, enables a variety of applications. The product is 4-Pin single row Pin package [10].

3) GY-30 Bh1750 Intensity Digital Light Sensor Module: BH1750FVI module GY-30 is a digital light intensity sensor integrated circuit for a two-wire serial bus interface [11].

B. Software Module

1) Web Application: Its a client-server computer program in which the client requests for a web page to the server. In our case the sensing node comprises of the components mentioned above, sends the recorded data to a local server from where its fetched, processed into a web page and is transmitted to the client as a HTML file. The HTML file is rendered into a web page at the client-side in the web browser. Django is a Python-based free and open-source web framework, which follows the model-view-template architectural pattern. It is maintained by the Django Software Foundation, an independent organization established as a 501 non-profit [12].

2) Android Application: An android application is a software application meant to run on a device with android operating system. Android operating system is usually run on mobile devices [13]. Our system provides an android application which is built using Android Studio an integrated development environment by Google and IntelliJ. The application essentially offers all the features of a website in mobile phones .

C. Hardware and Software Integration

This section covers the integrated model of above components along with the details of the connections with cloud, data handling, manipulation and calculation techniques.

The remote sensor node's circuit diagram is depicted in Fig. 2. The figure shows all the connections to the respective sensors and a motor driver. The details of each sensor used is given below:

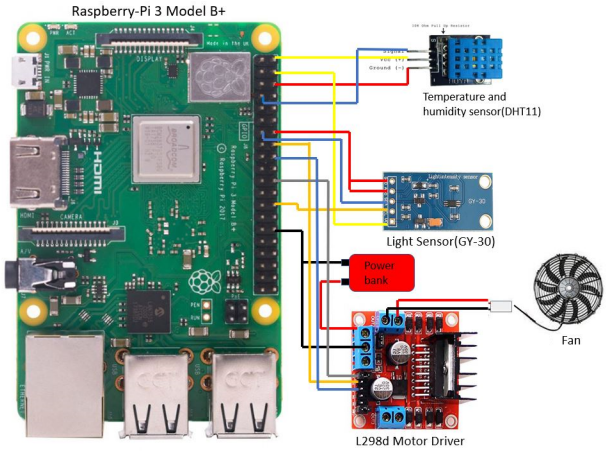


Fig. 2. Circuit Diagram for Sensor Node

1) Sensors and Micro-controller: DHT11 Temperature and humidity sensor with Raspberry Pi: Connections are made to Raspberry Pi according to Fig. 2 DHT11 consists of a surface mounted NTC thermistor and a resistive humidity sensor. An IC at the back of the module converts resistance changes in the thermistor and humidity sensor into digital temperature (in C) and relative humidity measurements and sends a digital signal on the data Pin of Raspberry Pi.

GY-30 Bh1750 Intensity Digital Light Sensor Module with Raspberry Pi: Connections are made to Raspberry Pi according to Fig. 2 BH1750FVI is a Digital ambient light sensor IC for IC2 interface. Detects light close to the spectral characteristics of visual acuity and gives digital output with 1 Lux high precision measurements.

2) Micro-controller to Cloud: Raspberry Pi executes python scripts for communicating with servers and sensors. For sending and receiving data from the servers several APis (Application program interface) are developed which allows to send and receive a particular data by hitting a URL. The server sends back a HTTP Response each time the URL gets hit. It would have required two URL hits, one for sending and one for receiving the data to the cloud but the URL was altered in according to the parameters and thus turning the number of required URLs to one.

The current parameters of the warehouse are captured by the sensors and sent to Raspberry Pi. The parameters are then encoded and encapsulated in a URL by the Raspberry Pi. Then, the Pi hits the URL and waits for the response. Server on receiving the request decodes the URL and extracts the parameter sent. It saves it into a csv file unique to a session. Server also sends a HTTP response corresponding to that request which contains the parameters that the warehouse should have. Same URL hit is also used to indicate the Raspberry Pi that the warehouse should be turned off or on.

3) Controlling: Warehouses environment parameters can be controlled and monitored from anywhere through its website or android application. These parameters can be initialized by the warehouse operator through the website/android application by entering the name of the fruit to be stored and the date till which fruits are to be stored. Website will then tell the operator optimum parameter values for the storage of that fruit for the specified number of days and send these parameter values to the warehouse temperature control unit. These optimum values are referred from [14]

4) Monitoring: The system lets the user to both visualize and analyze the data. As mentioned, the server saves the parameters sent by the micro-controller to the Raspberry Pi. When the user chooses to monitor the data in the website. The server draw plots using matplotlib (a plotting library in python) by reading the respective .csv file. Different plots are drawn corresponding to each parameter sent by the Raspberry Pi. In each case, a time vs value plot will be drawn for the parameter. The plots will update themselves corresponding to each new entry they receive from the Raspberry Pi. When the warehouse is stopped, the server saves the data for the session. The user can again see all the data of a particular session

in the history tab of the website. The user can also see the parameters that are currently being recommended to Raspberry Pi and change them accordingly. As the server stores the file in .csv format, the data can be read and visualized in any way possible. Drawing a plot of time vs value of the parameters is just one of the ways how the data can be visualized.

5) Generating Distress Signal: The warehouse can detect any malfunctioning in warehouse whatsoever and generates distress signal. The distress signal is generated in the following cases:

- Power Failure
- Sensing node Failure
- Circuit Failure
- Deteriorating quality of stored fruits
- Manual errors leading to a change in warehouse parameters

The detailed working model of the proposed system is shown in Fig. 3.

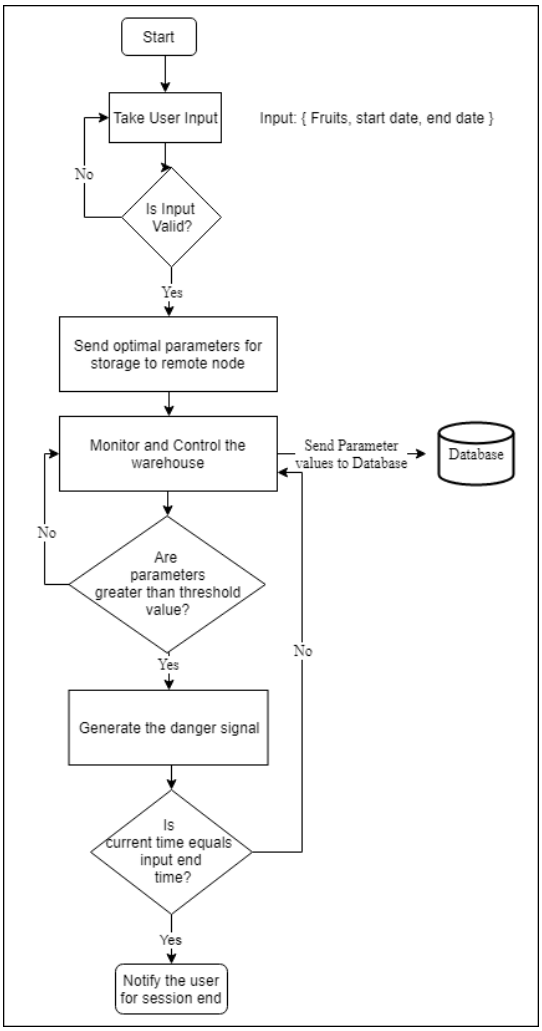


Fig. 3. Flowchart depicting the working of the model

III. Experiment

The warehouse prototype consists of a cardboard box, inside which are sensors, i.e. DHT11 and GY-30, Raspberry Pi, a fan, motor driver and a water heater as shown in Fig. 1 fan and water heater mirror the changes in environment of an actual warehouse i.e fan mirrors the refrigeration units and the heater brings variation in temperature.

An apple was stored in this warehouse prototype for 2 days and the temperature, humidity and light intensity changes were recorded by the sensors continuously. A local computer and Raspberry Pi were connected to the same network and the computer acts as a server to receive, process and send relevant data to the Raspberry Pi. User started the warehouse in the website specifying the type of fruit to be stored and the duration of storage. Raspberry Pi then received the optimum environment parameter values from the server, which is in this case 35, relative humidity of 93% and a storage limit of 12 months. User could also change these values as per his choice from the website, anytime he wants during the operation of the warehouse. Heater spikes the temperature values for a small amount of time, to which the Raspberry Pi reacts by switching on the fan to cool down the temperature to optimum values provided by the server. This replicates the refrigeration mechanism reacting to unforeseen temperature changes in the warehouse.

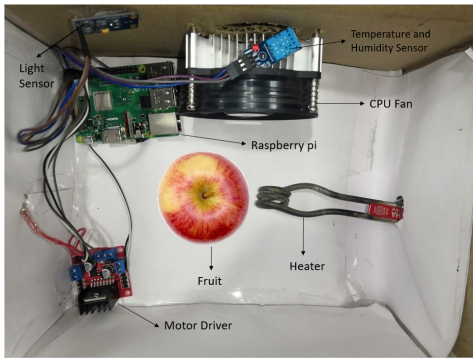


Fig. 4. A view of proposed prototype

Fig. 4 shows the prototype developed to monitor and control the warehouse. A heater is installed to bring variation in temperature. The fan responds to the variation and keep the temperature in acceptable levels. After 2 days, the warehouse server stopped sending parameter values to Raspberry Pi and all the recorded data is saved in .csv file. The recorded data is presented to the user as raw data as well as plotted in a parameter vs time graph, during the operation of the warehouse and is also stored in history log, for future reference.

IV. Result and Analysis

Fig. 5 shows the data stored in the server during the execution of the program i.e. for the time interval in which the fruit is stored in the warehouse. This history of data recorded can be used by the warehouse owner to monitor the warehouse parameters through a long-time interval.

	temperature	humidity	light	date
0	30	39	0	2019-04-04 01:00:48.818850
1	30	41	145	2019-04-04 01:00:50.028636
2	30	40	146	2019-04-04 01:00:51.258578
3	30	39	145	2019-04-04 01:00:53.541196
4	31	38	146	2019-04-04 01:00:56.829084
5	30	39	145	2019-04-04 01:00:57.959004
6	30	39	145	2019-04-04 01:00:59.090493
7	30	39	146	2019-04-04 01:01:00.214448
8	30	39	146	2019-04-04 01:01:04.404956
9	30	39	145	2019-04-04 01:01:05.611582
10	31	40	145	2019-04-04 01:01:06.826743
11	30	38	145	2019-04-04 01:01:10.125892
12	30	39	145	2019-04-04 01:01:11.330761
13	30	39	146	2019-04-04 01:01:12.551548
14	30	39	145	2019-04-04 01:01:13.792372
15	30	39	146	2019-04-04 01:01:17.165468

Fig. 5. Data collected from the experiment

The transmitted parameter values are visualized in our website using matplotlib. Matplotlib is a Python 2D plotting library which produces various kinds of plots that are used to reference raw data. Fig. 6, 7 and 8 shows the plot of temperature, humidity and light intensity values respectively recorded by the sensor nodes against time. These plots can be used, instead of the raw data, to have a better understanding of the trends in the warehouse parameters.

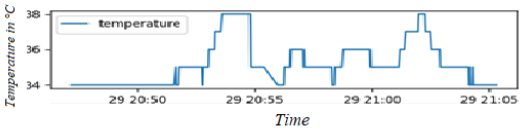


Fig. 6. Time vs Temperature graph

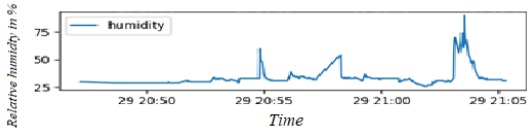


Fig. 7. Time vs Humidity graph

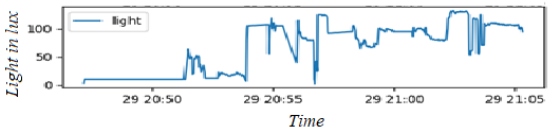


Fig. 8. Time vs Light graph

Fig. 6 shows time vs temperature graph of the warehouse during storage of apples. The time is in format DD HH:MM where DD represents the date, HH represents the hour and MM represents the minutes Temperature plays a major role in warehousing of fruits and so it should be maintained strictly in the acceptable range, to keep the fruits healthy and fresh. But due to various reasons such as climate change, power failure or human errors, warehouse temperature can change. As soon as this happens, refrigeration mechanism of the warehouse kicks in and reduces the temperature to acceptable range. To simulate this change and restoring, our prototype used a water heater which increases the temperature of the cardboard box, and the fan which brings it again to the optimum temperature range provided by the server. Temperature rising and falling can be seen in the graph.

Fig. 7 shows relative humidity vs time graph of the warehouse. Relative humidity influences growth of microorganisms on the fruits, and so it should be monitored. Relative humidity is the amount of water vapour in the air with respect to the saturation point of the water vapour in air. It is after this saturation point that water starts to condense over surfaces forming dew.

At higher temperatures, saturation point of water vapour in the air also rises which means air can hold more water vapour when its hot and thus will have lower relative humidity which can be seen in Fig. 6 and 7.

Fig. 8 shows light intensity vs time graph of the warehouse. Light intensity is being measured in Lux which equals to one lumen per square metre. Light intensity, while being of less importance in storage of apples, plays a major role in warehousing of some other fruits such as kiwis

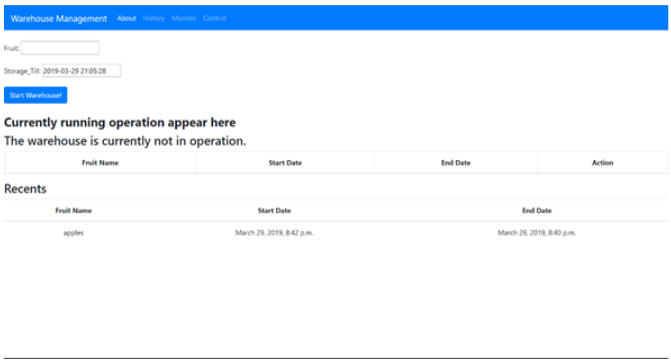


Fig. 9. View of web interface

Fig. 9 presents a web-page of the website that was created as an interface for controlling and monitoring the warehouse. There are four tabs in the website which are about, history, monitor and control. The about tab displays the information about the components and credits. The history tab shows the data that was stored in all previous sessions. The Monitor tab in website shows the graphs at real time for the parameters in the warehouse. The control tab lets the user to change the current parameters of the warehouse. It also shows the current parameters that is maintained in the warehouse.

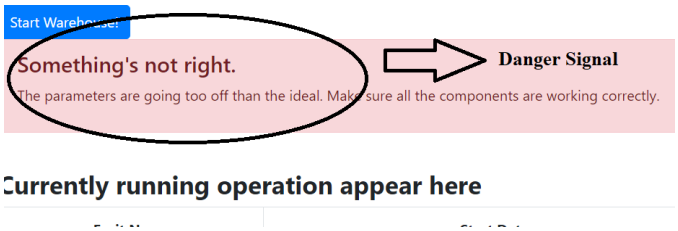


Fig. 10. Danger signal to notify malfunction

Fig. 10 shows the danger signal that was generated during the experiment. The signal notifies the user of any physical malfunctioning in the warehouse. The offset during the time of experiment was set to 15 centigrade for temperature i.e if the system detects that the current temperature of the warehouse is off by 15 centigrade than the ideal storage conditions then it will notify the user in the homepage by displaying the message: ”Something’s not right, The parameters are going too off than the ideal. Make sure all the components are working correctly” in a red coloured strip. The owner of the warehouse can take corresponding actions.

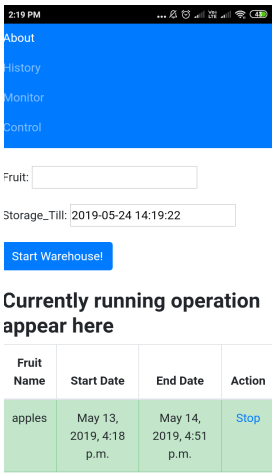


Fig. 11. View of Android Application

Fig. 11 shows the android mobile application screen-shot for the warehouse system. The application was tested on android sdk 27 (Android 8.1) and sdk 28 (Android 9). The mobile application offers the same functionality adding the comfort of operating the warehouse from a smart phone with internet connection.

The comparison of proposed model and existing model is given in Table I. It can be seen that proposed model surpasses other existing solutions by eliminating their various limitations.

TABLE I
COMPARISON OF PROPOSED SOLUTIONS AND EXISTING SOLUTIONS.

Factors	**[2]**	**[3]**	[4]	**Proposed Solution**
Automated Control	No	No	No	Yes
Automated Monitoring	Yes	Yes	Yes	Yes
Centralized	No	No	No	Yes
Interface Provided	No	No	No	Yes
Malfunction Detection	Somewhat	No	No	Yes
Manual Overwrite	Yes	Yes	Yes	Yes
Human Interaction	Yes	Yes	Yes	No

V. CONCLUSION

When the world is used to manually controlling automatic machines, our solution provides a robust, efficient and scalable alternative to the traditional warehousing techniques. Humans operating machines provides desired output at the cost of an increased chance of errors. We have introduced a way of automating the control of warehouses environment parameters thereby reducing chances of manual errors to zero. By providing the optimum number of days for storage of goods in the warehouse, were eliminating the probability of over-ripening of the fruits due to prolonged storage as well as making their storage economically feasible. Centralized monitoring and processing unit makes the supervision of the warehouse, easy without giving excessive processing weight on the local microcontrollers. That being so, the solution provides the modern-day alternative to traditional warehousing systems overcoming all their shortcomings. In future, the solution can benefit from artificial intelligence with machines learning on the recorded data. Machines will be learning warehousing patterns and thus can be used to estimate the number of days a fruit will remain of quality under certain environmental conditions.

REFERENCES

[1] Food and Agriculture Organisation. [online] Available: http://www.fao.org/india/fao-in-india/india-at-a-glance/en/

[2] Roy, Ananya, Prodipto Das, and Rajib Das. ”Temperature and humidity monitoring system for storage rooms of industries.” 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN). IEEE, 2017.

[3] Shubham Khumkar, Akshay Bhujbale, Sadanand Khandar, Shubham Deshmukh, and Mahendra Pund. ”IoT Based Monitoring And Control For Vegetables And Fruits Storage.” International Journal Of Advance Research And Innovative Ideas In Education 4.2(2018) : 4486-4490.

[4] Liu, Jun, et al. ”Design of fruits warehousing monitoring and control system based on WiFi.” 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2015.

[5] Sowmya T K, Shreya V Agadi, Saraswathi K G, Puneeth B Nirvani, Prajwal S, 0, Implementation of IoT based Smart Warehouse Monitoring System, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) ICRTT 2018 (Volume 06 Issue 15),

[6] The open data platform for the Internet of Things, [online] Available: https://thingspeak.com.

[7] Sabanc, Kadir, et al. ”Thingspeak Based Monitoring IoT System for Counting People in A Library.” 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE, 2018.

[8] Jain, Sarthak, Anant Vaibhav, and Lovely Goyal. ”Raspberry Pi based interactive home automation system through E-mail.” 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). IEEE, 2014.

[9] Frontpage-Raspbian, [online]. Available: https://www.raspbian.org.

[10] DHT11 Temperature and Humidity Sensor [online]. Available: https://www.robot-r-us.com/sensor-temp/humid/dht11-temperature-and-humidity-sensor.html.

[11] Pan, Yun. ”Application of Intelligent Lighting System in Sheep Farm Based on CC2530-WiFi.” 2017 7th International Conference on Applied Science, Engineering and Technology (ICASET 2017). Atlantis Press, 2017.

[12] Django (web framework), [online] Available: https://en.wikipedia.org/wiki/Django_(web_framework)

[13] Android software development. [online] Available: https://en.wikipedia.org/wiki/Android_software_development.

[14] The University of Maine. [online] Available: https://extension.umaine.edu/publications/4135e/

Abnormal Crowd Behaviour Detection Using 2-stream Deep Neural Networks

Muhammed Anees V
Department of Computer Science
Cochin University of Science and Technology
Kerala, India
anees@cusat.ac.in

Santhosh Kumar G
Department of Computer Science
Cochin University of Science and Technology
Kerala, India
san@cusat.ac.in

***Abstract*—The safety of human beings during a public event and crowded area is one of the major headaches of the concerned authorities. The authorities need to monitor the entire crowd continuously, and they are responsible for preventing all the abnormal activities in the crowd. To prevent abnormal behaviour, first they need to detect the abnormal crowd behaviour from the high-density crowd under observation. Detecting abnormal crowd behaviour has been one of the most important research areas in the intelligent video surveillance system field over the past few years. Numerous strategies for crowd abnormality detection with the assistance of computer vision and machine learning methods have been proposed in recent years. Many of those traditional approaches are using handcrafted features like optical flow, HoG, SIFT, SURF etc. Although most of these methods were able to produce a considerably good performance, however, these methods will take a lot of computational time to extract features and that eventually increases the whole computational time. In this paper, we propose a novel deep learning strategy for abnormal crowd behaviour detection. Rather than utilizing handcrafted features, deep neural networks naturally learn feature representations and which will help the system to detect abnormal behaviours. This method uses convolution neural networks which have been utilized as an integral tool for learning purpose in computer vision algorithms, to extract the features from the videos. We have used pre-trained 2 stream convolution neural networks to detect the crowd abnormality, which can consider both spatial and temporal information in the video. Our method is tested with available standard datasets and compared with state-of-the-art methods.**

***Index Terms*—Crowd flow;Surveillance;Optical flow;Crowd Abnormal Detection;Deep learning;Convolution neural network**

I. Introduction

The utilization of surveillance systems in public spaces has increased dynastically in last few years due to the security needs of the citizen [23]. The intelligent surveillance system is on the most significant advancement in the field of the surveillance system, which can make capable of detecting the abnormalities without human intervention. Crowd analysis needs to be applied in surveillance systems to enable its intelligence. Crowd analysis is one of the critical research areas in the field of computer vision which includes crowd monitoring, crowd tracking, action recognition, abnormality detection etc.

978-1-7281-2087-4/19/$31.00

Detect abnormal activities from the crowded scene is one of the fundamental challenges while developing an intelligent video surveillance system. In recent years, many computer vision techniques are employed for detecting abnormalities from the surveillance videos. Abnormality detection from the surveillance video is challenging and complicated because it is tough to define the abnormality. An ordinary event in a scene may be abnormal in some other views and because of this nature abnormality can be considered as a subjective behaviour [15].

Traditional methods usually modelled normal behaviour patterns in the video scene using mathematical models and detect abnormal behaviours by considering these normal patterns. These types of approaches cannot identify all the abnormalities present in a view due to its subjective nature. In the crowd scenario, the anomalies may be formed by some rare or non-frequent movements. Identifying this unusual or non-frequent movement from the entire image or video is a massive task for a surveillance system. Instead, the methods in the literature approaches divided the scene into the number of patches and tried to detect the abnormalities in each piece to create the model. These usual model act as the reference model for identifying normal behaviours, and if any patch does not follow this predefined reference, then it can be considered as an outlier and this outlier may be an abnormal event. Many methods in the literature are already pursuing this strategy which uses well-known methods like a histogram of gradients (HoG) and optical flows with some stability analysis methods [1].

In this paper, we have investigated the application of a convolutional neural network (CNN) for the detection of abnormalities from the surveillance video. Here, we have used two stream convolutional neural networks which runs in parallel to extract both spatial and temporal information from the crowd videos. Pre-trained VGG network was used in both streams to train the given data

This paper is organised as follows. Section 2 describes the related works, section 3 will discuss the architecture of the neural network. Section 4 describes the experimental set-up and the result analysis. Finally, the conclusion is given in section 5.

II. Related Work

The challenges in developing an intelligent surveillance system have been addressed by computer science researchers for the last few years with the evolution of high-performance computing device. Most of the traditional methods in the literature use both low level and high-level handcrafted features like optical flow and histogram of the gradient for the abnormality detection in crowd scenes [22] [3]. Some initial works in this field use these handcrafted features to develop some standard models to detect the crowd abnormality. Mehran et al. [10] proposed an abnormality detection method which uses a standard model called the social force model, and it is considered as one of the pioneering work in this field. Ali and shah [2] proposed another model which uses Finite-Time Lyapunov Exponent (FTLE) filed for the detection of abnormalities. Krausz et al. [8] use the histogram of optical flow to represent the crowd behaviour patterns, and the anomaly events are detected from these motion patterns using a heuristic approach. Ragavendra et al. [13] proposed an abnormal event detection method using interaction forces with the help of the particle swarm optimization algorithm. Xiong et al. [21] proposed an energy-based model for crowd abnormality detection using potential energy and kinetic energy.

Abnormal event detection from an unstructured crowd is more complicated than detecting abnormalities from the structured crowd. Wu et al. [20] first address the issue of abnormality detection from the instructed crowd using particle advection system. The abnormal event detection system needs to consider the spatiotemporal features for the adequate representation of unusual events. Mausavi et al. [11] propose a histogram based abnormal event detection method which considers spatio temporal movements of the crow. The methods we have listed till now considers only the handcrafted features for modelling the crowd flow. But recently proposed methods considers the combination of both features and textual information. Li et al. [6] use both the textual information and spatio temporal features for the detection of abnormal events. Kalsta et al. [7] proposed an interesting work in this field which uses Histogram of Swarms (Hos) as textual information and Histogram of Gradient as the feature (HoG). All these described methods are working on the features extracted from the videos. However, these methods have some limitations and drawbacks. Most of these features are designed to use in general purpose images, and we are trying to incorporate those features in our intelligent surveillance system. When we are applying those feature extraction mechanism in the videos, it may not perform well. The second biggest challenge is the selection of an appropriate, suitable feature for the scenario. If are selecting a wrong feature for developing the system for a particular scene, then the entire system may be a failure.

Recently Convolution neural network is proved to be effective in addressing of many challenging problems in various fields, such as problems in image processing based on classification like image classification [9], object detection [14] and activity recognition [17]. The approaches based on CNN can perform significantly better than traditional methods. Due to these improvements, we are using a convolution neural network for crowd abnormal detection. The first work that uses deep learning framework for the detection abnormality was proposed by Feng et al. [5]. But they haven't used a deep neural network for their method, instead, the method was implemented using deep Gaussian Mixture Model (GMM). Sabokrou et al. [16] introduced a crowd abnormality detection method that uses fully convolution neural networks. Wei [19] improved this method using a modified deep neural network called two stream fully convolution neural network. These are the latest work available in the literature.

In this proposed method we have implemented the crowd abnormality detection using Pre-trained two stream VGG16 network and the results are also compared with the mile stone works available in the literature.

III. Architecture

The architecture of the newly proposed crowd abnormality detection system is given in figure 1. The emerging deep learning techniques are used in our work to detect abnormal crowd videos. Instead of the widely used single stream convolution neural network, We have used 2-stream convolution neural networks in our work to detect the abnormalities. The single stream of layers is usually used in deep learning techniques for implementing convolution neural networks. But in this work, Two parallel streams of networks are used to learn more features than usual single stream networks. The raw frames taken from the input video is used for the first stream and optical flow frames are used for the second stream. Stream 1 is used to learn the appearance information or spatial information from the normal frames extracted from the input video and stream 2 is used to learn the motion information or temporal information from the corresponding optical flow frame.

The crowd videos are given to the proposed abnormal detection system as the input. Videos can be considered as the continues flow of image frames so that we can extract fixed number of frames from the input video. We have used the raw frames obtained from the input video for training the first stream for learning the spatial information. Before fed into the training module, some preprocessing steps need to be applied. In the preprocessing steps, the extracted frames are resized in to 299×299 pixels. These resized frames are fed into deep learning framework for for training. In this problem, we are using pre-trained VGG -16 network [18] for training purpose.

VGG is convolution neural network proposed by the Visual Geometry Group in the University of Oxford. VGG has two versions, VGG 16 and VGG 19. VGG 16 is a pre-trained model with 16 weight layers, whereas VGG 19 is a model with 19 weight layers.The size of input layer in VGG is $224 \times 224 \times 3$. Input layer to the last max pooling layer is the feature extraction part and the rest of the network is considered as classification part of the model. Softmax layers are usually used for the final classification with the respective number of

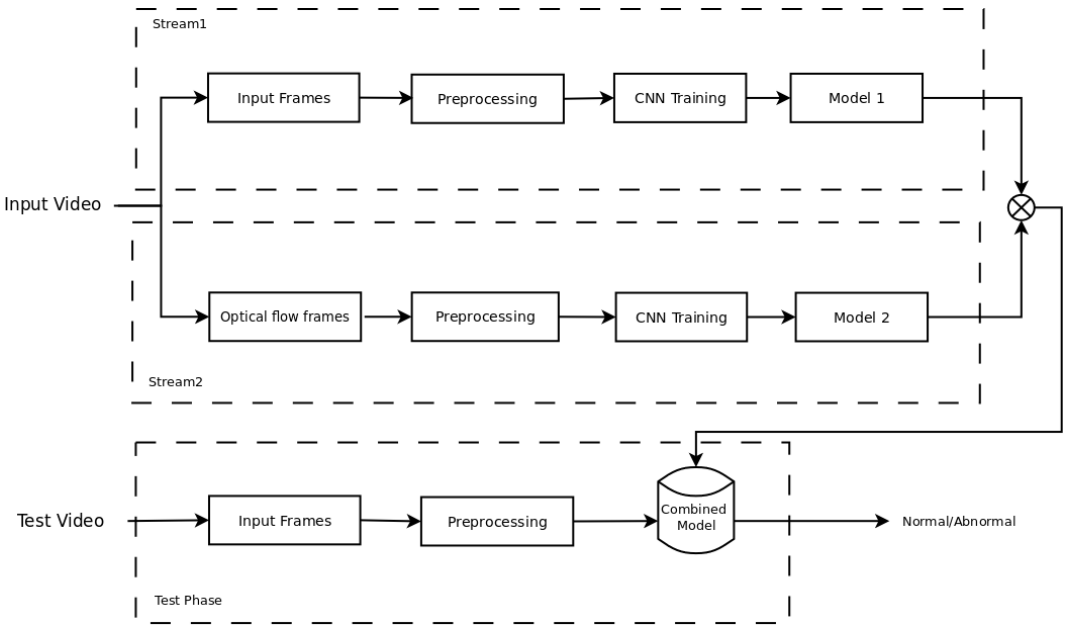


Fig. 1. Proposed System Architecture

classes. In our problem, the output layer is a softmax layer with two classes.

The first Convolution stream gives us the first convolution model and this model is created based on the real images extracted from the given video input. This model can represent only the spatial information in the video, but it cannot represent the temporal information in the scene. We need both spatial and temporal information of the data to analyse the exact crowd behaviour from the input crowd video. To represent the temporal information, we are using the second convolution stream which takes optical flow frames as input.

Optical flow is a 2D motion vector which shows the motion pattern resulted from the movements between two consecutive frames. We are using Gunner Farneback optical flow [4] algorithm to extract optical flow frames from the input crowd video. Gunner Farneback is used to calculate the dense optical flow present in the video. Sparse techniques only need to process some pixels from the whole image, dense techniques process all the pixels. Dense methods are slower then sparse but can be more accurate. We are passing these extracted optical flow frames to the same VGG-16 network that is explained in the above section to create the second stream. The second stream outputs the second convolution model and which can represent both the spatial and temporal contents of the video.

Now we have two convolution models obtained from two independent convolution streams running in parallel. To incorporate both spatial information and temporal information in a single convolution model, the spatial model and the temporal model are merged to create a new convolution model. This new model contains the properties of both the spatial model and the temporal model and this combined model can represent both spatial and temporal information of the data. So we get the combined model from the raw frames and optical flow frames, and the coupled model can detect the abnormal events present in the video.

IV. Experimental Results and Analysis

In this section,we will discuss the experimental set up. and analyis. We have used fight dataset [12] released by VISILAB. This dataset is released specifically for evaluating and measuring the crowd abnormality detection systems. This dataset considers both normal and abnormal events occurring in similar, but dynamic conditions. This dataset contains two sets of videos. The first video set is collected from National Hockey League (NHL) of Spain and the second set is collected from the various action sequence of different movies. This dataset contains total 246 videos from both normal and abnormal category. 123 videos are fall in the normal category and the remaining 123 videos are fall under abnormal category. Each video in this dataset contains 50 frames of size 720×576. 98 videos from each class is taken for training and the remaining 25 videos are taken for testing in both normal and abnormal class. The entire system runs for 200 epochs in the training phase.

We have tested our dataset with both single-stream network and double stream networks. We have passed the training data in to VGG network to create the first convolution model. Then the optical flow frames are extracted from the training videos and passed to the same VGG network to create the second model. Then the two models are combined to get more refined and accurate model for detecting the abnormalities from the crowd video. Then the model is tested with our test data set that contains 50 test videos. 24 out of 25 abnormal videos are classified as unusual, and 20 videos out of 25 normal videos are classified as normal using two-stream model. The

confusion matrix created from these results are given in the table I.

TABLE I
CONFUSION MATRIX

	Actual Normal	**Actual Abnormal**
Predicted Normal	TP = 24	FP = 5
Predicted Abnormal	FN = 1	TN = 20

We can calculate precision, recall, F1- score and Accuracy from the confusion matrix. We got precision as 85.71% , Recall as 96.00%, F1 -score as 90.57% and Accuracy as 90.00%. We have compared the two stream CNN results with both single stream networks. We have tested our dataset with both spatial and temporal models. The spatial model gives us the precision as 81.48% , Recall as 88.00%, F1 -score as 85.62% and Accuracy as 84.00%. The temporal model gives us the the precision as 85.19% , Recall as 92.00%, F1 -score as 88.46% and Accuracy as 88.00%. The results are tabulated in the table II.

TABLE II
COMPARISON OF SINGLE STREAM AND TWO STREAM NETWORKS

	Precision	**Recall**	**F1-Score**	**Accuracy**
Spatial Stream CNN	81.48	88.00	85.62	84.00
Temporal Stream CNN	85.19	92.00	88.46	88.00
Two Stream CNN	85.71	96.00	90.57	90.00

From the above table, we can understand that two stream convolution networks can outperform both single-stream networks. The entire system is implemented using OpenCV and Keras libraries in Python. The experiment is executed in a system with NVIDIA K80 GPU.

We have compared the proposed model with existing state-of-art models in the literature. Most of the works in the literature are based on single stream convolution networks that take raw input frames. All those methods are tested with our dataset and the comparison results are as shown in the table III. From this figure, we can understand that our method can outperform all the crowd anomaly detection methods in the literature.

V. CONCLUSION

In this paper, we have proposed a deep learning framework for abnormality detection from crowd videos. Instead of using single stream neural network, we have used two stream neural networks for abnormality detection. The proposed two stream CNN uses pre-trained VGG-16 model for creating the model. We have tested our model with fight dataset released by VISILAB. Experiment results provide around 90% accuracy for two stream CNN. Our proposed two-stream model for crowd abnormality detection outperforms all the existing crowd models in the literature, and we can apply this model in real-time applications. This system can be improved to get the type of crowd behaviour by adding those details in the learning phase.

TABLE III
COMPARISON STUDY

Precision	**Recall**
state of art methods	Performance
Sabokrou et al.	69 %
Dan Xu et al.	75 %
Spatial	84 %
Temporal	88 %
Wei's method	88 %
Proposed method	90 %

REFERENCES

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[2] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[3] Yang Cong, Junsong Yuan, and Ji Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, 2013.

[4] Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29 – July 2, 2003 Proceedings*, pages 363–370. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[5] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. Learning deep event models for crowd anomaly detection. *Neurocomputing*, 219:548–556, 2017.

[6] H. Guo, X. Wu, N. Li, R. Fu, G. Liang, and W. Feng. Anomaly detection and localization in crowded scenes using short-term trajectories. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 245–249, Dec 2013.

[7] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis. Swarm intelligence for detecting interesting events in crowded environments. *IEEE Transactions on Image Processing*, 24(7):2153–2166, July 2015.

[8] B. Krausz and C. Bauckhage. Analyzing pedestrian behavior in crowds for automatic detection of congestions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 144–149, Nov 2011.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[10] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, June 2009.

[11] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analyzing tracklets for the detection of abnormal crowd behavior. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 148–155, Jan 2015.

[12] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part II*, CAIP'11, pages 332–339, Berlin, Heidelberg, 2011. Springer-Verlag.

[13] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino. Optimizing interaction force for global anomaly detection in crowded scenes. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 136–143, Nov 2011.

[14] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[15] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. Real-time anomaly detection and localization in crowded scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2015-October:56–62, 2015.

[16] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. Deep-anomaly : Fully convolutional neural network for fast anomaly detection in crowded scenes Mohammad Sabokrou. *Computer Vision and Image Understanding*, (January):0–1, 2018.

[17] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[19] Hongtao Wei, Yao Xiao, Ruifang Li, and Xinhua Liu. Crowd abnormal detection using two-stream Fully Convolutional Neural Networks norm single frame norm optical flow norm. 2018.

[20] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2054–2060, June 2010.

[21] G. Xiong, X. Wu, Y. Chen, and Y. Ou. Abnormal crowd behavior detection based on the energy model. In *2011 IEEE International Conference on Information and Automation*, pages 495–500, June 2011.

[22] Y. Zhang, L. Qin, H. Yao, P. Xu, and Q. Huang. Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition. In *2013 IEEE International Conference on Image Processing*, pages 3572–3576, Sept 2013.

[23] Ying Zhang, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Combining motion and appearance cues for anomaly detection. *Pattern Recogn.*, 51(C):443–452, March 2016.

A Hybrid Binary Classifier for Pattern Classification

Kaumil Trivedi
AI Developer, IT Department
Nebula Infraspace LLP
Ahmedabad, India
kaumil.trivedi97@gmail.com

Tanujit Chakraborty
SQC & OR Unit
Indian Statistical Institute
Kolkata, India
tanujit_r@isical.ac.in

***Abstract*—In this work, we propose a hybrid binary classifier which combines a decision tree with a support vector machine. The proposed hybrid model has the advantages of improved accuracy and easy interpretability. The model will be useful for feature selection cum classification tasks in real-world supervised learning problems. Numerical evidence is also provided using 25 standard data sets from various fields to assess the performance of the model. Performance of the proposed hybrid binary classifier is quite better when compared to individual classifiers.**

***Index Terms*—Decision tree, Support vector machines, Hybrid model, Pattern classification.**

I. Introduction

Nonparametric models are very useful in the fields of machine learning and artificial intelligence for more than thirty years now. These prediction models can deal with high dimensional feature spaces and complex data structures. Two widely used classifiers are decision tree (DT) and Support vector machine (SVM). DT is more robust when limited data are available, and it has a built-in feature selection mechanism [1]. DT is defined by a hierarchy of rules which can be used to select essential features from the input feature space for any particular data set. But trees are high variance estimators and greedy algorithm [2]. SVM is a powerful supervised machine learning algorithms that use kernels to allow substantial flexibility for the decision boundaries, leading to better prediction performance [3]. But SVM has the drawback of being sensitive to the choice of kernel parameters [4]. Another major shortcoming of the SVM is the complexity of the algorithm when the number of support vectors increases. Due to that, SVM is sometimes considerably slower than the other techniques [5]. To utilize the positive aspects as well as to overcome the drawbacks of these two robust pattern classifiers, various hybrid models are often used to make decisions [6].

In recent years, several hybrid approaches have been evolved to combine DT and SVM. The first idea of hybridization in which SVM is used for each decision in the tree is popularly known as SVMtree model [7]. SVMtree results in a simple decision tree with multivariate linear and nonlinear decisions. The model has further been extended for multivariate classification problems [8] [9], [10]. Many other works in the field of hybridization of DT and SVM can also be found in previous literature, for example, see [11], [12], [13], [14]. Most existing methods overcame the issue of overfitting in the DT using SVM model or tried to build "optimal" tree model with the use of SVM. But this paper focuses on the reduction of the number of support vectors for the SVM using DT. Our goal is to improve the predictive behavior of both the models using a novel hybrid formulation. The resulting hybrid model can thus be used for feature selection cum classification tasks in a supervised set up. In this approach, the DT classifier is trained to extract important features from the data set. Further, the SVM classifier is trained using the essential features selected by DT along with the predicted results of the DT algorithm on the modified data set.

The rest of the paper is organized as follows. In section 2 and 3, the methods used in this study are discussed. Section 4 is entirely devoted to the experimental evaluation on 15 distinct publicly available data sets from various applied fields. Finally, Section 5 concludes the paper with a note on the future scope of research.

II. Background

In this section, we first review the fundamental ideas of DT and SVM and then present a brief details that are pertinent to the formulation of the proposed hybrid model.

A. Decision Tree (DT)

DT is a greedy divide-and-conquer algorithm that finds axis-parallel partitions via recursive partitioning the feature space into homogeneous regions. The construction of binary DT starts with assigning the total training data points in one group, named as the parent node. Parental nodes are split into two child nodes using one of the feature vectors. The selection of attributes is made based on any entropy based impurity function. DT usually uses Gini impurity to measure the quality of the splits, and the best split amongst all the splits at a given branch is selected. The splitting process ends when a full tree is grown. Also, different tree complexity measures are usually employed to prune the branches with very few data points to avoid data over-fitting. Finally, we can assign the class labels for each terminal or leaf nodes.

B. Support Vector Machine (SVM)

SVM was developed by Vapnik (1995) for solving binary classification problems. The rationale behind construction of a binary SVM is finding a separating hyperplane in the input

978-1-7281-2087- 4/19/$31.00 © 2019 IEEE

feature space that can divide the training sample into two classes with minimal error.

Suppose we have $X = \{X_1, X_2, ..., X_p\}$ be the p-dimensional feature space and Y be the binary response variable taking values $\{1, -1\}$ on n training samples. Then the linear separating hyperplane in the feature space is $\{x : f(x) = w^T x + b = 0\}$, where $w \in R^p$, b is a scalar and $f(x)$ is the decision function. We make the decision as follows: $\hat{f}(x) > 0$ implies $\hat{y}(x) = 1$ and $\hat{f}(x) < 0$ implies $\hat{y}(x) = -1$. Essential parameters (w, b) can be obtained by solving the following optimization problem:

$$\begin{aligned} &\text{Minimize} \quad L(w) = \frac{1}{2}||w||^2 \\ &\text{subject to} \quad y_i f(x_i) \geq 1, i = 1, 2, ..., n. \end{aligned} \tag{1}$$

Sometimes nonlinear hyperplanes are used for separating the data into two classes. Kernel functions $Ker(x_i, x_j)$ are used to define the nonlinear transformation. Therefore, the decision rule becomes:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i Ker(x_i, x) + b,$$

where α_i and b can be estimated by solving the optimization problem in (1). Gaussian kernel is one of the popular choice of kernel function, can be defined as:

$$Ker(x_i, x_j) = e^{-\beta|x_i - x_j|^2}, \text{ where } \beta \text{ is scale parameter.} \tag{2}$$

III. Proposed Hybrid Model

We propose a hybrid model based on DT and SVM that can utilize the positiveness of both the models and overcome their drawbacks. The model can prevent DT from overfitting and will be useful for limited data sets. The formulation of the proposed hybrid model is as follows (also see the flow diagram in Figure 1).

- First, we train the data set using a DT model and build a tree that calculates essential features from the input feature spaces.
- The prediction result of the DT algorithm is also recorded and kept as an additional feature in the input feature space of the SVM model.
- We build the SVM model with the essential input variables obtained from the DT algorithm along with the additional input variable.
- For nonlinear classification problems, run the SVM algorithm with Gaussian kernel function and record the classification outputs.

The proposed model can be considered as a two-step pipeline approach: (1) it selects important features using DT and record the predicted results of DT algorithm; (2) these are together used for further model building using SVM. The hybrid model will be useful for identifying essential features that can satisfy a specific goal of pattern classification. Further, the number of support vectors will be less in our model, an added advantage of the proposed model. The hybridization approach proposed here are expected to improve the prediction performance of the individual classifiers for feature selection and classification problems.

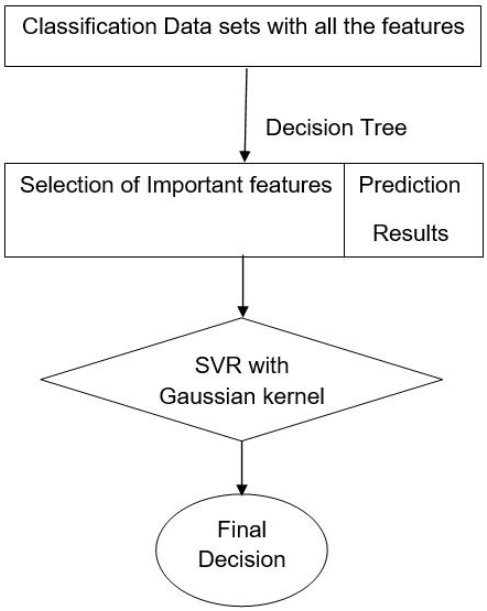


Fig. 1. Flow diagram of the proposed hybrid model

A. Merits

1) The proposed hybrid model uses the feature selection mechanism of DT. Thus, one needs to work with small set of features as inputs to SVM. This makes the model less complex and fast while actual implementation.
2) Incorporation of DT predicted results in the input feature space of SVM play an important role in the hybridization. It works like a guide for pattern classification in the SVM model and improves the performance of the hybrid classifier in a significant margin.
3) The propose hybrid pattern classifier is easily interpretable, simple, fast, and accurate.
4) The hybridization will be most useful for feature selection cum classification problems having small, medium, and large sized data sets.

B. Demerits

1) Situations when feature selection is not a task in pattern classification, like lab experiment data sets, NASA software defect prediction data sets, the proposed model may not be too useful.
2) For complex problems, where tree based model DT may become too large and less interpretable, the algorithm may not work work as expected. Even SVM takes long training time for large datasets which is another disadvantage of the proposed hybrid classifier.

TABLE I
CHARACTERISTICS OF THE BINARY CLASSIFICATION DATA SETS

Data	Total Objects (n)	Total number of Features	Number of ($+$)ve instances	Number of ($-$)ve instances
breast cancer	569	33	357	212
credit card	30000	25	23364	6636
hepatitis	155	20	123	32
htru-2	17898	9	16259	1639
hungarian-14-heart	294	14	188	106
hypothyroid	3163	26	151	3012
ionosphere	351	34	225	126
kr-vs-kp	3196	73	1669	1527
magic gamma	19020	11	12332	6688
musk-1	476	168	269	207
musk-2	6598	168	5581	1017
nomao	34465	120	24621	9844
occupancy	20560	7	15810	4750
ozone	2534	73	2374	160
parkinsons	195	21	48	147
pima diabetes	768	8	500	268
planning	182	12	130	52
sonar	208	60	97	111
spambase	4601	57	2788	1813
spect images	267	22	55	212
statlog (heart)	270	13	150	120
tictactoe	958	27	626	332
thyroid sick	3772	52	3541	231
twonorm	7400	28	3703	3697
vote	435	16	267	168

IV. EXPERIMENTAL EVALUATION

A. The datasets

The proposed model is evaluated using 25 publicly available standard binary classification data sets from Kaggle (https://www.kaggle.com/datasets) and UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets.html). These data sets have limited number of samples with reasonably large number of features [15]. Table 1 gives a summary of all these data sets.

B. Performance metrics

The performance metrics to be used in the experimental analysis are as follows:
Classification Accuracy = $\frac{(TP+TN)}{(TP+TN+FP+FN)}$;
F_1-score =$2 \times \frac{\left(\text{Precision}\times\text{Recall}\right)}{\left(\text{Precision}+\text{Recall}\right)}$;
where, Precision = $\frac{TP}{TP+FP}$; Recall = $\frac{TP}{TP+FN}$; and TP (True Positive): correct positive prediction; FP (False Positive): incorrect positive prediction; TN (True Negative): correct negative prediction; FN (False Negative): incorrect negative prediction. Higher the value of these metrics, the better the classifier is.

C. Data Preprocessing

The data sets are divided into a training set (70% of the training sample) and test set (30% of the training sample). We implemented a 10-fold split cross-validation strategy over all the data sets while evaluating the performance measures. The pre-processing of the data sets is kept to the minimal, other than handling the missing values. All the rows which contain *'NaN'* values were removed from the data set rather than filling the *'NaN'* values to create a bias towards a specific value. The columns for which the *'NaN'* values consisted of more than 60 percent of the values were dropped entirely. All the categorical variables are converted into binary variables for experimentation.

D. Results

The proposed hybrid model is applied to the 25 standard data sets from various applied fields having reasonably large number of input features. For these data sets, feature selection is a vital task while building a classifier. We have conducted experiments in a total of 25 data sets, each with the paradigm of binary classification. To ensure a consistent accuracy over the data sets, the metrics to compare the classifier are being calculated over a 10-fold cross validation split. The final score of the metric is the mean of all the metric values the classifier gives over all the cross-validation splits Table 2 compares the metrics of all the classifiers over all the data sets.

The data set if splitted into training set with 70% of the total sample size and remaining 30% as the test data set. DT is trained on the training set using the *'scikit-learn'* package in Python software, where all the parameters are kept at their default values. The classifier is trained only on the training set to ensure that there is no data leak viz. the classifier does not train on the test set. The importance of all the features are calculated using the trained classifier where the importance of a feature is computed as the normalized total reduction of the criterion, Gini impurity, brought by that feature. The mean of importance of all the features is selected as the threshold to determine if a feature is important or not, and the modified data set is formed by using the important features, as well as the output of the DT classifier. DT predicted values are used as an additional feature in the input feature space of SVM along with the other important data features obtained by DT. We define by k as the number of important features obtained using DT + 1, viz., number of input features in SVM model, are reported in Table 2.

The reason behind this is to reduce the number of support vector for the SVM, and thus improving the performance as well as reducing the time taken by the SVM for predictions. SVM is trained on the modified training set using the *scikit-learn* package in Python, where the Gaussian kernel is being used for the classification. The scale parameter defined in Eq (2) is given by the formula $\beta = \frac{1}{(n\times\sigma)}$, where n is the number of features in the data set and σ is the variance of the data set. The SVM is then trained on modified data set with the cross-validation strategy as mentioned above. The experimental results, reported in Table 2, show the performance of DT, SVM, and the proposed hybrid model on the test data sets. We can conclude from Table 2 that the proposed model outperforms individual models for more than 75% of the data sets in terms of accuracy metrics and remain competitive for other data sets as well. Therefore, it is worthy to mention that the proposed model is a 'good' choice for data scientists for feature selection cum classification problems in binary pattern classification problems.

DATA AND CODES

For the sake of reproducibility of this work, all codes and necessary data sets are made available at this link (https://github.com/kaumil/hybrid).

TABLE II
COMPARISON OF THE RESULTS OF DT, SVM, AND HYBRID MODEL ON STANDARD DATA SETS. BEST RESULTS ARE INDICATED AS BOLD IN THE TABLE.

Data	DT		SVM		Hybrid Model		
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	k
breast cancer	**92.81**	**0.928**	62.74	0.627	91.59	0.878	5
credit card	94.55	0.866	95.60	0.885	**99.12**	**0.978**	13
hepatitis	70.89	0.809	84.04	0.912	**84.05**	**0.913**	8
htru-2	96.80	0.827	97.24	0.831	**99.27**	**0.960**	6
hungarian-14-heart	69.99	0.613	76.26	0.601	**94.25**	**0.925**	33
hypothyroid	94.45	0.970	93.90	0.968	**99.00**	**0.995**	73
ionosphere	81.27	0.735	82.45	0.701	**86.72**	**0.784**	5
kr-vs-kp	97.84	0.976	92.43	0.917	**99.72**	**0.997**	9
magic gamma	97.84	0.976	92.43	0.917	**99.72**	**0.997**	9
musk-1	**71.31**	**0.713**	56.93	0.569	59.46	0.595	30
musk-2	78.06	0.588	85.08	0.850	**85.37**	**0.853**	38
nomao	94.54	0.866	95.60	0.884	**99.12**	**0.977**	23
occupancy	93.02	0.871	98.48	0.968	**98.71**	**0.973**	3
ozone	89.23	0.892	**93.08**	**0.931**	93.07	0.930	24
parkinsons	80.40	0.867	76.37	0.855	**82.37**	**0.896**	7
pima diabetes	70.18	0.702	65.11	0.651	**77.47**	**0.653**	4
planning	53.36	0.663	71.46	**0.833**	**87.95**	0.779	5
sonar	60.57	0.601	57.12	0.667	**92.85**	**0.931**	15
spambase	**89.93**	**0.875**	77.18	0.687	78.37	0.714	12
spect images	75.98	0.837	79.05	0.880	**81.54**	**0.881**	7
statlog (heart)	**74.44**	**0.701**	55.93	0.559	67.04	0.594	5
tictactoe	81.14	0.766	88.75	0.826	**98.23**	**0.975**	14
thyroid-sick	**98.70**	**0.987**	93.88	0.939	94.11	0.941	9
twonorm	83.77	0.837	**97.67**	**0.976**	96.58	0.965	8
vote	93.47	0.928	96.52	0.964	**98.28**	**0.989**	3

V. CONCLUSION

In this paper, a novel hybrid nonparametric classifier is proposed to achieve higher accuracy in classification performance with minimal computational cost (by working with a subset of input features). Our proposed hybrid approach is useful for feature selection cum binary classification problems. The proposed hybrid model is shown to be robust using experimental evaluation on various standard data sets. The proposed model, when applied to the data sets, performed better as compared to individual classifiers for most of the data sets. The model is beneficial when working with a considerable number of support vectors is a problem. It is worth noting that the proposed hybrid model is simple, easily interpretable, and fast during implementation. As a future scope of research of this paper, one can extend the model for multi-class classification problems and try to improve the model, especially for imbalanced data classification frameworks.

REFERENCES

[1] L. Breiman, *Classification and regression trees*. Routledge, 1984.

[2] T. Chakraborty, A. K. Chakraborty, and C. Murthy, "A nonparametric ensemble binary classifier and its statistical properties," *Statistics & Probability Letters*, vol. 149, pp. 16–23, 2019.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[4] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *Journal of machine learning research*, vol. 2, no. Nov, pp. 67–93, 2001.

[5] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research*, vol. 267, no. 2, pp. 687–699, 2018.

[6] T. Chakraborty, S. Chattopadhyay, and A. K. Chakraborty, "A novel hybridization of classification trees and artificial neural networks for selection of students in a business school," *Opsearch*, vol. 55, no. 2, pp. 434–446, 2018.

[7] K. P. Bennett and J. Blue, "A support vector machine approach to decision trees," in *IEEE International Joint Conference on Neural Networks (IJCNN) Proceedings*, vol. 3. IEEE, 1998, pp. 2396–2401.

[8] S. Cheong, S. H. Oh, and S.-Y. Lee, "Support vector machines with binary tree architecture for multi-class classification," *Neural Information Processing-Letters and Reviews*, vol. 2, no. 3, pp. 47–51, 2004.

[9] B. Fei and J. Liu, "Binary tree of svm: a new fast multiclass training and classification algorithm," *IEEE transactions on neural networks*, vol. 17, no. 3, pp. 696–704, 2006.

[10] G. Madjarov and D. Gjorgjevikj, "Hybrid decision tree architecture utilizing local svms for multi-label classification," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2012, pp. 1–12.

[11] V. Sugumaran, V. Muralidharan, and K. Ramachandran, "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing," *Mechanical systems and signal processing*, vol. 21, no. 2, pp. 930–942, 2007.

[12] M. A. Kumar and M. Gopal, "A hybrid svm based decision tree," *Pattern Recognition*, vol. 43, no. 12, pp. 3977–3987, 2010.

[13] K. Kim, "A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree," *Pattern Recognition*, vol. 60, pp. 157–163, 2016.

[14] T. Chakraborty, A. K. Chakraborty, and Z. Mansoor, "A hybrid regression model for water quality prediction," *OPSEARCH*, pp. 1–12, 2019.

[15] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

Next Word Prediction in Hindi Using Deep Learning Techniques

Radhika Sharma, Nishtha Goel, Nishita Aggarwal, Prajyot Kaur and Chandra Prakash
Indira Gandhi Delhi Technical University For Women, New Delhi, India 110006
Email: radhika2414@gmail.com, nishthagoel26@gmail.com, naggarwal97@gmail.com,
prajyotkaur@gmail.com, cse.cprakash@gmail.com

***Abstract*—Natural Language Generation (NLG) focuses on the generation of natural, human-interpretable language. This study proposes a novel methodology to predict the next word in a Hindi sentence. By predicting the next word in a sequence, the number of keystrokes of the user can be reduced. Two deep learning techniques namely Long Short Term Memory (LSTM) and Bi-LSTM have been explored for the task of predicting next word and accuracy of 59.46% and 81.07% was observed for LSTM and Bi-LSTM respectively. This approach may be used for various NLG tasks like story auto-completion, sentence autocompletion, etc.**

***Index Terms*—Next Word Prediction, Hindi, LSTM, Bi-LSTM**

I. Introduction

Natural Language Generation is the part of machine learning used to generate natural language from data input to it. It is used for translating data to natural language [1]. Various applications of NLG are text generation, summarization, auto-captioning of images, machine translation, etc. Automatic text generation involves generating new text by applying deep learning techniques on the text given . Next Word Prediction involves predicting the next word/s, which have a high probability of following the given sequence of words.

The model suggests a few words which are most likely to follow the current set of words from which the user can select the word of their choice. This helps in saving keystrokes of the user. This was extended to continue predicting the next few words for a given sequence of words. Word prediction is an elementary part of Natural Language Generation. It has various applications, such as the following :

- Helps in minimizing the keystrokes of users while typing.
- Helps save typing time of users.
- Helps in minimizing spelling mistakes of users. Especially useful for users who are not proficient or non-native to that language.
- Aids the non-native users in learning the language by suggesting new and correct words to them, thus expanding their vocabulary.

Hindi is a widely spoken language in India . Thus next word prediction in Hindi will be helpful to a vast majority of people living in India. The processing of the Hindi language is very difficult as it contains a lot of mantras and symbols. This can result in irrelevant results due to the confusion of spellings. Hence, processing the language at the word level gives better results and is less complex.

978-1-7281-2087-4/19/$31.00 © 2019 IEEE

The paper is organized as follows. Section 2 talks about the literature survey done about Next word prediction & Hindi language. Section 3 discusses the methodology proposed, including the specifications of the dataset used. Section 4 illustrates the results obtained. Section 5 discusses the conclusions drawn and the suggested future work.

II. Literature Survey

Natural Language Generation (NLG) is a systematic and significant approach to produce meaningful text that is understandable by humans. For generating the text, the data is collected from different sources or taken as input from the users. There has been a drastic change in the field of NLP over the past few years [2]. Previously, NLP techniques employed shallow machine learning models, which consisted of handcrafted features and very very time-consuming. Due to the increasing popularity of word embeddings, neural networks have achieved greater success in comparison to traditional machine learning models [3].

The English language is very widely used across the world. Hence, researchers have employed several statistical models and machine learning models at the character level and word level for the generation of text . Various statistical models that have been used for text generation in the English language include word2vec approach, a continuous bag of words, etc. which help in generation of text by creating a vocabulary. One limitation of these approaches is that they generate text without ensuring the syntactic correctness of the sentence. Whereas, Lemmatisation, Latent Semantic Analysis (LSA) [4], Parts of Speech (POS) Tagging [5] generate text that is free from any grammatical errors.

Neural Networks also gained popularity over the traditional methods due to their correctness in generating the text. Neural Networks are inspired by the functioning of the brain. RNN was popularly used for text generation because of their ability to process sequential data. But due to its limitation of vanishing gradients, it is being replaced by other neural networks. LSTM, and other versions of LSTM, i.e., Bi-LSTM, GRU are being popularly used nowadays for generating text in the English language. These models are also being used

for other NLP related tasks like Query auto-completion, story generation, etc.

Although English is the language being used worldwide in India, Hindi is the language used by the majority of the citizens for communication. A greater part of the population has little or no knowledge of Hindi language. Hence, a machine learning model which generates text in the Hindi language would help such citizens to connect with the world digitally.

Hindi is a very morphologically rich language consisting of 35 consonants, 11 vowels and 12 matras [6]. Thus, Hindi is a highly complex language. This makes the pre-processing of Hindi language a complex task . Several problems that are faced during pre-processing of Hindi text are -

- Large character set
  The character set of Hindi language is extremely vast. Along with the characters, it also consists of special symbols (Eg- "." , anuswar, halant, etc. This might decrease the no. of keystrokes saved by the model and lead to a decrease in its accuracy [7].
- Phonetically similar characters
  There are characters in the Hindi language that is near of the same shape and size or characters that sound similar. Thus, if such characters are present in the text, they can give rise to confusion.
- Typographical Variants
  A character can have more than one form of representation. This poses a problem while generating text as many alternate forms of a single character exist.

Research has been done to build several statistical and syntactical models that work at character level to generate text in the Hindi Language. But, the character level approach requires to break the sentence character by character . And, due to the complexity of Hindi Language, this approach is difficult and can give rise to ambiguities and generate irrelevant output. Thus, it is recommended to process the data at the word level, as firstly it is easier to understand and process, and secondly, it generates text with better accuracy.

Bi-gram and tri-gram statistical predictors have been used most commonly till date [7]. They predict the next word based on the previous two and three words, respectively. But, the drawback of these statistical models is that they fail to perform well in case of a large data corpus. Parts of Speech (POS) Tagging is a syntactical approach which has also been used by the researchers. But, this approach cannot handle words which are not present in the vocabulary. Various approaches like K-Nearest Neighbour (KNN) [8], Universal Networking Language (UNL) have been used earlier, which generate Hindi text from given English text corpus.

All the above approaches most commonly use character level architecture, i.e., the model helps in predicting the next character in the sequence. One advantage of character level models is that the size of the vocabulary is small. But such models suffer from the vanishing gradients problem [9]. Also, character level processing of a Hindi corpus is highly complex, and hence, word level architectural models provide better results and are less prone to errors. Models that follow word level architecture take a lesser amount of time to train, and generate more logical results by predicting the next complete word instead of only a character. Although, word level architecture requires to store all the individual words, which means the size of the vocabulary is large, hence more memory is required [10]. But, the accuracy given by word level architecture overcomes this drawback.

III. Methodology

The process starts with the cleaning of the dataset. Then, the dictionary of unique words is created by splitting the words in the dataset and filtering the unique words which constitute to a dictionary. The input file is parsed with the iterator, and unique words are collected. Succeeding, the unique words in the dictionary are mapped to indices. This indicates that the words are not easily processed by neural networks in machine learning, and it is important to map it to the indices, which are easy for neural networks to process.

Set sequence length by dividing the sentence into 6:1 ratio as input x and input y and mapping the first six words in input x. Divide the input file into tensor (dataX) based on the sequence length. Create another tensor (dataY) containing the next words of the sequence, which is in the input file which is the output of the project [11].

Assign probability to the content of dataY using softmax (activation) function. Select next word based on the probability of dataY, which will be the predicted word of the sequence. Since the aim of the process is to predict the next word, the predicted word completes the sequence. Each iteration user will be given three options in the output among which user will have to choose the most appropriate option in context to its application and the next word prediction will be based on the output chosen.

Repeat the process till the sequence is completed, i.e., the sequence of seven words are not completed, and our objective is not completed. Repeat the process for remaining sequences. Since two thousand sentences of length seven are remaining to be processed in the dataset, all the steps have to be repeated for all the sentences which length is equal to seven and the next word will be predicted.

A. Dataset

The IIT Bombay developed this dataset at the Indian Language Technology Center, IIT Bombay for several years. The dataset was collected from a plethora of already provided sources and corpora [12]. This dataset denotes English-Hindi corpus which can be processed for translation. The Hindi text from this corpus is used for building the model. The no of sentences present in the dataset is 15,61,841. The no of sentences used in the project is 2615. These are sentenced having more than six words in a sentence. The sentences which have the length of seven words are taken in the project for further computations. The dataset contains a large number of short sentences with sequences less than of length seven. The words which make up the sequence in the dataset are

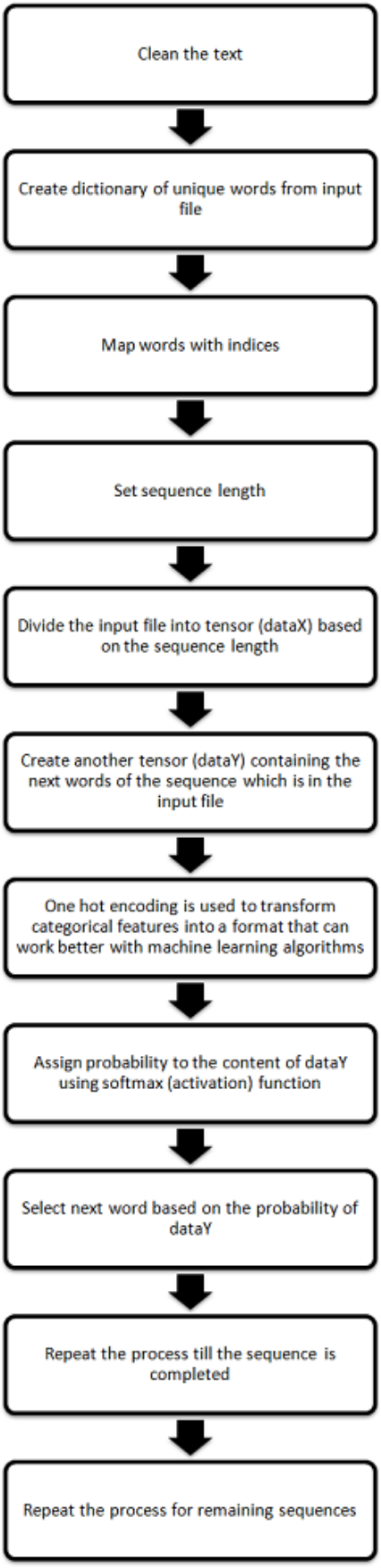


Fig. 1. Flowchart

of Devanagari Hindi script. This dataset is widely used for translation based objective [12]. A validation split of 80:20 is done. The number of sentences used for validation is 523, and that used for training is 2092.

B. LSTM (Long Short Term Memory)

Long-Term Short Term Memory (LSTMs) is a variant of RNNs which are capable of learning Long Term Dependencies, and thus are widely used for Natural Language Generation . LSTMs have memory and remember past data from the input for longer durations of time. They can selectively remember or forget things. They are well suitable for written data inputs, as any word in a sentence is related to words around it (previous and upcoming words) [13].

Each repeating module in LSTM consists of 4 neural network layers that interact with each other. The cell state in LSTMs decides when to read, write and what should be stored in memory. It has gates which conditionally let information pass through them, adding or removing it from cell [14].

Gates work using pointwise multiplication, pointwise addition, and sigmoid function. The output is from 0 to 1 signifying how much information should pass through, where 0 represents no information passes through, and 1 means all information passes through. LSTM has three gates: Forget gate, Input gate, and Output gate.

C. BI-DIRECTIONAL LSTM

Sequence classification problems can be improved by extending the bidirectional LSTMs to enhance model performance. LSTMs are trained twice in bidirectional LSTMs on the input given by the user [15]. The first iteration starts with the forward iteration and passing of sequence like the way LSTMs works and the second iteration is the reverse iteration on the input sequence.

Bi-LSTM processes the data from start-to-end in one iteration and from end-to-start in the other iteration. The prediction is done concerning the future and past of the data as the traversal is in both directions [16]. Bi-LSTMs are preferable over LSTMs because they provide feedback input to the successor layer. This functionality of Bi-LSTMs adds the advantage of complete and faster processing and learning on the input sequence [17].

On comparing the two algorithms, Bi-LSTM is a clear win as it processes the sequence in two iterations, one from front-to-back and other from back-to-front. The use of bidirectional LSTMs provide you with the added advantage of history and near future which makes the prediction accurate and faster. Bidirectional LSTMs (BiLSTM) has different layers which are fed by the learning algorithm to learn long-term dependencies of both the history and future data.

IV. Results

A model has been proposed in this paper to predict the next word in Hindi, given a sequence of at least six words using LSTM and Bi-LSTM Neural Networks. For both the

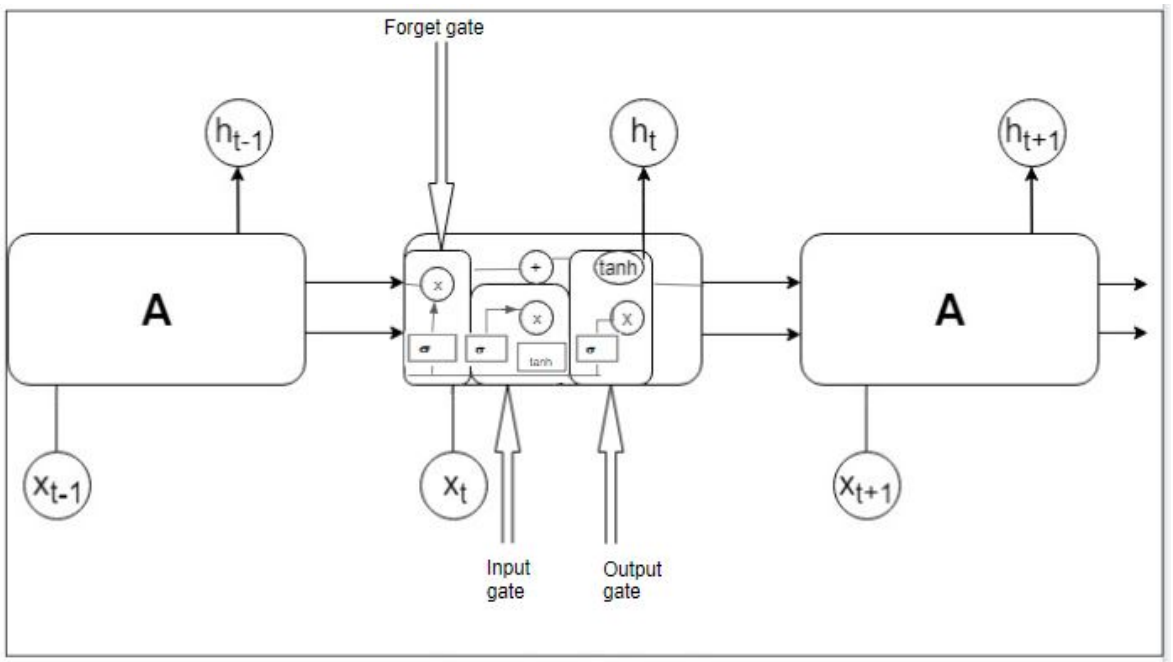


Fig. 2. LSTM Network Architecture

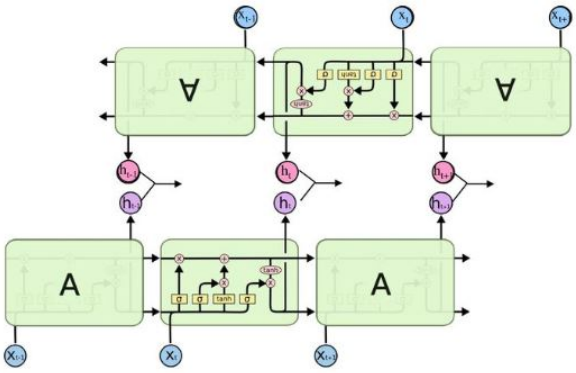


Fig. 3. Bi-LSTM Network Architecture [14]

models softmax activation function and categorical cross entropy loss function was used. Softmax activation function is a mathematical function that is used to obtain the probability that a word will be predicted on the basis of the score for each word obtained from the neural network model thus obtaining the probability distribution of all the words present in the dictionary. The function is defined as

$$\sigma(\mathbf{s})_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} \tag{1}$$

where $\sigma(\mathbf{s})_i$ is the probability and s_i is score for i^{th} word in the dictionary [18].
Since the task of predicting the next word is a case of multi class classification, Categorical Cross Entropy Loss is used for calculating the loss of model and hence for updating the weights. The function is defined as

$$CE = -\sum_{i}^{C} t_i log(\sigma(\mathbf{s})_i) \tag{2}$$

where CE is Categorical cross entropy loss, t_i denotes the actual truth value and $\sigma(\mathbf{s})_i$ is the probability for the i^{th} word of the dictionary.
A learning rate of 0.001 was used while training both the models. Through cross-validation the accuracy of both the models was calculated by comparing predicted word with actual word in the dataset.

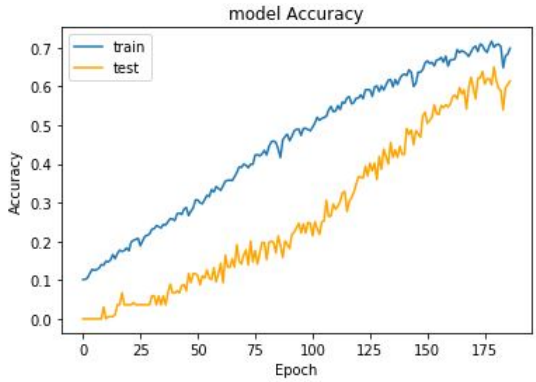


Fig. 4. Accuracy of LSTM Model

Recurrent Neural Networks(RNN) faced a major problem of vanishing gradient [9] due to which front layers of the network might learn slowly. LSTM and Bi-LSTM overcame

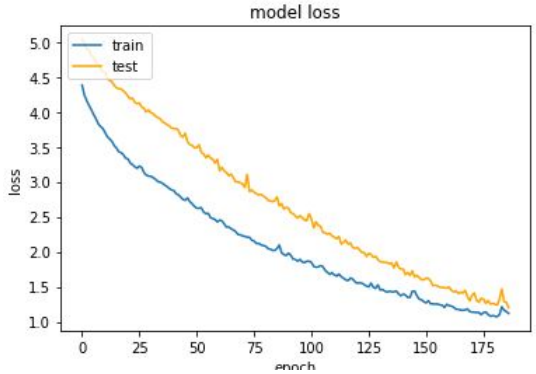


Fig. 5. Loss of LSTM Model

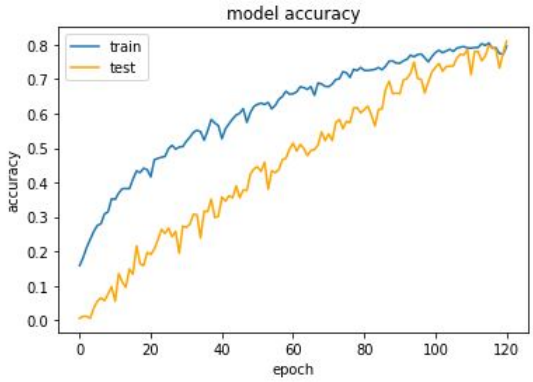


Fig. 6. Accuracy of Bi-LSTM Model

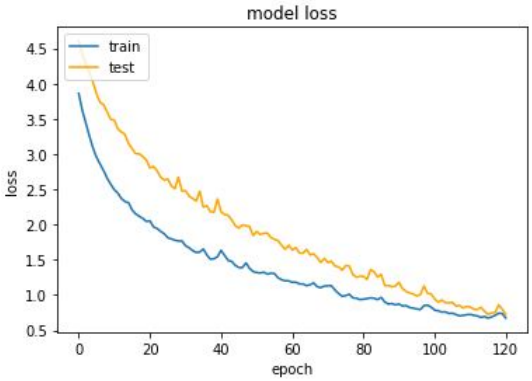


Fig. 7. Loss of Bi-LSTM Model

TABLE I
COMPARISON OF ML TECHNIQUES

	LSTM	***Bi-LSTM***
Number Of Epochs	197	121
Accuracy	0.7089	0.7954
Validation Accuracy	0.5946	0.8107
Loss	1.0730	0.6734
Validation Loss	1.2381	0.7260

this problem. They use gates to selectively forget information. The accuracy of LSTM model obtained is 70.89% and that of Bi-LSTM is 79.54%. However the validation accuracy for both of them is 59.46% and 81.07% respectively. The Bi-LSTM model gives better results as compared to LSTM.
The Bi-LSTM model also learns faster for the same dataset as compared to LSTM due to which the number of epochs was significantly less for the former model.

V. CONCLUSION

A machine learning model to predict the next word for a given sequence of words was built using LSTM and Bi-LSTM models. It was then extended to predict the next few words for the same sequence.
The model developed can be used for predicting the next word in the Hindi Language. This can effectively reduce the number of words that the user has to type, thus increasing the typing speed. It also helps in minimizing the spelling mistakes done by the user. In countries like India, where Hindi is the mother tongue of more than 25% of the population, this system can be a boon.

VI. FUTURE SCOPE

In the future, the system can be extended for other natural language generation tasks like story auto-completion, poem auto-completion, etc.
The model is limited to specific dataset and more randomness can be incorporated in the model by enhancing the scope. The system can be adapted to new words that are not a part of its vocabulary. This adaption will be done when model encounters a new word and adding the word to the vocabulary. This way the model becomes more generalized. The system can be personalized to predict words based on the user's history.

REFERENCES

[1] P. P. Barman and A. Boruah, "A rnn based approach for next word prediction in assamese phonetic transcription," *8th International Conference on Advances in Computing and Communication*, 2018.
[2] R. Perera and P. Nand, "Recent advances in natural language generation: A survey and classification of the empirical literature," *Computing and Informatics*, vol. 36, pp. 1–32, 01 2017.
[3] C. Aliprandi, N. Carmignani, N. Deha, P. Mancarella, and M. Rubino, "Advances in nlp applied to word prediction," *J. Mol. Biol.*, vol. 147, pp. 195–197, 2008.
[4] C. McCormick, *Latent Semantic Analysis (LSA) for Text Classification Tutorial*, 2019 (accessed February 3, 2019). http://mccormickml.com/2016/03/25/lsa-for-text-classification-tutorial/.
[5] Y. Wang, K. Kim, B. Lee, and H. Y. Youn, "Word clustering based on pos feature for efficient twitter sentiment analysis," *Human-centric Computing and Information Sciences*, vol. 8, p. 17, Jun 2018.
[6] N. N. Shah, N. Bhatt, and A. Ganatra, "A unique word prediction system for text entry in hindi," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, p. 118, ACM, 2016.
[7] M. K. Sharma and D. Smanta, "Word prediction system for text entry in hindi," *ACM Trans. Asian Lang. Inform. Process*, 06 2014.
[8] R. Devi and M. Dua, "Performance evaluation of different similarity functions and classification methods using web based hindi language question answering system," *Procedia Computer Science*, vol. 92, pp. 520–525, 2016.

[9] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
[10] D. Pawade, A. Sakhapara, M. Jain, N. Jain, and K. Gada, "Story scrambler - automatic text generation using word level rnn-lstm," *Modern Education and Computer Science*, 2018.
[11] R. L. Bishop and S. I. Goldberg, *Tensor analysis on manifolds*. Courier Corporation, 2012.
[12] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The iit bombay english-hindi parallel corpus," *Language Resources and Evaluation Conference*, 2018.
[13] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.
[14] C. Olah, *Understanding LSTM Networks*, 2019 (accessed March 7, 2019). http://colah.github.io/posts/2015-08-Understanding-LSTMs/.
[15] J. Brownlee, *How to Develop a Bidirectional LSTM For Sequence Classification*, 2019 (accessed April 10, 2019). https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/.
[16] Y. Huang, Y. Jiang, T. Hasan, Q. Jiang, and C. Li, "A topic bilstm model for sentiment classification," in *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, pp. 143–147, ACM, 2018.
[17] S. Stymne, S. Loáiciga, and F. Cap, "A bilstm-based system for cross-lingual pronoun prediction," in *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 47–53, 2017.
[18] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

Weighted Similarity Measure and Decision Making in ClinicalApplication of Neutrosophic Soft Set

Binu R.
Department of Mathematics
Rajagiri School of Enigineering and Technology
Kerala.

Paul Isaac
Department of Mathematics
Bharata Mata College, Thrikkakara
Kerala

***Abstract*—In this paper, we have proposed the normalized orthogonal distance between two neutrosophic soft sets and its properties. The similarity measure and weighted similarity measure are defined using the normalized orthogonal distance of neutrosophic soft set and the decision making procedure in the soft neutrosophic environment using normalized orthogonal distance is derived. Finally the multi attribute decision making is illustrated through clinical application. In order to investigate the clinical application of neutrosophic soft set, the best radiotherapy treatment method for the tumor of movable organs is estimated through the evaluation of some medical parameters.**

***Index Terms*—Soft set, Neutrosophic soft set, Normalized orthogonal distance, Similarity measure, Weighted similarity measure, Ideal attribute, Decision making**

I. Introduction

In 1965, Lotfi A. Zadeh introduced the concept of vagueness in mathematical modelling. As a generalization of fuzzy set [10] and intuitionistic fuzzy set [1], the neutrosophic set was defined with three different types of membership values by Smarandache in 1995 [7]. Neutrosophic set is a powerful tool and the appropriate frame work for dealing with incomplete,indeterminate and inconsistent information in the real world practical problems.

In 1999, Molodtsov introduced soft set theory as a general mathematical tool for dealing with uncertainty or imprecise boundaries. The algebraic structure of soft set theory dealing with uncertainty has been studied by some authors. Maji et al [4] defined the algebraic operations of soft sets for theoretical study. Some authors also extended the concept of soft algebraic structures to fuzzy soft or intuitionistic fuzzy algebraic structures. Combining neutrosophic set theory with algebra of soft set is an emerging trend in the area of mathematical research. Neutrosophic soft algebraic structures and its properties give us a strong mathematical background to explain applied mathematical concepts in engineering, data mining and economics.The main objective of this paper is to study the physical properties of normalized orthogonal distance and similarity relation in the context of neutrosophic soft set in decision making .

II. Preliminaries

Definition II.1. *[8] A neutrosophic set A on the universal set X is defined as*

$$A = \{(x, t_A(x), i_A(x), f_A(x)) : x \in X\}$$

978-1-7281-2087-4/19/$31.00 ©

where $t_A, i_A, f_A : X \rightarrow (^-0, 1^+)$. *The three components* t_A, i_A *and* f_A *represent membership value (Percentage of truth), indeterminacy (Percentage of indeterminacy) and non membership value (Percentage of falsity) respectively. These components are functions of non standard unit interval* $(^-0, 1^+)$.

Remark *[8] If the components of a neutrosophic set A,* $t_A, i_A, f_A : X \rightarrow [0, 1]$, *then A is known as single valued neutrosophic set(SVNS).*

Definition II.2. *[8] Let A and B be two neutrosophic sets on X. Then A is contained in B, denoted as* $A \subseteq B$ *if and only if* $A(x) \leqslant B(x)\ \forall x \in X$, *this means that*

$$t_A(x) \leq t_B(x), i_A(x) \leq i_B(x), f_A(x) \geq f_B(x),\ \forall x \in X$$

Definition II.3. *[5] Let* X *be an initial universe of objects and* E *be the set of parameters in relation to objects in* X*.Assume that* $P(X)$ *denotes the power set of* X *and* $A \subseteq E$. *A pair* $\langle F, A\rangle$ *is called a soft set over* X *where* F *is a mapping given by* $F : A \rightarrow P(X)$.
The soft set is a parametrized family of subset of the set X*.Parameters are often attributes or characteristics or properties of objects in* X. *For any parameter* $e \in A$, $F(e) \subseteq X$ *may be considered as the set of* e *- approximate elements of the soft set* $\langle F, A\rangle$ *and it is represented as*

$$\langle F, A\rangle = \{(e, F(e)) : e \in E,\ F(e) = \phi\ if\ e \in E - A\}$$

Example II.1. *[6] Let* $X = \{x_1, x_2, x_3, x_4\}$ *be the universal set which contains 4 houses under consideration by an agent and* $E = \{e_1 = cottage, e_2 = mansion, e_3 = terraced\}$. *A customer to select a house from the agent, he can construct a soft set* F_A *that describes the characteristics of the houses according to his own requirements.* $F(e_1) = \{x_1, x_2\}, F(e_2) = \{x_3\}, F(e_3) = \{x_3, x_4\}$. *Then the soft set* $\langle F, A\rangle$ *is represented as follows*

$$\langle F, A\rangle = \{(e_1, \{x_1, x_2\}), (e_2, \{x_3\}), (e_3, \{x_3, x_4\})\}$$

Definition II.4. *[2] For two soft sets* $\langle F, A\rangle$ *and* $\langle G, B\rangle$ *over a common universe* X, *we say that* $\langle F, A\rangle$ *is a soft subset of* $\langle G, B\rangle$ *if*

1) $A \subset B$
2) $\forall\ e \in A$, $F(e)$ *and* $G(e)$ *are identical approximations.*

We write $A \subset B$

Definition II.5. *[3] A similarity measure between two neutrosophic sets A and B on X is a function defined as $S : X \times X \to [0, 1]$ which satisfies the following properties*

1) $S(A, B) \in [0, 1]$
2) $S(A, B) = 1 \Leftrightarrow A = B$
3) $S(A, B) = S(B, A)$
4) $A \subset B \subset C \Longrightarrow S(A, C) \leq S(A, B) \wedge S(B, C)$

Definition II.6. *[6] Let X be an initial universe of objects and E be the set of parameters in relation to objects in X and $A \subseteq E$. $NS(X)$ denotes the set of all neutrosophic sets of X. A pair $\langle F, A\rangle$ is called a neutrosophic soft set over X where F is a mapping given by $F : A \to NS(X)$.*
A neutrosophic soft set $\langle F, E\rangle = \{(e, F(e)) : e \in E,\ F_A(e) \in NS(X)\}$ where $F(e)$is a neutrosophic set on X which is characterized by

$$F(e) = \{x, t_{F(e)}(x), i_{F(e)}(x), f_{F(e)}(x) : e \in E\ x \in X\}$$

and $F(e) = \phi$ i.e.

$$F(e) = \{t_{F(e)}(x) = 0, i_{F(e)}(x) = 0, f_{F(e)}(x) = 1 : e \in E - A\}$$

where $t_{F(e)}(x), i_{F(e)}(x)$ and $f_{F(e)}(x)$ represents the truth-membership degree, the indeterminacy -membership degree and the falsity -membership degree of an object x holds on parameter e respectively.Thus neutrosophic soft set is a parametrization tool.

Remark- We write $NSS(X)$ for neutrosophic soft set.

Example II.2. *Consider the example 2.1 and $NS(X)$ denotes the set of all neutrosophic subsets of X. Suppose that*

$$F(e_1) = \{\langle .4, .8, .3\rangle, \langle .3, .7, .1\rangle\ \langle 0, 0, .1\rangle, \langle 0, 1, 0\rangle\}$$
$$F(e_2) = \{\langle 0, 0, 1\rangle, \langle 0, 0, 1\rangle\ \langle .8, .2, .1\rangle, \langle .0, .0, 1\rangle\}$$
$$F(e_3) = \{\langle 0, 1, 0\rangle, \langle 0, 0, 1\rangle\ \langle .6, .1, .2\rangle, \langle .1, .6, .3\rangle\}$$

Then $F_A(E)$ is a parametrized family of $NS(X)$.

Definition II.7. *[6] Let E be the set of parameters and $A, B \subseteq E$. Suppose that $\langle F, A\rangle$ and $\langle G, B\rangle$ are two neutrosophic soft set over X. $\langle F, A\rangle$ is said to be a neutrosophic soft subset of $\langle G, B\rangle$ if $A \subseteq B$ and $t_{F_A(e)}(x) \leq t_{G_B(e)}(x)$,$i_{F_A(e)}(x) \leq i_{G_B(e)}(x)$ and $f_{F_A(e)}(x) \geq f_{G_B(e)}(x)\ \forall\ e \in A, x \in X$.*

Definition II.8. *[6] The complement $\langle F, A\rangle^C$ of a neutrosophic soft set*

$$\langle F, A\rangle = \{x, t_{F_A(e)}(x), i_{F_A(e)}(x), f_{F_A(e)}(x) : x \in X\}$$

is defined as

$$\langle F, A\rangle^C = \{x, f_{F_A(e)}(x), 1 - i_{F_A(e)}(x), t_{F_A(e)}(x) : x \in X\}$$

Definition II.9. *[6] The union of two neutrosophic soft set $\langle F, A\rangle$ and $\langle G, B\rangle$ over a common universe X is a neutrosophic set $\langle H, C\rangle$ where $C = A \cup B$ and for each $e \in C$ and $x \in X$*

$$t_{H(e)}(x) = \begin{cases} t_{F(e)}(x) & e \in A - B \\ t_{G(e)}(x) & e \in B - A \\ max\{t_{F(e)}(x), t_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

$$i_{H(e)}(x) = \begin{cases} i_{F(e)}(x) & e \in A - B \\ i_{G(e)}(x) & e \in B - A \\ max\{i_{F(e)}(x), i_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

$$f_{H(e)}(x) = \begin{cases} f_{F(e)}(x) & e \in A - B \\ f_{G(e)}(x) & e \in B - A \\ min\{t_{F(e)}(x), t_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

Definition II.10. *[6] The intersection of two neutrosophic soft set $\langle F, A\rangle$ and $\langle G, B\rangle$ over a common universe X is a neutrosophic set $\langle H, C\rangle$ where $C = A \cap B$ and for each $e \in C$ and $x \in X$*

$$t_{H(e)}(x) = \begin{cases} t_{F(e)}(x) & e \in A - B \\ t_{G(e)}(x) & e \in B - A \\ min\{t_{F(e)}(x), t_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

$$i_{H(e)}(x) = \begin{cases} i_{F(e)}(x) & e \in A - B \\ i_{G(e)}(x) & e \in B - A \\ min\{i_{F(e)}(x), i_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

$$f_{H(e)}(x) = \begin{cases} f_{F(e)}(x) & e \in A - B \\ f_{G(e)}(x) & e \in B - A \\ max\{t_{F(e)}(x), t_{G(e)}(x)\} & e \in A \cap B \end{cases}$$

Definition II.11. *[5] If $X = \{x_1, x_2, ..., x_m\}$, $E = \{e_1, e_2, ..., e_n\}$ and $A \subseteq E$ then the neutrosophic soft set $\langle F, E\rangle$ is uniquely characterised by the neutrosophic soft matrix $[a_{ij}]_{m\times n}$ where*

$$a_{ij} = (t_{F(e_j)}(x_i), i_{F(e_j)}(x_i), f_{F(e_j)}(x_i))$$

Example II.3. *In example 2.2, the neutrosophic set $\langle F, E\rangle$ is characterised by the following neutrosophic soft matrix $D = (\alpha_{ij})_{4\times 3} =$*

$$\begin{bmatrix} 0.4 & 0.8 & 0.3 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \\ 0.3 & 0.7 & 0.1 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 1.0 & 0.6 & 0.1 & 0.2 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.1 & 0.6 & 0.3 \end{bmatrix}$$

III. Normalized orthogonal distance between two neutrosophic soft set

Let $E = \{e_1, e_2, ..., e_n\}$ be set of parameters in relation to objects in $X = \{x_1, x_2, ..., x_m\}$ and $A \subseteq E$. $\alpha = \langle F, A\rangle$ and $\beta = \langle G, A\rangle$ are two neutrosophic soft sets over X where each $x_i \in X$

$$F(e_j) = \{(x_i, t_{F(e_j)}(x_i), i_{F(e_j)}(x_i), f_{F(e_j)}(x_i)) : e_j \in E\}$$
$$G(e_j) = \{(x_i, t_{G(e_j)}(x_i), i_{G(e_j)}(x_i), f_{G(e_j)}(x_i)) : e_j \in E\}$$

in which $i = 1, 2, ..., m, j = 1, 2, ..., n$ and the neutrosophic components are not a scalar multiple of each other.

Definition III.1. *Let $\alpha, \beta \in NSS(X)$. The normalized orthogonal distance between α and β can be denoted and defined as*

$$d^{\perp}(\alpha, \beta) = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} \frac{\sqrt{(T_{\alpha\beta}(x_i))^2 + (I_{\alpha\beta}(x_i))^2 + (I_{\alpha\beta}(x_i))^2}}{\max(|F(e_j)|, |G(e_j)|)}$$

where

$$T_{\alpha\beta}(x_i) = [t_{F(e_j)}(x_i)i_{G(e_j)}(x_i) - i_{F(e_j)}(x_i)t_{G(e_j)}(x_i)]$$

$$I_{\alpha\beta}(x_i) = [i_{F(e_j)}(x_i)f_{G(e_j)}(x_i) - f_{F(e_j)}(x_i)i_{G(e_j)}(x_i)]$$

$$\Gamma_{\alpha\beta}(x_i) = [f_{F(e_j)}(x_i)t_{G(e_j)}(x_i) - t_{F(e_j)}(x_i)f_{G(e_j)}(x_i)]$$

$$|F(e_j)| = \sqrt{(t_{F(e_j)}(x_i))^2 + (i_{F(e_j)}(x_i))^2 + (f_{F(e_j)}(x_i))^2}$$

$$|G(e_j)| = \sqrt{(t_{G(e_j)}(x_i))^2 + (i_{G(e_j)}(x_i))^2 + (f_{G(e_j)}(x_i))^2}$$

Definition III.2. *The normalised orthogonal distance $d^{\perp}(\alpha,\beta)$ between α and $\beta \in NSS(X)$ satisfies the following axioms*

1) $d^{\perp}(\alpha,\beta) \geq 0$
2) $\alpha = \beta \Rightarrow d^{\perp}(\alpha,\beta) = 0$
3) $d^{\perp}(\alpha,\beta) = d^{\perp}(\beta,\alpha)$
4) $d^{\perp}(\alpha,\gamma) \leq d^{\perp}(\alpha,\beta) + d^{\perp}(\beta,\gamma)$ *where γ is any third neutrosophic soft set over X.*

IV. Similarity measure using normalised orthogonal distance

A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. Similarity measure takes large values on similar objects and either zero or a negative value for very dissimilar objects. In this section the similarity measure between two neutrosophic soft sets is defined using the normalised orthogonal distance. Similarity measures are inversely proportional to distance between the sets.

Theorem IV.1. *A real valued function $S^{\perp}(\alpha,\beta) : NSS(X) \times NSS(X) \rightarrow [0,1]$ where α and $\beta \in NSS(X)$ which is defined as*

$$S^{\perp}(\alpha,\beta) = \frac{1}{1 + d^{\perp}(\alpha,\beta)}$$

is a similarity measure .

Proof: To prove the function $S^{\perp}(\alpha,\beta)$ is a similarity measure, it is enough to prove that the function $S^{\perp}(\alpha,\beta)$ satisfies the following properties:

1) $0 \leq S^{\perp}(\alpha,\beta) \leq 1$
2) $S^{\perp}(\alpha,\beta) = 1 \Leftrightarrow \alpha = \beta$
3) $S^{\perp}(\alpha,\beta) = S^{\perp}(\beta,\alpha)$
4) $\alpha \subseteq \beta \subseteq \gamma \Longrightarrow S^{\perp}(\alpha,\gamma) \leq S^{\perp}(\alpha,\beta) \wedge S^{\perp}(\beta,\gamma)$

Proof 1,2,3: It is clear from the given function
Proof 4: Given $\alpha \subseteq \beta \subseteq \gamma$ where $\forall\ x_i \in X$

$$\alpha = F(e_j) = \{(x_i, t_{F(e_j)}(x_i), i_{F(e_j)}(x_i), f_{F(e_j)}(x_i)) : e_j \in E\}$$

$$\beta = G(e_j) = \{(x_i, t_{G(e_j)}(x_i), i_{G(e_j)}(x_i), f_{G(e_j)}(x_i)) : e_j \in E\}$$

$$\gamma = H(e_j) = \{(x_i, t_{H(e_j)}(x_i), i_{H(e_j)}(x_i), f_{H(e_j)}(x_i)) : e_j \in E\}$$

Then,

$$t_{F(e_j)}(x_i) \leq t_{G(e_j)}(x_i) \leq t_{H(e_j)}(x_i)$$
$$i_{F(e_j)}(x_i) \leq i_{G(e_j)}(x_i \leq i_{H(e_j)}(x_i)$$
$$f_{F(e_j)}(x_i)) \geq f_{G(e_j)}(x_i)) \geq f_{H(e_j)}(x_i)$$

$$i_{H(e_j)}(x_i) \geq i_{G(e_j)}(x_i)$$
$$\Rightarrow t_{F(e_j)}(x_i)i_{H(e_j)}(x_j) \geq t_{F(e_j)}(x)i_{G(e_j)}(x_i) \quad (1)$$
$$t_{H(e_j)}(x_i) \geq t_{G(e_j)}(x_i)$$
$$\Rightarrow t_{H(e_j)}(x_i)i_{F(e_j)}(x_i) \geq t_{G(e_j)}(x_i)i_{F(e_j)}(x_i) \quad (2)$$

(1)-(2)

$$t_{F(e_j)}(x_i)i_{H(e_j)}(x_i) - t_{H(e_j)}(x_i)i_{F(e_j)}(x_i)$$
$$\geq t_{F(e_j)}(x_i)i_{G(e_j)}(x_i) - t_{G(e_j)}(x_j)i_{F(e_j)}(x_i)$$
$$T_{\alpha\gamma} \geq T_{\alpha\beta}$$

similarly $I_{\alpha\gamma} \geq I_{\alpha\beta}$
Consider

$$t_{G(e_j)}(x_i) \leq t_{H(e_j)}(x_i)$$
$$\Rightarrow f_{F(e_j)}(x_i)t_{G(e_j)}(x_i) \leq f_{F(e_j)}(x_i)t_{H(e_j)}(x_i) \quad (3)$$
$$f_{G(e_j)}(x_i) \geq f_{H(e_j)}(x_i)$$
$$\Rightarrow f_{G(e_j)}(x_i)t_{F(e_j)}(x_i) \geq f_{H(e_j)}(x_i)t_{F(e_j)}(x_i)$$
$$\Rightarrow -f_{G(e_j)}(x_i)t_{F(e_j)}(x_i) \leq -f_{H(e_j)}(x_i)t_{F(e_j)}(x_i \quad (4)$$

(3)+(4)

$$f_{F(e_j)}(x_i)t_{G(e_j)}(x_i) - f_{G(e_j)}(x_i)t_{F(e_j)}(x_i)$$
$$\leq f_{F(e_j)}(x_i)t_{H(e_j)}(x_i) - f_{H(e_j)}(x_i)t_{F(e_j)}(x_i)$$
$$\Gamma_{\alpha\beta} \leq \Gamma_{\alpha\gamma} \Rightarrow \Gamma_{\alpha\gamma} \geq \Gamma_{\alpha\beta}$$

similarly, $T_{\beta\gamma} \leq T_{\alpha\gamma}$, $I_{\beta\gamma} \leq I_{\alpha\gamma}$ and $\Gamma_{\beta\gamma} \leq \Gamma_{\alpha\gamma}$
Thus

$$d^{\perp}(\alpha,\beta) \leq d^{\perp}(\alpha,\gamma) \Rightarrow S^{\perp}(\alpha,\beta) \geq S^{\perp}(\alpha,\gamma)$$
$$d^{\perp}(\beta,\gamma) \leq d^{\perp}(\alpha,\gamma) \Rightarrow S^{\perp}(\beta,\gamma) \geq S^{\perp}(\alpha,\gamma)$$
$$\Rightarrow S^{\perp}(\alpha,\gamma) \leq S^{\perp}(\alpha,\beta) \wedge S^{\perp}(\beta,\gamma)$$

■

Definition IV.1. *Let $\alpha,\beta \in NSS(X)$ and $w_i \in [0,1]$ be the weight of each element x_i $(i = 1,2,...,m)$ with the property $\sum_{i=1}^{i=m} w_i = 1$. The weighted similarity measure using normalized orthogonal distance between two NSS(X)s α and β can be defined as follows*

$$WS^{\perp}(\alpha,\beta) = \frac{1}{1 + \sum_{i=1}^{i=m}\sum_{j=1}^{j=n} w_i \frac{\sqrt{(T_{\alpha\beta}(x_i))^2 + (I_{\alpha\beta}(x_i))^2 + (\Gamma_{\alpha\beta}(x_i))^2}}{\max(|F_A(e_j)|, |G_A(e_j)|)}}$$

Proposition IV.1. *The weighted similarity measure (WSM) using normalised orthogonal distance between two neutrosophic soft sets α and β satisfies the following properties*

1) $0 \leq WS^{\perp}(\alpha,\beta) \leq 1$
2) $WS^{\perp}(\alpha,\beta) = WS^{\perp}(\beta,\alpha)$
3) $WS^{\perp}(\alpha,\beta) = 1\ if\ \alpha = \beta$

V. Decision Making Using Weighted Similarity Measure

In this section, a multi criteria decision - making technique using neutrosophic soft set and weighted similarity measure, as a parametrization tool, is proposed. The following steps explain the fundamentals of decision making in neutrosophic soft environment.

1) Develop a model for the decision making :-It consists of building a hierarchy to analyse the decision.
   a) Choose the best alternative -The objects in the universal set $X = \{x_1, x_2, ..., x_m\}$
   b) Choose the criteria-Let $E = \{e_1, e_2, ..., e_n\}$ be the set of parameters or criteria in relation to objects in $X = \{x_1, x_2, ..., x_m\}$ and $A \subseteq E$.
2) Derive weights (Priorities) and consistency ratio for the criteria.
3) Derive local preferences for the alternatives.
4) Derive model synthesis:-Let w_j be the weight of criterion e_j where $j = 1, 2, ..., n$ fixed by the decision maker such that each $w_j \in [0, 1]$ and $\sum_{j=1}^{j=n} w_j = 1$. The characteristic of the alternatives x_i where $i = 1, 2, ..., m$ on criterion e_j where $j = 1, 2, ..., n$ is denoted by the following neutrosophic soft form defined on e_j. The neutrosophic soft set over X, $\langle F, A\rangle$ is defined as follows $\forall\ x_i \in X$.

$$F_A(e_j) = \{x_i, t_{F_A(e_j)}(x_i), i_{F_A(e_j)}(x_i), f_{F_A(e_j)}(x_i\} \tag{5}$$

A decision matrix $D = [a_{ij}]_{m\times n}$ of $\langle F, A\rangle$ represents evaluation of each object x_i $(0 \le i \le m)$ in X on each parameter e_j $(0 \le j \le n)$ in E, constructed from the above equation 5. In multi-criteria decision making neutrosophic environment, the concept of ideal point has been used to identify the best attribute in the decision set.

Definition V.1. *[9] In the decision making procedure, criteria are classified into two, according to their nature:*

a) *Benefit criteria:- Maximum operator is used for identifying ideal alternative in benefit criteria.*
b) *Cost criteria:-Minimum operator is used for identifying ideal alternative in cost criteria.*

Definition V.2. *In the multi attribute decision making process, the ideal attribute α_j $(j = 1, 2, ..., n)$ can be denoted and defined as follows for benefit criteria,*

$$\alpha_j = \langle \max_i\ a_{ij}, \max_i\ b_{ij}, \min_i\ c_{ij}\rangle \tag{6}$$

$$= \langle a_j, b_j, c_j\rangle\ where\ i = 1, 2, ..., m \tag{7}$$

For a cost criterion

$$\alpha_j = \langle \min_{\mathbf{i}} a_{ij}, \max_{\mathbf{i}} b_{ij}, \max_{\mathbf{i}} c_{ij}\rangle \tag{8}$$

$$= \langle a_j, b_j, c_j\rangle\ where\ i = 1, 2, ..., m \tag{9}$$

5) Perform sensitivity analysis:- Calculate the weighted similarity measure between the alternatives and criteria.

Definition V.3. *The similarity measure between an alternative x_i and the ideal attribute α_j can be defined as*

$$WS^{\perp}(x_i, \alpha_j) = \frac{1}{1 + \sum_{j=1}^{j=n} w_j d^{\perp}(x_i, \alpha_j)}$$

where

$$d^{\perp}(x_i, \alpha_j) = \frac{\sqrt{(T_{x_i\alpha_j})^2 + (I_{x_i\alpha_j})^2 + (F_{x_i\alpha_j})^2}}{\max(|x_i|, |\alpha_j|)}$$

$$T_{A_i\alpha_j} = a_{ij}b_j - b_{ij}a_j,\ I_{A_i\alpha_j} = b_{ij}c_j - c_{ij}b_j$$

$$F_{A_i,\alpha_j} = c_{ij}a_j - a_{ij}c_j, |x_i| = \sqrt{(a_{ij})^2 + (b_{ij})^2 + (c_{ij})^2}$$

$$|\alpha_j| = \sqrt{(a_j)^2 + (b_j)^2 + (c_j)^2}$$

Hence the similarity measure between each attribute and ideal attribute is calculated.

6) Making a final decision based on model synthesis and sensitivity analysis. The ranking order of all attributes can be determined using the relation $A_i^* = \sum_{j=1}^{i=n} WS^{\perp}(x_i, \alpha_j)$. Then the best decision can be selected easily.

VI. Clinical Application of Neutrosophic Soft Set

Let us consider a decision making problem of radio therapy treatment in oncology. The degree of success of external beam radiation therapy using high energy X-ray for the treatment of tumor cells in movable organs like lungs or chest wall is completely dependent on accurate information of tumor position and tracing of tumor position during treatment. There is a set $X = \{x_1, x_2, x_3, x_4\}$ with four elements(alternatives) which represent treatment methods available for tracking and delivering radiation for movable organs.

1) Target tracking treatment in cyberknife (x_1):- The device cyberknife combines a compact linear accelerator mounted on a robotic manipulator and an integrated image guidance system. The image guidance system acquires images during treatment, tracks the tumor motion and guides the robotic manipulator to precisely and accurately align the treatment beam to the moving tumor.
2) Automatic breath control device (ABC) (x_2):-It provides immobilization of internal organs affected by respiratory motion. The ABC device enables the patient to maintain a breath hold at a predetermined volume and length of time.
3) Cone beam computerized tomography (CBCT) (x_3):- Three dimensional imaging technique using KV-X rays.
4) Fluro portal imaging (x_4):- Two dimensional imaging process using MV-X rays

The oncologist must take a decision according to three criteria or parameters $E = \{e_1, e_2, e_3, e_4\}$. (1) Dosimetry (e_1) (2) Prognosis (e_2) (3) Environmental impact (e_3) In the above mentioned criteria, e_1 and e_2 are benefit criteria and e_3 is the cost criterion. The weight $w = (0.35, 0.25, 0.40)$ of the criteria is given by analytic hierarchical process of experts

in decision making. All alternatives are evaluated under the available criteria and the neutrosophic soft decision matrix $D = (a_{ij})_{4\times 3}$ is constructed as follows.

$$\begin{bmatrix} .45 & .25 & .35 & .50 & .20 & .30 & .80 & .25 & .45 \\ .65 & .15 & .25 & .65 & .15 & .25 & .45 & .40 & .45 \\ .45 & .25 & .35 & .55 & .25 & .35 & .45 & .30 & .80 \\ .75 & .5 & .15 & .65 & .15 & .20 & .65 & .35 & .85 \end{bmatrix}$$

From the above neutrosophic decision matrix D, the ideal attribute α_j $(j = 1, 2, 3)$ can be defined as follows using the definition V.2. α_j is constructed from the evaluation of alternatives x_i with respect to the criteria e_j by the decision maker or experts.

$$\alpha_j = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0.75 & 0.05 & 0.15 \\ 0.65 & 0.15 & 0.20 \\ 0.45 & 0.40 & 0.85 \end{bmatrix}$$

TABLE I: Calculation of weighted similarity measure using the criteria C_1

Attribute	distance $(d^{\perp})$	similarity measure	WSM
A_1	0.5211	0.6574	0.84574
A_2	0.6655	0.6004	0.81107
A_3	0.4116	0.7084	0.86534
A_4	0.0000	1.0000	0.74741

TABLE II: Calculation of weighted similarity measure using the criteria C_2

Attribute	distance $(d^{\perp})$	similarity measure	WSM
A_1	0.0836	0.9228	0.97953
A_2	0.1721	0.8532	0.95875
A_3	0.0.0016	0.9984	0.99960
A_4	0.0000	1.0000	0.80000

TABLE III: Calculation of weighted similarity measure using the criteria C_3

Attribute with	distance $(d^{\perp})$	similarity measure	WSM
A_1	0.5483	0.6076	0.82013
A_2	0.0772	0.9283	0.97004
A_3	0.08813	0.9190	0.73936
A_4	0.2030	0.8313	0.92489

From Table I, II and III,

$x_1^* = \sum_{j=1}^{3} WS^{\perp}(x_1, \alpha_j) = 2.6454$
$x_2^* = \sum_{j=1}^{3} WS^{\perp}(x_2, \alpha_j) = 2.7399$
$x_3^* = \sum_{j=1}^{3} WS^{\perp}(x_3, \alpha_j) = 2.6043$
$x_4^* = \sum_{j=1}^{3} WS^{\perp}(x_4, \alpha_j) = 2.4723$

Also

$$x_2^* > x_1^* > x_3^* > x_4^*$$

So it can be concluded that, the best choice of attribute is x_2. This method is very simple and effective to take an intelligent decision in neutrosophic soft set environment.So the final decision is the best radio therapy treatment method is Automatic breath control device (ABC).

TABLE IV: Ranking order of alternatives x_i

Attribute	x_1	x_2	x_3	x_4
Rank	2	1	3	4

A. Algorithm for decision making using weighted similarity measure in neutrosophic soft environment

Step 1 Develop a model for the decision and break down the model in to hierarchy of goals,criteria and alternatives.

Step 2 Define the weight w_j of each criterion.

Step 3 Derive model synthesis i.e. construct a decision matrix $D = (\alpha_{ij})_{m\times n}$.

Step 4 Calculate the ideal attribute α_j using the evaluation of each attribute A_i on each criteria C_j.

Step 5 Calculate weighted similarity measure (WSM) $WS^{\perp}(x_i, \alpha_j)$.

Step 6 Determine ranking order of all alternatives corresponding to each criterion using $A_i^* = \sum_{j=1}^{i=n} WS^{\perp}(A_i, \alpha_j)$.

VII. Conclusion

The proposed normalized orthogonal distance and weighted similarity measure in neutrosophic soft set are one of the most generalized notions of classical theories to explain vague or uncertain or indeterminate environment. The decision making process using weighted similarity measure can be extended to different fields like engineering and medicine and other highly complex decision making situations. The procedure proposed in this paper for decision making is convenient and simple to adopt for practical purposes.The application of normalized similarity measure in multi attribute decision making in neutrosophic soft environment is illustrated through an example in medical field.

References

[1] Atanassov, K. T. More on intuitionistic fuzzy sets. *Fuzzy sets and systems 33*, 1 (1989), 37–45.

[2] Khameneh, A. Z., and Kiliçman, A. Multi-attribute decision-making based on soft set theory: A systematic review. *Soft Computing* (2018), 1–22.

[3] Liang, Z., and Shi, P. Similarity measures on intuitionistic fuzzy sets. *Pattern Recognition Letters 24*, 15 (2003), 2687–2693.

[4] Maji, P. K., Biswas, R., and Roy, A. Soft set theory. *Computers & Mathematics with Applications 45*, 4-5 (2003), 555–562.

[5] Molodtsov, D. Soft set theoryfirst results. *Computers & Mathematics with Applications 37*, 4-5 (1999), 19–31.

[6] Şahin, R., and Küçük, A. On similarity and entropy of neutrosophic soft sets. *Journal of Intelligent & Fuzzy Systems 27*, 5 (2014), 2417–2430.

[7] Smarandache, F. A unifying field in logics: Neutrosophic logic. In *Philosophy*. American Research Press, 1999, pp. 1–141.

[8] Smarandache, F. Neutrosophic set-a generalization of the intuitionistic fuzzy set. *International journal of pure and applied mathematics 24*, 3 (2005), 287.

[9] Ye, J. *Vector similarity measures of simplified neutrosophic sets and their application in multicriteria decision making*. Infinite Study, 2014.

[10] Zadeh, L. A. Fuzzy sets. *Informtaion and control 8*, 3 (1965), 338–353.

Replay attack detection with raw audio waves and deep learning framework

Shikhar Shukla[1], Jiban Prakash[2] and Ravi Sankar Guntur[3]
IoT R&D
Samsung Research Institute Bangalore,
Bengaluru-560037, India
E-mail: [1]shikhar.0077@samsung.com, [2]p.jiban@samsung.com and [3]ravi.g@samsung.com

***Abstract*—Replay attacks present a significant threat to Automatic Speaker Verification systems as they can be easily mounted using everyday smart devices by any non-professional imposter. The ASVspoof 2017 challenge was an initiative to develop solutions to counteract such replay attacks. The proposed solution builds on the fact that all the distinguishing features between genuine and spoofed audio are not effectively captured by conventional feature extraction techniques. Hence we propose a 1D ConvNet system with raw audio waves as features to it. This approach is able to achieve an EER of 0.41% on development set and 5.29% on evaluation set and hence outperforming best submission to ASVspoof 2017 challenge which had EER of 3.95% and 6.73% on development and evaluation sets respectively.**

Keywords—asvspoof2017, raw audio, CNN, deep learning, replay attack detection

I. Introduction

The Automatic Speaker Verification (ASV) systems are becoming significant and more commonplace due to their numerous commercial applications. As such, these systems need to be robust to spoofing attacks which are categorized into the following types: impersonation, voice synthesis, voice conversion (VC) and replay attacks [1].

Voice synthesis and voice conversion attacks require significant audio domain knowledge. Impersonation attacks can be handled by ASV system itself [2]. Replay attacks can be mounted using speaker's recordings and hence ASV systems are most susceptible to them due to numerous recording devices present around us in this day and age.

ASVspoof 2017 [3] focused on developing standalone countermeasures which can detect replay attacks in highly-varying acoustic conditions. This challenge setup a standard dataset and protocol to benchmark all replay attack detection solutions. This paper presents one such anti-spoofing solution developed by us.

The main aim of this research is to study the effectiveness of raw waveforms as features to CNN for spoofing detection.

II. Related Work

There has been a lot of research on feature extraction using Fast Fourier Transform (FFT), Constant Q Cepstra Coefficients (CQCC) [4] MFCC [5], etc. In [6], authors used FFT spectrograms as feature set to Light CNN which outperformed all the other submissions made to ASVspoof 2017 challenge. The average EER of all submissions made to the challenge was 25.91% [7]. This shows the difficulty of distinguishing between genuine and spoofed audio samples.

The baseline system provided by the organizers of ASVspoof 2017 uses CQCC (constant Q transform cepstral coefficients) as front-end and a standard GMM (Gaussian Mixture model) as classifier to distinguish between genuine and spoofed speech. This solution achieves an EER of 24.77% on the evaluation dataset.

There have been attempts to run audio classification and speech recognition tasks with raw waveforms as features [8]. In [9] authors have argued that most of the distinguishing characteristics between genuine and spoofed speech lie in high frequency region which prompted us to feed filtered signals to our CNN. Inspired by [10] we attempted to combine high-level features obtained from our 1D and 2D models. To the best of our knowledge there has been no analysis of using raw waveforms as feature set for spoofing detection which motivated us to go forward with this approach. The rest of the paper is organized as follows. Section 3 describes the dataset. Section 4 describes the experiment setup. In Section 5 we present the results.

TABLE I. Review of Previous Studies along with our Results

Year	Model basis	Feature Used	EER (Dev Set)	EER (Eval. set)	Ref
2017	LCNN	FFT	4.53	7.37	[6]
2017	LCNN,SVM,RNN	FFT, i-vect	3.95	7.63	[6]
2017	GMM,BLSTM	SFCC	2.21	17.82	[19]
2017	GMM	CQCC	5.13	17.31	[9]
2017	GMM	Cepstrum	3.38	22.24	[9]
2018	ResNet 18 with Attention	Group Delay grams	0.0	0.0	[20]
2019	ConvNet 1D	Raw Audio	1.28	5.29	self
2019	ConvNet 1D + ConvNet 2D	Raw Audio + FFT	0.41	9.45	self

III. Dataset

We used the dataset provided by the organizers of ASVspoof Challenge 2017. Training and dev sets were provided to the participants for developing solutions which were tested on the evaluation set. The dataset consists of 179 replay sessions with 42 speakers recorded in 61 unique replay configurations involving different recording and replay devices. A breakdown of the dataset is as follows:

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

TABLE II. ASVSPOOF 2017 DATABASE

Set:	Genuine	Spoofed	Total
Training	1508	1508	3016
Development	760	950	1710
Evaluation	1298	12008	13306

IV. METHODOLOGY

A. Feature extraction

- *Raw waveforms*

We used raw waveforms of audio files provided as a part of ASVspoof 2017 database as our features. The audio files were sampled at 16 kHz. We could have sampled at a higher rate but that would have provided us with excess data making it hard for CNN model to classify it. Also, we are interested only in frequencies till 8 kHz which are preserved with a sampling rate of 16 kHz. Pre-emphasis was applied on the sampled audio data. This is done to amplify the higher frequencies since they usually have smaller magnitudes as compared to lower frequencies. These raw waveforms were then passed through a bandpass filter [11] as a result of which the resulting audio data only contained specific frequencies. We experimented using different frequency ranges with bandpass filter. The audio lengths in ASVspoof database vary from 1 second to 10 seconds. Since CNN accepts only fixed length data, we repeated the audio data if it was shorter and truncated the data if it was longer than a certain length. We experimented with various lengths of data to gain a better understanding of how it affects the performance of our model. After this the data was normalized which is another necessary step for improving classification accuracy of CNN model.

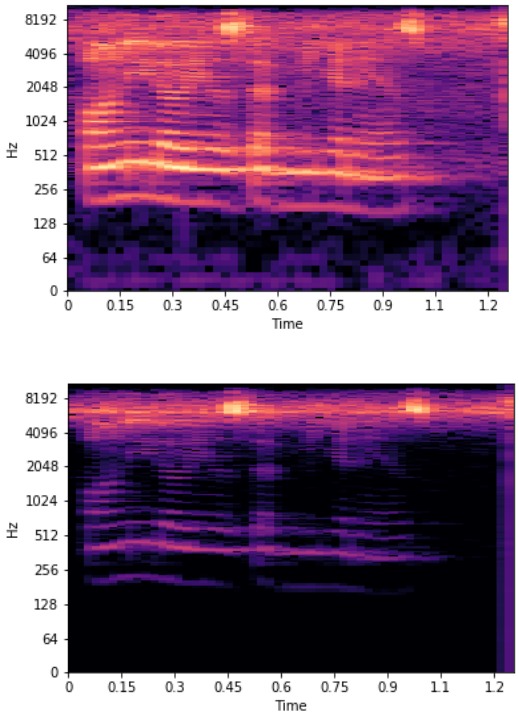


Fig. 1. 1) Magnitude spectrum of raw audio file (above), 2) magnitude spectrum of the same file after bandpass filtering (below) generated using 'librosa'.

- *Spectrogram generation and data augmentation*

We generated magnitude spectrograms of audio data using Fast Fourier transforms. Audio data was sampled at 16 kHz and FFT with window size of 2048 and hop size of 256 using Hanning window was computed. Spectrogram generation was done using 'librosa'[12] library in python. The following data augmentation techniques were applied to spectrograms:

i) Mixup augmentation [13]: Using this augmentation technique the model is trained on a linear mixing of the images in the training dataset.

$$X = X_1*t + X_2*(1-t)$$
$$Y = Y_1*t + Y_2*(1-t)$$

Here, 't' is a random variable of Beta(0.45,0.45). 'X_1' and 'X_2' are images with their labels as 'Y_1' and 'Y_2'. We randomly select two images from the training dataset and after linear mixing, we feed the generated image 'X' with label 'Y' to the model for training.

ii) Random erasing [14]: This selects a random region of the spectrogram image and erases its pixels with random values. This technique makes it more robust to occlusions that might be present in images.

Both these augmentation techniques ensure that overfitting is minimized during training.

B. CNN Models

We experimented with ConvNet 1D and 2D models and also by combining the high level features of both these models for classification.

- *ConvNet 1D*

TABLE III. CONVNET 1D ARCHITECTURE

Type	Filter size/Stride/ # of filters	Output	# of Params.
Conv1a	9 / 1 / 32	80000 x 32	0.3K
Conv1b	9 / 1 / 32	80000 x 32	9.2K
MaxPool1	16 / 16	4250 x 32	0
Conv2a	3 / 1 / 64	4250 x 64	6.2K
Conv2b	3 / 1 / 64	4250 x 64	12.3K
MaxPool2	4 / 4	1062 x 64	0
Conv3a	3 / 1 / 64	1062 x 64	12.3K
Conv3b	3 / 1 / 64	1062 x 64	12.3K
MaxPool3	4 / 4	265 x 64	0
Conv4a	3 / 1 / 128	265 x 128	24.7K
Conv4b	3 / 1 / 128	265 x 128	49.2K
MaxPool4	4 / 4	66 x 128	0
Conv5a	3 / 1 / 256	66 x 256	98.5K
Conv5b	3 / 1 / 256	66 x 256	196.8K
GlobalAver-agePool	-	256	0
Dense1	-	512	131.5K
Dense2	-	2	1.0K
Total	-	-	554.8K

'Relu' is applied after each convolutional layer. Dropout with rate of 0.1 is applied after each convolutional block. The initial learning rate is set to be 0.001 and decays by a factor of 0.3 if validation loss doesn't improve for 5 epochs. Minimum learning rate is set to 1e-6. Adam optimizer[15] is used to optimize the categorical cross-entropy loss. Early stopping is done if loss doesn't improve after a patience of 10 epochs.

- *ConvNet 2D*

The ConvNet 2D architecture [16] is described in Table 3. BatchNormalization and Rectified Linear Unit activations were applied after each convolutional layer. Batch Normalization layer is used to normalize the outputs from the previous layer. It is used to keep the mean of outputs from activation layer close to 0 and standard deviation near 1. This increases the stability of the neural network and speeds up the convergence during training, Dropout rate of 0.6 is used to avoid overfitting on training data. Dropout randomly selects a fraction of units specified by the rate and sets it to 0 during training time. We experimented with grayscale images as inputs to CNN but the performance tends to degrade considerably with them as compared to RGB images. Using Inception [17] modules in our architecture we can get wider networks which enhance the performance of overall architecture. We experimented with 'ResNet' modules instead of Inception modules in our architecture but they led to a degradation in the performance. Deeper networks without 'ResNet' or Inception modules also do not lead to any improvement in the performance. The inception modules are described in the figures following the architecture.

TABLE IV. CONVNET 2D ARCHITECTURE

Type	Filter size/Stride/ # of filters	Output	# of Params.
Conv1 MaxPool1	3x3/2/32 3x3/2/ -	143x143x32 71x71x32	0.9K 0
Conv2 MaxPool2	3x3 /1/32 3x3/2 / -	69x69x32 34x34x32	9.2K 0
Conv3 MaxPool3	3x3/1/ 64 3x3/2/ -	32x32x64 15x15x64	18.4K 0
Inception block A			
Inception block B			
Inception block C			
GlobalAveragePool	-	128	0
Dense1 Dense2	- -	200 2	25.8K 0.4K
Total	-	-	235K

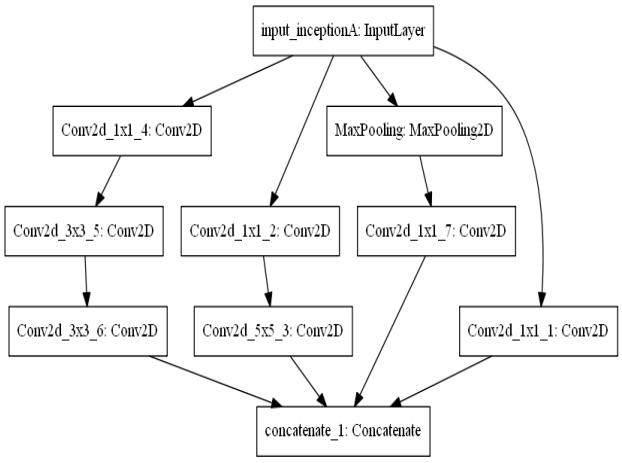


Fig. 2. Inception Block A

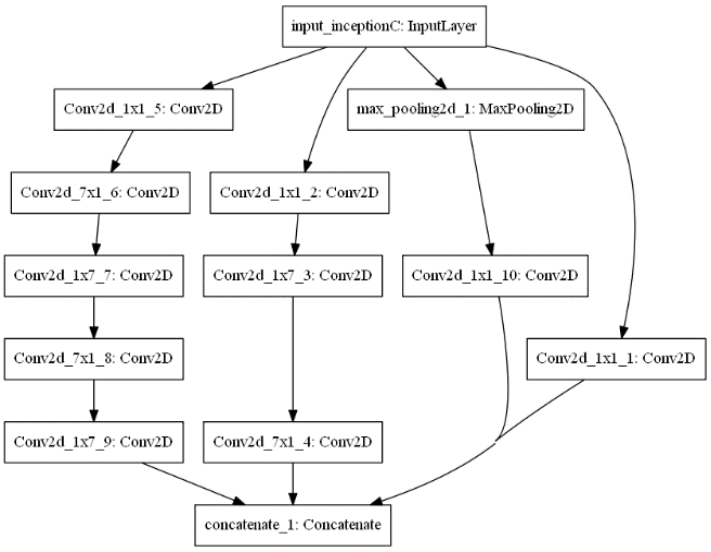


Fig. 3. Inception Block B

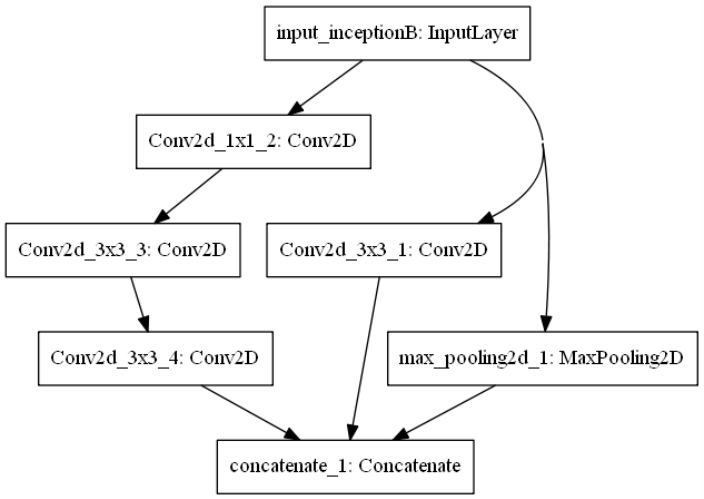


Fig. 4. Inception Block C

In order to combine the high level features extracted by both of our models we removed the last dense layer from our models and concatenated the outputs of the remaining model. The combined features were then fed to a dense layer with 512 hidden units and then passed to 'softmax' layer for classification. Both of the models are trained and their weights are then frozen before concatenating their outputs. After this only the added dense layers are trained.

V. RESULTS AND DISCUSSION

In Fig. 5 and Fig 6, we visualize the feature maps after the third convolutional layer of 2D CNN architecture. Fig. 5 has a genuine sample as input whereas Fig.6 has a spoofed sample as input.

All Equal Error Rates (EER) mentioned below were computed using EER ROCCH of bob.measure library [18]. It is the python equivalent of the official 'Bosaris' toolkit used by the organizers of ASVspoof 2017 for EER calculation.

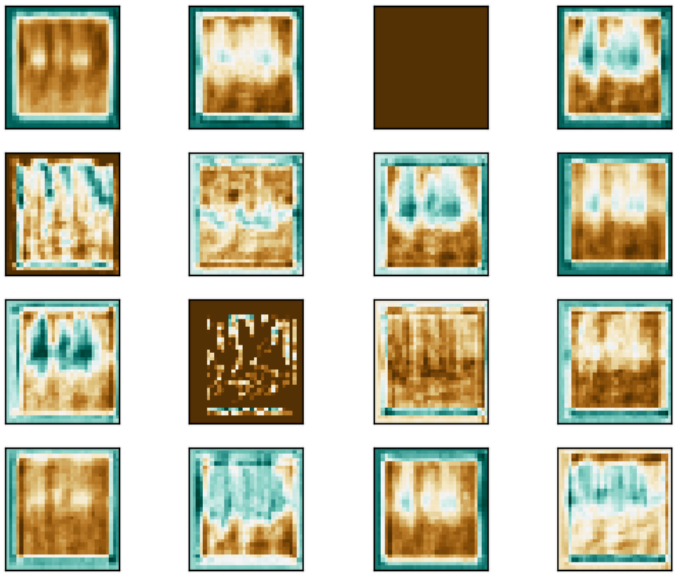


Fig. 5. Activation maps for a genuine sample

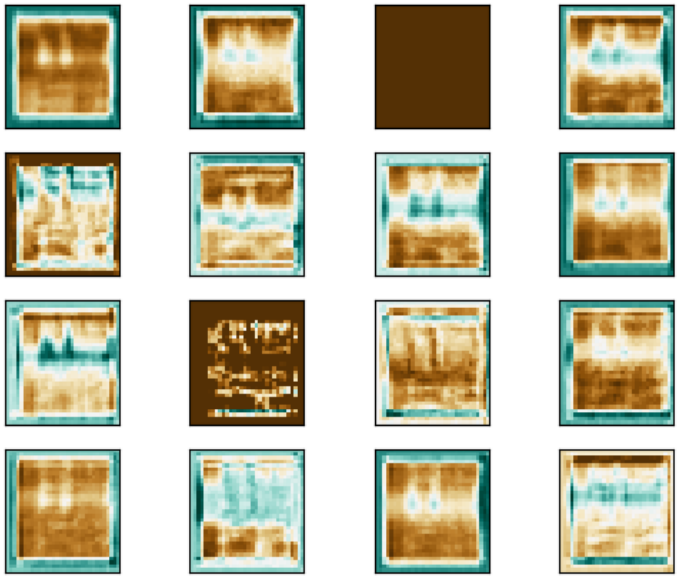


Fig. 6. Activation maps for a spoofed sample

We experimented with different passband frequency ranges in our bandpass filter. The results are as in the table shown below:

TABLE V. VARIATION OF EER WITH PASSBAND FREQUENCY

Passband Frequency	EER
0 - 8000 Hz	29.53%
0 - 4000 Hz	8.80%
4000 - 8000 Hz	3.60%
5000 - 8000 Hz	1.28%
6000 - 8000 Hz	1.80%

From this we observe that the distinguishing features between genuine and spoofed audio are present in the frequency range of 5-8 kHz. The training and inference was done on 5 sec audio samples.

TABLE VI. EFFECT OF AUDIO LENGTH ON EER

Length of audio	EER
2 sec	2.33%
3 sec	3.71%
4 sec	1.69%
5 sec	1.28%

We tried using different lengths of audio as input to our 1D CNN model. If the audio was shorter than the mentioned length then the audio data was replicated and truncated if it was longer. We present our results in Table V. We observe that there is a slight improvement in the EER as length of audio data is increased. Increasing the length of audio data also leads to higher training and inference time. We conclude that there is a trade-off between accuracy and the computation time of our system.

TABLE VII. RESULTS

Training on:	Train set	Train + Dev set
Testing on:	**Dev Set**	**Evaluation set**
EER (CNN 1D)	1.28%	**5.29%**
EER (combined)	**0.41%**	9.45%

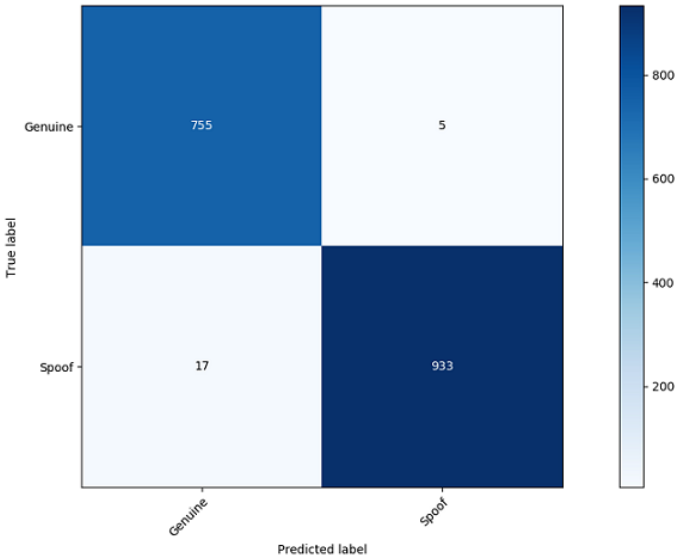


Fig. 7. Confusion matrix for dev set

F1 score on dev set is 0.988 with 933 out of 950 spoofed samples identified correctly and 755 out of 760 genuine samples.
It is observed that in spoofing detection raw waveforms can be effectively used as features for classification. Combining high level features of ConvNet 1D and 2D leads to a further improvement in EER on dev set but it doesn't generalize well to evaluation set.

VI. CONCLUSION

The main contributions of this paper towards the development of ASVspoof systems can be summarised as follows: firstly, we performed an exhaustive analysis of using raw audio waveforms as features to ConvNet. We put forward our results demonstrating how length of audio and the frequency range present in audio effect the performance of the system. Using raw audio as features

to ConvNet can be beneficial since the cost of computing spectrograms proposed by previous works is eliminated without much compromise in accuracy of the system.

Next, we analysed how the combination of high level features of ConvNet 1D and ConvNet 2D impacts the performance of the system. It is observed that although this approach performs better on development set, it does not generalize well to the evaluation set.

Future Scope: In future, the performance of ConvNet 1D with raw audio features can be further improved by applying silence removal in the audio. Also, data can be augmented using techniques like addition of random noise and pitch shifting. These help in making the system more robust to noisy real world scenarios. Besides, mix-up augmentation can also be applied to raw audio to reduce overfitting in ConvNet 1D.

Further, there can be exploration on how ConvNet 1D and ConvNet 2D can benefit by using skip connections in their architectures. Also, 'attention' mechanism — which can selectively focus on distinguishing visual features — can be used to further enhance the performance of Convolutional Neural Networks.

REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun*., vol. 66, pp. 130–153, Feb. 2015.

[2] R. G. Hautamaki, T. Kinnunen, V. Hautam ¨ aki, T. Leino, and A.- ¨ M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proc. Interspeech*, 2013

[3] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee. "The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge" *in IEEE Journal of Selected Topics in Signal Processing.* IEEE, 2017 PP(99):1-1

[4] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Speaker Lang. Recognit. Workshop*, Bilbao, Spain, Jun. 21–24, 2016, pp. 283–290.

[5] R. Font, J. M. Espín, and M. J. Cano, "Experimental Analysis of Features for Replay Attack Detection — Results on the ASVspoof 2017 Challenge," in *Interspeech*, 2017, pp. 7–11.

[6] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Interspeech*, 2017, pp. 82–86.

[7] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, 2017.

[8] Jongpil Lee, Taejun Kim, Jiyoung Park, and Juhan Nam, "Raw waveform-based audio classification using sample-level CNN architectures". *CoRR*, abs/1712.00866, 2017

[9] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio Replay Attack Detection Using High-Frequency Features," in *Interspeech*, 2017, pp. 27–31.

[10] M. Lederle and B. Wilhelm, "Combining high-level features of raw audio waves and mel-spectrograms for audio tagging," *arXiv preprint arXiv: 1811.10708*, 2018.

[11] Jones E, Oliphant E, Peterson P, *et al.* SciPy: Open Source Scientific Tools for Python, 2001-, http://www.scipy.org/ [Online; accessed 2019-05-06].

[12] McFee, Brian & Raffel, Colin & Liang, Dawen & Ellis, Daniel & Mcvicar, Matt & Battenberg, Eric & Nieto, Oriol. (2015). librosa: Audio and Music Signal Analysis in Python. 18-24. 10.25080/Majora-7b98e3ed-003.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization", *arXiv preprint arXiv:1710.09412,* 2017.

[14] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: "Random erasing data augmentation". *arXiv preprint arXiv: 1708.04896*, 2017.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[16] Qingkai Wei, Yanfang Liu & Xiaohui Ruan, "A report on audio tagging with deeper CNN, 1D-CONVNET and 2D-CONVNET" in *Detection and Classification of Acoustic Scenes and Events 2018.*

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich et al., "Going deeper with convolutions", *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015.

[18] André Anjos, El L Shafey, Roy Wallace, Manuel Günther, Chris Mccool, & Sébastien Marcel. "Bob: a free signal processing and machine learning toolbox for researchers" in *20th ACM Conference on Multimedia Systems*. 10.1145/2393347.2396517. 2012.

[19] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[20] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *INTERSPEECH*, Hyderabad, India, 2018.

Frontal Gait Recognition based on Hierarchical Centroid Shape Descriptor and Similarity Measurement

Anusha R and Jaidhar C D
Department of Information Technology
National Institute of Technology Karnataka, Surathkal, India
it16fv01.anusha@nitk.edu.in; jaidharcd@nitk.edu.in

***Abstract*—Gait recognition is an expanding stream in biometrics, intended to recognize individuals through the investigation of their walking pattern. This pattern is obtained from a distance, without the active participation of the people. One of the difficulties of the appearance-based gait approach is to enhance the performance of frontal gait recognition, as it carries less spatial and temporal data when compared with other view variations. As a result, to increase the performance of the frontal gait recognition, this paper presents a method which uses two-step procedure; the Hierarchical centroid Shape descriptor (HCSD) and the similarity measurement. The proposed method was assessed on the broadly used CASIA A, CASIA B, and CMU MoBo gait databases. The experimental outcomes showed that the proposed method gave promising results and outperforms certain state-of-the-art methods in terms of recognition performance.**

***Index Terms*—Classification, feature extraction, gait recognition, human identification**

I. Introduction

Gait recognition is a type of behavioral biometrics which has achieved significant consideration in the last decade. The advantage of this method is that the input data can be collected at a distance with minimum cooperation from the subjects when compared with other physiological and behavioral biometric modalities. This characteristic makes it primarily attractive for investigation of criminals, surveillance, and security applications [1].

The existing gait recognition methods can be typically divided into appearance-based and model-based methods. The modeling of the motion of the human body from the gait sequences is carried out in the model-based methods [2]. These methods illustrate the kinematics and kinetics of a person's joints to compute gait characteristics such as directions, hip, knee and ankle movements, hand movement, etc. These methodologies are usually scaling invariant and are computationally expensive in general, as they need the tracking and modeling of the subject's body. Moreover, they also need high-resolution images.

In contrast, the operations are directly carried out on the processed gait images without using any explicit model in the appearance-based methods. The various gait templates used in the appearance-based techniques provides a mostly successful solution to the gait detection problem in the literature. Han and Bhanu [3] proposed a concept called GEI, where a gait cycle is converted into a single greyscale image. Roy et al. represented a gait sequence using a set of poses and extracted pose energy image as the average image of all the key poses. Later on, some methods based on Gait Flow Image (GFI) [4], Gait Entropy Image (GEnI) [5], and Chrono Gait Image (CGI) [6] were also proposed.

The details of the several works which focused exclusively on frontal gait recognition are as follows. Sivapalan et al. [7] extended the concept of GEI to 3D, resulting in the creation of gait energy volume and frontal depth images. A feature descriptor called curve spread was proposed by Soriano et al. [8] for front view gait videos. Here, the Freeman code is used to represent the minute time-variations of the silhouette outline of a moving body as a 2D vector. Sivapalan et al. [9] proposed a backfilled GEI obtained from both side-view silhouettes and frontal depth images for frontal gait recognition.

A Spatio-Temporal Interest Point (STIP) based method was proposed by Huang et al. [10], in which the histogram of oriented gradients was extracted directly from the frontal gait videos without the extraction of gait silhouettes. Barnich and Droogenbroeck [11] proposed an intra-frame description of the silhouettes approach consisting of a set of rectangles, which fitted into any closed silhouette. Chattopadhyay et al. [12] reconstructed partial volume of the surface of each silhouette obtained from the frontal view and derived feature called as Pose Depth Volume (PDV). He further extracted the features of the back view from the depth data and the edge of the silhouette [13]. Even though all these methods are designed to provide solutions for frontal gait recognition, most of them use the gait dataset containing less number of subjects obtained from Kinect for performance analysis.

II. Motivation and Contribution

The performance of the gait recognition systems is predominantly influenced by various covariate aspects, such as view angle variations, walking surface conditions, elapsed time, shoe type, walking speed, and carrying conditions. Among all these various covariates, one of the most vital factors is the view angle variation. The generally challenging of all the view variations is the frontal view as the spatial and temporal changes are very less noticeable in frontal view, while compared to other view variations. Hence this paper

978-1-7281-2087- 4/19/$31.00 ©

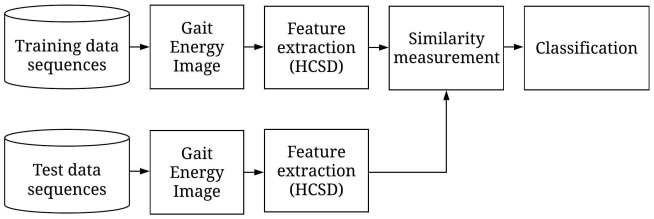


Fig. 1. Framework of the proposed approach.

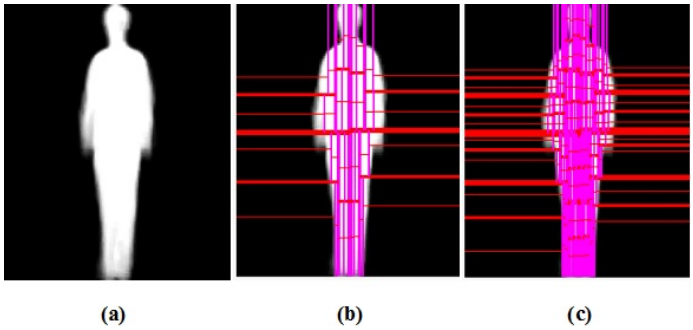


Fig. 2. Sample image representing (a) frontal GEI (b) HCSD extraction from a frontal GEI for level 6 of kd-tree decomposition (c) HCSD extraction from a frontal GEI for level 8 of kd-tree decomposition.

proposes a method which increases the performance of the frontal gait recognition system.

Among appearance-based methods, most of the existing individual identification systems are based on GEI, as it preserves and represents the spatial and temporal variations of a gait cycle. The front view GEI contains very less temporal information when compared with other lateral views. Hence, the frontal gait detection performance can be enhanced by building a solid representation for spatial data, which increases the inter-class variance. To accomplish this goal, the proposed work uses hierarchical centroid descriptor and similarity measurement.

The contributions of this paper can be summarized as follows.

1. The GEI is a greyscale image, which can be represented by the spatial distribution of pixels. Here, the frontal GEI contains the maximum number of pixels representing the same information. Hence, the HCSD is used to extract accurate shape information.
2. The steps of similarity measurement presented here increase the recognition performance, as it increases the inter-class variance even when different subjects shape information is much similar.
3. The performance of the proposed method is evaluated on three benchmark gait databases, and the results are compared with the state-of-the-art gait recognition approaches.

The rest of the paper is organized as follows: The proposed methodology is explained in Section 3. Experimental results are illustrated in Section 4, and the conclusions are given in Section 5.

III. Proposed Method

An outline of the proposed method is shown in Figure 1. The pre-processing of the gait video is done by steps such as background subtraction, de-noising, and normalization to obtain gait silhouettes, which were later combined over a gait cycle to form a GEI. The proposed method consists of three different steps: HCSD feature extraction, similarity measurement, and classification. Algorithm 1 demonstrates the entire process of recognition.

Algorithm 1 Process of recognition

Input: Collection of all gallery samples, a probe sample.
Output: A subject.

1: Begin
2: Preprocess input to extract GEI's.
3: Extract HCSD features from GEI's of all gallery samples G and probe sample p.
4: Perform similarity measurement using G and p which results in the list of distances D^i.
5: Perform classification to identify the subject belongs to p.
6: End

A. GEI Extraction

A greyscale image acquired by averaging the gait silhouette images extracted over a gait cycle is called GEI. It is computed as follows:

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B(x,y,t) \tag{1}$$

Where, N represents the total number of frames of a gait cycle, B specifies the silhouette image, x and y signify the spatial co-ordinates of the image, and t shows the frame number in the gait cycle [14].

B. Hierarchical Centroid Shape Descriptor

The shape descriptor formed using the centroid coordinates obtained from a greyscale image is called HCSD [15]. This descriptor is extracted by the decomposition of an image into sub-images recursively by using the kd-tree representation, where the data is divided with reference to the center of gravity at each level of decomposition as shown in Figure 3. This process gives sub-images at each stage, which usually differs in size. For each sub-image, the centroid coordinates for the local region are extracted and are stored in the corresponding level of a kd-tree. As a result, a descriptor, whose size is determined by the depth of decomposition is obtained. The sample image representing frontal GEI and HCSD for different levels of kd-tree decomposition is shown in Figure 2 and 4.

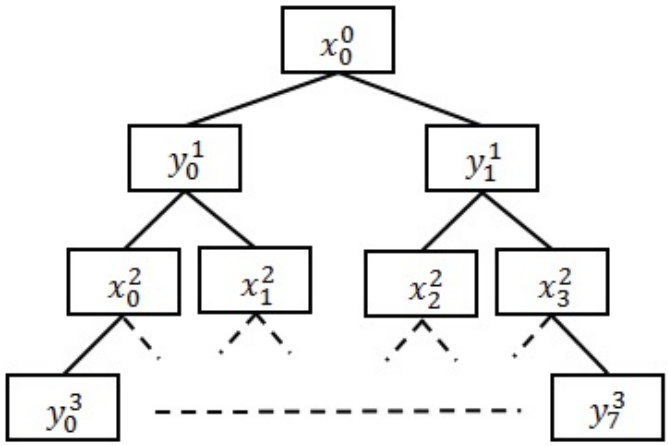


Fig. 3. Structure of kd-tree decomposition [15].

The length of the descriptor is given by $2 \times (2^d - 2)$, where d represents the depth of the feature extraction process. Let I be the greyscale image of size $A \times B$, with background I_b and foreground I_f, the HCSD is obtained by the following steps.

1. Consider an input image I and its transpose I^T,
2. Compute the centroid $C(x_{ct}, y_{ct})$ for each input, at the root level by using (2) and (3).

$$x_{ct} = \frac{m_{10}}{m_{00}} \tag{2}$$

$$y_{ct} = \frac{m_{01}}{m_{00}} \tag{3}$$

where m_{10}, m_{01} represents the first order moment along the x-axis and y-axis, m_{00} signifies the area of I_f. The moment of order $(r+s)$ of an image with pixel intensities $I(i,j)$ is defined as

$$m_{rs} = \sum_{i=0}^{A} \sum_{j=0}^{B} i^r j^s I(i,j) \tag{4}$$

3. Recursively split the image into two sub-images based on the centroid until the desired depth of decomposition is reached. The axis of coordinates obtained is altered at each consecutive level.
4. Normalization of the descriptor in the range -0.5 to 0.5 is done where 0 denotes the centroid of the root level. The positive values represent the features from the tree decomposition of the right part while the negative values portray the left side of the image.
5. Concatenate the extracted features from the image I and I^T.

C. Similarity measurement

The similarity measurement section compares each probe sample p to all features from the gallery samples. The result of similarity measurement is a list of distances for each p. Namely $dist^i$ containing the distances of p to each gallery feature. This vector is obtained by finding the distance d between two normalized feature vectors, probe p, and gallery g as follows:

$$d = \sum_{i=1}^{f} (p(i) - g(i))^2 \tag{5}$$

where $i = 1, 2, 3, ..., f$ be the total number of HCSD features.

The steps followed in this module is shown in Algorithm 2. In similarity measurement, the minimum difference between the summation of the distances obtained for a particular subject and probe is considered for classification, instead of the minimum distance between a probe and a gallery sample. This is because when a probe sample and a gallery sample belonging to different subjects have largely the same spatial data, by considering individual gallery samples, the difference between many gallery samples and a probe sample may give the same value. This element decreases the recognition performance, whereas the summation of the distances belonging to a particular subject gives different values so that the minimum value can be selected, thus increasing the recognition performance.

Algorithm 2 Similarity measurement

Input: Gallery features, a probe feature.
Output: List of distances.

1: Begin
2: Let the number of subjects present in the gallery dataset be $S = \{s_1, s_2, ..., s_n\}$.
3: Each subject persent in the gallery dataset consists of t number of GEI's, $s_i = \{g_i^1, g_i^2, ..., g_i^t\}$.
4: Extract HCSD features from GEI's of all gallery samples, $G = \{g_1^1, g_1^2, .., g_1^t, g_2^1, g_2^2, .., g_2^t, ..., g_n^1, g_n^2, .., g_n^t\}$.
5: Extract HCSD feature from GEI of a probe sample, p.
6: **for** each $g_i^j \in G$ **do** where $i = \{1, 2, ..., n\}$ and $j = \{1, 2, ..., t\}$
7: Compute the distance d between HCSD features of p and g_i^j.
8: **end for**
9: Let $dist^i = \{d_i^1, d_i^2, ..., d_i^t\}$ be the distances obtained between p and gallery samples belonging to subject s_i.
10: Perform the addition of the distances obtained between p and $dist^i$ to obtain $D^i = \{d_i^1 + d_i^2 + ... + d_i^t\}$
11: The above step results in the list of distances $D^i = D^1, D^2, ..., D^n$.
12: Perform classification using D^i to identify the subject S which p belongs to.
13: End

D. Classification

Let $S = s_1, s_2, ..., s_n$ be the number of subjects present in the gallery dataset. Consider a vector consisting of a list of distances $D^i = D^1, D^2, ..., D^n$ which represent the distance between the subjects $s_1, s_2, ..., s_n$ and probe sample p after performing the similarity measurement. Let the minimum value present in vector D^i is given by $D^m = min(D^i)$. If

TABLE I
COMPARISION OF RECOGNITION ACCURACIES OF THE PROPOSED METHOD WITH EXISTING METHODS ON CMU MOBO DATABASE.

Method	B/F	B/I	B/S	F/B	F/I	F/S	I/B	I/F	I/S	S/B	S/F	S/I
Huang et al. [10]	88	88	**96**	92	84	96	75	**88**	92	**96**	96	92
Proposed method	**90**	**90**	**96**	**94**	**88**	**98**	**82**	**88**	**96**	92	**98**	**94**

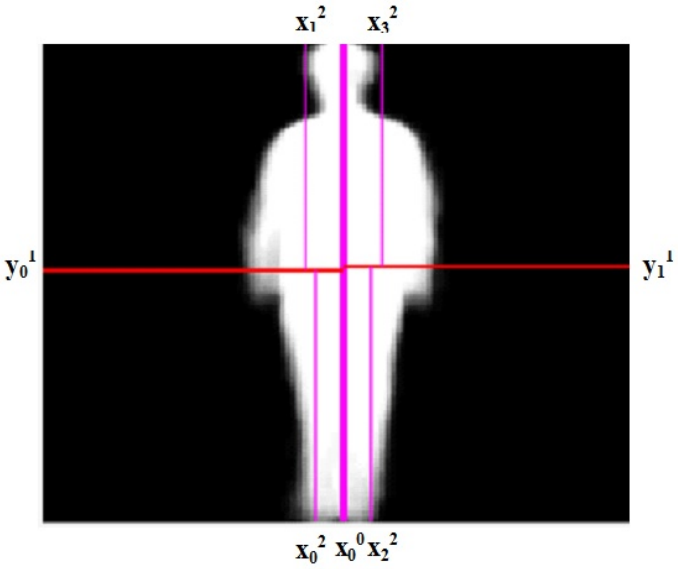


Fig. 4. HCSD extraction from a frontal GEI for level 2 of kd-tree decomposition.

D^m is the distance between probe p and subject s_i, then the probe sample p is assigned to s_i. This is because the minimum distance signifies that there is the least difference between the shape and static data between the probe and gallery sample.

IV. EXPERIMENTAL SETUP, RESULTS, AND DISCUSSION

A. Experimental setup

This experiment was carried out in a Windows 10 operating system with Intel Core i5-7200U CPU @ 2.50 GHz processor. The software tool used for the implementation of the proposed work is Matlab R2017b.

In this paper, the depth d of the features extraction process to extract HCSD is set to 8. As a result, the length of HCSD obtained for each GEI is of size 1×508. The correlation of the feature vector extracted is illustrated in the feature space diagram. Figure 5(a) shows the features extracted from the same subject for three different GEI's. The features extracted from three different subjects for a single GEI is shown in Figure 5(b). It is apparent from the feature space diagram that extremely minute differences are observed within three GEI's of the same subject (GEI1_S1, GEI2_S1, GEI3_S1), but a considerable difference is formed for GEI of three different subjects (GEI1_S1, GEI1_S2, GEI1_S3). Hence the value of difference $dist$ when probe p and gallery g belongs to the same subject is around 0 to 0.004, whereas the value of $dist$ when probe p and gallery g belongs to the different subject is around 0.084 to 0.097.

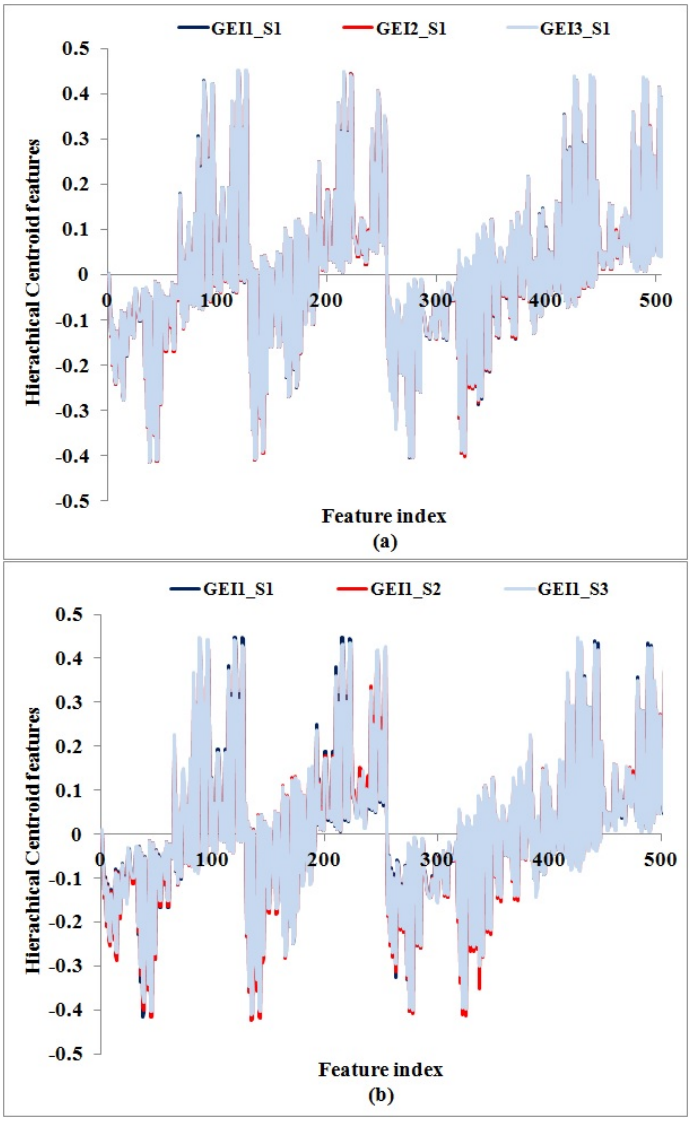


Fig. 5. Feature space diagram representing HCSD features of three (a) GEI's belonging to same subject (b) GEI's belonging to different subjects.

Initially, the experiment was conducted by assigning different values to depth d, such as 4, 5, 6, 7, 8 and 9. However, the encouraging results were obtained when its value is set to 8. The performance of the gait recognition system was measured by the Correct Classification Rate (CCR) [1] on the testing dataset.

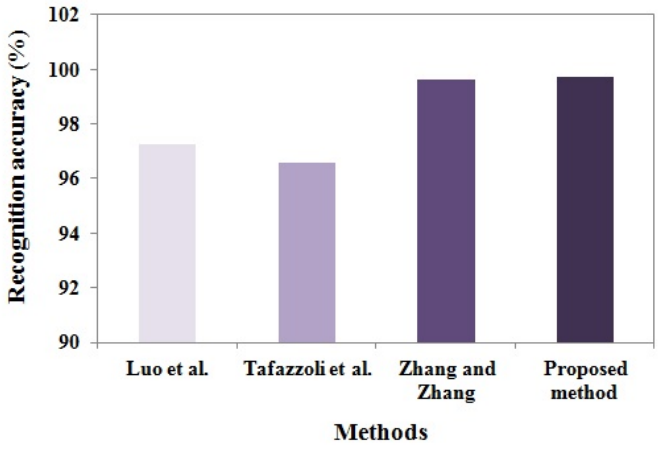


Fig. 6. Recognition accuracy (%) of the proposed method on CASIA A gait database and comparision of it with other existing methods.

B. Experimental results and discussion

The performance of the proposed method is verified on three extensively used gait databases. They are (1) CASIA A gait database (2) CASIA B gait database and (3) CMU MoBo database.

The CASIA A gait database [16] consists of 20 different subjects. Each subject walks in three different directions, i.e., 0^0, 45^0, and 90^0 to the image plane. Four sequences are obtained from all individuals in each direction. Therefore, for the 0^0 view, 80 sequences were obtained from 20 subjects. The performance of the proposed approach on CASIA A database for 0^0 view is reported in Figure 6. Three sequences out of 4 sequences are considered for training, and the remaining sequence is used for testing. The proposed method is compared with other existing methods such as Luo et al. [17], Zhang and Zhang [18] and Tafazzoli et al. [19]. Figure II shows that the proposed method gives higher recognition accuracy.

The CMU MoBo database [20] has videos of 25 individuals walking on a treadmill. The videos of gait were obtained in different types of walking such as fast walk (F), slow walk (S), walking with a ball in one hand (B), and walking on an inclined plane (I). All the four forms of walking were used for both gallery and probe datasets, where F/B represents the gallery F and the probe B. The results of the proposed method on the CMU MoBo database for the frontal view are represented in Table I. It is obvious from the Table I that the proposed method gives high CCR in most of the cases. It outperformed the other method in almost all cases of gallery/probe combinations. The experiment was performed on the most demanding cases which involve walking on an inclined plane and ball in hands, such as F/I, B/I, I/B, I/F, B/F, and F/B. The proposed method offered good performance in more challenging cases.

"The CASIA B gait database [16] is a multiview gait database which consists of 124 subjects captured from 11 distinct angles starting from 0^0 to 180^0. Each subject has six normal walking sequences (NM), two carrying conditions sequences (BG) and two clothing variations sequences (CL) [21]."

TABLE II
PERFORMANCE OF THE PROPOSED METHOD AND EXISTING METHODS ON CASIA B GAIT DATABASE.

	0^0	0^0	0^0	180^0	180^0	180^0
Methodology	NM/NM	NM/BG	NM/CL	NM/NM	NM/BG	NM/CL
Dupuis et al. [22]	97.17	73.15	81.64	99.60	74.56	82.70
Choudhury et al. [23]	100.0	93.00	67.00	99.00	89.00	66.00
Rida et al. [24]	97.97	72.76	80.49	97.58	76.11	83.06
Alotaibi et al. [25]	90.67	91.98	88.77	83.99	87.76	90.00
Isaac et al. [21]	98.50	95.00	97.00	**98.99**	94.44	93.94
Proposed method	**98.70**	**97.00**	**97.56**	98.97	**95.06**	**94.58**

The two experiments performed on this database are as follows. At first, four sequences of NM were used for training and the remaining two sequences of NM, CL, and BG was used for testing. These testing sequences are employed to measure the performance of normal, clothing, and carrying variations, respectively. Given that the GEI's of 0^0 and 180^0 view is similar to a large extent, the experiments were conducted on both 0^0 and 180^0 views to assess the performance of the proposed method. The results presented in Table II shows that the proposed method gives considerably high CCR when compared to the other methods reported in the literature. They also demonstrate the capability of the proposed method in managing the carrying and clothing variations.

Secondly, a training dataset consisting of NM gait sequences of 0^0 and 180^0 viewpoint were constructed. For the training process, four NM gait sequences of each subject were used, which leads to 496 GEI's in the training dataset. The testing dataset consists of different number of GEI's at each experiment as shown in Table 2. The performance of the proposed method for 16 experiments is illustrated in Table III. The results show that the proposed method efficiently captures the statistical information present in the frontal gait images.

V. CONCLUSIONS

In this paper, the method which increases the performance of the frontal gait recognition is proposed using the two-step procedure. The first step makes use of a shape descriptor based on hierarchical centroid to extract gait features. The second step called similarity measurement is used to assign the probe sample to a set of gallery samples. Extensive experimentation on the three gait databases shows the efficiency of the proposed method as it performs better than several existing approaches in the literature. The overall experimental results demonstrate that the proposed method is susceptible to significant spatial variations in gait sequences and thus increases the inter-class variance. Further, the experiments on more extensive and varied database need to be done, and the research can be directed at extending the proposed gait recognition method to

TABLE III
THE RECOGNITION RATES OF THE PROPOSED METHOD ON CASIA B GAIT DATABASE.

Experiment	Gallery set	Probe set	Gallery Size	Probe Size	CCR (%)
1	0^0(NM)	180^0(NM)	124×4	124×4	90.97
2	0^0(NM)	180^0(NM)	124×4	124×3	91.19
3	0^0(NM)	180^0(NM)	124×4	124×2	92.79
4	0^0(NM)	180^0(NM)	124×4	124×1	92.50
5	0^0(NM)	180^0(BG)	124×4	124×2	93.15
6	0^0(NM)	180^0(BG)	124×4	124×1	93.60
7	0^0(NM)	180^0(CL)	124×4	124×2	91.60
8	0^0(NM)	180^0(CL)	124×4	124×1	91.41
9	180^0(NM)	0^0(NM)	124×4	124×4	89.59
10	180^0(NM)	0^0(NM)	124×4	124×3	90.76
11	180^0(NM)	0^0(NM)	124×4	124×2	90.39
12	180^0(NM)	0^0(NM)	124×4	124×1	91.81
13	180^0(NM)	0^0(BG)	124×4	124×2	94.56
14	180^0(NM)	0^0(BG)	124×4	124×1	94.96
15	180^0(NM)	0^0(CL)	124×4	124×2	93.79
16	180^0(NM)	0^0(CL)	124×4	124×1	94.60

obtain high recognition accuracy with variations in clothing and carrying conditions.

ACKNOWLEDGMENT

The authors are extremely thankful to the team behind CASIA [16] and CMU MoBo [20] gait databases. This work is supported by Visvesvaraya Ph.D. Scheme, Ministry of Electronics and Information Technology, Government of India.

REFERENCES

[1] V. B. Semwal, M. Raj, and G. C. Nandi, "Biometric gait identification based on a multilayer perceptron," *Robotics and Autonomous Systems*, vol. 65, pp. 65–75, 2015.

[2] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE transactions on cybernetics*, vol. 46, no. 7, pp. 1602–1615, 2016.

[3] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 2, pp. 316–322, 2006.

[4] T. H. Lam, K. H. Cheung, and J. N. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern recognition*, vol. 44, no. 4, pp. 973–987, 2011.

[5] A. Nandy, A. Pathak, and P. Chakraborty, "A study on gait entropy image analysis for clothing invariant human identification," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9133–9167, 2017.

[6] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *European Conference on Computer Vision*, pp. 257–270, Springer, 2010.

[7] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *Biometrics (IJCB), 2011 International Joint Conference on*, pp. 1–6, IEEE, 2011.

[8] M. Soriano, A. Araullo, and C. Saloma, "Curve spreads-a biometric from front-view gait video," *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1595–1602, 2004.

[9] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "The backfilled gei-a cross-capture modality gait feature for frontal and side-view gait recognition," in *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pp. 1–8, IEEE, 2012.

[10] C.-C. Huang, C.-C. Hsu, H.-Y. Liao, S.-H. Yang, L.-L. Wang, and S.-Y. Chen, "Frontal gait recognition based on spatio-temporal interest points," *Journal of the Chinese Institute of Engineers*, vol. 39, no. 8, pp. 997–1002, 2016.

[11] O. Barnich and M. Van Droogenbroeck, "Frontal-view gait recognition by intra-and inter-frame rectangle size distribution," *Pattern Recognition Letters*, vol. 30, no. 10, pp. 893–901, 2009.

[12] P. Chattopadhyay, A. Roy, S. Sural, and J. Mukhopadhyay, "Pose depth volume extraction from rgb-d streams for frontal gait recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 53–63, 2014.

[13] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from occluded scenes," *Pattern Recognition Letters*, vol. 63, pp. 9–15, 2015.

[14] I. Rida, S. Almaadeed, and A. Bouridane, "Gait recognition based on modified phase-only correlation," *Signal, Image and Video Processing*, vol. 10, no. 3, pp. 463–470, 2016.

[15] E. Ilunga-Mbuyamba, J. G. Avina-Cervantes, D. Lindner, J. Guerrero-Turrubiates, and C. Chalopin, "Automatic brain tumor tissue detection based on hierarchical centroid shape descriptor in tl-weighted mr images," in *2016 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pp. 62–67, IEEE, 2016.

[16] S. Zheng, *CASIA Gait Database*, accessed July 27, 2017.

[17] J. Luo, J. Zhang, C. Zi, Y. Niu, H. Tian, and C. Xiu, "Gait recognition using gei and afdei," *International Journal of Optics*, vol. 2015, 2015.

[18] S. Zhang and L. Zhang, "Combining weighted adaptive cs-lbp and local linear discriminant projection for gait recognition," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 12331–12347, 2018.

[19] F. Tafazzoli, G. Bebis, S. Louis, and M. Hussain, "Genetic feature selection for gait recognition," *Journal of Electronic Imaging*, vol. 24, no. 1, p. 013036, 2015.

[20] R. Gross and J. Shi, "The cmu motion of body (mobo) database," Tech. Rep. CMU-RI-TR-01-18, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[21] E. R. Isaac, S. Elias, S. Rajagopalan, and K. Easwarakumar, "View-invariant gait recognition through genetic template segmentation," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1188–1192, 2017.

[22] Y. Dupuis, X. Savatier, and P. Vasseur, "Feature subset selection applied to model-free gait recognition," *Image and vision computing*, vol. 31, no. 8, pp. 580–591, 2013.

[23] S. D. Choudhury and T. Tjahjadi, "Robust view-invariant multiscale gait recognition," *Pattern Recognition*, vol. 48, no. 3, pp. 798–811, 2015.

[24] I. Rida, X. Jiang, and G. L. Marcialis, "Human body part selection by group lasso of motion for model-free gait recognition," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 154–158, 2016.

[25] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Computer Vision and Image Understanding*, vol. 164, pp. 103–110, 2017.

Human Activity Recognition using Deep Neural Network

Piyush Mishra*, Sourankana Dey†, Suvro Shankar Ghosh‡,
Dibyendu Bikash Seal§, Saptarsi Goswami¶
*International Institute of Information Technology, Bhubaneswar
†Neotia Institute of Technology, Management and Science, Kolkata
‡MUST Research Club, India
§¶University of Calcutta, Kolkata
Email: *piyushmishra1999@gmail.com, †sourankanadey30@gmail.com, ‡suvroshankar@gmail.com,
§dibs.adi@gmail.com, §saptarsi007@gmail.com

***Abstract*—The smartphone has become quite ubiquitous and an indispensable part of our lives in the modern day. It has many sensors which capture several minute details pertaining to our activities. So, it is but inevitable that human desire creeps in to augment and improve one's own actions by studying such behaviour captured through the instrumentalities of the smartphone. In this context, study of data on human activities captured through accelerometer and gyroscope get primal significance. In this paper, we have attempted to apply machine learning and deep learning techniques on a publicly available dataset. Initially, classification algorithms like K-Nearest Neighbours and Random Forest are applied. The classification accuracies observed are 90.46% and 92.97% respectively. Using benchmark feature selection and dimensionality reduction techniques does not improve the model accuracies to a large extent – with reported accuracies of 91.48% and 92.56% respectively. However, on employing deep neural network techniques, an accuracy of 97.32% is achieved, which indicates suitability of deep learning techniques over traditional machine learning techniques for the task of human activity recognition using mobile sensor data.**

***Index Terms*—Human activity recognition, machine learning, feature selection, deep neural network**

I. Introduction

The dependence on smartphones for an average human is increasing day by day. A social experiment conducted by Sumathi et al. [1], with academic students as subjects, found that 34% of the respondents spent 5 to 7 hours of their day on smartphones on academic and social communication itself. Out of those respondents, 55% responded positively towards the use of such devices. In another study, 72% of the respondents stated that they are comfortable and familiar with the basic knowledge and functioning of smartphones [2]. In an independent survey by Gupta et al. [3], 18% of respondents admitted to using smartphones in the classrooms. Due to this evidently increased dependence on smartphones, it would be beneficial, if patterns of human activity can be studied using smartphones as the source of data. Generally, a smartphone has many movement tracking devices, such as accelerometers and gyroscopes, as part of its hardware structure. So, it becomes quite natural that the data generated from these devices can be used for analysis and for reaching conclusions. In [4], Ranasinghe et al. mention that the task of activity recognition has traditionally been carried out by human operators. However, an increase in the number of technical monitoring devices has been able to eliminate the involvement of these human operators for more efficiency. The results of human activity recognition can have multiple applications like anti-terrorism security, surveillance, anti-crime security, among others [5]. Many healthcare professionals can be benefitted through the use of a mechanism that is trained using this smartphone generated data [6], to gain insights on the patterns of behaviour that one shows with regard to one's day-to-day activities. Moreover, detecting and identifying these activities is an economical and effective way of monitoring health and fitness.

Keeping this in mind, we have worked, using machine learning and deep learning techniques, towards human activity recognition, whose data is provided through smartphone accelerometers and gyroscopes. Since these in-built devices collect data constantly, they act as very helpful sources to classify and study the basic activities that humans do: walking, sitting, standing etc. In this study, we compare and contrast between many classical machine learning methods, as well as a relatively novel approach of the neural network architecture, to classify various activities effectively.

Since this is a classic case of multi-class classification, we use machine learning algorithms like KNN and Random Forest classifier [17]. We also use feature selection and dimensionality reduction methods to reduce the computational cost while increasing accuracies [12]. We further delve into deep learning approaches and develop a deep neural network (DNN) architecture to tackle the problem at hand.

In particular, the main contributions of this paper are summarised as follows.

(1) Exploratory data analysis (EDA): Initially, EDA is carried out on the dataset to explore different relationships between variables, as well as to check the distribution of the data.

(2) Feature selection and dimensionality reduction: The use of feature selection and dimensionality reduction is explored, using RFE and PCA respectively, to improve the recognition rate and reduce the time cost in prediction.

978-1-7281-2087- 4/19/$31.00 © 2019 IEEE

(3) Classification: After reducing the dimension of the feature vector, K-Nearest Neighbours and Random Forest (classical machine learning) algorithms and deep neural network approach are employed for classification.
(4) Experimental evaluation: Extensive experiments are conducted and several valuable results are obtained that can compare the improvement in recognition rate between classical machine learning approaches and deep neural network approach along with different feature selection approaches.
The rest of this paper is organised in the following way. Section II discusses the related work that has already been carried out for human activity recognition. Section III presents the proposed methodology. Section IV elaborates about materials and methods i.e. the data collection, environment of computation and analysis of competing algorithms. Section V presents results of EDA, simulation results of different machine learning algorithms and the class specific precision and execution time in detail. Section VI concludes this study with a brief summary and points out future research work.

II. Related Work

Human Activity Recognition (HAR) is generally evaluated based on accuracy and computational cost. In order to effectively recognise different activities, previous works have tried to extract hand-crafted features from the signals of the accelerometer and gyroscope [7], and applied different classical machine learning approaches like Support Vector Machine [8], Random Forest [9] etc. for classification. Sharma et al. have used neural networks (ANN) [10], while Khan has used decision trees to classify basic activities [11]. In order to reduce the computational cost, some researchers have used feature selection [12] before applying classification model. However, most of these works have employed methodologies that use certain selected and hand-crafted features which not only increase the computational cost but also make them difficult to compare. Some research has also been carried out for HAR using deep learning. Duffner, Berlemont, Lefebvre and Garcia (2014) have used deep convolutional neural network (DCNN) which uses accelerometer and gyroscope data together to automatically extract features for activity recognition [13].

In this paper, instead of using DCNN, we use a deep neural feed forward network (DNN) on the benchmark data, collected from 30 volunteers and show how DNN, which automatically extracts features, outperforms classical machine learning algorithms that use hand-crafted features.

III. Proposed Methodology

The dataset has 561 features and 7352 observations, the entirety of which is used for our initial analysis for generating conclusions using K-Nearest Neighbour (KNN) and Random Forest classifier. KNN runs through the entire dataset finding the distance "d" between the unobserved point and each point of the training set. The K is the number of training points that are close to the unobserved point. In our model we have set the value of K as 15. The Random Forest classifier

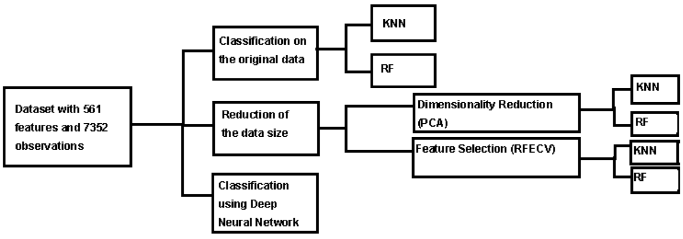


Fig. 1: Block Diagram of the proposed method

is a collection of decision trees from a randomly selected subset of training data and the final class is predicted by aggregating all the votes from each decision tree. In our model we have taken the number of decision trees i.e. n_estimators = 500. Then the data is reduced, using both feature selection and dimensionality reduction techniques for better recognition accuracy, as well as to reduce the computational cost and the results are compared. Dimensionality reduction is done using Principal Component analysis (PCA) retaining 95% of the total variance – which yields 67 effective features. In feature selection technique, Recursive Feature Elimination with Cross Validation (RFECV) is chosen which uses random forest classifier in the background to recursively eliminate the weaker feature – which yields 374 effective features with a little over 90 features after which the f1-score stops increasing. Both the methods are discussed in elaborate detail in Section V. Finally, a Deep Neural Feed Forward Network is used with the first layer as a dense layer having 96 nodes and "tanh" as an activation function. It is followed by a dropout layer with 50% dropping rate, another dense layer with 30 nodes and "softmax" as an activation function followed by a 20% dropout layer, a final hidden dense layer with 24 nodes having "tanh" as an activation function and an output layer of 7 nodes with "softmax" activation function. The DNN uses the "adam" optimiser function and "categorical cross-entropy" as the loss function. We have used 100 epochs with a batch size of 50 and validation split = 0.1 to reach conclusions and the results are compared. The schematic diagram of the proposed methodology is shown in Fig. 1

IV. Materials and Methods

A. Collection of Data

The techniques and experimental setup to collect data are elaborated in significant detail in [8]. Anguita et al. selected 30 volunteers aged from 19 to 48 years and asked them to perform six basic activities: walking, walking upstairs, walking downstairs, sitting, standing and laying whilst wearing a "waist-mounted Samsung Galaxy S II smartphone". The data collection took place in two stages: first the smartphone was attached on the left side, and then on the right side. Moreover, the performance of each task is separated by a time-gap of 5 seconds.
17 primary signals were collected using the in-built accelerometer and gyroscope at a 50 Hz sampling rate, and preprocessed with a median filter and a third order low-pass Butterworth

filter with a 20 Hz cut-off frequency for effective noise management.
These 17 primary signals are used to obtain a vector of other features–mean, correlation, energy of frequency bands, frequency skewness etc.

B. Computation Environment

The computation is light enough to be performed on a system with 3.1 GHz Intel Core i5 processor, 8GB 2133 MHz LPDDR3 memory and Intel Iris Plus 650 1536 MB graphics. The programming language in which computation is carried out is Python (version 3.5 and above). The deep learning model is made using a Tensorflow back-end with Keras libraries but the computation can be programmable in a Theano back-end as well.

C. Competing Algorithms Analyses

We have employed a number of classical as well as novel methods for training the data to render a prediction with as high an accuracy as possible. Notably, we have used the K-Nearest Neighbours, Random Forest Classifier, and mainly the neural network architecture.

1) K-Nearest Neighbours (KNN): The algorithm is a non-parametric and a lazy learning algorithm. In KNN, similarity is defined according to a distance metric between two data points. We use the Euclidean distance as this metric.
$E(x, y) = \sqrt{\Sigma_{i=1}^{n}(x_i - y_i)^2}$
In this study, the algorithm proceeds to find $K = 15$ nearest neighbours based on their similarities.

2) Random Forest Classifier: The Random Forest Classifier builds a forest, which is an ensemble of decision trees trained with the 'bagging' method: the idea that the overall result is increased by using a combination of learning models.
In the experiment, the algorithm builds 500 trees. It considers a number equal to the square root of the number of features to split a node. The minimum number of leaves that are required to split an internal node is taken to be 1.

3) Deep Neural Network (DNN): We employ the use of three hidden layers in the network h_1, h_2 and h_3. The following equations describe the mapping of the model from input x to output prediction $\tilde{y}$ given the ground truth value y.
$h_1 = tanh(W_1^T x)$
$h_2 = softmax(W_2^T h_1) = \frac{e^{a_i}}{\Sigma_{j=1}^{k} e^{a_j}}$
where $a = W_2^T h_1$
$h_3 = tanh(W_3^T h_2)$
$\tilde{y} = softmax(W_4^T h_3)$
Here, W_1 is the weight matrix from the input layer to the first hidden layer, W_2, from the first to the second hidden layer, W_3, from second to third hidden layer, and W_4 from third hidden layer to the output layer. We also use dropouts at the first and second hidden layers to prevent overfitting [14]. We use gradient descent techniques to back-propagate for 100 epochs taking categorical cross-entropy as the loss function for multi-class classification.

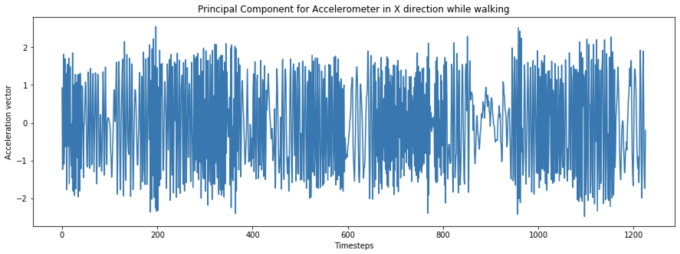


Fig. 2: Principal Component for Accelerometer in X direction while walking

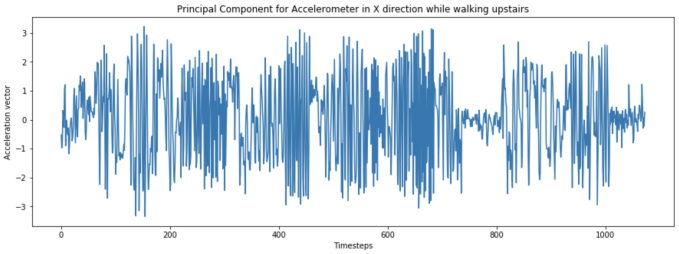


Fig. 3: Principal Component for Accelerometer in X direction while walking upstairs

V. Simulation Results and Discussion

A. Exploratory Data Analysis (EDA)

1) Studying Patterns: In many cases, a candidate's pattern of behaviour in regard to his/her activities can be predicted. For instance, a person leading a rather sedentary lifestyle will show pronounced effects while sitting or laying whereas an objectively active candidate will show fluctuations while performing activities like walking or standing. It is important to clarify that both of the above candidates are extremes of a spectrum. So, to study a more or less general behaviour, the patterns of behaviour in the principal components for the activities are looked at.

A general candidate shows a large amount of fluctuations while doing activities that are objectively, more physically demanding than others like walking (Fig. 2), or walking upstairs (Fig. 3). With this intuition, the candidate's behaviour while doing passive activities, like laying or sitting, can be predicted.

The accelerometer does not show much fluctuation for activities that do not require a lot of movement (Fig. 4 and Fig. 5). It is also observed that the gyroscope shows high fluctuations even for a passive activity.

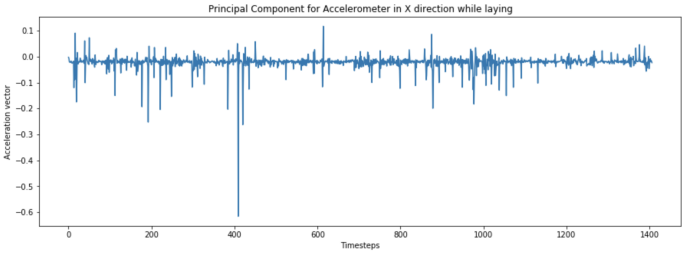


Fig. 4: Principal Component for Accelerometer in X direction while laying

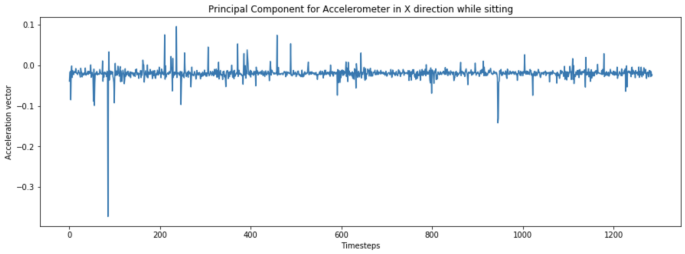


Fig. 5: Principal Component for Accelerometer in X direction while sitting

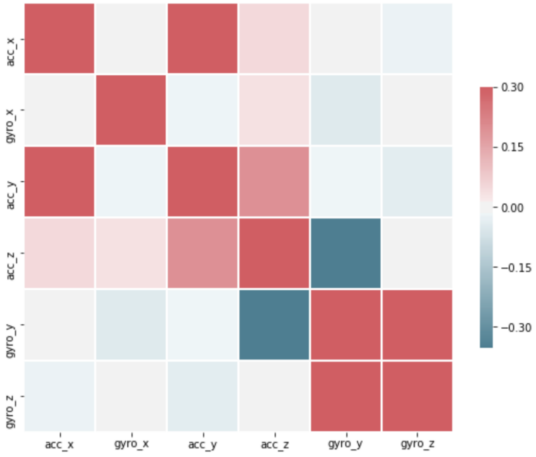


Fig. 6: Correlation Matrix

2) Studying Correlations: There could be factors that influence the similarity or dissimilarity seen in the behaviour patterns for various candidates while performing different activities. However, it would be incorrect and absurd to state that if two features are correlated, then one causes the other. This is because correlation does not imply causation.

From the correlation matrix (Fig. 6), it can be inferred that accelerometer readings in the x and y directions, and gyroscope readings in the y and z directions, are significantly positively correlated. Moreover, the gyroscope reading in the y direction and the accelerometer reading in the z direction are significantly negatively correlated. This could be because of many factors: positioning of the smartphones, a general fashion of performing activities among the subjects, movement in one direction affecting that in another direction etc. Deciphering these causes is beyond the scope of the present work since the objective here is to solely study the patterns of behaviour in data.

B. Simulation Results of Machine Learning Algorithms on the Entire Dataset

After doing some EDA, we applied two classical machine learning techniques Random Forest (RF) and K-Nearest Neighbour (KNN) on the entire dataset to find the classification accuracy score shown in Table I

1) Recursive Feature Elimination (RFE): Since the dimension of the dataset is huge (561 number of features), in order to find the optimal number of features, cross-validation is used

TABLE I: Comparison of Accuracies (%) of the classical ML models

KNN	Random Forest
92.97	90.46

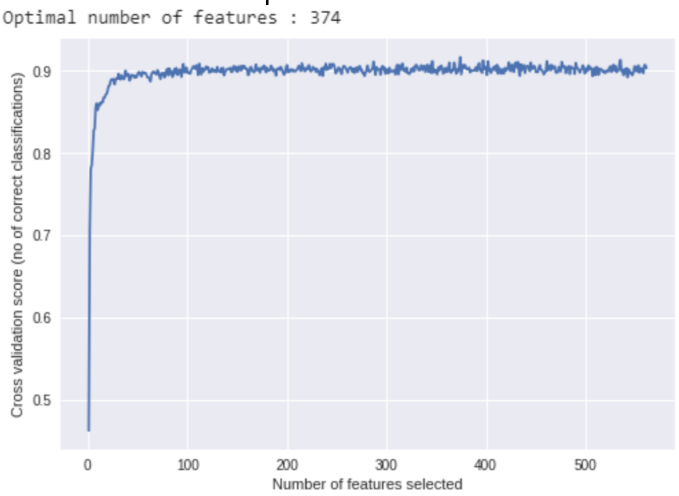


Fig. 7: RFECV using Random Forest Classifier

along with RFE to score different feature subsets and select the best scoring subset. We have used Random Forest Classifier at the background of Recursive Feature Elimination with cross-validation (RFECV) which eliminates n features from a model by fitting the model multiple times, and at each step, removing the weakest feature.

From Fig. 7, it is evident that the curve jumps to an excellent accuracy when the informative features are captured, then gradually decreases in accuracy as the non-informative features are added into the model. In this process, 374 features are selected, although, the f1 score of the model does not seem to improve after the inclusion of around 90 features.

2) Principal Component Analysis (PCA): PCA is employed to convert the set of observations of possibly correlated features, into a set of values of linearly uncorrelated variables [15], [16]. Each of the principle components is chosen in such a way that it describes most of the available variance. All these components are orthogonal to each other. The dataset used for this study has 561 variables which is quite a high dimensionality. It can be reduced to 2 or 3 dimensions using PCA so that we can plot and hopefully understand the data better. The activities are characterised in numbers as

1. Walking
2. Walking Upstairs
3. Walking Downstairs
4. Sitting
5. Standing
6. Laying

The explained variance tells how much information can be attributed to each of the principal components. This is important since while converting high dimensional space to 2-dimensional space, we are bound to lose some variance.

In Fig. 8, the first principal component (PC1) contains 62.54% of the variance, and the second principal component

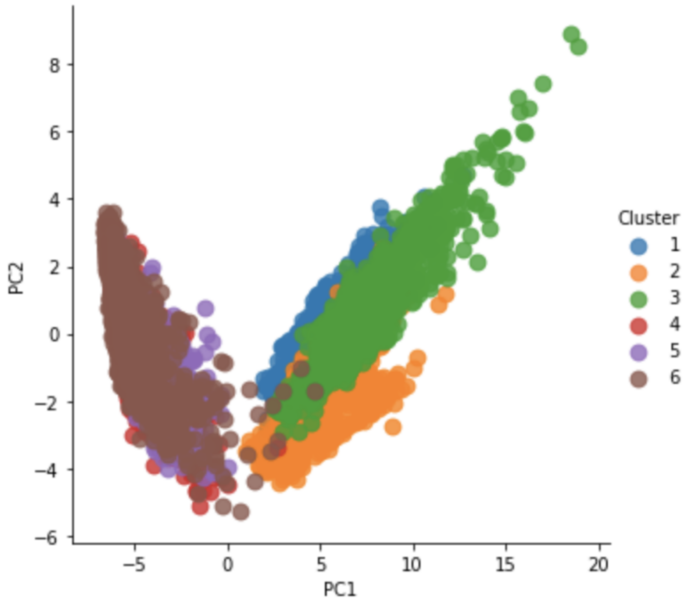


Fig. 8: 2-Dimensional Plot of the PCA

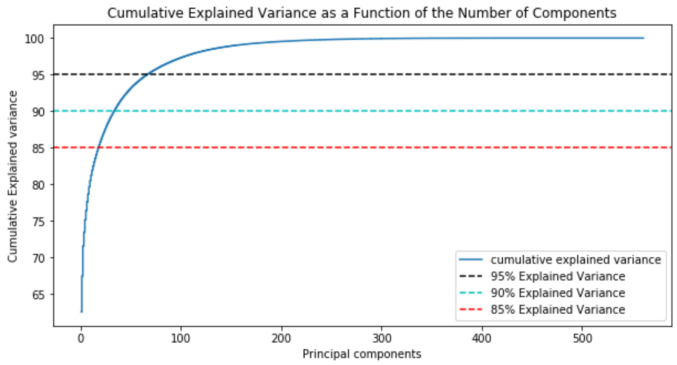


Fig. 9: Cumulative Explained Variance as a Function of the Number of Components

(PC2) contains 4.9% of the variance. Together, the two components contain only 67.44% of the total information.

Through experiment, it is found that the minimum number of principal components, such that 95% of the variance is retained, is 67.

C. Simulation Results of Machine Learning Algorithms on the Reduced Dataset

Table II gives a snapshot of the accuracy of different models on the Reduced Dataset. It shows that the percentage of classification accuracy is more if we use RFECV rather than PCA i.e., feature selection technique outperforms the dimensionality reduction technique.

TABLE II: Comparison of Accuracies (%) of the classical ML model with RFECV and PCA

PCA + KNN	RFECV + KNN	PCA + Random Forest	RFECV + Random Forest
90.49	91.48	91.68	92.56

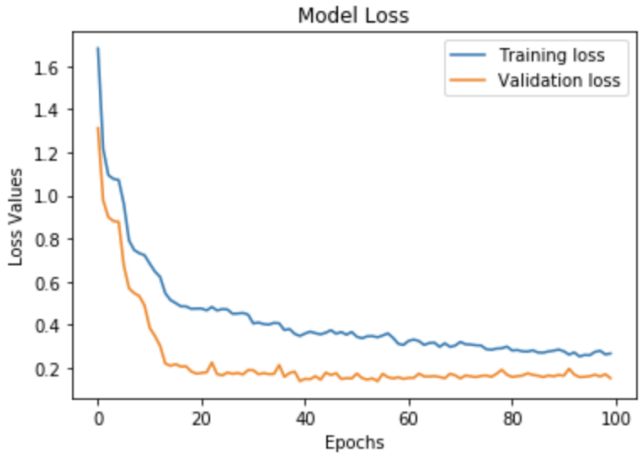


Fig. 10: Training vs Validation Loss

TABLE III: Precision, Recall, f1−score of KNN with RFECV

	Precision	Recall	f1-score	Support
Walking	0.87	0.98	0.93	496
Walking_Upstairs	0.89	0.93	0.91	471
Walking_Downstairs	0.97	0.79	0.87	420
Sitting	0.92	0.82	0.87	491
Standing	0.85	0.94	0.90	532
Laying	1.00	0.99	1.00	537

D. Simulation Results of Deep Learning Algorithm

From the behaviour of the loss function in Fig. 9, it is observed that the validation loss converges to be significantly less than the training loss. This shows that the model did not over-fit (training loss is not less than validation loss) or under-fit (training loss is not equal to validation loss).

We compare the proposed DNN with four other techniques: Random Forest and KNN, both with and without feature selection and dimensionality reduction. The former uses the entire dataset whereas the latter aims to select the most important features required in order to reduce the computational cost.

Table III, Table IV and Table V show the precision values, recall values and the f1-scores on training through KNN, Random Forest and DNN respectively and Fig. 11 gives the comparison between them.

The quantitative comparison is summarised in Table VI.

The proposed DNN method achieves the best performance overall, in terms of accuracy and computational cost. Although Random Forest with RFECV gives the second best accuracy,

TABLE IV: Precision, Recall, $f1-$score of Random Forest with RFECV

	Precision	Recall	f1-score	Support
Walking	0.89	0.97	0.93	496
Walking_Upstairs	0.91	0.91	0.91	471
Walking_Downstairs	0.96	0.86	0.91	420
Sitting	0.90	0.89	0.89	491
Standing	0.90	0.91	0.90	532
Laying	1.00	1.00	1.00	537

TABLE V: Precision, Recall, $f1-$score of DNN

	Precision	Recall	f1-score	Support
Walking	0.95	0.99	0.97	496
Walking_Upstairs	0.93	0.96	0.95	471
Walking_Downstairs	0.99	0.90	0.94	420
Sitting	0.93	0.81	0.87	491
Standing	0.83	0.98	0.90	532
Laying	1.00	0.96	0.98	537

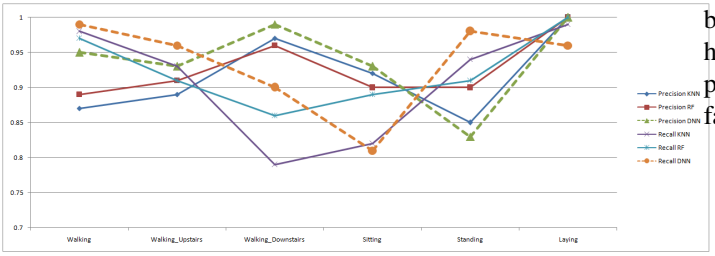


Fig. 11: Comparative study of precision and recall

its computational cost turns out to be the highest. Thus, there is a trade-off between accuracy and computational cost and the proposed DNN achieves a perfect balance between them.

The random forest classifier model gives an accuracy of 92.56% on the testing data whereas the KNN model gives an accuracy of 91.48%. However, the neural network approach renders an accuracy of 97.32% on the testing data, which is a significantly better score than the classical models.

One can argue that this occurs because of the fact that the neural network is formulated so as to mimic the human brain [18] in the sense that it essentially learns to give the best predictions. Mathematically, it means that the model is constantly updating its weights based on the effectiveness of each node to minimise the loss function.

VI. Conclusion

In this study, data extracted from the accelerometers and gyroscopes of smartphones is used to derive patterns and formulate results pertaining to the basic physical activity of an individual. We propose a deep learning methodology to reach the conclusions, and find that the proposed deep learning method gives better results compared to the other methods, when computed with the same data, and under the same environment of computation. The study, of course, is not free of flaws in many ways. For instance, it is currently beyond the scope of this study to find the reasons as to why an individual shows such a highly specific behaviour while performing a particular activity. Moreover, the accuracy for the training models can be much better, despite having undergone parameter and hyper-parameter pruning, as well as an optimum neural network architecture development.

The results and insights inferred from this study can prove to be very fruitful in providing assistance to many individuals and organisations, that require reliable and accurate information about the basic healthcare and maintenance of the human body. It could also be used for further studies related to human behaviour. It could aid an individual in preventing (or promoting) certain habits or behaviours that deteriorate (or facilitate) the functioning of the body.

TABLE VI: Comparison of Accuracies (%) and Computational Cost of the Classical ML Model with RFECV and DNN

Accuracy (%)			Computational Cost (min.)		
RFECV + KNN	RFECV + Random Forest	DNN	RFECV + KNN	RFECV + Random Forest	DNN
91.48	92.56	97.32	26.78	27	1.41

References

[1] K. Sumathi, N. S. Lakshmi, and S. K. Kundhavai, "Reviewing the impact of smartphone usage on academic performance among students of higher learning." in *International Journal of Pure and Applied Mathematics*, vol. 118, 2018.

[2] M. A. Osman, A. Z. Talib, Z. A. Sanusi, T. Shiang-Yen, and A. S. Alvi, "A study of the trend of smartphone and its usage behavior in malaysia," in *International Journal on New Computer Architectures and Their Applications*. IEEE, 2012.

[3] N. Gupta, S. Garg, and K. Arora, "Pattern of mobile phone usage and its effects on psychological health, sleep, and academic performance in students of a medical university." in *National Journal of Physiology, Pharmacy and Pharmacology*, vol. 6, 2016.

[4] S. Ranasinghe, F. A. Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation." in *International Journal of Distributed Sensor Networks*, 2016.

[5] J. T. Sunny and S. M. George, "Applications and challenges of human activity recognition using sensors in a smart environment." in *International Journal for Innovative Research in Science and Technology*, vol. 4, 2015.

[6] P. R. Woznowski, R. King, W. Harwin, and I. Craddock, "A human activity recognition framework for healthcare applications: ontology, labelling strategies and best practices." in *International Conference on Internet of Things and Big Data*, 2016.

[7] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *Esann*, 2013.

[8] ——, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012, pp. 216–223.

[9] T. Peterek, M. Penhaker, P. Gajdoš, and P. Dohnálek, "Comparison of classification algorithms for physical activity recognition," in *Innovations in bio-inspired computing and applications*. Springer, 2014, pp. 123–131.

[10] A. Sharma, Y.-D. Lee, and W.-Y. Chung, "High accuracy human activity monitoring using neural network," in

2008 Third International Conference on Convergence and Hybrid Information Technology, vol. 1. IEEE, 2008, pp. 430–435.

[11] A. M. Khan, "Recognizing physical activities using wii remote," *International Journal of Information and Education Technology*, vol. 3, no. 1, p. 60, 2013.

[12] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2011, pp. 289–296.

[13] S. Duffner, S. Berlemont, G. Lefebvre, and C. Garcia, "3d gesture classification with convolutional neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5432–5436.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting." in *Journal of Machine Learning Research*, 2014.

[15] H. Abdi and L. J. Williams, "Goals of pca." in *Principal Component Analysis*, vol. 2. John Wiley & Sons, Inc., 2010.

[16] L. Fu, "The discriminate analysis and dimension reduction methods of high dimension." in *Open Journal of Social Sciences*, 2015.

[17] C. A. Ronao and S.-B. Cho, "Human activity recognition using smartphone sensors with two-stage continuous hidden markov models," in *2014 10th International Conference on Natural Computation (ICNC)*. IEEE, 2014, pp. 681–686.

[18] S. Haykin, "The brain." in *Neural Network and Learning Machines*. Pearson Education Inc., 2009, pp. 5–15.

Speech Recognition Learning Framework for Non-Native English Accent

Mihir Thakkar
School of Electronics Engineering
Vellore Institute of Technology
Chennai, India
mukeshkumar.thakkar2016@vitstudent.ac.in

Dr. Susan Elias
School of Electronics Engineering
Vellore Institute of Technology
Chennai, India
susan.elias@vit.ac.in

Dr. Ashwin Ashok
Dept. of Computer Science
Georgia State University
Atlanta, USA
aashok@gsu.edu

***Abstract*—Accent is a distinctive way of pronouncing a language, especially one associated with a particular country, area or social class. While dialects are usually spoken by groups united by geography or class and are a variety of same language differing in vocabulary and grammar as well as pronunciation, the accent is how the same language is spoken differently by people of different ethnicities. Accent plays a vital role when it comes to speech recognition by voice assistance systems. The modern-day voice assistant systems tend to misinterpret the words spoken by a person with a strong accent influenced by his native language. Machine learning algorithms, especially Support Vector Machine (SVM) and Random Forest when applied on a proper training set can play a vital role in the classification of accent. We propose in this paper the learning framework for speech recognition of Indian accent by analysing the features of Indian accented English and classifying based on sounds that are typical to Indian accented English.**

***Index Terms*—SVM, Random Forest, Accent, Pitch, Acoustic, Formant Frequency**

I. Scope of Work

The proposed work presents a learning framework for accent characterization by analysis and classification of features typical to Indian accent by focusing on specific vocabulary that consists of different combinations of consonants and vowels that affect the pronunciation of the language. The work differs from the previous work in the area of accent analysis as it incorporates the difficult vowel-consonant combination, thereby helping better understand the features that can be used for accent identification. The relevant features obtained can be considered for incorporation into the speech recognition and voice based digital assistant systems for improving the understandability of accented speech and thereby producing an appropriate response.

II. Introduction

The voice assistance systems have been in a constant phase of improvement and will soon enough become a major technology that will be a part of all the appliances and devices ranging from handheld devices to cars and speaker systems. A major problem with these voice assistance systems is the inability to understand the command speech with a strong accent. It is often observed that these systems tend to misinterpret some words when spoken by a set of people with a particular linguistic background. For example, the word Wednesday is pronounced as wed-nus-day instead of wenz-day in Indian accented English. As many widely spoken languages lack the sound of z present in the word Wednesday, people with a strong accent of their native language tend to mispronounce the word. Due to such variations in the pronunciation of words, the voice assistance systems often tend to misunderstand the command and produce an undesirable response. In order to make these systems more efficient and useful for all types of users, it is required to have these systems understand the accent and thereby produce a desirable response to the command. In this paper, we attempt to classify the Indian accent (Specifically, Gujarati accent) of English by analysing and extracting the features that are unique to the Gujarati accent and also considering a negative class of British accented English. Using this labelled dataset of British and Indian accent, a classification model is trained to classify the accent. State of the art machine learning algorithms like Support Vector Machine and Random Forrest classifier are used to train a model and test it on the test set. Since it is a binary classification problem of whether the accent is Indian or not, SVM which is designed to maximize the distance of separating boundary between two classes and also by maximizing the distance of separating planes is used. Choosing a kernel function for SVM is also an important factor. Linear kernel proved to be best suitable for accent classification. Random Forest classifier also proved to be good for accent classification. This paper is arranged with Related work in Section III, Characteristics of Speech in Section IV, Methodology in Section V and Result Analysis and Discussion in Section VI followed by Section VII which concludes the paper.

III. Related Work

Accent classification has been a major research problem. For building state of the art speech recognition systems there has been constant research going on that concentrates on modelling the speech variances among different speakers such as dialects and accents. Linguistic differences in pronunciation have led to an increasing need to improve the modern speech recognition systems to understand the voice commands of non-native users. There has been research in the areas of accent recognition for Indian English but they lack the relevant

978-1-7281-2087-4/19/$31.00 © 2019IEEE

details considering the large span of native languages spoken in India. The prior work in speech accent recognition for Indian language focussed on Marathi and Arabic language comparison and recognition of Indian accent (specifically, the Marathi accent). The research by K.V.Kale et al.[1] is focussed on the recognition of Indian English by taking the speech sample of speeches of numbers from one to nine and comparing the values of acoustic features of the speeches of speakers of Arabic and Marathi native language. While the paper classifies accent based on acoustic features, it does not consider the many combinations of consonants and vowels that affect the pronunciation of a word. Difficult clusters of consonants like P-S and S-N have better influence on the pronunciation of the word as many combinations of sound used in English may not be used in the native language of the speaker or some like the sound Z may not even exist in the native language, giving a distinguishable sound for accent comparison. Moreover, the words with more syllables have more than one parts pronounced differently in accented language. In this paper, we attempt to classify the accent based on some of the words that are spoken uniquely by the speakers of Gujarati origin.

IV. Characteristics of Speech

A. Information Contents of Speech Signal

The information present in any speech signal can be divided into three categories: linguistic, paralinguistic and nonlinguistic information [2]. It is this information that defines the phonetic behaviour of a signal and thus the accent which thereby makes it important to discuss these characteristics here.

1) Lingusitic Information: The linguistic information describes the symbolic information which is represented by different symbols and rules for their combination. The syllables, phonemes and collective pronunciation of the syllables to utter a word which leads to the generation of a unique way of pronunciation comprises the linguistic information in any speech signal. The linguistic information of a speech is important to understand the accent of a speaker.

2) Paralinguistic Information: The information that cannot be inferred from the written counterpart of the speech but is added by the speaker knowingly to modify the linguistic information is called paralinguistic information. For example, changing a declarative sentence into a question by emphasizing certain words. Such information is local to a particular part of the speech. Paralinguistic information comprises information related to the intensity of the spoken words in a speech. After extracting the intensity values using Praat tool from a raw speech of a British and an Indian speaker speaking same sentence taken from Speech Accent Archive, it is inferred that the average intensity of the speech of Indian origin speaker is high when compared to that of British origin speaker. The Intensity values extracted are given in Table I.

From the table, it can be noted that the Average intensity values of the Indian accent are higher than that of the British accent.

TABLE I
Difference in intensities of speech of Indian Accent (IA) and British Accent (BA)

Word	Avg. Intensity (IA)	Avg. Intensity (BA)	Difference
Ask	71.069615	70.9041954	0.165419
Six	70.44072	68.93406	1.50666
Fresh	72.02530446	71.18237452	0.84292994
Peas	70.2092024	67.7521151	2.4570873
Snack	71.11910322	68.83099607	2.28810715
Scoop	71.84884046	70.89110178	0.95773868
Wednesday	73.0850407	70.0149962	3.0700445
Snake	71.73715	68.27099	3.46616
Red	75.64376	71.89398	3.74978
Slabs	73.05599	70.99402	2.11579
Maybe	76.52977	72.84851	3.68126

B. Acoustic Characteristics

1) Frequency Characteristics: Frequency plays an important role when it comes to speech and speaker recognition. Frequency characteristics help in differentiation of not only gender and age group but also the accent of a speech. Table II shows the values of corner frequency for the utterance of each word for British and Indian accented English. The corner frequencies of words when spoken by a person of British accent and Indian accent was extracted using Praat tool and is listed in Table II. From the table, it is observed that the value of the corner frequency of British speaker is less than that of an Indian speaker.

TABLE II
Corner Frequency (in Hz) of words when spoken in Indian Accent (IA) and British Accent (BA)

Word	Frequency (IA)	Frequency (BA)
Ask	1322.48	826.39
Six	1493.01	798.24
Fresh	1710.05	20.73
Peas	2190.65	35.74
Snack	1493.02	221.78
Scoop	965.92	97.39
Wednesday	1399.13	1180.95
Snake	1431.01	376.81
Red	2919.28	267.67
Slabs	2423.19	314.79
Maybe	826.39	810.89

2) Pitch Characteristics: Pitch is the quality of sound governed by the rate of vibrations producing it. It represents the degree of highness or lowness of a tone when a speech is delivered. In Indian and European languages the change of pitch has a major role to play however it does not change the meaning of a spoken word. There are many research works that have contributed to the investigation of pitch as feature for speech based systems[7,8]. By manipulating the pitch of a speech, it is observed that the speech still remains comprehensible but it affects the tone of the speech. While the pronunciation may not be affected much by the pitch it may have a significant impact when considered along with other features.

3) Formant Frequency: A formant can be described as a concentration of acoustic energy over a specific frequency

in a speech wave. There exists various formants, each at a different frequency which is approximately one in each 1000 Hz band. Each formant corresponds to a resonance in the vocal track. The formant frequencies are due to the shaping of frequencies by the vocal tract. The specific arrangement of the articulators for every phoneme creates resonance at a particular frequency. This frequency is called the formant frequency. There are many research works that have contributed to the investigation of formant frequencies as features for speech recognition systems [9,10,11,12]. Table III shows the average formant frequency values of British and Indian accents of English extracted using Praat tool.

TABLE III
AVERAGE VALUES OF FORMANT FREQUENCIES OF BRITISH AND INDIAN ACCENTS

Class	F1	F2	F3	F4
IA	670.18	1961.67	2963.29	4091.12
BA	779.44	2046.77	2963.53	4099.45

From the table, it is observed that the average formant frequency values of the speaker of British accent are higher than that of an Indian accent.

V. METHODOLOGY

The audio samples of accented English were taken from Speech Accent Archive which is an online database of accent samples compiled by George Washington University's Linguistic Department (Reference link: http://accent.gmu.edu).
The speech is of the passage which reads as follows:
Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.
This paragraph consists of difficult clusters of consonants like P-S and S-N and almost all the vowels and just about every consonant which makes it the best choice for analysis of the speakers accent. 14 British samples and 10 Gujarati samples of audio of the above passage were taken and 9 different features were extracted for each of the 10 selected words that were found suitable for taking into consideration based on the ease of accent comprehensibility when spoken by a Gujarati speaker making 90 different parameters per speaker. For extracting the features from the speech signal, Praat which is a free computer software package for the scientific analysis of speech in phonetics was used (Reference link: http://www.fon.hum.uva.nl/praat). The generated dataset consisted of the value of each feature corresponding to different words spoken by the same person. It was later scaled by considering the features and class label for each word, generating a dataset of 250 rows. The features and words considered are listed in Table IV.

These words were selected keeping in mind the pronunciation and misinterpretation by the voice assistance systems. For example, the speech recognition systems tend to misinterpret the word Snacks as Snakes also Peas is often misinterpreted as Piece/Peace when a person with a strong Gujarati accent speaks.

TABLE IV
WORDS AND ACOUSTIC FETURES

Words	Features
Ask	Maximum Pitch
Six	Minimum Pitch
Fresh	Median Pitch
Peas	Pitch Standard Deviation
Slabs	Mean Pitch
Maybe	Formant Frequency (F1, F2, F3, F4)
Snack	
Snake	
Scoop	
Red	
Wednesday	

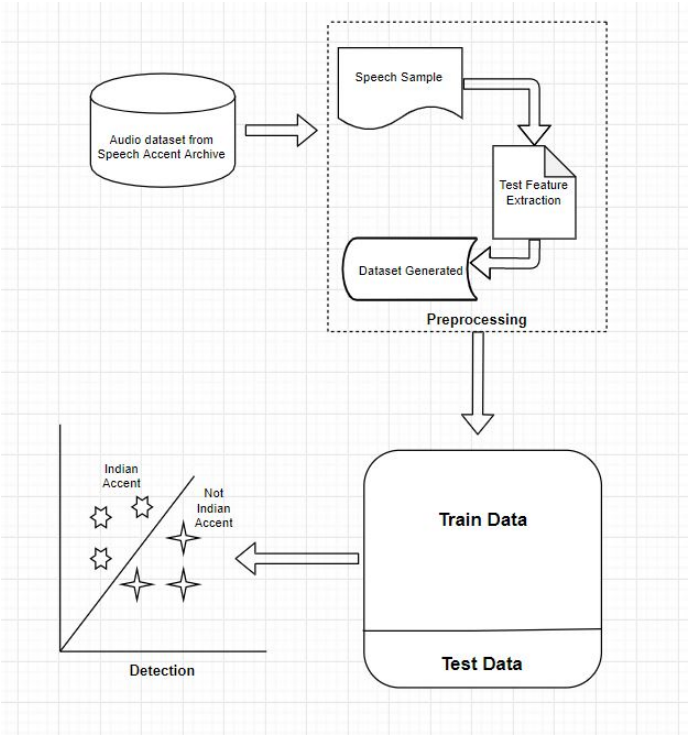


Fig. 1. Approach for accent classification

A. Acoustic Features

Based on the past research it is observed that Pitch and formant frequencies are effective acoustic features for accent recognition [3, 4, 5, 6, 7]. The phonetic background of the speech can be well determined with the help of the formant frequency. The pronunciation of the word is a result of emphasis laid on particular sound while utterance, Table V presents the phonetic transcription of the words taken into consideration along with the phonetic transcription of the mispronounced form of the same word.

The dataset generated was arranged initially in a matrix form for each speaker with features on the vertical and words

TABLE V
PHONETIC TRANSCRIPTION AND ACCENTED PHONETIC TRANSCRIPTION OF WORDS

Words	Phonetic Transcription	Accented Phonetic Transcription
Ask	Ask	Aax
Six	sIks	sIcks
Fresh	frEsh	frAEsh
Peas	pEEz	pIs
Slabs	slAbz	slaEbz
Maybe	mAY-bee	may-bEE
Snack	snAk	snakE
Snake	snAYk	snAYke
Scoop	skOOp	scOpe
Red	rEd	rAId
Wednesday	wEnz-day	wEd-nUs-day

on the horizontal axis (Figure 2). This dataset was later scaled feature-wise to generate a dataset of 250 rows with the aforementioned features. From the raw dataset generated, different sets of features were used to train a model to determine accent and the model was evaluated with F1-Score, Precision and Recall values obtained from the confusion matrix.

g2m	Ask	Six	fresh
Intensity	72.79136599201776 dB	74.73341801	75.5305296
Pitch MAX	152.586 Hz	169.308	311.212
Pitch MIN	130.072 Hz	153.689	124.872
Pitch Median	147.821 Hz	159.302	135.994
Pitch SD	7.222 Hz	4.568	68.926
Pitch Mean	145.910 Hz	159.836	170.651
Frequency	1322.4872750734824 Hz	1493.019213	1710.059861
F1	792.1643389601619 Hz	1135.663817	790.6964862
F2	1870.4452689677837 Hz	2522.494733	1993.909995
F3	3013.781473711476 Hz	3493.768634	2821.763018
F4	4017.81638217269 Hz	4357.809155	3737.256805

Fig. 2. Sample from the dataset generated

B. Evaluation Metric

The performance evaluation parameters such as precision, recall and F1-Score are analyzed in this section. Table VI shows the symbolic representation of Indian and British accent positives and negatives in order to explain the Recall, Precision and Accuracy.

TABLE VI
DESCRIPTION TABLE

	Indian Accent	British Accent
Classified as Indian	I_1	B_1
Classified as British	I_2	B_2

Recall(R):

$$R = \frac{I_1}{I_1 + I_2}$$

Precision(P):

$$P = \frac{I_1}{I_1 + B_1}$$

Accuracy(A):

$$A = \frac{I_1 + B_2}{T}$$

Where T is the total number of samples. T= $I_1 + I_2 + B_1 + B_2$

C. Classification of Accent

1) Random Forest: Random Forest is a classifier which is constructed by combining several different independent base classifiers. The randomness introduced by selecting the best split features from a random subset of available features helps to achieve independence to a certain degree thereby improving performance.

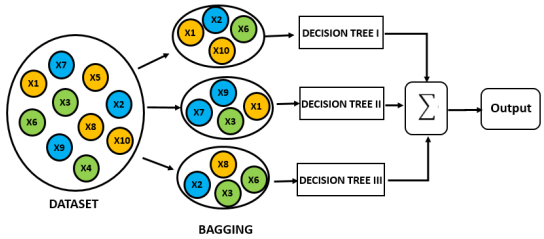


Fig. 3. Random Forest Algorithm

Since the dataset generated had mild correlation among variables, random forest algorithm was chosen as it works well in case of correlation among features. With a different set of features taken different accuracy values were obtained which are listed in Table VII.

TABLE VII
CLASSIFICATION RESULT WITH RANDOM FOREST CLASSIFIER

		Precision	Recall	F1-Score
Pitch SD, Pitch Mean, F3, F4	NI	0.74	0.96	0.84
	IN	0.93	0.61	0.74
Pitch SD, Pitch Mean, F4	NI	0.71	1	0.83
	IN	1	0.59	0.69
Pitch SD, Pitch Mean	NI	0.70	0.76	0.73
	IN	0.74	0.68	0.71
Pitch Mean, F4	NI	0.69	0.72	0.71
	IN	0.71	0.68	0.69

Here, IN denotes INDIAN, NI denotes NOT INDIAN accent and SD denotes Standard Deviation.
From table VII, the following observations are made:

- F1-Score obtained was highest when Pitch Standard Deviation, Pitch Mean and Formant frequencies 3 and 4 are considered as features for classification.
- When the Formant frequency 3 was eliminated and the model was trained, the F1-Score of NI did not get affected significantly while the F1-Score of IN reduced by of 5%. making F3 an important feature with respect to the recognition of Indian accent.
- Also, when only pitch related parameters were considered, the accuracy decreased.

- When Pitch Mean and Formant frequency 4 were considered, the accuracy went down as compared to when Pitch Standard Deviation, Pitch Mean, F3 and F4 were considered altogether.

Parameters for the model were configured as given in Table VIII.

TABLE VIII
CLASSIFICATION PASRMETERS FOR RANDOM FOREST CLASSIFIER

Parameter	Value
Bootstrap	True
class_weight	None
criterion	gini
max_depth	2
max_features	auto
max_leaf_nodes	None
min_impurity_decrease	0.0
min_impurity_split	None
min_samples_leaf	1
min_samples_split	90
min_weight_fraction_leaf	0.0
n_estimators	220
n_jobs	None
oob_score	False
warm_start	False
vebose	0

2) Support Vector Machine: Support Vector Machine generally deals with the problem of binary classification for linearly separable data. Data points obtained in the hyperplane of each class are called support vectors and its number plays an important role in classification. Since here we are dealing with a binary classification problem SVM is one of the preferred choices of the classifier.
The optimal solution for SVM is obtained as:

$$f(x) = sgn[(w^*.x) + b^*]$$

here, ω* is the orientation of hyperplane and b* is the position of hyperplane and x denotes the input and f(x) represents the hyperplane of maximum margin.

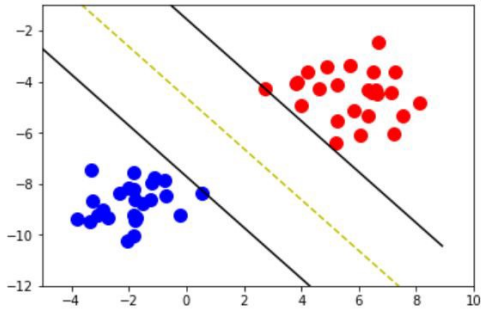


Fig. 4. Hyperplane seperating two classes

Figure 4 describes the hyperplanes separating two classes in Support Vector Machine. Here, the solid black line denotes the support vector and the yellow line denotes the best fit hyperplane.

Each time with a different set of parameters taken, GridSearchCV method was implemented to fine tune the hyperparameters. GridSearchCV implements fit and predict method and its parameter are optimized by cross-validation. The result obtained is presented in Table IX.

TABLE IX
CLASSIFICATION RESULT WITH SVM CLASSIFIER

		Precision	Recall	F1-Score
F3, F4	NI	0.69	0.82	0.75
	IN	0.45	0.29	0.36
Pitch SD, Pitch Mean	NI	0.64	0.88	0.74
	IN	0.79	0.46	0.58
Pitch SD, Pitch Mean, F4	NI	0.66	0.96	0.78
	IN	0.92	0.46	0.61
Pitch Mean, F4	NI	0.71	0.96	0.82
	IN	0.93	0.58	0.72

The results obtained when the model was trained using Support Vector Machine are listed in Table VIII. From the results obtained when SVM was used as a classifier, the F1-Score obtained was the highest when Pitch Mean and Formant frequency 4 were used for classification. The tuned parameters for this case are listed in Table X.

TABLE X
CLASSIFICATION PASRMETERS FOR SVM CLASSIFIER

Parameter	Value
C	10
class_weight	None
cache_size	150
class_weight	None
coef0	0.0
decision_function_shape	ovr
degree	1
gamma	0.01
kernel	linear
max_iter	-1
Probability	False
shrinking	True
tol	0.001
verbose	False

VI. RESULT ANALYSIS AND DISCUSSION

The experiment resulted in the following observations:

1) Pitch Mean, Pitch Standard Deviation, Formant frequency 3 and Formant frequency 4 together made effective parameters for accent classification of the Indian accent.
2) Average intensity values of Indian accent were found to be higher than that of British accent.
3) The average values of formant frequencies were found to be higher for British accent than that of the Indian accent.

While Pitch Mean and Formant Frequency 4 gave better results when the classifier used was Support Vector Machine whereas Pitch Mean, Pitch Standard Deviation and formant

TABLE XI
EVALUATION MATRIX OF SVM AND RANDOM FOREST CLASSIFIERS

Algorithm	Features	F1-Score	Recall	Precision
SVM	Pitch Mean, F4	0.82	0.96	0.71
		0.72	0.58	0.93
RF	Pitch SD, Pitch Mean, F3, F4	0.84	0.96	0.74
		0.74	0.61	0.93

frequencies 3 and 4 gave better results when classifier used was Random Forest as listed in Table XI. From the results obtained, it is inferred that when Pitch Mean, Pitch Standard Deviation and formant frequencies 3 and 4 were considered and random forest classifier was used, the F1-Score obtained was highest, having a value of 0.84 for Not Indian class and 0.74 for Indian class with precision and recall values of 0.74 and 0.96 for Not Indian class and 0.93 and 0.61 for Indian class respectively. Also, the average intensity values of the Indian accent were found to be greater than that of the British accent. These parameters are found to play an important role in determining the accent and thus can improve learning models of speech recognition systems.

VII. CONCLUSION

In this paper, we have proposed a learning framework for classification of non-native English accent by experimenting with a different set of acoustic features such as Pitch Mean, Pitch Standard Deviation, Formant Frequencies and Intensity. From the results obtained it is noted that Pitch and higher Formant Frequencies were found to be better features for classification of non-native English accent. Also, it is important to note that the average intensity value of speaker of Indian accent is higher than the speaker of a British accent and Formant Frequency values of a British accent is higher than that of an Indian accent. The corner frequency of a specific part of speech confined to a word was found to be higher in value for Indian accented English than that of British accented English. These parameters, when incorporated into the learning models of speech recognition systems can effectively enhance the comprehensibility of these systems when an accented command is produced.

REFERENCES

[1] Santosh Gaikwad, Bharti Gawali, KV Kale. Accent Recognition for Indian English using Acoustic Feature Approach, IJCA (0975-8887)

[2] Handbook of Neural Networks for Speech Processing (Artech House Signal Processing Library) Shigeru Katagiri

[3] Tang, H., Ghorbani, A. A. (n.d.). Accent Classification Using Support Vector Machine and Hidden Markov Model, (1), 34.

[4] Arslan, L. M., Hansen, J. H. L. (1996). Language accent classification in American English. Speech Communication, 18(4), 353367. doi:10.1016/0167-6393(96)00024-6 .

[5] Hansen, J. H. L., Arslan, L. M., Carolina, N. (1997). Frequency characteristics for foreign accented speech, Duke University Department of Electrical Engineering, 11231126.

[6] Kat, L. I. U. W., Fung, P., Bay, C. W., Kong, H. (1999). Fats accent identification and accented speech recognition, 36.

[7] Xuejing Sun . Pitch accent prediction using ensemble machine learning, Department of Communication Sciences and Disorders, Northwestern University 2299 N. Campus Dr., Evanston, IL 60208, USA.

[8] Levow, G. (2009). Investigating Pitch Accent Recognition in Non-native Speech, (August), 269272.

[9] John N. Holmes, Wendy J. Holmes and Philip N. Garner Using formant frequencies in speech recognition, Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH, U.K.

[10] P. Schmid and E. Barnard, "Robust, N-Best Formant Tracking", Proc. EUROSPEECH'95, pp. 737-740, Madrid, 1995

[11] Stantic, D., Jo, J. (2012). Accent Identification by Clustering and Scoring Formants, 232237.

[12] Ishi, C. T., Hirose, K., Minematsu, N. (2003). Mora F0 representation for accent type identification in continuous speech and considerations on its relation with perceived pitch values. Speech Communication, 41(2-3), 441453. doi:10.1016/S0167-6393(03)0014-1

Surface Remeshing using Quadric based Mesh Simplification and Minimal Angle Improvement

Dakshata M.Panchal
Department of Computer Engineering
St.Francis Institute of Technology
Mumbai,India
dakshatapanchal@sfit.ac.in

Dr.Deepak J.Jayaswal
Department of Electronics & Tele Communication,
St. Francis Institute of Technology
Mumbai, India.
djjayaswal@sfit.ac.in

***Abstract*— Surface remeshing intends to yield high quality, high regularity and low complexity meshes that are geometrically faithful to original models and free from the low-quality elements. Unfortunately, attaining balance between quality, regularity, complexity and approximation error becomes tedious during remeshing. In the work presented here, authors propose a surface remeshing techniques based on quadric based simplification and high-quality approximation that tries to attain balance of remeshing goals. Given a triangular mesh and user assigned approximation error, the output mesh (remesh) achieves a higher minimal interior angle and low mesh complexity with implicit feature preservation. The proposed approach recapitulates in two-steps. First, the mesh complexity is optimized to lessen the extent of vertices, and then followed by enhancement of the quality of the elements using local operators. This approach can be efficiently incorporated in preprocessing for numerous applications. Investigations have demonstrated that the proposed approach is efficient and robust. The results of proposed approach attain high quality triangles along with preservation of the features of the original geometry.**

Keywords—surface remeshing, remesh, quadric error, local operators, feature preservation, minimal angle improvement

I. INTRODUCTION

Recently, increased use of triangular meshes in many industrial applications, graphical applications, mechanical engineering, medical imaging and virtual reality have immensely amplified the practice of 3D scanners to generate them. Frequently these triangular meshes assimilated by scanning devices are treated as raw meshes due to oversampling where the meshes contain redundant vertices as well as poor quality triangles. Using these meshes for surface representation turn out to be a burden for exhibiting, loading and further processing of the meshes in various applications. There arises need to recuperate the quality of mesh with regards of mesh complexity, element quality, mesh regularity and vertex sampling. The technique of improving the mesh for its further use in numerous applications is the key component termed to be surface remeshing. Despite the fundamental nature of the problem, surface remeshing is considered as NP-hard problem. During the remeshing process, there rise some serious issues such as validity, quality, fidelity, uncertainty, correspondence, discrete inputs, large data sets. Support for the continuous level of details, the balance between output mesh quality and remeshing speed along with guarantees on the bounds of shape elements, distortion error and topology, are the desirable algorithmic functionalities of remeshing algorithms. Remeshing algorithms are classified in several ways depending upon their end goals and applications for which they are used. Based on the goals, they are classified as structured, compatible, high quality, error-driven and feature remeshing. Remeshing algorithms are further classified as isotropic surface remeshing and explicit surface remeshing based on techniques. Even though many surface remeshing practices are classified based on their goals; their prevalent goal is to attain balance between mesh quality and complexity, with geometric fidelity; while preserving features of the mesh. The key requirement of many applications is, the Geometric fidelity studied as an approximation error. For effectual representation of complex surfaces, the mesh complexity i.e. number of elements is also well-thought-out. Usually a conflict exists between the complexity of the mesh with the geometric fidelity and element quality of the mesh. Thus, the ultimate target of remeshing is to find a satisfactory resolution of the mesh for required quality and regularity while maintaining geometric fidelity.

(a) (b)

Fig. 1. Remeshing example of a joint model (a) Original mesh (b) Remeshed result [1]

To the best of authors' studies, only limited approaches till today have fulfilled the goal mentioned here. Most of the structured remeshing techniques replace the unstructured mesh with a structured (regular) mesh that has all the internal vertices enclosed by a fixed number of elements i.e. panels and vertices. The output mesh generated can be semi regular/ regular based on the curvature information of the original mesh. High quality remeshing techniques discretize the original surface or shape to generate a new mesh that is numerically stable with well-shaped elements that are uniformly sampled and smoothly graded. Feature remeshing techniques treat the triangular mesh as an object with curvature having sharp features alike tips, edges, corners. With this triangular mesh, a new tessellation of a surface is generated for preserving the features of the original shape of the model by reducing the delusion between the original and the afresh- approximated mesh. Many remeshing techniques use automated tools or predefine the features in meshes, tag them and process further. The focus of error driven remeshing is to achieve a close balance between mesh complexity and geometric fidelity i.e. accuracy. Complexity is counted as no. of mesh elements. Accuracy is related to the deviation between the original shape and the remeshed shape. Many of the remeshing techniques work on global

978-1-7281-2087-4/19/$31.00©2019 IEEE

and local parameterization of the surface, but these approaches become computationally costly and need to maintain the proper correspondence between the input and the output mesh.

In this paper, authors propose a surface remeshing approach that directly works on the input mesh, and does not need any parameterization that may lead to metric distortion. The proposed approach tries to control and find a balance between approximation error and element quality simultaneously.

II. Related work

In [1], the author defined of remeshing as," Given a 3D triangular mesh, compute another mesh, whose elements satisfy some quality requirements, while approximating the input." They referred quality to properties like vertex regularity, element size and shape. Even though there exists a variety of remeshing techniques, our discussion is limited to the most pertinent remeshing techniques such as high quality, error driven and feature preserving. The detailed survey of various remeshing algorithms is presented in [1][2].

High quality remeshing techniques generate a new remesh (i.e. output mesh) that exhibits well-shaped elements, isotropic/ uniform and smooth gradation sampling. Earlier approaches of high remeshing apply 2D CVT (Centroidal Voronoi Tessellation) in a parametric domain. Unfortunately, indirect methods which are based on parameterization suffer metric distortions. A direct approach was performed on an input surface in [3] but the resultant mesh was inappropriate due to imprecise calculations of discrete version of CVT. In [4], 3D CVT was computed constrained to the input surface to evade parameterization. They additionally proposed many techniques [5] to enhance the quality of the mesh. Still these approaches could not bound geometric fidelity and minimal angle. Sharp feature identification was done in advance. An isotropic surface remeshing approach was done on the mesh directly in [6]. In this approach, length of the edge is taken as an input and the edges that are long resulting in maximal angles are split repeatedly and short edges resulting in minimal angles are collapsed. It performs vertex relocation and valence equalization until all the edges are roughly of the fixed projected edge length. The above-mentioned approach generates well-shaped elements (close to equilateral) while respecting the geometric fidelity and features of the original mesh. The approach was extended in [7] where the predefined edge length was replaced by an adaptive sizing field, sensitive to the local curvature. Since both the methods mentioned here do not require any parameterization or density function, they are simple for implementation and robust for high genus models. The proposed approach falls in this category where these algorithms are efficient for real-time applications. These methods use local operators like edge collapse, edge split, vertex relocation etc. to guarantee geometric fidelity and high-quality results with implicit feature preservation. An optimal remeshing approach should be able to handle arbitrary meshes of any genus. Unfortunately, the current remeshing techniques do not achieve the fundamental requirements. In [8], a two-step linear sparse system which iteratively alters the connectivity optimization and generates high quality meshes with feature preservation was incorporated. The presented approach gained high quality isotropic elements but lacked in improvement of minimal angle and vertex regularity.

Error-driven remeshing techniques attempt to achieve a close control between mesh complexity and variation between the original mesh and remesh. In [9], the original mesh was coarsened by variational geometric partitioning for approximation. Lloyd's iteration was used for optimization of the set of planes to minimize predefined approximation error.

Mesh simplification and approximation is another set of procedures that transforms an input mesh to output mesh with less faces, edges and vertices [2][10]. These simplification and approximation techniques in most techniques are administered by predefined user quality norms favoring meshes to maintain distinct properties of the original mesh. In literature, various simplification algorithms are categorized. Vertex clustering techniques are competent and vigorous with linear computational complexity, but at the cost of the unsatisfactory quality of the resulting mesh. In incremental algorithms, iterative decimation is performed based on the user defined criteria for the next vertex removal operation. The concern with this category of simplification algorithms is the computational complexity which can go upto $O(n^2)$ in the worst case.

Various mesh optimization strategies are devised to lessen the apparent error metrics. The authors in [11] use vertex pair contraction iteratively based on a quadric error metric to simplify the model. In [12], a new surface mesh is guaranteed to be generated from scratch to be inside the given tolerance volume. With the focus mesh complexity, algorithm overlooks mesh quality and regularity.

A 5-6-7 remeshing approach is presented in [13] where a closed triangular mesh of an arbitrary genus is converted into a closely approximated 5-6-7 mesh with a comparable vertex count. The algorithm works well with sharp features preservation of different topological and complex models but compromises on the quality of the elements. However, in [14] a good balance between mesh complexity and element quality is achieved within the user specified bounds. A series of local operators like edge split, collapse, flip and vertex relocate are applied on the input mesh with controlled approximation error. The proposed remeshing approach is inspired by the direct approach presented in [14] where they have achieved impressive results with a balance between approximation error, element quality simultaneously while preserving the features. They have only been deficient in considering the balance of the vertex regularity with the other quality parameters of the mesh.

In feature preserving remeshing, the entire focus is on feature preservation where the triangular mesh is treated as a curved object with sharp features. Automatic identification of the sharp features based on local shape, global circumstances and semantic information of the curvature is a crucial problem. A variety of approaches have addressed this problem; however, none of them consistently drove for all kinds of meshes. In [6], the authors address this problem using CVT optimization embedded by the quadric error metric approximation. With high dependency on the quality of input mesh and vertex budget, the algorithm could preserve sharp features to a limited extent. Many approaches considered sharp features to be detected and mentioned in advance or implicitly preserve features which yields to

difficulty and errors. In an isotropic remeshing [15], a local unified smoothing operator is used to align edges to sharp features. However, very little consideration was towards to the approximation error, quality, and regularity of the mesh.

In this presented work, authors emphasis only on the combined approach of high quality, error driven and feature preserving remeshing techniques and attempts to achieve a satisfactory balance between approximation error, quality of the mesh elements, mesh complexity and mesh regularity.

III. PROPOSED METHODOLOGY

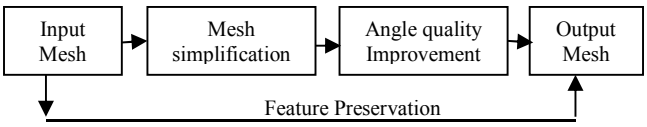


Fig. 2. Flow diagram of the proposed surface remeshing approach

Fig.2. demonstrates the flow diagram of the proposed remeshing approach. The raw mesh generated from the scanner is given as the input to the remeshing algorithm. The mesh is first simplified using a quadric based mesh simplification technique [11] within a specified user defined error bound. Further, the mesh element quality is improved using minimal angle improvement [14]. While simplifying the mesh, the features are preserved implicitly using feature intensity functions. The mesh is enhanced in terms of the element quality and regularity by performing local operators on the mesh directly. Since the proposed approach operates directly on the mesh and does not involve any parameterization and stitching, it is simple and efficient to implement.

A. Initial Mesh Simplification

The initial mesh simplification is desirable due to dense and oversampled input meshes with many redundant vertices because of the scanning process. This preprocessing in the algorithm is desirable to remove the redundant vertices of the mesh and reduce the time complexity of further processing of the mesh. Even though there are several techniques mentioned in the literature, mesh simplification using quadric error metric [11] provides an efficient and fast way to control the process of simplification with quite less space and time complexity. The simplification results in high visual fidelity in the output mesh.

The authors use a mesh simplification approach based on a quadric error. It works on a concept of "edge collapse" for mesh simplification. Given two vertices v1 and *v2*, a pair *(v1,v2)* is contracted to the new position v. This contraction is possible iff *(v1,v2)* is an edge or $||v1\text{-}v2|| < \mathcal{E}$, where $\mathcal{E}$ is the user specified threshold parameter. If $\mathcal{E} > 0$, then a pair of close vertices can be contracted, resulting in a non-manifold mesh. Vertex contraction can be eliminated by setting $\mathcal{E}$ to 0. A 4x4 error metric Q is used to calculate error in terms of the distance measure. Each vertex v is held by error metric Q_v. The error at a vertex $v= [v_1,v_2,v_3,1]^T$, $\Delta(v)$ is described as v^TQ_vv. For a valid pair *(v1, v2),* new position can be either of the two vertices (*v1* or *v2*) or the midpoint of the pair i.e. *(v1+v2/ 2)*. This selection of the new position is not always suitable. Hence, a better approach of selecting new position of v is that minimizes the error $\Delta(v) = (v^TQ_1v+v^TQ_2v)/2 =[v^T(Q_1+Q_2)v]/2$ where Q_1 and Q_2 are the error metric matrices of the vertices *v1* and *v2*. After computation of *v*, an edge (*v1,v2*) is collapsed. The new vertex v obtains the error value $\Delta(v)$ and error metric (Q_1+Q_2). The algorithm proceeds by placing all the nominated edges in a heap by means of $\Delta(v)$ as a key. While the heap is non-empty, the top edge is removed, collapsed to the calculated *v*. Finally, the mesh is updated along with the keys.

To find initial Q_v for *v*:

a. Given: a plane *P: ax+by+cz+d=0,* where $a^2+b^2+c^2=1$ (i.e. normalized) and a point $v=(x_1,y_1,z_1)$, the error i.e. the distance from vertex *v* to plane *P* is $\Delta_p(v)=ax_1+by_1+cz_1+d$.

b. If $P=<a,b,c,d>$ and $v=<x_1,y_1,z_1,1>$, then $\Delta_p(v)=ax_1+by_1+cz_1+d=P.v$

c. The distance i.e. the error at vertex *v* with reference to the plane *P* is evaluated by persevering coordinates of *v* coordinates into an equation of plane *P*. If *P.v* is zero, then *v* is inside *P*. Else *P.v* estimates the signed "distance" from *v* to *P*.

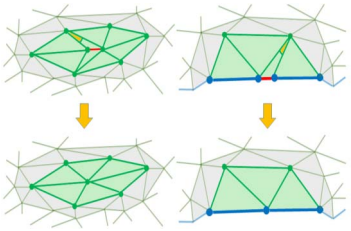
$$\begin{aligned}(P^T.v)^2 &= (P^T.v)^T\,(P^T.v)\\ &= (v^T.P)\,(P^T.v)\\ &= v^T(PP^T)v\\ &= vT\begin{bmatrix} aa & ab & ac & ad\\ ab & bb & bc & bd\\ ac & bc & cc & cd\\ ad & bd & cd & dd\end{bmatrix}v\\ &= v^T\,Q_p(v)\,v \qquad (1)\end{aligned}$$

where $Q_p(v)$ is the error metric matrix of vertex v with respect to P, instead of the error value itself.

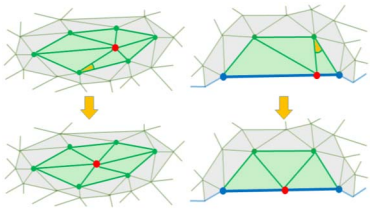
d. The error metric for each vertex v in the mesh is calculated as sum total of all $Qp(v)$, where *P* is the plane containing all the incident triangles of v: $Q_v=\sum_{for\ all\ p's\ incident\ to\ v} Qp(v)$.

B. Minimal Angle Improvement

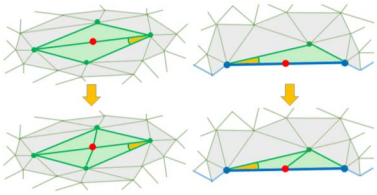
The second step involves the use of local operators [2][14] to enhance the quality of mesh by monitoring the minimal angles of the mesh to the looked-for angle bound. The need of improvement in the minimal angle is due the skewness that these angles lead in the meshes resulting in the low quality of the mesh. The limit to the minimal angle is decided by the user, but generally to avoid low quality in the mesh; it should be more than 30 degrees. The local operators such as edge collapse, edge split, vertex relocate and edge flip are used for mesh angle improvement. Generally, edge flip operation is the least preferred as it is a combined action of edge split subsequently succeeded by an edge collapse. Pure edge collapse and edge split do not contribute much in mesh quality improvement, hence are being associated with the vertex relocation operation.



(a) Edge collapse



(b) Vertex Relocation



(c) Edge split

Fig. 3: Local operators [2][14]

The process of local operators is as shown in fig. 3. The local operators are executed if the resultant mesh does not lead to another interior angle which is less than the predefined threshold and maintains 2-D manifoldness in the mesh. The important task is to schedule these local operators for mesh quality improvement in terms of vertex count and minimal angle. The edge collapse operation is first performed as it reduces the vertex count. If in case, the edge collapse is not possible, then vertex relocation is done to improve the mesh angles. This phase tries to lift the interior minimal angles to a predefined angle threshold. In case, both edge collapse and vertex relocation are not possible, then only edge split is performed. Edge split is performed on the principle of longest-side propagation path. The longest edge in the region is searched and then split. While performing these operations, some constraints such as topology, geometry and fidelity must be satisfied. The operators should not lead to the fold-overs by flipping the orientation of faces and the approximation error between the two meshes should remain below the threshold. Link condition is checked for each edge collapse to prevent the topology changes. A series of vertex relocation operations are being performed to optimize vertex locations and ensured that further improvement in the interior angles cannot be done.

C. Feature Preservation

Preservation of features is the most critical task in remeshing which is a hard and unsolved problem. Instead of explicit tagging of the features, the authors have tried to implicitly identify and preserve the features [14]. This reduces the time-consuming manual intervention and enables the retrieval of the features mislaid during the process of inadequate sampling by advance 3D laser scanners. The edge features (boundaries and creases) and the vertex features (tips, darts, cusps and corners) are distinguished in implicit feature identification and preservation. Feature handling is incorporated during the operations performed by local operators. Gaussian curvature $K(v)$ is considered to identify angle defects i.e. feature vertices.

$$K(v) = \begin{cases} \pi - \theta sum(v) & v \text{ is on the boundary} \\ 2\pi - \theta sum(v) & otherwise \end{cases} \quad (2)$$

where $\theta sum(v)$ is the total sum of all interior angles enclosed by v.

Large dihedral angles contribute in prediction of feature edges. Feature edge intensity *E(v)* of a vertex is defined as the maximal unsigned dihedral angle of the edges adjacent to *v*.

$$E(v) = \max_{e \in Ne(v)} |D(e)| \quad (3)$$

where $Ne(v)$ are the adjoining edges to *v* and $D(e)$ is the dihedral angle at the edge.

Feature intensity is eventually termed as the combination;

$$F(v) = (\tau(|K(v)|) + 1).\left(\tau\left(E(v)\right) + 1\right) - 1 \quad (4)$$

with the transfer function $\tau(x) = \min\{\pi, 2.x\}$. This rescales' *x* by a factor of 2 and clamps at π. This leads to the feature intensity is a value between 0 and $((\pi + 1)^2 - 1)$.

IV. Performance Parameters

There exist limited standard performance parameters for remeshing techniques. The output mesh is evaluated based on the following performance parameters as mentioned in literature.

a. $\#V$: No. of vertices in the mesh.

b. $Qmin$: The minimal quality of the triangles in the remesh. Triangle quality is evaluated as:

$$Qt = 2\sqrt{3}\frac{St}{Pt.ht} \quad (5)$$

where S_t: an area of the triangle, P_t: in-radius of the triangle and h_t: longest edge length of the triangle.

c. $Qavg$: The average triangle quality of the overall mesh.

d. θmin: The minimum of all the interior angles formed in the triangles in the mesh.

e. θmax: The maximal of all the interior angles formed in the triangles in the mesh.

f. $Hdist$: Hausdroff distance is the approximation error between the original mesh and the remesh.

g. RMS: Root mean square distance between the original mesh and remesh.

h. $\theta < 30$: The percentage of angles in the mesh smaller than 30 degrees.

i. $V567$: The percentage of regular vertices in the mesh i.e. percentage of vertices with valence 5,6 or 7 in the mesh.

V. Experimental Setup

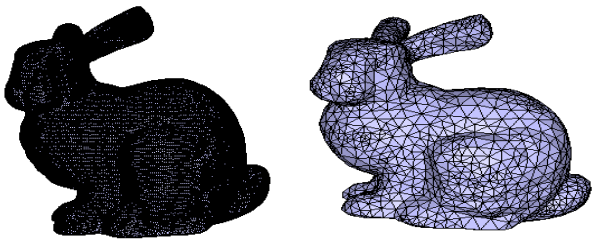
The proposed approach was implemented in C++ on windows 10 operating system of 64-bit. The CGAL library [16] was used for different operations involving basic data structures. The proposed technique was tested on a variety of triangular surface meshes and analyzed them with respect to the above-mentioned performance parameters. The results of the proposed methodology were compared with MAI (Minimal Angle Improvement) algorithm [14] that was

mentioned as an attempt to achieve the balance between the three criteria of remeshing. The author [14] has compared their results with various other remeshing techniques and proved that their technique is one of the best methods to optimize the balance between various criteria.

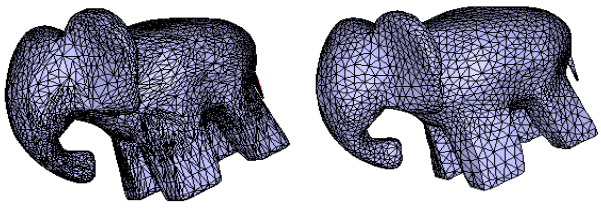
The approximation error limit was restricted to 0.2% of the diagonal length of the bounding box i.e. on any operation performed by local operators; the approximation error should not be beyond the pre-defined limits. The statistics evaluated of the above-mentioned experiment are presented in table 1.

As observed in table 1, the mesh complexity of the original mesh is high. This generally leads to high time complexity for processing of many applications. The scanned input meshes have the low quality of the triangles. Generally, the high quality of the triangles (close to 1) is expected. The small interior angles lead to skewness and sharpness in the meshes. High interior angles in the mesh lead to inferior quality with an occurrence of long edges. Overall quality of the input meshes is less as the percentage of triangles less than 30 degrees is more in the input meshes. These interior angles less than 30 degrees are major concern for many applications. Because of high mesh complexity, low triangle quality and irregularity, there arises a need of remeshing the input mesh.

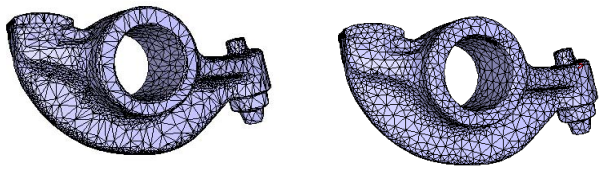
As seen in table 1, the results of the proposed approach and MAI have achieved comparatively a smaller no. of vertices in the remesh, improved mesh quality and interior minimal and maximal angles. The proposed approach has a limitation of reduction in the number of vertices but on the contrary, has better triangle quality, enhanced interior angles and vertex regularity compared to MAI. The limitation of more no. of vertices could be due to the quadric error approach of simplification. Since the quadric error approach reduces only the most inadequate vertices while maintaining the smoothness and regularity in the mesh, it is acceptable as a simplification approach.



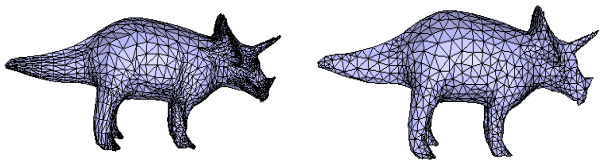
(a) Input Bunny (b) Remesh Bunny



(a) Input Elephant (b) Remesh Elephant



(a) Input Rockerarm (b) Remesh Rockerarm



(a) Input Triceratops (b) Remesh Triceratops

Fig. 4. The original input meshes and the output remesh models of the proposed approach

A. Limitations

The major limitation of the proposed approach is that it produces a greater number of vertices during remeshing as compared to MAI. Secondly, even though the user can decide and control the threshold on the minimal angle (in the proposed approach to above 30^o), there are no theoretical guarantees to prove any upper bound of the minimal angle. The approach does not work well while remeshing the noisy models as feature handling and preservation becomes difficult sometimes. Hence, some appropriate technique for feature extraction and preservation for noisy models needs to be worked on. It also has a limitation of handling only 2-mainfold meshes. The method does not support an adaptive/ area based remeshing based on the smoothness and the curvature of the surface.

VI. Conclusion

In this paper, an approach for surface remeshing based on quadric error mesh simplification and minimal angle improvement using local operators is presented. The proposed technique improves the minimal and average quality of mesh at the cost of simplification. The regularity and RMS have shown improvement in the comparative results while minimal angles, maximal angles, Hausdroff distance are not enhanced significantly. As evaluated with the other reported state-of-art methods, the achieved results are claimed to be better with respect to regularity as well as the quality.

Moreover, the existing methods do not consider any nature inspired algorithms to perform remeshing that could provide significant results in mesh simplification which could be the future scope of the proposed work. Also, different clustering techniques can be incorporated to perform an area-based surface remeshing.

References

[1] P.Alliez,G. Ucelli, C. Gotsman, and M. Attene, “Recent advances in remeshing of surfaces," in *Shape analysis and structuring, Mathematics and Visualization.* Springer, 2008, pp. 53-82.

[2] M. Botsch, L. Kobbelt, M. Paulty, P. Alliez, and B. Levy, *Polygon Mesh Processing*. AK Peters, 2010.

[3] S. Vallete, J-M. Chassery, and R. Prost, “Generic remeshing of 3D triangular meshes with metric-dependent discrete Voronoi diagrams”,

in *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 2, pp. 369-381, 2008.

[4] D.M.Yan, and P. Wonka, "Non-obtuse remeshing with centroidal Voronoi tesellation", in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 9, pp. 2136-2144, 2016.

[5] A.G.M.Ahmed, J. Guo, D.M.Yan, X. Zhang and O. Deussen, "A simple push-pull algorithm for blue-noise sampling", in *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 12, pp. 2496-2508, 2016.

[6] M.Botsch and L.Kobbelt, " A remeshing approach to multiresolution modeling", in *Proceedings of Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, pp. 185-192, 2004.

[7] M. Dunyach, D. Vanderhaeghe, L. Barthe and M. Botsch," Adaptive remeshing for real-time mesh deformation", in *Eurographics short papers. Eurographics Association*, pp. 29-32, 2013.

[8] J. Du, Y. Jin, and R. Tong," As-equilateral-as-possible surface remeshing", Journal of Advanced Mechanical Design, Systems, and Manufacturing 9(4), pp. JAMDSM0052- JAMDSM0052, 2015.

[9] D. Cohen-Steiner, P.Alliez, and M. Desbrun, "Variational shape approximation,"in *ACM Transaction on Graphics*, vol. 23, no. 3, pp. 905-914, 2004.

[10] Luebke, and P. David, "A developer's survey of polygonal simplification algorithms", in *IEEE Computer Graphics and Applications*, vol. 21, no. 3, pp. 24-35, 2001.

[11] M. Garland and P.S. Heckerbert, "Surface simplification using quadric eror metric", in *SIGGRAPH*, pp. 209-216, 1997

[12] M. Mandad, D. Cohen-Steiner, and P. Allez, " Isotropic approximation within a toleranc evolume," in *ACM Transaction on Graphics*, vol. 34, no. 4, pp. 64:1-64:12, 2015.

[13] N. Aghdaii, H. Younesy, anf H. Zhang, "5-6-7 meshes: Remeshing and analysis", in *Computers and Graphics*, vol. 36, no. 8, pp. 1027-1083, 2012.

[14] K. Hu, D.M. Yan, P. Alliez, and B. Benes, "Error bounded and feature preserving surface remeshing with minimal angle improvement", in *IEEE Transcations on Visualization and Computer Graphics*, vol. 23, no. 12, pp. 2560-2573, 2016.

[15] W. Jakob, M. Tarini, D. Panozzo, and O. Sorkine-Hornung "Instant field-aligned meshes", in *ACM Transaction on Graphics (Proceedings of SIGGRAPH ASIA)* , vol. 34, no. 6, 2015.

[16] "CGAL, Computational Geometry Algorithms Library," http://www.cgal.org

TABLE I. THE STATISTICS AND COMPARISON OF THE EXPERIMENTAL RESULTS

Input	Method	#V	Qmin	Qavg	Θ min	Avg minimal angle	Θ max	H dist (% bb)	RMS (% bb)	Θ < 30	V-567
Bunny	Original	34834	0	0.712713	0.000114	36.7525	180	-	-	7.033	97.7723
	Proposed	2228	0.477228	0.777612	30.004	43.2884	118.347	20.86	0.0608	0	92.5404
	MAI	1228	0.468763	0.791394	30.0035	43.6992	119.422	19.99	0.0655	0	90.228
Elephant	Original	6859	0.01973	0.59281	0.657049	29.7776	169.866	-	-	53.522	82.1694
	Proposed	3860	0.473302	0.81327	30.0116	44.983	118.865	19.97	0.0268	0	94.0155
	MAI	1088	0.469766	0.780501	30.0252	43.2619	119.298	19.98	0.0656	0	89.7059
Homer	Original	6002	0.044138	0.662495	2.14959	34.6365	173.302	-	-	36.95	96.0513
	Proposed	4046	0.482262	0.816367	30.0071	45.2081	117.763	19.59	0.0241	0	95.5017
	MAI	890	0.467194	0.790059	30.0608	43.6896	119.168	19.91	0.0680	0	90.6742
Tricertops	Original	2832	0	0.58044	0.00024	29.5743	180	-	-	52.0495	95.1624
	Proposed	1277	0.483848	0.788057	30.0149	43.5649	117.567	19.99	0.0487	0	89.2717
	MAI	977	0.46704	0.779515	30.0021	43.0549	119.637	19.96	0.0667	0	86.694
Rockerarm	Original	3431	0.000983	0.64922	0.042329	34.07	179.817	-	-	37.5838	83.8823
	Proposed	2506	0.474429	0.804625	30.0235	44.5905	118.705	19.99	0.0409	0	93.8994
	MAI	1116	0.471303	0.777638	30.0019	43.0085	119.106	20.29	0.0635	0	92.0251
Helmet	Original	496	0.051351	0.644992	2.33577	34.5162	171.445	-	-	36.8	82.8629
	Proposed	604	0.480555	0.778058	30.0481	43.215	117.938	19.97	0.050	0	87.9139
	MAI	665	0.492993	0.786964	30.0065	43.5146	116.409	19.99	0.055	0	89.6241
Close Hemisphere	Original	8068	0.159498	0.552874	5.56891	26.5123	92.8214	-	-	57.9365	99.9504
	Proposed	2961	0.520327	0.757047	30.0027	40.3853	112.878	19.21	0.019	0	98.0412
	MAI	607	0.487504	0.813586	30.0029	45.3437	117.119	19.99	0.057	0	95.2224
Blade	Original	5002	0.146393	0.58051	5.76437	29.9701	160.207	-	-	54.95	87.1451
	Proposed	1948	0.475519	0.798279	30.0071	44.2468	118.582	19.97	0.039	0	92.6591
	MAI	838	0.471504	0.776225	30.0039	43.0594	119.079	19.96	0.056	0	91.7661
Disk	Original	210	0.025917	0.348819	1.71385	16.447	176.571	-	-	49.5192	98.0952
	Proposed	130	0.499594	0.782429	30.1076	43.27	115.436	19.59	0.024	0	98.4615
	MAI	116	0.549452	0.77043	30.0494	41.9238	109.517	17.72	0.0214	0	100
U-shape	Original	86	0.001124	0.291133	0.073896	12.3893	179.851	-	-	88.0952	73.2558
	Proposed	133	0.496768	0.791348	30.0046	43.3639	115.962	19.85	0.0378	0	93.985
	MAI	119	0.513472	0.766387	30.3834	41.147	113.539	18.91	0.0392	0	89.916
Smooth crease	Original	6177	0.053251	0.715344	1.7898	44.7449	153.149	-	-	0.46153	99.886
	Proposed	2710	0.479821	0.776389	27.5514	44.7007	118.038	13.94	0.0042	0.03	93.87
	MAI	1963	0.554547	0.765848	30.356	43.5621	108.925	19.78	0.0236	0	93.65

Capturing Contextual Influence in Context Aware Recommender Systems

Vandana A.Patil
Information Technology Department
St.Francis Institute of Technology
Mumbai,India
vandanapatil@sfit.ac.in

Dr.Deepak J.Jayaswal
Electronics & Tele Communication Department,
St. Francis Institute of Technology
Mumbai, India.
djjayaswal@sfit.ac.in

***Abstract*—In the present evolving phase of information technology, Recommender systems (RSs) have been established as widely accepted platform for handling & managing the information overload problem. In order to facilitate an individual user in decision making for selection of any product or service, user preferences are first captured implicitly or explicitly and a predictive model is built to derive personalized recommendations. Data Sparsity, malicious / biased data by users, diversity in recommendations, temporal dynamics, etc are few of the key challenges experienced by new age recommendation systems. Dimensionality reduction techniques help to decompose the rating matrix in the form of latent factors with lower ranks and attempts to solve the data sparsity problem. Matrix Factorization (MF) is the well-accepted technique in this aspect. In order to generate more accurate and meaningful recommendations, Context aware RS (CARS) is the new emerged technique in recommender systems. Tensor Factorization or generalized matrix factorization facilitates the most suitable and generic way of integrating contextual information into RSs. The relevant contextual information will always improvise the performance of the recommender system but irrelevant contextual information could degrade it drastically. The work presented here, provides the discussion over comparison of Recommender System's performance at various scenarios such as generating recommendations without considering any contextual information and with considering the contextual information. There are few techniques available to find out the context relevancy in CARS. Their impact on the performance of RS is also discussed.**

Keywords— Recommender System, Context aware Recommendation Systems, Tensor Factorization, Context relevancy, etc.

I. INTRODUCTION

Nowadays, the internet is growing to such a large extend and a wide range of diversified information is accessible to its users. Every individual is surfing through this information to find relevant resources which will satisfy his/her needs. But the information available and the rate of producing new information on the internet are really very high for an individual to handle it efficiently. The resources available are ranging from movies, news articles, blogs, music, documents, various e-commerce products, etc.

This situation gives rise to the serious problem of information overload as individual user cannot dedicate too much effort to browse through all the information available on the internet before reaching to the actual target. Thus, the overflow of information drastically degrades user's ability to choose the most appropriate option as per their needs. Also a lot of time is required to search through all the available information and arrive at a conclusion with required resources. In long run users may loose interest in this kind of process. Therefore, recommender systems are developed to solve this critical issue of information overflow and assist the individual user in decision making. The recommender systems are proved as a way superior to the traditional information retrieval mechanisms as they develop a long term predictive model for every individual user in accordance with his preferences and also tactfully combine other user's opinions to provide each user personalized recommendations [1].

Recommender Systems (RSs) provide the integrated framework of various tools and techniques to identify items from a huge collection of data that could be of interest to users. Generally, the personalized recommendations are presented as the structured group of recommended items. During the process of deriving the recommendations, RSs identify the products and/or services which are most relevant, in context to the user's preferences. These preferences can be gathered from the users either explicitly or implicitly. Getting the ratings for items and information through feedbacks are explicit means of gathering preferences, whereas gathering the preferences by tracking the user activities is the implicit way. The application of recommender systems in an e-commerce environment can drastically impact the financial performance of any business as well as the level of the trust with customers. The scope of the presented work is limited to evaluation of the performance of Tensor factorization based context aware RS in three scenarios namely without context, with context and context with relevancy. The evaluation has been done with varying number of features and context combination.

II. CLASSIFICATION OF RS

The process of deriving personalized recommendations through RS varies depending on the quality, size and sparsity of sample rating data available. It is further influenced by the filtering techniques used, the recommendation model selected and the expected quality of the recommendations. Considering all these parameters the Recommender systems are broadly classified as follows:

A. Content-based RS

Content-based recommendation also known as the information filtering method first identifies the descriptive features of the items and users' ratings for those items. It then signifies the recommendations to a target user based on the same. This technique rely on the simple assumption that items possessing similar features are likely to be rated similarly.

The real challenge in the context aware RSs is to distinctly identify and extract the item attributes/features. On the basis of these features extracted from the items the user has rated, his profile is built. In the next step the user profile is compared with item profiles of new items. Those items

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

whose features are matching with the features in user profile are recommended to the intended target user. The main shortfall of the filtering technique based on content is that, if the items do not contain exactly the same features mentioned in the user's interest profile, they may not be recommended to the user even if they are almost similar [2].

B. Collaborative filtering based RS

In the collaborative Filtering (CF) based technique [2], the items preferred by likeminded users in the past are recommended to active users. The co-relation between two users is termed as similarity and is evaluated on the basis of the ratings given to the respective items by the users. Hence, this technique is also referred as "P2P correlation".

Collaborative filtering techniques require explicit user ratings of an item to generate meaningful recommendations. It does not require content or features of the items. Collaborative filtering facilitates evaluation of item features not directly, but indirectly, because users are doing the evaluation as well as analysis on their part and then based on that they are rating the items.

C. Hybrid RS

Hybrid RSs are the systems, built by combining two or more recommendation techniques by complementing the benefits provided by each of them. For example, the new-user/ item problem is encountered in CF methods but does not exist in case of content-based filtering techniques because here the prediction of rating for new items is based on attributes of item and preferences by user. Several mechanisms have been discussed in the literature for combining various RS techniques in order to develop a new hybrid system.

Besides the above mentioned RS classification, one more kind of taxonomy for recommendation techniques viz. memory-based and model-based RS is widely accepted.

Memory-based approach [3] relies mainly on the matrix of item, user and ratings. They usually operate on the user/item similarity principle. Similarity metrics such as Pearson Correlation coefficient, Cosine Similarity coefficient, Jaccard Similarity coefficient etc. are used to derive the nearest neighbors.

Model-based approach [3] uses rating information to create a model and based on which the recommendations are generated. Bayesian classifiers, fuzzy systems, neural networks, genetic algorithms and matrix factorization are few well-adopted techniques under this paradigm.

As most of the recommendation techniques rely on the rating data and the scarcity of it happens to be the key challenge in the process of generating meaningful recommendations.

Dimensionality reduction techniques can proficiently handle the severe data sparsity problems raised in real world recommender systems [4]. The dimensionality reduction methods are best realized through Matrix Factorization (MF). In this approach the missing data interpretation is considered as the matrix completion problem. The Singular Value Decomposition (SVD) method provides good prediction results, but proves to be very expensive in terms of computational complexity; SVD method is deployed on static known rating data which does not alter with time.In succession to SVD, several MF-based factorization techniques have been evolved in the model based CF recommender systems. CF based on Regularized MF (RMF) has shown high accuracy and scalability at the Netflix Prize competition. Few of the widely used models of this kind are biased SVD model, biased regularized MF model, SVD++ model, Probabilistic MF model, etc [5].

III. Context Aware RS

The recent advancements in the personalized recommender systems gave rise to the context aware RSs. They aim at generating more accurate, relevant and personalized recommendations. Context is any information class used to characterize/ model the state of an entity. For a user, there can be various context attributes that can be considered, such as time, location, weather, companion, mood etc. The recommender systems analyze the contextual information of the reviewers as well as current user to generate relevant recommendations. The incorporation of contextual information in the recommender system makes it more personalized and efficient [6].

The context aware recommender systems can be implemented through any of the following techniques:

A. Context aware pre-filtering:

In this technique, initially the relevant ratings are selected considering the desired specific context information and then the predicted ratings are calculated by applying suitable 2D recommendation algorithms. Thus, the irrelevant records with respect to specified context information are first removed out from the dataset and then the recommendations are generated with respect to it [6]. Few examples of contextual pre-filtering technique are found in the literature which is discussed further.

In [7], authors have suggested a reduction-based method, which first reduces the given dataset based on the specific context. Thus, the multidimensional (MD) contextual recommendation problem is shifted to standard two dimensional *User* × *Item* recommendation space. Once the reduction is done, all the existing research on 2D recommendation algorithms can be easily applied to context aware recommender systems.

A technique in-line with contextual pre-filtering is used in [8] to derive recommendations for advertisements to mobile users. The contextual variables considered are location, interest of user and time. In [9] authors have used contextual pre-filtering technique to analyze the effect of contextual information on an online retailer data. The item splitting technique which is quite different approach to contextual pre filtering is adopted in [10] to evaluate the performance of context aware RSs.

B. Context aware post-filtering:

In Context post-filtering technique, the predicted ratings are obtained using two-dimensional recommendation algorithms on overall data initially. Then considering the context values, the resultant list of generated recommendations is filtered out.

The contextual post-filtering method is further classified as heuristic technique and model-based technique. Heuristic post-filtering technique aims at searching for item characteristics, which are common for the specified user in some specified context. Whereas, model based post-filtering technique builds a predictive model, which calculates the

probability of choosing a specific item in a given context by a specific user. In short, it is termed as the probability of relevance, as it suggests the likelihood of choosing the item in a given context and generates the recommendation list in accordance with this probability [6].

In [11], authors have compared two different post-filtering methods – Weight and Filter with the help of several real-world e-commerce datasets. In weight post filtering, certain weight is assigned to the predicted rating considering its relevance in the specific context. The list of recommended items is then reordered in accordance with the weighted predicted rating values. The second method, i.e. filter post filtering, directly filters out the items having less relevance with the specific context and then generates the recommended item list.

The major advantage of both contextual pre-filtering as well as post filtering approach is that all the traditional two dimensional recommendation techniques can be applicable as it is to them; however, the computational complexity will be increased to considerable extent [12].

C. Contextual modeling:

In Contextual Modeling, the context information is used directly into the recommendation function as a key predictor for item rating by a user. Therefore, a multidimensional recommendation model can be formulated by this approach. With contextual modeling, the rating function now becomes a multi-dimensional function ($R: User \times Item \times Context \rightarrow Rating$), where *User* and *Item* is the set of users and items respectively. Rating is the set of ratings and Context is the set of contextual information to be considered within the application.

In [13], authors have adopted contextual modeling technique by incorporating additional contextual dimensions directly into the recommendation model. Further, machine-learning techniques are used to generate meaningful recommendations in a restaurant recommender system. Similar approach is also adopted in [14] by introducing tensor factorization based method. Authors have modeled the $Users \times Items \times Contexts$ space as an n-dimensional tensor. This tensor is further factorized to generate personalized context aware recommendations.

Contextual modeling is most widely used techniques nowadays as it outperforms over contextual pre and post filtering if computational complexity is taken into consideration.

Contextual modeling is best realized through Tensor factorization, which is the N-dimensional generalization of traditional two-dimensional matrix factorization [14]. CP decomposition and Tucker decomposition are the two most widely used tensor factorization algorithms, which are discussed in detail further.

Tensors can be visualized as multidimensional arrays of numerical values. They are used in order to generalize the matrices to higher dimensions. Data cube is the simplest high-dimensional case which is a three-dimensional array of numerical values [15].

Tensor factorization serves as effective mechanism to generate a predictive model which reveals patterns from the data. The multifaceted nature of user-item interactions are taken into consideration by tensor based factorization approach [15].

- Candecomp/ Parafac (CP) decomposition:

This type of tensor decompositions are also considered as rank decompositions. In this approach, a tensor is expressed as the sum of a finite number of rank-one tensors. The CANonicalDECOMPosition (CANDECOMP) and the PARAllelFACtors (PARAFAC) decomposition are the most prominent rank decompositions. Though both of them belong to different knowledge domains and have been independently discovered, they both rely on the same principles. Therefore, this type of decomposition is referred as the CANDECOMP/ PARAFAC or canonical polyadic decomposition (CPD). The 3-way CPD function can be formalized as follows:

$$\min_{\hat{x}} \|x - \hat{x}\| \ where\ \hat{x} = \sum_{r=1}^{R} a_r \circledcirc b_r \circledcirc c_r = [\![A, B, C]\!] \quad (1)$$

The CPD of given tensor can be computed with the aid of many algorithms. Jennrich's algorithm and Alternating Least Squares (ALS) Algorithm are the most popular once amongst them.

- Tucker decomposition:

In Tucker decomposition, a tensor is decomposed into one core tensor and multiple matrices, each one belonging to individual mode in tensor. The Tucker decomposition can also be considered as a higher-order Principal Component Analysis.

For the 3-way tensor $X \in R^{I \times J \times K}$ with $G \in R^{P \times Q \times R}$, $A \in R^{I \times P}$, $B \in R^{J \times Q}$, $C \in R^{K \times R}$ the Tucker decomposition can be expressed as follows:

$$\min_{\hat{x}} \|x - \hat{x}\| \ with\ \hat{x} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}\, a_r \circledcirc b_r \circledcirc c_r = \mathcal{G} \times_1 A \times_2 B \times_3 C = [\![\mathcal{G}; A, B, C]\!] \quad (2)$$

This definition of Tucker decomposition can be extended beyond 3 modes; however, the storage requirements are likely to grow exponentially based on the number of modes. Tucker decomposition is provided by the Higher Order Orthogonal Iteration (HOOI) and Higher Order Singular Value Decomposition (HOSVD) methods.

We may come across the scenario wherein the irrelevant or less important contextual information gets incorporated into the modeling of RS which tends to generate irrelevant recommendations. This eventually degrades the accuracy and hence the performance of RSs. Thus, identification of relevant contextual information will definitely lead to improved recommendations as well as the reduction in the real data acquisition cost.

IV. Context Relevancy in CARS

Utilizing context information for recommendations has been always challenging to identify the meaningful and relevant contextual information. As, a single irrelevant context selected may act as noise and could degrade the performance of CARS.

There are few examples in the literature where an attempt to capture the contextual influence is made are discussed

below. They all have used LDOS-CoMoDa dataset for context relevancy assessment. It contains ratings registered by 121 users for about 1232 movies. The dataset having overall 2296 ratings contains 12 contextual variables, namely time, daytype, endEmo, dominantEmo, season, location, weather, social, decision, interaction, mood, physical.

In [16], information gain algorithm is used to identify the most relevant context variables. The algorithm calculates the entropy of contextual variables. If it falls above the predefined threshold value, the contextual variable is considered as relevant. The threshold value considered is 0.014. Among the 12 contexts available, contexts endEmo, dominantEmo, interaction, mood and social are derived as the most relevant contexts.

In [17], two distinct approaches of context relevancy determination are discussed. The context relevancy assessment method is based on user survey and the context relevancy detection method is based on statistical testing approach. Although both the methods aim at deciding the context relevancy they differ in the information needed to decide the relevancy, when to use which method and whether they are based on real or hypothetical condition.

- Assessing Context Relevancy (survey method):

Online survey was conducted, in order to obtain user's opinion on whether the contextual information is relevant or not. As LDOS-CoMoDa dataset contained 12 contextual variables, 12 distinct questions were presented in the online questionnaire.

The probable answers for various questions were one amongst: "Yes", "No", "May be", "Probably Yes" and "Probably No". For the convenience of participants all the questions were presented and well explained to them in their own mother tongues. Assessment score for each context variable can be calculated as:

$$S = \sum_{i=1}^{5} w_i n_i \quad (3)$$

Where, n is the amount of answer *i* (*i* goes from 1(NO) to 5 (YES)) and w is weight assigned to answers (-2 for NO to +2 for YES). If the derived score $S \geq 0$, then the contextual information under consideration is relevant. The relevant contextual variables found according this technique are time, location, social, endEmo, dominantEmo, mood and interactions.

- Detecting Context Relevancy:

In this approach, hypothesis testing is used to determine the degree of association between the contextual information and ratings. Experimentation was conducted on LDOS-CoMoDa dataset. The null hypothesis for the test was stated as the contextual information and rating values are independent and in contradiction to this, the alterative hypothesis was stated as they depend on one another. If the null hypothesis is rejected it can be concluded that contextual information influences the rating value and hence it can be considered as relevant.

LDOS-CoMoDa database contains all the contextual variables which are categorical in nature. Therefore, the Freeman-Halton test is used. The authors have decided the significance level of the test as $\alpha = 0.05$. The list of relevant contexts derived includes location, endEmo, dominantEmo, mood, physical, decision, social and interaction.

The next section explains the methodology adopted to implement Tensor factorization based RS and to incorporate context relevancy assessment method into it.

V. METHODOLOGY ADOPTED

In our approach, we have adopted HOSVD tensor decomposition technique [14]. In this technique, the 3-dimensional tensor is factorized into three matrices $U \in R^{n \times dU}$, $I \in R^{m \times dI}$ and $C \in R^{c \times dC}$ and one central tensor $G \in R^{dU \times dI \times dC}$ [23]. Therefore, the decision function for a single user *i*, item *j* and context *k* can be stated as follows:

$$Y_{ijk} = G \times_U U_{i*} \times_M I_{j*} \times_C C_{k*} \quad (4)$$

This technique is well suitable for large sized real world datasets. Its main advantage is that if allows for the complete control over the dimensionality of user, item and context factors. Following are the steps to obtain the objective function *R[U,I,C,G]*.

- In analogy to MF approaches, the loss function is defined as:

$$L(Y, \hat{Y}) := \frac{1}{||G||} \sum_{i,j,k} D_{ijk}\, l\left(Y_{ijk}, \hat{Y}_{ijk}\right) \quad (5)$$

Where $l\ (Y,\hat{Y})$ is point wise loss function and calculated as difference between predicted and actual rating value for a single user *i*, item *j* and context *k*. *L* is total loss over *Y* and $\hat{Y}$

- Minimizing the above loss function may lead to over-fitting. Therefore, the regularization term is added as follows. For regularization l2 norm is adopted.

$$\Omega[U, I, C] := \frac{1}{2}\left[\lambda_U ||U||^2_{Frob} + \lambda_I ||I||^2_{Frob} + \lambda_C ||C||^2_{Frob}\right] \quad (6)$$

- Similarly, l2 norm penalty is also imposed on the central tensor *G*:

$$\Omega[G] := \frac{1}{2}[\lambda_G ||G||^2_{Frob}] \quad (7)$$

- Finally, the objective function is generated which is a combination of $L(Y,\hat{Y})$, $\Omega[U,I,C]$ and $\Omega[G]$.It is formulated as:

$$R[U, I, C, G] = L(Y, \hat{Y}) + \Omega[U, I, C] + \Omega[G] \quad (8)$$

For minimizing the above loss function the stochastic gradient descent (SGD) optimization strategy is used.

VI. EXPERIMENTAL SETUP

D. Experimental Setup:

LDOS-CoMoDa dataset is used for experimentation as it is the most context rich real dataset available for the same. The various context relevancy techniques mentioned in section IV are studied and their list of relevant contexts is compared along with one more technique named feature based context selection.

This technique derives context relevancy based on the Pearson correlation coefficient between output variable which is rating here, along with the context variable. Higher is the correlation more is the relevance of that context. The following table summarizes the lists of relevant contexts obtained from this method along with all earlier discussed methods.

TABLE I. Listing of relevant contexts for various techniques

Relevancy Ranking	Information Gain	Context Relevancy Assessment Method	Apriory Power Analysis Method	Correlation based Feature Selection
1	**endEmo**	Time	**endEmo**	**endEmo**
2	**DominantEmo**	Location	**DominantEmo**	**DominantEmo**
3	**Interaction**	**social**	**mood**	**Interaction**
4	**Mood**	**endEmo**	Physical	**Mood**
5	**Social**	**DominantEmo**	Decision	Location
6		**Interaction**	**Social**	**Social**
7		**Mood**	Location	
8			**Interaction**	

From the above comparison in TABLE- I, it is observed that endEmo, dominantEmo, Interaction, Mood and Social are the most relevant contexts (5 nos.) found in all of the techniques. Therefore, they are used as relevant contexts in further experimentation. The data set is divided into training and testing dataset in 2:1 proportion. The performance for following three models is evaluated over 20 iterations. The number of features and the number of contexts both are varied in the range of 3-5. The system performance is evaluated considering following three approaches:

- Recommendations generated with the factorization based approach without considering any context
- Recommendations generated with Tensor factorization based approach with considering context
- Recommendations generated with Tensor factorization based approach with considering contextual relevancy while selecting the context.

A. Evaluation Protocol:

$$RMSE = \sqrt{\frac{1}{K}\sum_{ijk}^{n,m,c}\left(Y_{ijk} - F_{ijk}\right)^2} \quad (9)$$

Where, Y_{ijk} is actual rating value and F_{ijk} is predicted rating value for all K ratings in the test dataset [28].

VII. RESULTS AND DISCUSSION

- Recommendations generated with factorization based approach without considering any context

TABLE II. Results without considering any context

Features	RMSE	Time Elapsed(s)
3	1.99	74.8
4	1.98	80.5
5	1.95	88.4

The above TABLE- II discusses the output of factorization based Recommendation system when no context is considered. The RMSE value decreases as the number of features increases, but the time required to perform the execution is increased.

- Recommendations generated with Tensor factorization based approach with considering context

TABLE III. Results considering context without relevancy assessment

Features	Contexts	RMSE	Time Elapsed (s)
3	3	2.11	80.4
	4	1.95	82.2
	5	1.89	91.1
4	3	1.98	84.3
	4	1.95	86.6
	5	1.83	111.2
5	3	1.91	100.4
	4	1.88	106.1
	5	1.82	156.4

The TABLE- III above, discusses the results of the Tensor Factorization based Context aware recommendation system when contexts are considered but no method for finding context relevancy is applied.

The results clearly show the improvement in RMSE value when compared to the previous no context case. Moreover, though the time required for the computation is increased as the number of features and the numbers of contexts are increased from 3 to 5, the performance of recommender system is further elevated.

- Recommendations generated with Tensor factorization based approach with considering contextual relevancy while selecting the context.

TABLE IV. Results considering context with relevancy assessment

Features	Contexts	RMSE	Time Elapsed (s)
3	3	1.83	79.4
	4	1.78	87.2
	5	1.74	94.3
4	3	1.86	89.3
	4	1.73	91.6
	5	1.68	111.2
5	3	1.72	104.3
	4	1.64	110.6
	5	1.48	164.7

The TABLE- IV above, discusses the results of the Tensor Factorization based Context aware recommendation system when contexts are considered in the order of their context relevancy as mentioned in TABLE- I.

The results clearly show the improvement in RMSE value when compared to the previous no context case as well as the case considering contexts without any contextual relevancy technique. Moreover, as the number of features as well as number of contexts is increased from 3 to 5 the performance of recommender system is further elevated. However, the time elapsed will also be higher than the earlier discussed cases.

Following graphs are obtained when we compare the results of all three approaches i.e. No context, context without applying context relevancy technique and applying context with context relevancy technique

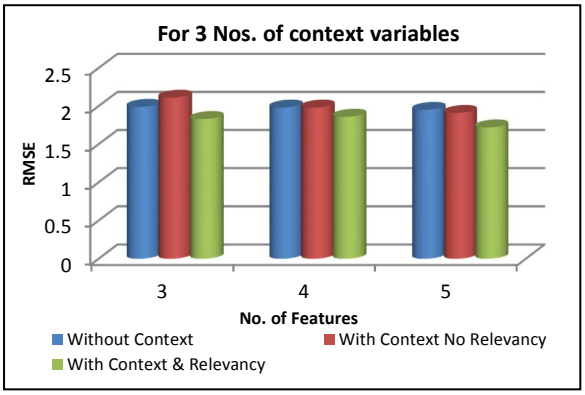


Fig. 1. Considering 3 context variables

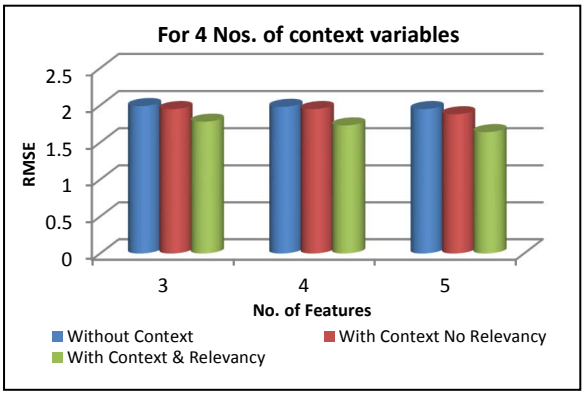


Fig. 2. Considering 4 context variables

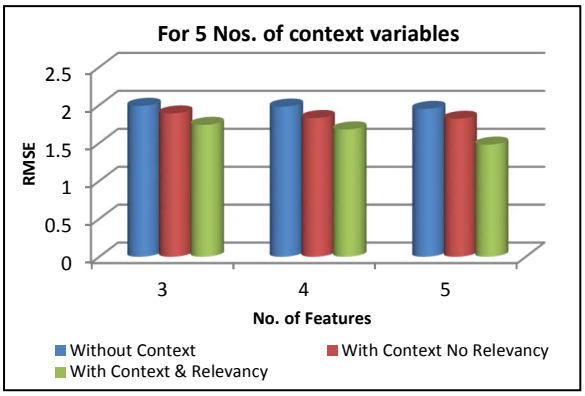


Fig. 3. Considering 5 context variables

The graphs also show clear improvement in the performance form no context case to context without context relevancy mechanism to the case which considers context after applying the context relevancy mechanism. Also, it is observed that as the number of features used in factorization increase the RMSE values show drop which means that the performance of the system is getting elevated.

VIII. Conclusion

Matrix Factorization is most widely used approach in the model based Collaborative filtering to handle the data sparsity issue. It adopts the dimensionality reduction technique and considers the data sparsity problem as Matrix Completion Problem. To add contextual dimension and extent the factorization model to N-dimensions, the Tensor Factorization based approach is adopted in the proposed work. HOSVD based tensor decomposition which is the technique under Tucker Tensor Decomposition is used as it calculates point wise loss function and utilizes squared loss function to calculate the overall loss which is most suitable for large real world datasets. Incorporating meaningful and relevant contexts always alleviate the performance of RS. This is proved by comparing the Root Mean Squared Error (RMSE) for three cases viz. no context case, incorporating context without any context relevancy mechanism and considering the context after applying context relevancy mechanism. Experimentation is conducted on LDOS-CoMoDa dataset which contains 12 contextual variables. Results clearly show gradual improvement in the performance over the above mentioned three cases.

As a future scope, the fusion based context relevancy assessment techniques shall be experimented in order to evaluate the improvement in the performance gain. Other tensor factorization algorithms can also be evaluated to study performance as well as computation overload impact. The performance relevancy of the model in the present scope shall also be verified against other available datasets such as Yahoo and Yelp in order to generalize the model applicability.

References

[1] F. Ricci, L. Rokach, and B. Shapira, “Recommender Systems Handbook”, Springer, New York, 2015

[2] Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez, "Recommender systems survey.", Knowledge-based systems 46, pp.109-132.2013

[3] P. Vijayakumar, and V. Reddy, “A survey on recommender systems (RSs) and its applications”, Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no. 8, pp. 2320-9798, Aug 2014

[4] Koren, Yehuda, Robert Bell, and Chris Volinsky, "Matrix factorization techniques for recommender systems.", Computer8, pp. 30-37,2009.

[5] X. Luo, Y. Xia, Q. Zhu, “Incremental collaborative filtering recommender based on regularized matrix factorization”, Knowledge-Based Systems 27, pp.271–280,2012.

[6] G. Adomavicius, and A. Tuzhilin, “Context-aware recommender systems”, Recommender systems handbook, Springer, pp. 217-253, 2011

[7] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, “Incorporating contextual information in recommender systems using a multidimensional approach”, ACM Transactions on Information Systems, vol. 23, no. 1, pp. 103–145, Jan 2005

[8] Ahn, Hynuchul, Kyoung-jae Kim, and Ingoo Han, "Mobile advertisement recommender system using collaborative filtering: MAR-CF.", KGSF-Conference, vol. 2006, pp. 709-715. The Korea Society of Management Information Systems, 2006.

[9] Lombardi, S., S. S. Anand, and M. Gorgoglione, "Context and customer behaviour in recommendation.”, Workshop on Context-Aware Recommender Systems (CARS 2009), 2009.

[10] Baltrunas, L., and Ricci, F.,“Context-Based Splitting of Item Ratings in Collaborative Filtering.”, Proceedings of the 2009 ACM Conference on Recommender Systems,pp. 245– 248,2009

[11] Panniello, Umberto, Alexander Tuzhilin, Michele Gorgoglione, CosimoPalmisano, and AntoPedone., "Experimental comparison of pre-vs. Post-filtering approaches in context-aware recommender systems.", Proceedings of the third ACM conference on Recommender systems, pp. 265-268, ACM, 2009.

[12] Adomavicius, Gediminas, and Alexander Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.", IEEE Transactions on Knowledge & Data Engineering, pp.734-749, 2005.

[13] Oku, Kenta, Shinsuke Nakajima, Jun Miyazaki, and ShunsukeUemura, "Context-aware SVM for context-dependent information recommendation." Proceedings of the 7th international Conference on Mobile Data Management, IEEE Computer Society, 2006.

[14] Karatzoglou, Alexandros, Xavier Amatriain, LinasBaltrunas, and Nuria Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering." Proceedings of the fourth ACM conference on Recommender systems, pp. 79-86. ACM, 2010

[15] Rabanser, Stephan, OleksandrShchur, and Stephan Günnemann, "Introduction to Tensor Decompositions and their Applications in Machine Learning.", arXiv preprint arXiv:1711.10781,2017

[16] Li, Jiyun, Rongyuan Yang, and Linlin Jiang, "DTCMF: Dynamic trust-based context-aware matrix factorization for collaborative filtering." Information Technology, Networking, Electronic and Automation Control Conference, pp. 914-919, IEEE, 2016.

[17] Odić, Ante, Marko Tkalčič, Jurij F. Tasič, and Andrej Košir, "Predicting and detecting the relevant contextual information in a movie-recommender system.", Interacting with Computers25, no. 1 pp.4-90,2013

Analyzing Smart Meter Data using a Two-stage Competitive Learning Method

Ankit Mahato
Dept. of Mechanical Engineering
IIT Kanpur
Kanpur, India
ankmahato@gmail.com

Ashita Prasad
PGP 2015-17
IIM Ahmedabad
Ahmedabad, India
ashitaprasad92@gmail.com

***Abstract*—Smart energy management is a major area of interest to meet the rising energy demand for which several countries are deploying smart meters. Presently, there is a need to better visualize the high-volume of data captured by smart meters to provide a means to effectively gather various analytical insights which can help in better understanding the energy usage patterns. This article presents a cascade application of two competitive learning algorithms - Self-organizing Map (SOM) and K-means clustering, to discover knowledge from smart meter data. A SOM is applied to construct a 2-D topologically preserving map which is useful in understanding and visualizing the consumer load profiles. Then K-means is applied on the codebook vectors of the SOM to determine the clusters containing consumers with similar energy consumption patterns. The identified consumer clusters enable the utility firms in preparing segment-specific tariffs to efficiently shape the future energy usage patterns.**

***Index Terms*—Smart Meter Analytics, Self-organizing Map, Clustering, Machine Learning**

I. Introduction

The use of smart systems has opened up new avenues for energy security and management, like the application of analytics to effectively manage the smart grid infrastructure. This infrastructure forms a key component of the IoT (Internet of Things) framework in which devices are inter-connected across various industries through the internet and peer-to-peer networks to provide valuable insights to both customers and service providers. It has been estimated that by 2020 the number of IoT devices will grow to 50 billion [1], [2]. IoT provides energy utility firms with unprecedented capabilities to efficiently manage the smart grids by forecasting future demands, reducing losses, shaping consumer usage patterns, meeting the energy demand distribution and increasing the overall efficiency of the smart grid [3]. Large scale deployment of smart meters is the key to building these efficient smart grids.

A smart meter is a device which registers the units of energy consumed at intervals of a specific period [3], [4]. It includes various features like communication, quantitative measurement, power management, calibration and control, synchronization and display. The data logged by a smart meter is communicated back to the data lake of the utility firm in real-time, where it is used for analyzing the energy usage behavior in order to predict and shape the future demand [5]. Smart meter adoption is on a rise since the past decade and its global market shows a promising growth of 400% from $4 billion (2011) to $20 billion (2018) [6]. Today, one of the major challenges in this industry is to visualize and analyze the growing amount of data that is being generated by these smart meters. It is important for the utility firms to understand the importance of knowledge discovery and data mining as the systems are smart only to an extent to which the data is being used to derive actionable insights. In the case of smart meters, knowledge discovery can lead to a good understanding of the energy consumption and load profiles which will allow a deeper micromanagement of the smart grid.

Competitive learning methods are a class of algorithms with non-linear computations, assisting in unsupervised learning. They are called competitive learning methods as for each input observation the processing elements (PEs) or nodes or centroids compete, and the winner PE is updated [7]. In the case of hard clustering methods, like generalized Lloyd's algorithm, k-means, etc., for each input there is strictly a single winning PE which adapts [7]. In the case of soft learning methods, like neural gas, self-organizing map, etc., the winning PE adapts along with the PEs in its neighborhood [7]. These methods are very useful in the knowledge discovery process and work towards achieving goals like error minimization, feature mapping, clustering, etc [7]. Smart meter data is a time series with high dimensionality which makes feature mapping an attractive goal for visualization of the entire data in a 2-D discrete map which is preserved topologically. Also, the goal of clustering is useful in partitioning the data into sub-groups of similar load profile.

K-means clustering is the most prominent method used for exploring energy profiles registered by smart meters for residential load patterns to design targeted incentive schemes and Time of Use based tariff [4], [8]–[10]. Self-organizing maps are also a great tool for data exploration where the input data is used to competitively learn and arrive at a topologically preserved map of code-book vectors [11], [12]. This projected low-dimensional (2D) grid can be utilized to visualize and explore the data effectively. Although researchers [13]–[16] have worked in the domain of cascading SOM with clustering in the past, the focus of the current article is to highlight

978-1-7281-2087-4/19/$31.00 © 2019 IEEE

the importance of cascading SOM and k-means clustering for the process of knowledge discovery in smart meter data via visualization and unsupervised learning.

II. Methodology

A two-stage competitive learning method is applied on the data where it is subjected to a SOM, followed by the K-means clustering performed on the obtained codebook vectors of the SOM. The algorithm steps are provided in the following subsections.

A. Self-organizing Map

SOM is an artificial neural network with a single computational layer where each neuron is connected to all the input nodes. The computational steps as given by [17]:

1) Initialization: Initial weights are assigned randomly to all codebook vectors.
2) Competition: All units compete and the one with minimum distance from the input data wins.
3) Cooperation: The winning unit excites its neighboring units.
4) Adaptation: The winning unit and its neighbors are adjusted towards the input with the neighbors having lesser adjustments with respect to the winning unit.
5) Iterate steps 2-4 until maximum iterations are reached.

B. K-means Clustering

Let the entire set of obtained codebook vectors of the SOM above be $X \subset R^d$ and $D_i(x)$ denote the distance of codebook vectors from the center c_i which is already chosen. K-means algorithm as given by [18]:

1) Pick k centers c_i, chosen randomly from X.
2) For each point in X, calculate $D_i(x)$.
3) Assign each point to cluster i where i is given by the minimum $D_i(x)$.
4) Calculate new cluster centers c_i.
5) Proceed with Step 2-4 until the within cluster sum-squares converges or a maximum number of iterations is reached.

III. Data & Implementation

To demonstrate the application of the suggested approach, the input data is simulated from the energy profiles provided in literature [19]–[21] along with commercial data. It contains time series data of around 3,000 smart meters with 96 energy consumption values (in kW) captured at every 15-minute interval beginning from 00:15 hrs to 24:00 hrs for a particular day.

The SOM grid size is determined as 16×17 (272) as given by Equation 1 [22]:

$$S = 5 \cdot \sqrt{N} \tag{1}$$

where S is the optimal size of the SOM and N is the total number of observations.

The R kohonen package [23] is used for SOM implementation. The codebook vectors are randomly initialized and the som function parameters are set as 16×17 hexagonal grids and 400 iterations. The number of iterations is determined based on the convergence of sum-of-squared errors to a stable value across multiple simulations.

The k-means algorithm [18] is implemented and used to cluster the codebook vectors into clusters of similar energy usage patterns. The number of clusters (k) is determined via an elbow plot and is set as 12 for the current analysis.

IV. Results

A. Self-organizing Map

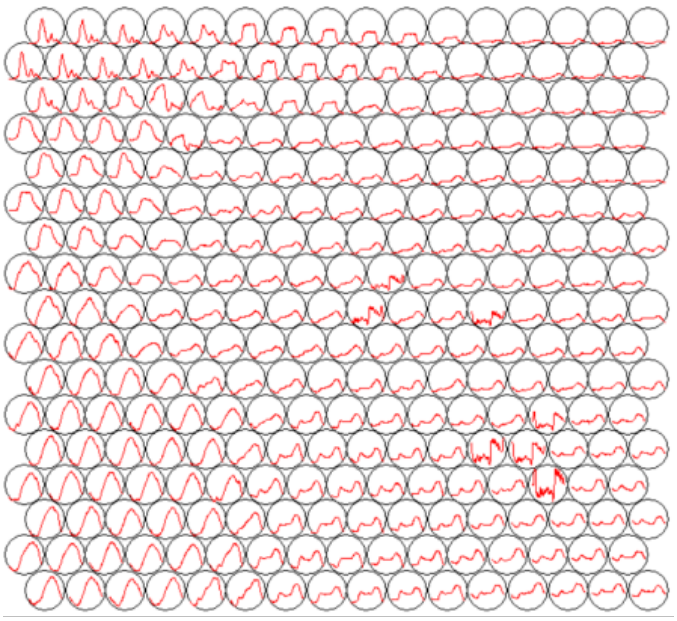


Fig. 1. Resulting SOM topology displaying the codebook vectors of all the units.

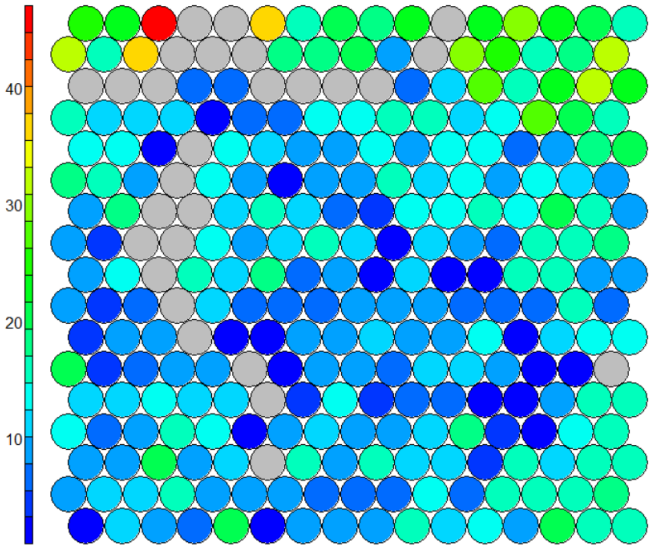


Fig. 2. Counts plot for the SOM.

The resulting codebook vectors plot of the 16-by-17 mapping is shown in Fig. 1. The codebook vector profile in each unit represents the corresponding energy consumption profile of the customers belonging to that unit.

The observation count for each unit is shown via a color plot in Fig. 2. Empty units are colored in gray. It can be observed that the smart meter readings are well distributed across the map with a median count of 11 observations per unit and an interquartile range of 15 – 6.75 = 8.25.

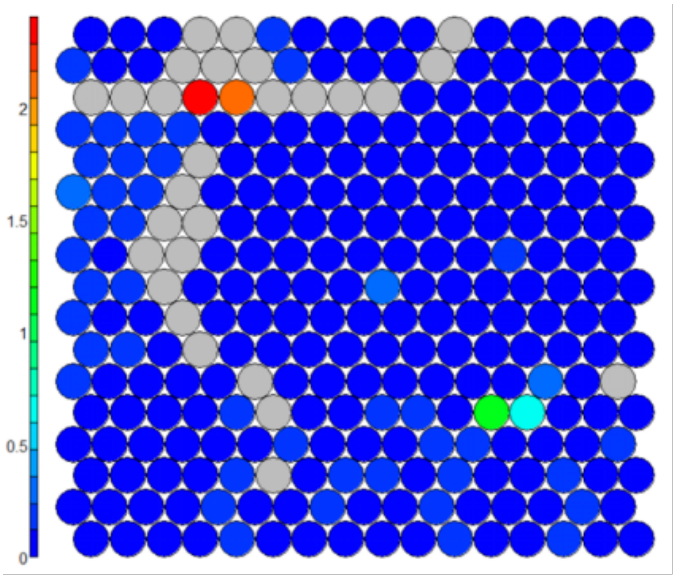


Fig. 3. Quality (mean distance of observations from the codebook vectors) of the SOM.

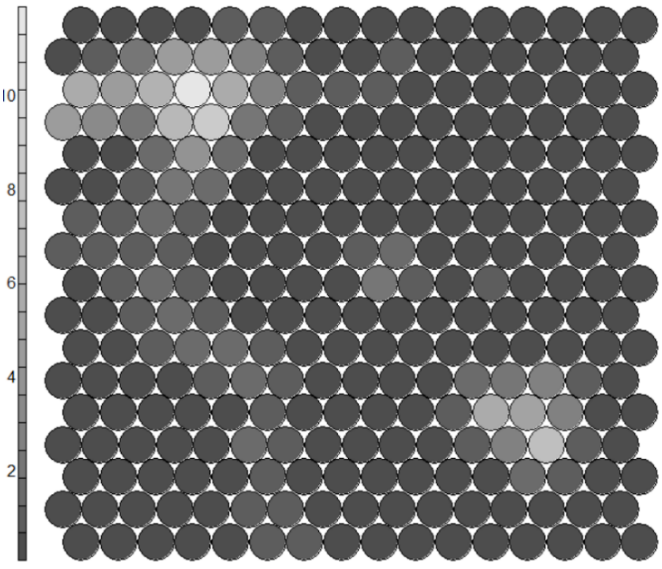


Fig. 4. U-matrix plot of the SOM.

Fig. 3 shows the mean distance of observations mapped to a unit from the codebook vector of that unit. This value is small for a major portion of the map, which demonstrates that the observations are better represented by the codebook vectors. In the plot, four units can be located having a mean distance greater than 0.5. At times smart meter data contains information on energy leaks and irregular load profiles. Using this plot we can identify such units containing (anomalous) observations which are not well represented and can be further investigated to detect energy leaks or theft.

The U-matrix plot of the SOM is presented in Fig. 4 using a grayscale color palette. The values are calculated as the sum of the distances to all immediate neighbors. As seen in the figure, there are units with higher average distances denoted by lighter color, whereas the codebook vectors of darker units are close to each other. This plot is useful in determining clusters in input data as the darker regions can be treated as clusters with the lighter areas denoting the presence of class (cluster) boundaries. When k-means is applied on the codebook vectors as described in the next section, it is observed that some of the cluster boundaries (in Fig. 5) lie in the lighter regions shown in Fig. 4.

B. K-means Clustering

As demonstrated in the previous section, SOM provides a 2-D visualization of the entire dataset with topology preservation. The resulting codebook vectors from SOM are clustered to identify similar consumer energy usage patterns and divide the map into regions having similar load profiles. Using an elbow plot, we identify the optimal number of clusters k = 12. Fig. 5 shows the 12 clusters obtained after applying K-means algorithm.

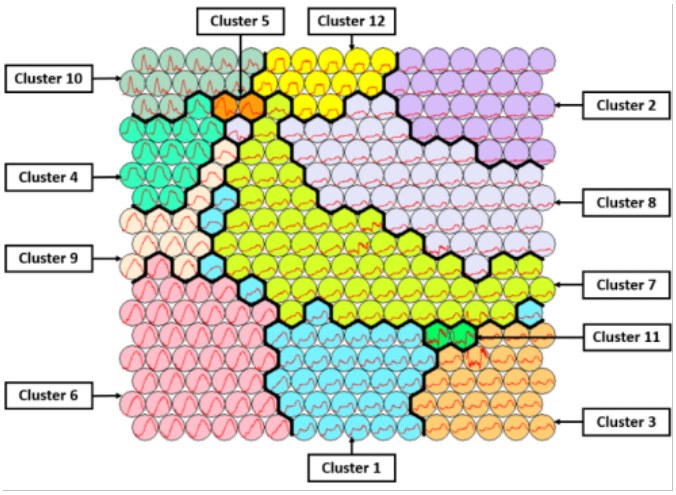


Fig. 5. Clusters obtained after applying K-means.

The centroids obtained for the 12 clusters are shown in Fig. 6. It reveals various energy usage patterns which can be used to derive insights and design incentive schemes for customers to mitigate peak loads and manage the energy demand efficiently. The energy consumption patterns are also identified for these clusters:

- Cluster 1 – Energy consumption increases in the morning, sustains midday and peaks in the late-evening.

- Cluster 4 – High morning to evening usage.
- Cluster 10 – Peak usage in the morning, evening and night with decreasing magnitude of the peaks.
- Cluster 12 – Moderate sustained morning to evening usage.

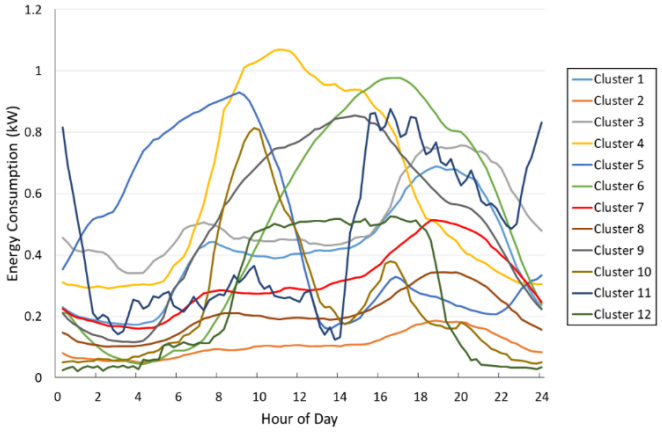


Fig. 6. Average energy load profiles of the obtained clusters.

V. Conclusion

The article demonstrates a novel usage of competitive learning algorithms – Self-organizing Maps and K-means as a two stage method for knowledge discovery from smart meter data. Using the proposed methodology, the data can now be visualized in the form of a meaningful 2D map which can be further analyzed to identify the various energy usage patterns useful for managing the smart grid. It is demonstrated how SOM can be used to present a topological map of the entire dataset along with insightful statistics which is useful in determining the quality of mapping and detecting anomalies in the input data. The application of k-means algorithm on the obtained SOM provides useful information regarding the energy usage patterns. Some of the clusters obtained are presented which demonstrate the effectiveness of the proposed method. The proposed knowledge discovery method will enable energy utility firms in better understanding their customers and their energy consumption behavior which will help in preparing segment-specific tariffs and campaigns to modify the usage patterns in order to distribute the peak loads. Thus, energy supply-demand in the grids will be met more efficiently making them smarter.

References

[1] D. Evans, "The internet of things: How the next evolution of the internet is changing everything," Whitepaper, CISCO IBSG, 2011.

[2] O. Monnier, "A smarter grid with the internet of things," Whitepaper, Texas Instruments, 2013.

[3] X. Liu and P.S. Nielsen, "A hybrid ICT-solution for smart meter data analytics," Energy, vol. 115, 2016, pp. 1710–1722.

[4] P. Arora, Deepali and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," Procedia Computer Science, vol. 78, 2016, pp. 507–512.

[5] S. Rastogi, M. Sharma and P. Varshney, "Internet of things based smart electricity meters," International Journal of Computer Applications, vol. 133(8), 2016, pp. 13–16.

[6] L. Alejandro, C. Blair, L. Bloodgood, M. Khan, M. Lawless, D. Meehan, P. Schneider and K. Tsuji, "Global market for smart electricity meters: Government policies driving strong growth," Technical Report, Office of Industries - U.S. International Trade Commission, 2014.

[7] B. Fritzke, "Some competitive learning methods," Institute for Neural Computation, Ruhr-Universität Bochum, 1997.

[8] A. Lavin and D. Klabjan, "Clustering time-series energy data from smart meters," Energy Efficiency, vol. 8(4), 2015, pp. 681—689.

[9] A. Al-Wakeel and J. Wu, "K-means based cluster analysis of residential smart meter measurements," Energy Procedia, vol. 88, 2016, pp. 754–760.

[10] A. Al-Wakeel, J. Wu and N. Jenkins, "K-means based load estimation of domestic smart meter measurements," Applied Energy, vol. 194, 2017, pp. 333–342.

[11] T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, vol. 43(1), 1982, pp. 59–69.

[12] J. Walter, H. Riter and K. Schulten, "Nonlinear prediction with self-organizing maps," IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 1990, pp. 589–594.

[13] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," IEEE Transactions on Neural Networks, vol. 11(3), 2000, pp. 586–600.

[14] J. Llanos, R. Morales, A. Núñez, D. Sáez, M. Lacalle, L.G. Marin, R. Hernandez and F. Lanas, "Load estimation for microgrid planning based on a self-organizing map methodology," Applied Soft Computing, vol. 53, 2017, pp. 323–335.

[15] X. Wang, K.A. Smith and R.J. Hyndman, "Dimension reduction for clustering time series using global characteristics," Computational Science - ICCS 2005: 5th International Conference, V.S. Sunderam, G.D. van Albada, P.M.A. Sloot and J. Dongarra Eds., Atlanta, GA, USA, May 22-25 2005, Proceedings, Part III, Springer, Heidelberg, 2005, pp. 792-795.

[16] L. Hernández, C. Baladrón, J.M. Aguiar, B. Carro and A. Sánchez-Esguevillas, "Classification and clustering of electricity demand patterns in industrial parks," Energies, vol. 5(12), 2012, pp. 5215–5228.

[17] A.A. Akinduko, E.M. Mirkes and A.N. Gorban, "SOM: Stochastic initialization versus principal components," Information Sciences, vol. 364, 2016, pp. 213–221.

[18] S.P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, 1982, pp. 129—137.

[19] B. McDonald, P. Pudney and J. Rong, "Pattern recognition and segmentation of smart meter data," Australian and New Zealand Industrial and Applied Mathematics Journal, vol. 54, 2014, pp. 105–150.

[20] J.A. Jardini, C.M.V. Tahan, M.R. Gouvea, S.U. Ahn and F.M. Figueiredo, "Daily load profiles for residential," Commercial and Industrial Low Voltage Consumers, IEEE Transactions on Power Delivery, vol. 15(1), 2000, pp. 375–380.

[21] D. Vercamer, B. Steurtewagen, D.V.D. Poel and F. Vermeulen, "Predicting consumer load profiles using commercial and open data," IEEE Transactions on Power Systems, vol. 31(5), 2016, pp. 3693–3701.

[22] A. Shalaginov and K. Franke, "A new method for an optimal SOM size determination in neuro-fuzzy for the digital forensics applications," Advances in Computational Intelligence, 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, June 10–12 2015.

[23] R. Wehrens and L.M.C. Buydens, "Self- and super-organizing maps in R: The kohonen package," Journal of Statistical Software, vol. 21(5), 2007.

From Light to Li-Fi : Research Challenges in Modulation, MIMO, Deployment Strategies and Handover

Sanket Salvi
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
sanketsalvi.salvi@gmail.com

Geetha V
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
geethav@nitk.edu.in

***Abstract*—With an increasing number of network-connected devices, there is a need for new and innovative ways of providing communication. Due to the over-usage of existing Radio Frequency(RF) spectrum, it is essential to explore other alternatives. A possible solution for reducing the load over RF spectrum for communication could be the usage of alternative 380 to 740 nanometers wavelength, i.e., Visible Light spectrum. It satisfies multiple objectives, such as provisioning of illumination and communication. Characteristics of channel and propagation medium have been studied to provide the required preliminary understanding of the problem domain. The primary research focuses on Li-Fi communication is in the area of modulation techniques and handover management. Various existing modulation techniques have been studied and compared with respect its performance parameters such as attainable bandwidth, spectral efficiency, noise ratio, and illumination. Different MIMO techniques, design requirement, deployment patterns, and Handover techniques have also been highlighted. This work provides open research challenges related to modulation techniques, MIMO, Handover, and Deployment Strategies for Li-Fi.**

***Index Terms*—Li-Fi, Visible Light Communication, Modulation Techniques, MIMO, Handover.**

I. Introduction

Recently, an increasing growth in number of internet connected devices has been observed. During same time, various new technologies and methods have emerged to provide faster communication and efficient utilization of available bandwidth. It is because light communication is faster than any other wired or wireless communication, a smaller division of Optical Wireless Communication (OWC) also know as Visible Light Communication (VLC), is being strongly considered for 5G technology. (Light Emitting Diode)LEDs which are traditionally used solely for illumination purpose can also be used for communication under VLC.

Thus increasing the spectral usage and providing opportunities to revisit and innovate various navigation, positioning, and data communication related applications under indoor and outdoor scenarios. As VLC is a form of wireless communication, its underlying backbone architecture remains the same with significant modifications at the access points and receiver designs and working. VLC systems are a mostly simplex form of communication which is typically used for broadcasting information. It is because, traditionally, a light fixture is mounted over ceiling which works as transmitters and receivers are non-illuminating devices under the ceiling facing towards the transmitter. However, bidirectional communication can be accomplished by using Visible Light Communication for data downlink and Infra-Red (IR) Communication for data uplink. This type of system is called as "Li-Fi," a term coined by Prof. Herald Haas for Light Fidelity[1].

Fig.1 shows the underlying architecture of a Li-Fi system which consists of backbone internet, LED driver, IR Reciever, LED Lamp, Photo-sensor, IR transmitter, Amplification and Processing Unit, and any connecting device such as laptop, desktop or mobile. Here, Light fixture is capable of sending and receiving data using LEDs and IR, respectively. Such a light fixture is called Li-Fi Access Point. An LED driver is used to convert back-end network data to corresponding light pulses by incorporating suitable modulation and encoding scheme. Photo-diodes, Image sensors, and/or Cameras are used as receivers with respective demodulation and decoding scheme to reproduce data.

As existing LEDs are commonly used only for illumination in order to support communication, it has to overcome several research challenges. These challenges majorly revolve around the adoption of existing RF modulation schemes for Li-Fi limited intensity modulated / direct detection (IM/DD) scheme while maintaining illumination standards. This paper explores and provides a concise understanding of the variety of work carried out in the area of Li-Fi, highlighting the issues and challenges, modulation schemes, design methods, and handover techniques.

The remaining part of the paper is arranged as follows: Section II, provides a broad insight over the various challenges in Li-Fi communication. Classification and performance with respect to basic parameters of various modulation techniques is discussed in Section III by considering factors such as illumination requirements, complexity of implementation, spectral efficiency, noise and attainable data rates. Various MIMO enable Li-Fi receiver designs, its limitations, and techniques

978-1-7281-2087- 4/19/$31.00

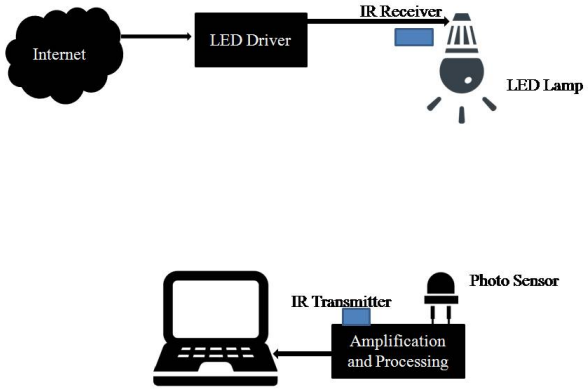


Figure 1. Basic LiFi Architecture

for achieving MIMO in Li-Fi is discussed under section IV. Section V highlight effects of inter-cell interference with respect to Li-Fi Access Point deployment patterns. Methods for achieving handover, its effects and factors to be considered are discussed under section VI. Finally, in conclusion the paper provides prospective open research challenges under Section VII.

II. Challenges in Building Li-Fi System

LEDs are the most suitable candidates for implementation of Li-Fi system mainly due to its fast response time, long operational lifetime, and low cost[2]. However, it also imposes few limitations. This section provides brief insight on various issues and challenges faced while building a Li-Fi system.

- **Transmitter (Tx) and Modulation Technique:** The current commercial Li-Fi communication[3][4] can offer maximum working distance range of 1 to 50 meters. However, since Li-Fi is inherently Line-of-Sight(LOS) communication, this distance is significantly less compared to Radio Frequency counterpart. Also, the achievable data rates by using Li-Fi is limited to a few Mbps. However, active research is going on in this field to improve data rates by using different materials and techniques on LED and Photo-sensors. By using a modulation scheme which has high Spectral Efficiency(SE) and bandwidth, high-speed communication can be achieved even in Li-Fi[5]. By using tri-chromatic LEDs instead of phosphorescent LEDs has shown improvement in throughput by a factor of three[6]. Fig.2, shows basic building blocks of Wavelength Division Multiplexing(WDM) which provides better spectral efficiency by using multi-color LEDs over phosphorescent LEDs. It consists of Modulator, Signal Generator, LEDs, color filters, Photo-diodes, Oscilloscope, and Demodulator. In this system, the symbol to be transmitted is be encoded and modulated using modulator, which is then mapped to a specific colored LED. Using a signal generator specific LED will send data. This allows simultaneous operation of three transmitters allowing three channels for sending data. On the receiver side, by using chromatic filter light intensity across each channel will be gathered, and data is obtained after performing the corresponding demodulation.

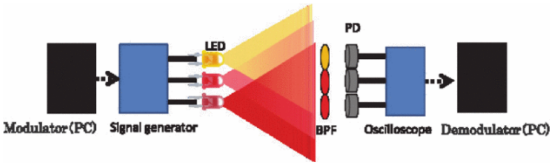


Figure 2. OFDM Based WDM-MIMO VLC System[5]

- **Optical Wireless Indoor Communication Channel:** The Li-Fi communication system is restricted to LOS type of communication, and any misalignment may degrade the performance drastically. However, the shadowing effect caused by blocking direct ray path between transmitter and photo-sensor can be used for reconstruction of data using non-line-of-sight(NLOS) communication channels. In[7], Channel Impulse Response(CIR) is obtained by studying delays between direct/reflected rays, which is affected by wall pain material, furniture, and size of the room. Obtained CIR can be used for re-calibrating the receiver to work in NLOS communication channel.
- **Receiver Device (Rx) and its Properties:** The pattern of change in light intensities is detected using photo-sensors at the receiver side. These photo-sensors could be image sensors or photo-diodes. Photo-diode converts detected light intensity into photo-current. By comparing the performance of various photodiodes, it was found that avalanche photodiodes can be used for Li-Fi communication due to its smaller size, low cost, and faster response time[8]. However, due to interference from sunlight and other light fixture, the performance of Li-Fi system may get compromised. Thus, by using specific filters and selecting proper receiver design, the effects of interference can be reduced significantly. It was observed that photo-diodes for stationary receivers and image sensor for mobile receivers are well suited in Li-Fi Communication system[9]. Image sensors can provide faster data rates by exploiting the rolling shutter effect. However, Li-Fi using image sensors is resource inefficient because it requires more complex computation and power compared to photo-diode. Thus, a suitable trade-off between the speed and complexity should be considered while using photo-diodes or Camera[10].
- **MIMO optical wireless communications:** Applying traditional RF MIMO techniques for Li-Fi has major challenge due to design limitation of narrow beam-width at receivers.It degrades the communication quality even with slight misalignment.Therefore, efficient receivers and intelligent transmitter deployment based modulation schemes to support MIMO is potential research area.[11].

- **Cross-layer load balancing:** As the network traffic has to be converted either from Li-Fi to RF or vice-versa suitable load balancing should be created. A central unit design of load balancing by using the two-tier buffer structure requires the optimization of cross-layer (MAC layer and Physical Layer) parameters. Efficient cross-layer load balancing will improve the performance in terms of data rates[12].
- **User movement modelling:** While designing Li-Fi network, user movement and device orientation should be considered and modelled to study and provide seamless connectivity. The random way point model is popularly used to simulate user mobility, however, it is impractical in real scenarios. Users in a shopping mall[13] definitely have different moving performance from those in the office scenarios. In addition, the random orientations of Li-Fi receiver will affect the user data rate, which should be carefully modelled[14]. In [15][16], performance metrics such as Bit-Error-Ratio (BER) and Signal-to-Noise Ratio (SNR) is evaluated based on characterization of device orientation. An orientation-based random waypoint (ORWP)[17] mobility model was specifically modelled for Li-Fi mobile devices and it is assessed based on the handover rate under realistic scenarios.
- **Illumination Requirements:** An ideal Li-Fi system must have provision for supporting various dimming levels. There two techniques which can be used to achieve dimming in indoor Li-Fi systems either by Continuous Current Reduction (CCR) or by Pulse Width Modulation (PWM)[18]. In CCR, the ON/OFF levels of the LED are redefined based on the selected dimming level and maintain the same data rate. However, it was studied that it not only affects the reliability of communication at low dimming levels but also color rendering property of LED becomes non-linear and unpredictable under continuous low power[19]. In PWM based dimming, additional ON/OFF pulses are inserted according to desired dimming level. More OFF pulses provide dimmer light and more ON pulses provide brighter light. It is considered that the frequency of combined pulses for transmission of a symbol is higher than Flicker Free Frequency. As studied in paper[20][21] 50% dimming level provides maximum data rate, further decreasing or increasing the brightness of the LED results into decrease in data rate. Thus, it shows that communication efficiency with PWM based dimming in Li-Fi is a triangular function with peak performance at 50% brightness. It also suggests that the modulation scheme to be used should have an equal distribution of bits to avoid flickering.

III. Modulation Techniques

At the core of any communication, technique lays exploitation of different properties of the concerned medium. These properties are modulated and demodulated to send information across the channel. This section discusses modulation techniques used in Li-Fi, some of which are adapted from its RF counterparts and some unique to its nature. The modulation techniques under Li-Fi are classified in Single Carrier Modulation(SCM), Multi-Carrier Modulation(MCM) and Li-Fi Specific Color based Modulation.

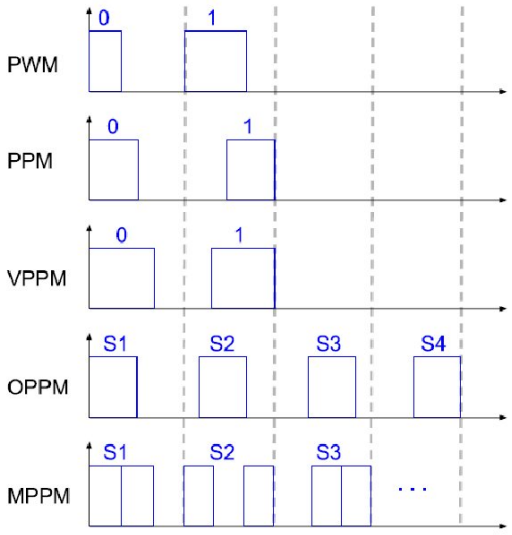


Figure 3. Various Pulse Position Modulation Techniques[25]

A. Single Carrier Modulation(SCM)

In Single Carrier Modulation the information in transmitted using a single channel. Widely used SCM schemes for Li-Fi include on-off keying (OOK), pulse position modulation (PPM) and pulse amplitude modulation (PAM).

1) On-Off Keying (OOK): In OOK, for transmitting data bits, the LED is switched ON and OFF for 1 and 0 respectively. In some variations of OOK, the LED is dimmed instead of completely turning it OFF. Generally, this level of detectable reduced light intensity is determined by the sensitivity of the photo-sensor. Thus, if sensitivity is high, the reduced light intensity can be closer to the full brightness of the LED. It also gives a better wider range of PAPR and thus supports various dimming levels. Although its simple and easy to implement the major issue is with the limited bandwidth as the response time of White LED is slow.

The proposed Non-Return-to-Zero OOK (NRZ-OOK)[22] demonstrated 10Mbps VLC link using White LEDs and P-I-N Photodiode. Based on the same premise, further improvements in performance was achieved by using a blue filter and a combination of a blue filter with analog equalization[23] to gain bandwidth of 40 Mbps and 125 Mbps respectively. However, by using Avalanche Photodiode at receiver better performance in terms of bandwidth, i.e., 230 Mbps was achieved. It was due to high sensitivity and fast response time offered by avalanche photodiode [8]. It was also observed that by using RGB LED for modulation at transmitter and Avalanche Photodiode at receiver bandwidth of up to 477 Mbps is achievable[24].

2) Pulse Position Modulation (PPM): Fig. 3 represents various Pulse Modulation techniques. In PPM, symbol duration is kept constant and divided into a certain number of slots. The symbol is decoded depending on the presence of a pulse in a particular slot. Although PPM is more power efficient compared to OOK, the bandwidth consumed is higher to provide the same data rates. Thus, due to limitations of lower data rates and spectral efficiency, several other modifications have been suggested over time. The proposed Differential PPM (DPPM)[26] provides improved power and SE gains; however, due to the unequal distribution of power for different symbols, considerable flickering was observed affecting the illumination performance. An improvement over DPPM was proposed by Sevincer, Bhattarai, Bilgi, *et al.* [27] by adding a layer of bit encoding to ascertain even distribution of duty cycle across various symbols. Although it solves the flickering problem, the attainable data rate was lesser compared to DPPM.

An alternative method to implement Variable PPM (VPPM)[28] showed better data rate and support for dimming by using PWM methods for dimming. Due to these advantages of VPPM, it is considered as one of the standard modulations for VLC. In [25] Multiple PPM was compared with VPPM to observe that MPPM achieves better spectral efficiency in comparison with VPPM. Other PPM methods such as Overlapping PPM(OPPM)[30] and Multipulse PPM(M-PPM)[31] allowed transmission of symbols at higher data rate due to the inclusion of multiple pulse levels and a group of pulses respectively. The combination of best of OPPM and MPPM was also considered called Overlapped Multipulse PPM[32] which performed better compared MPPM in terms of spectral efficiency under the specific condition of fewer pulse slots and more pulse per symbol duration. Another particular modulation scheme named Trellis-Coded OMPPM [33] was explicitly used to combat error rate.

If the symbol is transmitted using the amplitude of the pulse, it is called Pulse Amplitude Modulation (PAM) which provide higher spectral efficiency. A combination of PAM and VPPM was proposed by Yi and Lee[34], which on comparison with other dimmable modulation schemes like OPPM, VPPM, and RZ-OOK was found to provide better bandwidth efficiency at the cost of more power. Another variation proposed by Zeng, Chen, Zhao, *et al.* [35] provided and more power efficient model. Under controlled environment proposed in [36] and [37] 40Gbps and 56Gbps data, the rate was achieved by using M-ary variation of PAM[38]. It was also studied that by converting Multicarrier OFDM to SCM by using DFT and IFFT, [39][40] nonlinearity tolerance is improved at the receiver.

B. Multi-carrier Modulation (MCM)

SCM techniques requires intricate equalization methods at faster data rates[41][23]. Whereas, on the other hand, MCM techniques such as OFDM can be applied in IM/DD domain by adding computationally efficient single tap equalizer. Here, OFDM transmitter first modulates incoming bits into modulation format such as M-QAM. Fig .4, shows 8-QAM Signal

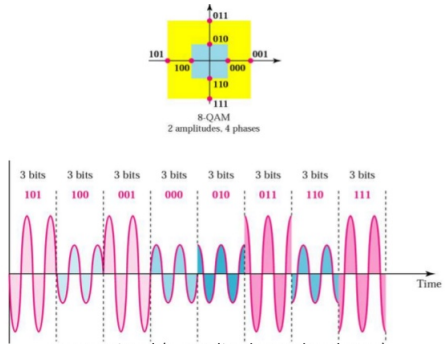


Figure 4. 8-QAM Signal 4 Phases 2 Amplitudes[42]

with 2 Amplitudes and 4 Phases. The M-QAM[42] modulated bits are then loaded over orthogonal sub-carriers and by using Inverse Fast Fourier Transform multiple symbols are then multiplexed into the time domain. However, as the output of OFDM is bipolar and complex, it cannot be directly applied to the IM/DD system. Hence, these signals are first converted into unipolar real values by using Hermitian Symmetry property. Depending on methods used for exploiting this property few variants with its advantages and disadvantages are discussed as follows,

1) DC-biased Optical OFDM (DCO-OFDM): An addition of DC bias to the real part of the bipolar OFDM signal offers better BER and SE[43]. However, high PAPR of OFDM Signal results in clipping distortion leading to undesirable electrical and optical spikes.

2) Asymmetrically Clipped Optical OFDM (ACO-OFDM): In ACO-OFDM[45], even sub-carriers of the original OFDM frame is skipped and information is added only to the odd sub-carrier. This creates a symmetry allowing negative sample to pass through without any distortion. However, this reduces the available bandwidth by half as compared to DCO-OFDM.

3) Asymmetrically Clipped DC biased Optical OFDM (ADO-OFDM): Combination of best of ACO-OFDM and DCO-OFDM is considered in ADO-OFDM[46]. It applies DCO-OFDM on the even sub-carriers and odd sub-carriers are modulated by using ACO-OFDM. This provides utilization of even as well as odd sub-carriers for data transmission. ADO-OFDM provides better power efficiency compared to DCO-OFDM and better bandwidth compared to ACO-OFDM at the cost of increased computational complexity.

4) Pulse Amplitude Modulation-Discrete Multitone (PAM-DMT): PAM-DMT[47] is similar to ACO-OFDM in terms of transmission of the positive part of DMT and asymmetric clipping at zero. However, it provides better spectral efficiency as it uses all the sub-carriers of OFDM. It was also observed that the detection performance at the receiver could be improved by exploiting the nonlinear distortions of PAM-DMT[48]. Islim, Tsonev, and Haas[49] proposed a modulation technique which overlaps multiple unipolar streams of PAM-DMT. In Hybrid ACO-OFDM[50], PAM-DMT is used for transmission of data on the even sub-carriers and ACO-OFDM is transmitted on the odd sub-carriers. In PAM based hybrid optical OFDM(PHO-

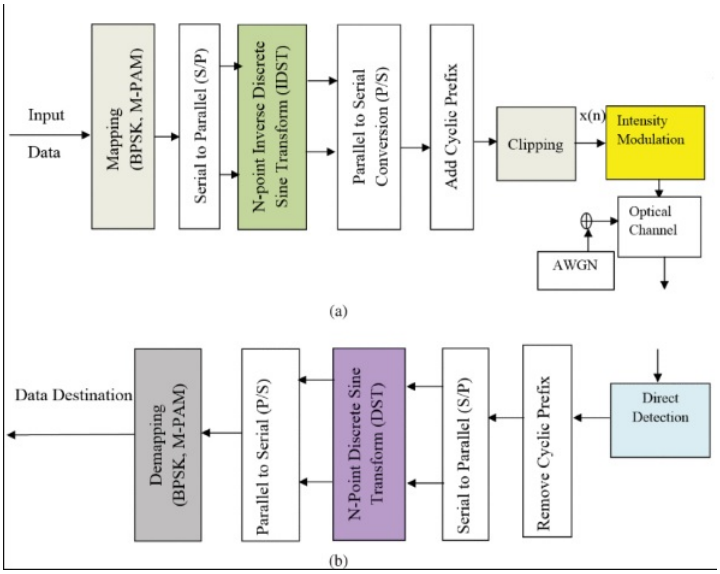


Figure 5. DST-based DCO-OFDM/ACO-OFDM system model for VLC. (a) Transmitter of DST-DCO/ACO-OFDM system for VLC. (b) Receiver of DST-DCO/ACO-OFDM system for VLC.[44]

OFDM)[51] a hybrid high order QAM is used to replace 1-dimensional PAM in order to compensate for data capacity of PAM-DMT providing better BER at the cost of reduced PAPR.

In[52] a hybrid OFDM-PTM (pulse time modulation) scheme is proposed, where a bipolar Optical OFDM signal is converted into digital PCM formats(PWM, PPM, Digital Pulse Interval Modulation(DPIM)). It is done by generating a PWM signal of varying width or PPM signal corresponding to input discrete OFDM sample. It showed improved BER compared to ACO-OFDM. In Reed Solomon-OFDM[53] after encoding data using RS codeword, the redundant part which crosses the clipping range is filtered and the redundancy left is used at the receiver to recreate data. In Layered ACO-OFDM[54], multiple ACO-OFDM signals are mapped to different layers of time divided input data signal. It was observed that LACO-OFDM gives better performance in terms of SNR for two-layer compared to more layers.

Hybrid DC-biased asymmetrically-clipped PAM OFDM (HDAP-OFDM)[55] and discrete sine transform (DST)-based OFDM[44] were proposed to achieve power efficiency. HDAP-OFDM is a combination of three OFDM formats (DCO-OFDM, ACO-OFDM, PAM-DMT).It uses higher order sub-carriers to carry ACO-OFDM on the odd-index, PAM-DMT on the even and DCO-OFDM on remaining lower order sub-carriers. DST-OFDM provides computationally efficient signal transmission as real-valued signals are achieved without the need to comply with Hermitian symmetry criteria. Fig. 5, shows the basic architecture of DST based OFDM modulation. It can be used with the DCO or ACO OFDM technique.

5) Unipolar OFDM: The Unipolar OFDM (U-OFDM)[56][57] applies Hermitian symmetry over input signal of M-QAM symbols. Bipolar signal thus obtained is unfolded into two time-domain frames. The first frame represents positive values, and the second represents the negative value from original bipolar signal. DC biasing is not needed as the resultant signal consists of only non-negative values. At the receiver side, subtraction of second frame from the first one results into reconstruction of original bipolar signal. However, since it uses two frames for conveying the same amount of data as in conventional DCO-OFDM, its bandwidth is half as that of DCO-OFDM.

In an improvement over U-OFDM, a Variable Pulse Width-UOFDM[58] was proposed. It uses variable width pulses for conveying magnitude and phase information which are adjusted based on clipping distortion and noise. Fig.6(a) and

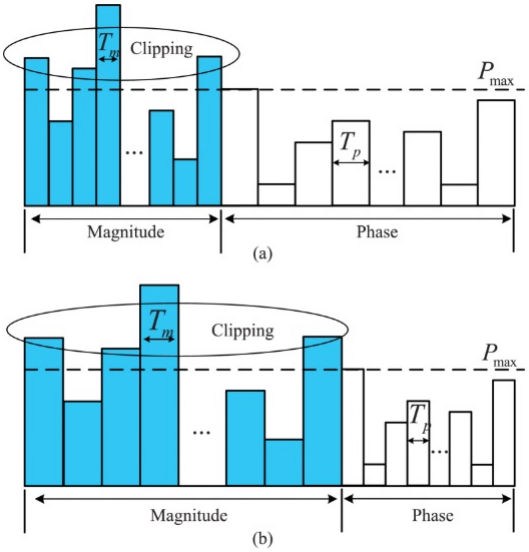


Figure 6. An illustration of VPW-OFDM[58]

6(b), shows examples of two different PW signals. The width of magnitude component of pulse in 6(b) is wider than that

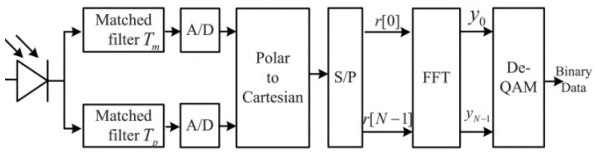


Figure 7. A block diagram of VPW-OFDM Receiver[58]

in 6a). The OFDM symbol duration can be calculated as $Ts{=}N(Tm + Tp)$. In Fig. 6 (a), the signals exceeding the maximum transmitted power must be hard clipped for the magnitude part. However, as the phase part is normalized and limited between 0 to $Pmax$, there is no clipping distortion while transmitting the phase component. The received optical signal is converted into electrical signal and matched filters with pulse widths Tm and Tp are used to detect the magnitude and phase components respectively as shown in Fig.7. After sampling and conversion from polar to cartesian, serial data is converted to parallel and passed to FFT followed by M-QAM decoder. However, the required bandwidth is more, and it underperforms severely in low bandwidth conditions.

In another U-OFDM based modulation technique, the clipped information due to peak power constraint is transmitted via extra time slots. This Clipping-enhanced Optical OFDM[59] achieves better SNR at the cost of bandwidth. In an improvement over CEO-OFDM was proposed in the form of L-slot CEO-OFDM[60] as illustrated in Fig. 8 where $Pmax$ is peak transmitted power constraint. Fig. 8 (a) shows an illustration of a bipolar real OFDM signal which uses Hermitian symmetry. Fig. 8 (b) shows one-slot CEO-OFDM signal, in which positive, negative, and its respective clipped parts are transmitted in time slots 1,2 and three, respectively. It reduces clipping distortion while retaining important information. However, for higher modulation index, the third time slot can also experience clipping. Thus by generalizing the idea for L-slots, the effects of clipping distortion can reduce by delaying the slot containing clipped signals. As shown in Fig. 8 (c) where is L is several slots. It was observed that L-slot CEO-OFDM techniques provide better BER at higher data rates and better support for dimming compared to ACO-OFDM and DCO-OFDM.

C. Other Multicarrier Multiplexing

OFDM uses FFT based transformation for providing Multi-Carrier Multiplexing. Few other transformations such as Hadamard coded modulation (HCM), wavelet packet division multiplexing (WPDM), Discrete Hartley Transformation (DHT) and Spatial Modulation(SM) are also considered for Li-Fi channels.

1) Hadamard Coded Modulation: In HCM[61] fast Walsh-Hadamard transformation (FWHT) is used as an alternative to traditional FFT used in OFDM. Although HCM is reported to achieve better performance gains at higher illumination levels compared to ACO and DCO-OFDM, an alternative DC

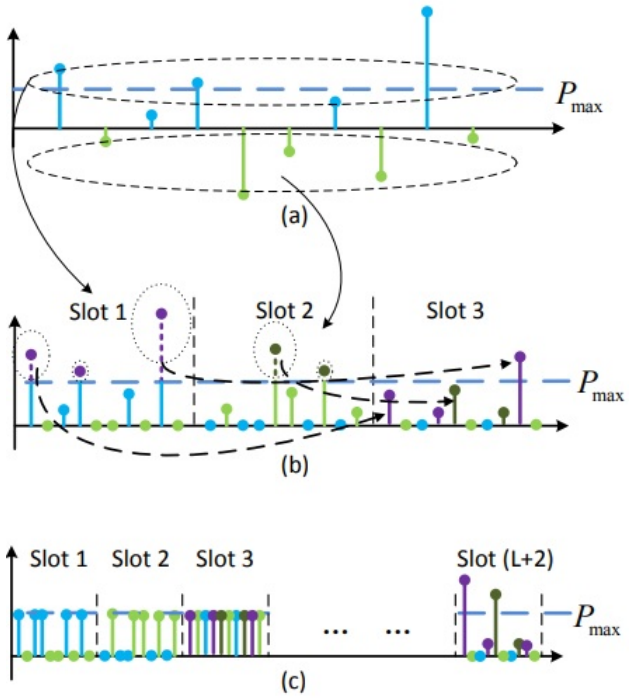


Figure 8. Principle of multi-slot CEO-OFDM. (a) bipolar OFDM signal, (b) One-slot CEO-OFDM signal, (c) L-CEO-OFDM signal[60]

reduced HCM(DCR-HCM)[62] was proposed to reduce the power consumption and support dimming.

2) Wavelet Packet Division Multiplexing(WPDM): WPDM[63] transmits multiple signals which are encoded into a waveform using wavelet packet basis functions. It provides channel capacity improvement compared to FDM and TDM as it allows time and frequency overlapped signals. Also, it maintains orthogonality of the resultant wave, which can be separated by using correlator at receiver. However, clipping distortion is observed [64].

3) Discrete Hartly Transform: DHT is one of the Fourier related transforms, which is Discrete version of Hartley Transform. It produces real-valued output for real-valued input, unlike FFT. By applying this transform, a multicarrier IM/DD system was proposed[65]. To achieve unipolar output, Asymmetrical Clipping and DC-biasing can be applied. As it does not require Hermitian Symmetry to be satisfied, it is computationally less intensive and improves SE compared to ACO-OFDM. However, since it accepts only real-valued input like M-PAM, the SE is lower than DCO-OFDM.

4) Spatial Modulation: In traditional Spatial Modulation (SM)[66] only one light fixture will be transmitting symbols using any underlying modulation scheme at a given instance, while others will provide regular illumination. The selection of fixture for transmission would be based on the symbol to be transmitted. At the receiver, side depending on the direction and received data, the symbol will be decoded. In the Spatially Modulated version of LACO-OFDM[67] generalized spatial

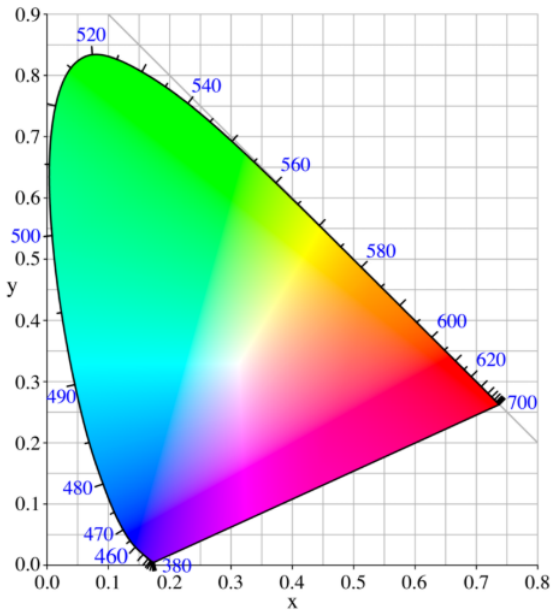


Figure 9. International Commission on Illumination CIE 1931[71]

modulation (GenSM) technology is combined with LACO-OFDM. In [68][69][70] a generalized LED Index Modulation(IM) for OFDM-based VLC systems is discussed with focus on minimizing losses incurred due to time and frequency shaping of OFDM. It integrates location information of single or group(constellation) of LEDs along with spatial modulation to transmit data to single or multiple users.

D. Color Domain Modulation

Color is one of the unique characteristics of the light which can be exploited for achieving extra dimension while bit encoding, thus, making color domain modulation techniques specific to Optical Wireless Communication. This section highlights a class of modulation techniques based on chromatic characteristics of LED.

1) Color Shift Keying: In Li-Fi modulation frequency does not represent carrier frequency, since carrier frequency is dependent on the LED. All previously mentioned modulation methods were baseband, which makes it difficult to modulate the carrier frequency of LEDs. However, the limitation of baseband communication can be overcome by using the color property of light. In fig. 9 x and y-axis determine chromaticity of the color. Chromaticity is an objective specification of the quality of a color regardless of its luminance. Multiple combinations of colors can be used to keep the overall intensity of output color constant. As shown in fig. 10, each color LED transmits the data, and it is controlled independently depending on the symbol to be transmitted. On the receiver side, chromatic filters are used before photodiode to detect color, and the respective signal is demodulated to set of bits[72]. However, deciding the constellation for the symbol to be transmitted requires optimization in which the distance between symbols should be maximum with minimum inter-symbol interference. This problem was discussed in [73][74] and possible solution in terms of use of quad-LED was proposed, which allows simple symbol mapping due to quadrilateral constellation shape. Also, in order to remove the limitation imposed by amplitude dimming, Okumura, Kozawa, Umeda, *et al.* [75] proposed PWM based modulation, which ensures reduced brightness without a change in color.

2) Color Intensity Modulation (CIM): In another color based Modulation intensity of the color is used as a modulating parameter. CIM[76] was proposed to overcome the complexity of CSK to achieve dimming for efficient illumination. A hybrid of CSK and CIM called CISK[77] was proposed which uses CSK and Non-Zero Level-PAM to achieve power efficiency with freedom of varied brightness. By maintaining a constant average transmit power, it relaxes the requirement of constant power constraint for CSK which fulfills the illumination requirement of non-flickering.

However, due to the cost and complexity involved in Color Domain Modulation concerning the separation of the colors, there is a lack of specific design standard for receivers compared to OFDM. Hence, this area still has sufficient scope for designing efficient modulation techniques, which are computationally less intensive.

In [78], an experimental evaluation of a Li-Fi system with 15.73 Gb/s data rate and 16.m distance was performed. Four low-cost monochromatic LEDs were selected to modulate four wavelengths in the visible light spectrum. Wavelength Division Multiplexing(WDM) was used for higher spectral efficiency in addition to Orthogonal frequency division multiplexing (OFDM) with adaptive bit loading to provide higher data rate.

Apart from modulation techniques discussed in this section, many other techniques are always under modification. Out of these techniques, one using the polarization of light as modulation was also proposed in[79]. This method provides dimming support and low data rate support without flickering as it uses polarity of light for modulation.

IV. MIMO

In order to efficiently utilize the available Visible Light spectrum multiple LEDs can be considered as multiple transmitters but due to limited spatial diversity the interference would be high. To eliminate this interference suitable MIMO Receiver Designs can be considered. Design of such MIMO receiver is a challenge and issues of the same are discussed in following section.

A. MIMO Receiver Design

MIMO Receivers can be of two types depending on what component is used at receiver, i.e., Photo-diode or image sensor. The performance of receiver varies based on the type of sensor and modulation technique used. Photo-diode provides narrow Field of View (FOV) thus providing high gain but requires constraint alignment and performance degrades drastically with small misalignment. On the other hand, image sensor uses array of photodiodes and has large FOV which relaxes alignment requirement. It uses phenomenon called

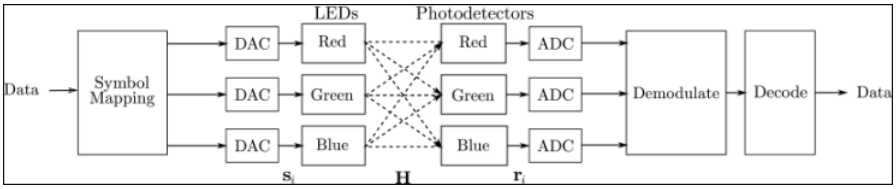


Figure 10. General block diagram for CSK modulation in Li-Fi[72]

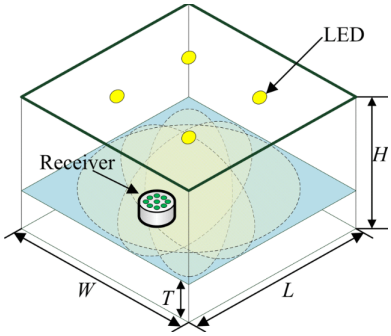


Figure 11. Multiple Photo-diodes for Spatial Diversity to achieve MIMO[80]

Rolling Shutter Effect to achieve high data rate. But individual photodiodes have low gain and complex image processing it needed at receiver, hence, it cannot be used for resource constraint devices.

A hybrid of image sensor and photo-diodes was proposed[11] which had advantages of both the techniques. In[80], a design was proposed as shown in fig.11 which spherically mounted photodiodes, which had improved spatial diversity due to non-LOS channels but at the cost of additional hardware. For efficient utilization of spatial diversity not only design of suitable receiver but also sufficient precoding at the transmitter is also required. This aspect was studied in[81], where the proposed optical adaptive precoding (OAP) scheme exploits the information of transmitted symbols to enhance the effective SNR and S/I.

B. Li-Fi MIMO Techniques

This section highlights MIMO techniques that are exploited in Li-Fi [82].

- Repetition Coding (RC): In this technique same signal is transmitted over all transmitters. This provides increased gain as the signals from all transmitters is reconstructed at receiver. However, this technique provides flexibility for transmitter-receiver alignment but provides restricted spectral efficiency.
- Spatial Multiplexing (SMP): On the contrary to RC in SMP, each transmitter sends different data which acts like multiple parallel SISO. This increases spectral efficiency but compared to RC SMP requires tighter alignment of transmitter-receiver to avoid interference however it still provides better data rates compared to RC.
- Spatial Modulation (SM): In this type of MIMO technique spatial dimension is used for transmitting data as only one transmitter transmits data at any arbitrary time. Each transmitter is activated based on the assigned symbols. Based on the received signal, estimation of which LED was activated is done which is used to reconstruct respective symbol at receiver. Thus it achieves higher SE compared to RC and SMP.

It was observed that image sensor provides better SNR compared to photodiodes for SM or SMP due its inherent feature of wider FOV and relaxed alignment[82].

V. Cell Design and Deployment

In [83], effects of inter-cell interference based on Li-Fi access point arrangements have been discussed. As the modulation techniques used show drastic degradation in the quality of signal for edge users, it is important to address the issue of inter-cell interference. Thus, a device which does not receive sufficient light due to room arrangement may experience low data rate compared to device which is aligned to the transmitter. In order to avoid this, several strategies are proposed like partitioned clusters [84], dynamic resource allocation based on uplink[85], Joint Transmission (JT)[86] using delayed transmission and reconstruction. However, it is important to design better interference avoidance techniques to ensure high data rate communication under close deployment of receivers as the coverage of normal VLC cells is small.

VI. Handover in Li-Fi

An important advantage of wireless communication is freedom of movement. However, in order to serve communication for devices under motion, the connected device should be able to transfer data session from one access point to others. This process of switching an access point while ensuring device connectivity is called handover. In this section, handover techniques for Li-Fi and related issues are discussed. This will provide an insight on which parameter for handover should be considered. Following mentioned are few parameters that are derived from RF handover and which can also be applied to Li-Fi cell-to-cell handover.

- **Receive Signal Strength (RSS):** Received Signal Strength (RSS) is the amount of power received by the receiver. In Li-Fi, Signal Strength is dependent on the received light intensity. Hence, the RSS for Li-Fi will not only change based on mobility but also device orientation. Using RSS value device and Li-Fi access point can undergo handover from one access point to others.
- **Signal-to-Interference Ratio (SIR):** The handover can also be initiated based on Signal-to-Interference Ratio

(SIR). If the interference is high compared to the received signal, then handover from one Li-Fi access point to other is initiated.
- **Speed of MN:** Handover occurs due to the mobility of the device. The handover can be initiated by estimating future position depending on the current speed and direction of mobility. The work in[87] provides insight into this problem. A mobile user can be provided with a dynamic quality of service when it is moving across heterogeneous networks.

In [87], a basic technique of handover in Li-Fi cells is proposed by considering the effects of device orientation and mobility. It calculates the probability of handover based on RSS by estimating it for simulated receiver movement and orientation. It uses a geometric model for receiver orientation and random way-point mode (RWP) for receiver movement. However, this handover is studied for horizontal handover, i.e., Li-Fi cell to Li-Fi cell. Whereas, in Vertical Handover[88] connectivity type is changed, i.e., Li-Fi to Wi-Fi or vice versa. It was observed that this technique requires additional overhead for adopting change in data frames. However, it provides more mobility and freedom of orientation compared to horizontal handover. In Dynamic load balancing (LB)[12], a solution is proposed for better utilization of bandwidth by providing RF-based connectivity for mobile device and Li-Fi based connectivity for almost stationary devices. An improvement over the same was proposed by using fuzzy logic for determining mobility and predicting handover with dynamic load balancing[89], which reduces handover overhead.

Our paper highlights the multi-faceted progress in the field of LiFi communication and potential challenges. Addressing these challenges will be key to the practical deployment of LiFi as future communication technology. Following are the open areas for research based on the conducted study:

- **Efficient Design of Transmitters and Modulation Techniques**: As light is the fastest mode of communication, the speed of communication is only limited by the processing speed of Transmitters and Receivers. However, by tapping into different properties of light such as color, polarity, and direction of transmission, the faster and efficient LiFi communication system can be achieved. Design of such a system of coordinated transmitters and implementation of suitable modulation scheme which could exploit these properties to enable faster and robust communication is active research in this area.
- **Receiver Designs to Enable MIMO and Mobility**: Receivers such as Camera or Multi-directional Photo-diodes with wider FOV have been observed to provide better support for MIMO and Device Mobility. However, the issues of inter-cell interference and active noise is prevalent are such systems. Different spatial modulation techniques and 3-dimensional localization and positioning techniques have been observed to improve the performance of LiFi system concerning device mobility. Hence, there is need of designing hybrid sensors which could leverage the benefits of Camera and Photo-diodes to develop faster, MIMO enabled and Mobility supported Receivers.
- **Modulation and Receiver Design dependent Transmitter Deployment Strategies**: The performance of various LiFi modulation techniques and receiver designs concerning transmitter deployment patterns have not been studied extensively. Thus providing an opportunity to explore communication performance concerning various modulation schemes and receiver designs to come up with a unified model for enhancing the overall performance.
- **Robust Load Balancing Techniques to Handle Effects of Handover** With respect to the practical deployment of LiFi enabled devices, it is important to understand the nature of change in network traffic and flow control when the user shifts between different networks. User mobility modeling and network load balancing is current active research in the field of handover in LiFi.

VII. Conclusions

Li-Fi is a promising new domain for enhancing the existing communication system. With this paper, we have tried to understand this growing domain in terms of its various challenges pertaining to design, modulation, deployment, handover, illumination standards and MIMO. It was observed that, hybrid design of receiver with multiple smaller FOVs can serve better support for MIMO and Handover. It was also observed that adoption of various OFDM techniques in IM/DD, have can be computationally complex but it can also serve faster data rates. Adoption of Spatial Modulation techniques paired with Color Domain Modulation techniques can provide with better Spectral Efficiency and Mobility support.

References

[1] H. Haas, L. Yin, Y. Wang, and C. Chen, "What is lifi?", *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1533–1544, Mar. 2016.

[2] S. Haruyama, "Visible light communication using sustainable led lights", in *2013 Proceedings of ITU Kaleidoscope: Building Sustainable Communities*, Apr. 2013, pp. 1–6.

[3] *Oledcomm products*. [Online]. Available: https : / / lifilighting.com/oledcomm-products/.

[4] *Home - purelifi - connectivity is evolving*. [Online]. Available: https://purelifi.com/.

[5] N. Omura, A. Higashi, J. Yabuuchi, T. Iwamatsu, and S. Oshiba, "Experimental demonstration of ofdm based wdm-mimo visible light communication system", in *2018 Asia-Pacific Microwave Conference (APMC)*, Nov. 2018, pp. 872–874.

[6] K. Ahn and J. K. Kwon, "Color intensity modulation for multicolored visible light communications", *IEEE Photonics Technology Letters*, vol. 24, no. 24, pp. 2254–2257, Dec. 2012.

[7] E. Sarbazi, M. Uysal, M. Abdallah, and K. Qaraqe, "Indoor channel modelling and characterization for visible light communications", in *2014 16th International Conference on Transparent Optical Networks (ICTON)*, Jul. 2014, pp. 1–4.

[8] O. Kharraz and D. Forsyth, "Performance comparisons between pin and apd photodetectors for use in optical communication systems", *Optik - International Journal for Light and Electron Optics*, vol. 124, pp. 1493–1498, Jul. 2013.

[9] S. Singh, G. Kakamanshadi, and S. Gupta, "Visible light communication-an emerging wireless communication technology", in *2015 2nd International Conference on Recent Advances in Engineering Computational Sciences (RAECS)*, Dec. 2015, pp. 1–3.

[10] L. I. Albraheem, L. H. Alhudaithy, A. A. Aljaser, M. R. Aldhafian, and G. M. Bahliwah, "Toward designing a li-fi-based hierarchical iot architecture", *IEEE Access*, vol. 6, pp. 40 811–40 825, 2018.

[11] L. Zeng, D. C. O'Brien, H. L. Minh, G. E. Faulkner, K. Lee, D. Jung, Y. Oh, and E. T. Won, "High data rate multiple input multiple output (mimo) optical wireless communications using white led lighting", *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 9, pp. 1654–1662, Dec. 2009.

[12] Y. Wang and H. Haas, "Dynamic load balancing with handover in hybrid li-fi and wi-fi networks", *Journal of Lightwave Technology*, vol. 33, no. 22, pp. 4671–4682, Nov. 2015.

[13] A. Galati and C. Greenhalgh, "Human mobility in shopping mall environments", *2nd International Workshop on Mobile Opportunistic Networking, MobiOpp 2010*, pp. 1–7, Jan. 2010.

[14] N. Garg and J. Parikh, "Wireless transceiver design for visible light communication", in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Nov. 2017, pp. 509–511.

[15] M. Dehghani Soltani, A. A. Purwita, I. Tavakkolnia, H. Haas, and M. Safari, "Impact of device orientation on error performance of lifi systems", *IEEE Access*, vol. 7, pp. 41 690–41 701, 2019.

[16] A. A. Purwita, M. Dehghani Soltani, M. Safari, and H. Haas, "Impact of terminal orientation on performance in lifi systems", in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018, pp. 1–6.

[17] M. D. Soltani, A. A. Purwita, Z. Zeng, H. Haas, and M. Safari, "Modeling the random orientation of mobile devices: Measurement, analysis and lifi use case", *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2157–2172, Mar. 2019.

[18] R. D. Roberts, S. Rajagopal, and S. Lim, "Ieee 802.15.7 physical layer summary", in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec. 2011, pp. 772–776.

[19] F. Zafar, D. Karunatilaka, and R. Parthiban, "Dimming schemes for visible light communication: The state of research", *IEEE Wireless Communications*, vol. 22, no. 2, pp. 29–35, Apr. 2015.

[20] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges", *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2047–2077, Fourthquarter 2015.

[21] A. Jovicic, J. Li, and T. Richardson, "Visible light communication: Opportunities, challenges and the path to market", *IEEE Communications Magazine*, vol. 51, no. 12, pp. 26–32, Dec. 2013.

[22] S. P. et al., "Information broadcasting system based on visible light signboard", *Proc. Wireless Opt. Commun.*, pp. 311–313, 2007.

[23] J. V. et al., "125 mbits over 5 m wireless distance by use of ook-modulated phosphorescent white leds", *Proc. 35th ECOC*, pp. 1–2, 2009.

[24] N. Fujimoto and H. Mochizuki, "477 mbit/s visible light transmission based on ook-nrz modulation using a single commercially available visible led and a practical led driver with a pre-emphasis circuit", 2013.

[25] K. Lee and H. Park, "Modulations for visible light communications with dimming control", *IEEE Photonics Technology Letters*, vol. 23, no. 16, pp. 1136–1138, Aug. 2011.

[26] S. V. D. Tsonev and H. Haas, "Light fidelity (lifi) towards all optical networking", *PProc. SPIE OPTO*, 2013.

[27] A. Sevincer, A. Bhattarai, M. Bilgi, M. Yuksel, and N. Pala, "Lightnets: Smart lighting and mobile optical wireless networks — a survey", *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1620–1641, Fourth 2013.

[28] S. Rajagopal, R. D. Roberts, and S. Lim, "Ieee 802.15.7 visible light communication: Modulation schemes and dimming support", *IEEE Communications Magazine*, vol. 50, no. 3, pp. 72–82, Mar. 2012.

[29] "Ieee draft standard for local and metropolitan area networks - part 15.7: Short-range optical wireless communications", *IEEE P802.15.7/D3, August 2018*, pp. 1–412, Jan. 2018.

[30] J. E. Gancarz, H. Elgala, and T. D. Little, "Overlapping ppm for band-limited visible light communication and dimming", *Journal of Solid State Lighting*, vol. 2, no. 1, p. 3, May 2015. [Online]. Available: https://doi.org/10.1186/s40539-015-0022-0.

[31] A. B. Siddique and M. Tahir, "Bandwidth efficient multi-level mppm encoding decoding algorithms for joint brightness-rate control in vlc systems", in *2014 IEEE Global Communications Conference*, Dec. 2014, pp. 2143–2147.

[32] T. Ohtsuki, I. Sasase, and S. Mori, "Overlapping multipulse pulse position modulation in optical direct detection channel", in *Proceedings of ICC '93 - IEEE International Conference on Communications*, vol. 2, May 1993, 1123–1127 vol.2.

[33] S. Long, P. Tsai, Y. Huang, and I. Lai, "Trellis coded generalized spatial modulation with spatial multiplexing", in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2017, pp. 832–837.

[34] L. Yi and S. G. Lee, "Performance improvement of dimmable vlc system with variable pulse amplitude and position modulation control scheme", in *2014 International Conference on Wireless Communication and Sensor Network*, Dec. 2014, pp. 81–85.

[35] Y. Zeng, Y. Chen, H. Zhao, and X. Wang, "Multiple pulse amplitude and position modulation for optical wireless channel", in *2015 Seventh International Conference on Ubiquitous and Future Networks*, Jul. 2015, pp. 132–134.

[36] C. Li, H. Lu, T. Lu, W. Tsai, B. Chen, C. Chu, C. Wu, and C. Liao, "A 100m/40gbps 680-nm vcsel-based lifi transmission system", in *2016 Conference on Lasers and Electro-Optics (CLEO)*, Jun. 2016, pp. 1–2.

[37] H. Lu, C. Li, H. Chen, C. Ho, M. Cheng, Z. Yang, and C. Lu, "A 56 gb/s pam4 vcsel-based lifi transmission with two-stage injection-locked technique", *IEEE Photonics Journal*, vol. 9, no. 1, pp. 1–8, Feb. 2017.

[38] S. Nishikawa, "Sequential m-ary pam system", *IEEE Transactions on Communications*, vol. 21, no. 1, pp. 22–33, Jan. 1973.

[39] M. Shi, C. Wang, H. Guo, Y. Wang, X. Li, and N. Chi, "A high-speed visible light communication system based on dft-s ofdm", in *2016 IEEE International Conference on Communication Systems (ICCS)*, Dec. 2016, pp. 1–5.

[40] Y. Li, J. Han, and X. Zhao, "Performance investigation of dft-spread ofdm signal for short reach communication systems beyond ng-pon2", *IEEE Access*, vol. 7, pp. 27 426–27 431, 2019.

[41] S.-z. Zhang, C.-t. Zheng, Y.-t. Li, W.-l. Ye, and Y. Liu, "Design and experiment of post-equalization for ook-nrz visible light communication system", *Optoelectronics Letters*, vol. 8, no. 2, pp. 142–145, Mar. 2012. [Online]. Available: https://doi.org/10.1007/s11801-012-1106-3.

[42] *M-qam*, https://www.slideshare.net/jhordyencarnacionnunez/m-qam/, Accessed: 2018-11-30.

[43] M. Zhang and Z. Zhang, "An optimum dc-biasing for dco-ofdm system", *IEEE Communications Letters*, vol. 18, no. 8, pp. 1351–1354, Aug. 2014.

[44] S. Vappangi and V. V. Mani, "Performance analysis of dst-based intensity modulated/direct detection (im/dd) systems for vlc", *IEEE Sensors Journal*, vol. 19, no. 4, pp. 1320–1337, Feb. 2019.

[45] Namei Yin, C. Guo, Y. Yang, P. Luo, and Chunyan Feng, "Asymmetrical and direct current biased optical ofdm for visible light communication with dimming control", in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 23–28.

[46] H. Chen, S. Hu, J. Ding, S. Bian, H. Wu, P. Hua, S. You, X. Li, Q. Yang, and M. Luo, "Performance comparison of visible light communication systems based on aco-ofdm, dco-ofdm and ado-ofdm", in *2017 16th International Conference on Optical Communications and Networks (ICOCN)*, Aug. 2017, pp. 1–3.

[47] S. C. J. Lee, S. Randel, F. Breyer, and A. M. J. Koonen, "Pam-dmt for intensity-modulated and direct-detection optical communication systems", *IEEE Photonics Technology Letters*, vol. 21, no. 23, pp. 1749–1751, Dec. 2009.

[48] C. K. H. Vasconcelos, A. N. Barreto, and D. A. A. Mello, "Quadrature diversity combining for pulse-amplitude modulated dmt in im/dd channels", in *2014 IEEE Global Communications Conference*, Dec. 2014, pp. 2090–2095.

[49] M. S. Islim, D. Tsonev, and H. Haas, "Spectrally enhanced pam-dmt for im/dd optical wireless communications", in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Aug. 2015, pp. 877–882.

[50] T. Wang, Y. Hou, and M. Ma, "A novel receiver design for haco-ofdm by time-domain clipping noise elimination", *IEEE Communications Letters*, vol. 22, no. 9, pp. 1862–1865, Sep. 2018.

[51] T. Zhang, Y. Zou, J. Sun, and S. Qiao, "Design of pam-dmt-based hybrid optical ofdm for visible light communications", *IEEE Wireless Communications Letters*, pp. 1–1, 2018.

[52] Z. Ghassemlooy, F. Ebrahimi, S. Rajbhandari, S. Olyaee, X. Tang, and S. Zvanovec, "Visible light communications with hybrid ofdm-ptm", in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Jun. 2017, pp. 894–898.

[53] N. Taherkhani and K. Kiasaleh, "Reed solomon encoding for the mitigation of clipping noise in ofdm-based visible light communications", in *2018 International Conference on Computing, Networking and Communications (ICNC)*, Mar. 2018, pp. 285–289.

[54] Q. Wang, C. Qian, X. Guo, Z. Wang, D. G. Cunningham, and I. H. White, "Layered aco-ofdm for intensity-modulated direct-detection optical wireless transmission", *Opt. Express*, vol. 23, no. 9, pp. 12 382–12 393, May 2015. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-23-9-12382.

[55] R. Islam and M. R. H. Mondal, "Hybrid dco-ofdm, aco-ofdm and pam-dmt for dimmable lifi", *Optik*, vol. 180, pp. 939–952, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402618318746.

[56] D. Tsonev, S. Sinanovic, and H. Haas, "Novel unipolar orthogonal frequency division multiplexing (u-ofdm) for optical wireless", in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, May 2012, pp. 1–5.

[57] J. Zhou and W. Zhang, "A comparative study of unipolar ofdm schemes in gaussian optical intensity channel", *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1549–1564, Apr. 2018.

[58] J. Lian, Y. Gao, and D. Lian, "Variable pulse width unipolar orthogonal frequency division multiplexing for visible light communication systems", *IEEE Access*, vol. 7, pp. 31 022–31 030, 2019.

[59] J. Lian and M. Brandt-Pearce, "Clipping-enhanced optical ofdm for im/dd communication systems", in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[60] J. Lian and M. Brandt-Pearce, "Clipping-enhanced optical ofdm for visible light communication systems", *Journal of Lightwave Technology*, vol. 37, no. 13, pp. 3324–3332, 2019.

[61] M. Noshad and M. Brandt-Pearce, "Hadamard coded modulation: An alternative to ofdm for wireless optical communications", in *2014 IEEE Global Communications Conference*, Dec. 2014, pp. 2102–2107.

[62] M. Noshad and M. Brandt-Pearce, "Hadamard-coded modulation for visible light communications", *IEEE Transactions on Communications*, vol. 64, no. 3, pp. 1167–1175, 2016.

[63] K. M. Wong, J. Wu, T. N. Davidson, and Q. Jin, "Wavelet packet division multiplexing and wavelet packet design under timing error effects", *IEEE Transactions on Signal Processing*, vol. 45, no. 12, pp. 2877–2890, Dec. 1997.

[64] W. Huang, C. Gong, and Z. Xu, "Visible light communication based on wavelet packet division multiplexing", in *2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2014, pp. 1–5.

[65] M. S. Moreolo, R. Munoz, and G. Junyent, "Novel power efficient optical ofdm based on hartley transform for intensity-modulated direct-detection systems", *Journal of Lightwave Technology*, vol. 28, no. 5, pp. 798–805, Mar. 2010.

[66] J. Jeganathan, A. Ghrayeb, and L. Szczecinski, "Spatial modulation: Optimal detection and performance analysis", *IEEE Communications Letters*, vol. 12, no. 8, pp. 545–547, Aug. 2008.

[67] T. Wang, F. Yang, L. Cheng, and J. Song, "Spectral-efficient generalized spatial modulation based hybrid dimming scheme with laco-ofdm in vlc", *IEEE Access*, vol. 6, pp. 41 153–41 162, 2018.

[68] A. Yesilkaya, E. Basar, F. Miramirkhani, E. Panayirci, M. Uysal, and H. Haas, "Optical mimo-ofdm with generalized led index modulation", *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3429–3441, Aug. 2017.

[69] T. Mao, Z. Wang, Q. Wang, S. Chen, and L. Hanzo, "Dual-mode index modulation aided ofdm", *IEEE Access*, vol. 5, pp. 50–60, 2017.

[70] C. Rajesh Kumar and R. K. Jeyachitra, "Power efficient generalized spatial modulation mimo for indoor visible light communications", *IEEE Photonics Technology Letters*, vol. 29, no. 11, pp. 921–924, Jun. 2017.

[71] *International commission on illumination*, May 2019. [Online]. Available: https://en.wikipedia.org/wiki/International_Commission_on_Illumination#/media/File:CIExy1931.png.

[72] E. Monteiro and S. Hranilovic, "Design and implementation of color-shift keying for visible light communications", *Journal of Lightwave Technology*, vol. 32, no. 10, pp. 2053–2060, May 2014.

[73] R. Singh, T. OFarrell, and J. P. R. David, "An enhanced color shift keying modulation scheme for high-speed wireless visible light communications", *Journal of Lightwave Technology*, vol. 32, no. 14, pp. 2582–2592, Jul. 2014.

[74] X. Liang, M. Yuan, J. Wang, Z. Ding, M. Jiang, and C. Zhao, "Constellation design enhancement for color-shift keying modulation of quadrichromatic leds in visible light communications", *Journal of Lightwave Technology*, vol. 35, no. 17, pp. 3650–3663, Sep. 2017.

[75] J. Okumura, Y. Kozawa, Y. Umeda, and H. Habuchi, "Hybrid pwm/dpam dimming control for digital color shift keying using rgb-led array", *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 1, pp. 45–52, Jan. 2018.

[76] K. Ahn and J. K. Kwon, "Color intensity modulation for multicolored visible light communications", *IEEE Photonics Technology Letters*, vol. 24, no. 24, pp. 2254–2257, Dec. 2012.

[77] W. Chen, Z. Li, and M. Jiang, "Color-and-intensity shift keying for visible light communication", *IEEE Communications Letters*, vol. 22, no. 9, pp. 1790–1793, Sep. 2018.

[78] R. Bian, I. Tavakkolnia, and H. Haas, "15.73 gb/s visible light communication with off-the-shelf leds", *Journal of Lightwave Technology*, vol. 37, no. 10, pp. 2418–2424, May 2019.

[79] M. A. Atta and A. Bermak, "A polarization-based interference-tolerant vlc link for low data rate applications", *IEEE Photonics Journal*, vol. 10, no. 2, pp. 1–11, Apr. 2018.

[80] T. Q. Wang, C. He, and J. Armstrong, "Performance analysis of aperture-based receivers for mimo im/dd visible light communications", *Journal of Lightwave Technology*, vol. 35, no. 9, pp. 1513–1523, May 2017.

[81] M. M. Cespedes and A. G. Armada, "On the optimality of multiple photodiode receivers using precoding schemes for visible light communications", in *2018 Global LIFI Congress (GLC)*, Feb. 2018, pp. 1–4.

[82] T. Fath and H. Haas, "Performance comparison of mimo techniques for optical wireless communications in indoor environments", *IEEE Transactions on Communications*, vol. 61, no. 2, pp. 733–742, Feb. 2013.

[83] L. Feng, R. Q. Hu, J. Wang, and Y. Qian, "Deployment issues and performance study in a relay-assisted indoor visible light communication system", *IEEE Systems Journal*, pp. 1–9, 2018.

[84] S. Jung, D. Kwon, S. Yang, and S. Han, "Reduction of inter-cell interference in asynchronous multi-cellular vlc by using ofdma-based cell partitioning", in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Jul. 2016, pp. 1–4.

[85] M. Kashef, M. Abdallah, K. Qaraqe, H. Haas, and M. Uysal, "Coordinated interference management for visible light communication systems", *IEEE/OSA Journal of Optical Communications and Networking*, vol. 7, no. 11, pp. 1098–1108, Nov. 2015.

[86] C. Chen, D. Tsonev, and H. Haas, "Joint transmission in indoor visible light communication downlink cellular networks", in *2013 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 1127–1132.

[87] M. D. Soltani, H. Kazemi, M. Safari, and H. Haas, "Handover modeling for indoor li-fi cellular networks: The effects of receiver mobility and rotation", in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[88] H. Han, L. He, and Q. Li, "Vertical handover in optical wireless heterogeneous networks", in *2015 IEEE International Conference on Communication Software and Networks (ICCSN)*, Jun. 2015, pp. 427–431.

[89] Y. Wang, X. Wu, and H. Haas, "Fuzzy logic based dynamic handover scheme for indoor li-fi and rf hybrid network", in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

Selection of sub-optimal feature set of network data to implement Machine Learning models to develop an efficient NIDS

Jashanpreet Singh Sadioura
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala, India
jashansadioura@gmail.com

Satbir Singh
Mechanical Measurement Instrumentation
CSIR-CSIO, Chandigarh, India
satbir2008@ymail.com

Amitava Das
Mechanical Measurement Instrumentation
CSIR-CSIO, Chandigarh, India
adas@csio.res.in

***Abstract*—With the rapid increase in the dependency on technology and the internet in our personal and professional life, the computer networks have become very congested, and the frequency of presence of an intrusion in a network has also increased. An active IDS (Intrusion Detection System) protects the network from intrusions and provide security to the system. Machine learning techniques are the most efficient technologies to develop IDS as substantial network data can be easily trained and tested using ML models. Any general machine learning models work in three phases: Data pre-processing, Feature selection and training, and testing the developed models. The major contribution of this paper is the extraction of sub-optimal feature set from NSL-KDD data set having 41 features and then implementing different ML models to find the best suitable model using this set of features. It is observed that the ML models SVC and MLPClassifier performed better as compared to CNN in terms of complexity, accuracy and training time when trained and tested using the selected optimal feature set. CNN is an excellent deep learning algorithm that gives good results for image data perform better in comparison to simple text data machine learning models like SVC and MLP Classifier. MLP Classifier gave a higher accuracy of 98.19%.**

***Index Terms*—KDD 99, NSL-KDD, Information Gain, SVC, MLP Classifier, CNN.**

I. Introduction

With the rapid growth in technology and expansion of the internet the use of computer systems has increased in the last few decades. Large companies, startups and even individuals use the internet for their personal and professional purposes. With such a high dependency on the internet for data storage, it becomes essential to have a safe and secure computer network on which we can rely for data storage. This fact is a big challenge in the field of network security and has been accepted by many researchers who are working on it. Almost all the computer-based systems are always in connection with internet which creates a serious security threat to the smart working of these systems. An effective intrusion detection system needs to be embedded in systems such as servers and data receiving devices.

978-1-7281-2087- 4/19/$31.00 © 2019 IEEE

There are many security systems like firewall and antimalware, but the attackers can easily exploit the vulnerabilities to these by injecting intrusion in the system. An intrusion is unauthorized activity on a computer network. Intrusion Detection System is a device or software application that enhances the network security and safeguard the data by monitoring comprehensive network and analyzing the network requests to identify the malicious or abnormal requests or attacks in a network.

There are two methods of intrusion detection: Signature based and Anomaly-based ID. Machine learning and deep learning algorithms can be used to develop and train the models for intrusion detection. Machine learning and deep learning techniques are being widely used to develop the classification models for the intrusion detection system. These techniques give effective results in detecting an intrusion in a real-time IDS.

With the large size of the high dimensional network data set it becomes computationally expensive to implement the machine learning models on such data sets. This is where the major contribution of this paper in selection of optimal and informative features from the network come into play. In feature selection, we select the features that are relevant and informative for our classification and eliminate the features that are redundant and that do not give much information about the classification. Eliminating redundant and less informative features makes our model simple and computationally fast without having much effect on the accuracy. After getting the optimal subset of features we train and test our developed models for accuracy.

In research that is proposed in this paper, we have trained and tested three ML models MLP, SVC and CNN using the selected feature set and based on the accuracy we have selected the best model which can be further used to develop

an efficient NIDS.

II. Related Literature

Much work has already been done in the area of network security. This section explains the work that is more or less similar to our area of work. Intrusions present in the network are a significant threat to the data, and intrusion detection systems play a vital role in providing security to the system and dealing with the network attacks. The purpose of IDS is to deal with the attack by collecting information within the computer system and compare it with existing patterns to discriminate between an attack and a genuine hit request [2]. Majorly there are two basic approaches for ID, signature-based approach in which attack pattern of the intruder is modelled, if the match is detected the system signals it as an intrusion and an anomaly based approach the normal behavior is modelled and the system detects the intrusion if the behavior does not matches the normal behavior [3].

Many ID techniques have been already developed such as techniques like SNORT for signature-based IDS in [6],[5] and for anomaly-based IDS models like auto-associative kernel regression (AAKR) model coupled with the statistical probability ratio test (SPRT) is developed which is applied to Supervisory Control and Data Acquisition Systems (SCADA) system for anomaly based detections [4]. The speed, integrity and availability of service are the critical features of any IDS, and this is where machine learning and data mining comes into the picture.

For developing, testing and evaluating the various machine learning models for IDS development, various data sets described in [9] are available such as KDD 99, NSL-KDD and KYOTO 2006+. These data sets are standard data sets and a lot of research work in the field of ID is done on these data sets. KDD 99 is a very standard data set for research in this field.

The available network data sets are very robust having a large volume of data and a large number of features which makes it very difficult to work on these data sets. Various research work has been done on these data sets to select the optimal subset of features to get the better performance of the IDS. In [14] Shahrzad Zargari and Dave Voorhis have proposed an optimum subset of four features in KDD data set with feature number (3, 5, 6, and 39) for reduction of the dimensionality of the input data and simplification of the modelling process to achieve faster and better accuracy.

In [8] Mouhammad Alkasassbeh and Mohammad Almseidin performed the intrusion detection accuracy test on KDD data set for J48, MLP, and Bayes Network where the calculated values of accuracy are 93.1%, 91.9% and 90.73% respectively. The researchers Zhu Xiaoliang, Wang Jian, YanHongcan and Wu Shangzhuo performed the decision tree based classification on KDD data set for intrusion detection where they got the detection accuracy of 97.8% for C4.5 decision tree [13]. Using traditional ANN the detection rate obtained is 81.2% and 79.9% for intrusion detection and attack type classification task respectively for NSLKDD dataset [19].

In [16] researchers Vinayakumar R, Soman KP and Prabaharan Poornachandran proposed different CNN models on KDD 99 data set including CNN models with 1 Layer, 2 Layers, 3 Layers and other combinations of models like CNN+RNN, CNN+LSTM and CNN+GRU and showed that CNN model performed better as compared to its variants.

In the previous works done various models and algorithms are proposed for this problem such as data set analysis and feature selection [14],[12], comparison of performance of various machine learning models [11], [13] and even deep learning models have also been developed for this problem [16]. There has been no such article that aims to develop the computationally most simple model we in this research have the aim to develop and find out the model that is computationally simple and that uses the minimum optimal feature set and then integrate this subset with the proposed models.

III. Dataset Description

A. KDD 99

It is a benchmark data set to research network intrusion detection. It is the subset of 1998 DARPA dataset that was collected by simulation of the operation of a typical US Air Force LAN with multiple attacks and acquired nine weeks of TCP dump data. The dataset was collected and distributed at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory [9]. It has 41 features that are labelled as normal or an attack with precisely 22 different attack types which are divided into four main categories as shown in Table 1. KDD 99 data set has several versions available, out of which we have used KDD 10.

As KDD 99 data set is derived from DARPA'98 there, exist some inherited problems in KDD 99 data set. Some of the Issues with KDD 99 are as:

1) Work load of the synthesized data is not similar to the real network traffic [17].
2) There are large number of redundant records which cause the learning algorithm to be biased towards the more frequently appearing records [17].
3) Unrealistic relationship between the individual categories of the attack. [9],[17].
4) R2L instances have same or similar value of features as normal instances.

B. NSL-KDD

NSL-KDD is a refined version of the KDD 99 data set designed to fix the issues that were there in the original KDD 99 data set [18]. NSL-KDD does not include redundant and duplicate instances in training or testing data set, which improves the intrusion detection rate. Similar to KDD 99 it also has various versions, but we have used NSL-KDD 20 percent for our research in this paper. In [18] L.Dhanabal and Dr S.P. Shantharajah analysed the performance of this data set on various machine learning algorithms and stated that this is the best data set to simulate and test the performance of IDS.

TABLE I: Description of four categories of the attack type and the list of attacks that fall into that categories

Attack Category	Description	Attacks
Probe	Attempt to collect system information for ulterior motive	ipsweep, nmap, portsweep, satan
DoS (Denial of Service)	Attacker does not allow the computer to respond to the request by overloading the resources	back, land, neptune, pod, smurf, teardrop
U2R (User to Root)	Attacker tries to gain the root access to the system by having the access to the normal user account	U2R (User to Root) buffer$_overflow, loadmodule, perl, rootkit$
R2L (Remote to Local)	Attacker tries to send the data packets to the victim system without having access to the remote system	ftp$_write, guesspasswd, imap, multihop, phf, spy, warezlient, warezmaster$

IV. Methodology

The methodology proposed for this research works in three steps: Feature selection followed by pre-processing of data and then training and testing the models for accuracy using the selected optimal feature set.

A. Feature selection

Due to the size and complexity of the NSL-KDD 99 data set, it becomes computationally expensive to process the data and get effectively trained the model. To reduce the dimensional space of the data set, we need to identify the subset of features that are useful and informative. Feature selection for the network intrusion detection dataset is essential for the performance and high accuracy of the system. Irrelevant and redundant and reduce performance. The goal of feature selection is to find the minimum set of features such that the probability of correct classification using the reduced feature set is nearly same as the probability of correct classification using the original set of all features. We have used Information Gain for feature selection because it helps us to identify the attributes that are informative for the models and the attributes that are not very informative and that can be ignored without affecting the classification accuracy of the model. This algorithm ranks the attributes on basis of information gain value from higher to lower making it easy for us to select the suitable attributes. attributes of the system can lead to high complexity.

1) Information Gain Algorithm for feature selection: Information Gain (IG) is an entropy [21] based feature selection technique in machine learning for classification. In general IG measures how much "information" a feature gives about the class. IG is frequently employed as the term goodness criterion in machine learning [23]. Therefore, the calculation of the entropy is fundamental for calculating IG. The formulae for the calculation of IG is:

$$Entropy(T) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (1)$$

$$Entropy(T, X) = \sum_{c \epsilon x} p(c) E(c) \quad (2)$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (3)$$

where T = test class, X = attribute whose gain value has to be found.

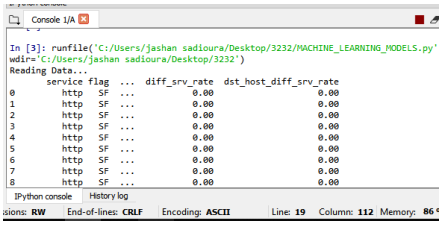


Fig. 1: Data attributes for pre-processing before training

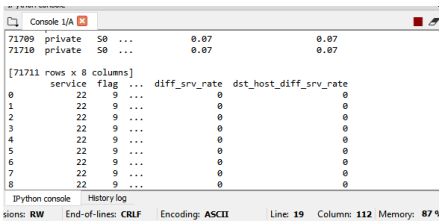


Fig. 2: Pre-processed data to be used for training model

B. Pre-processing

Attributes in any dataset have all forms continuous, discrete and also there are of different data types; hence, pre-processing is required before the development of any machine learning model. There are some features in our selected feature set that have String values that need to be encoded for developing the classifiers as the neural network classifier will not work with text data, and for that matter, we have used sklearn pre-processing module. In figure 2 the service and flag attributes have String values which are encoded as integer values giving a unique value to each String as shown in figure 3. In figure 2 the service and flag attributes have String values which are encoded as integer values giving a unique value to each String as shown in figure 3.

C. Training and development of machine learning models

The proposed machine learning models are developed using sci-kit learn machine learning library in Python 3.7. There are multiple classifiers in scikit-learn that can be used to solve various machine learning problems. We are two classifiers for our problem namely Support Vector Classification(SVC) and Multi-Layer Perceptron (MLP).

1) SVC: SVC is a supervised learning algorithm in machine learning for classification. Main idea behind SVC is the construction of optimal hyperplane, which can be used for

```
def train_model(x_train,y_train,x_test,y_test,classifier,**kwargs):
    "fit the chosen model and print out the score"

    #model -instantiate
    model=classifier(**kwargs)

    #training model
    model.fit(x_train,y_train)

    #checking accuracy and printing out the results
    fit_accuracy=model.score(x_train,y_train)
    test_accuracy=model.score(x_test,y_test)

    print(f"train accuracy: {fit_accuracy:0.2%}")
    print(f"test accuracy: {test_accuracy:0.2%}")
    return model
```

Fig. 3: Code for development of ML model for classification

classification for linearly separable patterns by maximizing the margin of the hyperplane i.e. the distance of the hyperplane to the nearest point of each pattern and for non-linearly separable patterns we classify by mapping the original data to higher dimensional space using kernel function. For SVC classification in sklearn, we need to import support vector machine module using from "sklearn.svm import SVC". SVC in sklearn uses a polynomial kernel with degree 3.

2) MLP: MLP is an essential model in neural networks which uses a back-propagation training algorithm. There are various hidden layers embedded between input and output layer in the MLP model. Neurons are interconnected, and the connections are always from lower to the higher layer with no neurons connections within the same layer. The number of neurons in the input layer is equal to the number of features the model is to be trained with, and the number of neurons in the output layer is equal to the number of classes. We use the neuralnetwork library in $sklearn$ for implementing MLP classification $\backslash fromsklearn.neural_network importMLPClassifier$". The proposed MLP classifier uses '$relu'$ activation function, '$adam'$ for weight optimization and constant learning rate of 0.001.

The algorithm for the proposed model is as follows:

1) Define a function that takes 5 parameters which includes training and testing data sets, classifier name and $**$ $kwargs$ keyword.
2) Initiate the model
3) Start model training using $model.fit(x_train, y_train$
4) Calculate and print train and test accuracy using $model.score(train, test)$ function.

D. Training and developing CNN (Convolutional Neural Networks)

Convolutional Neural Network (CNN) is one of the famous and influential deep learning models that can process and handle a huge volume of data. One of the most crucial features of CNN is that it reduces the number of parameters in traditional ANN's. CNN works with a multi-layer architecture having convolutional, non-linearity, pooling, dense and activation layers [15]. Depending on the dimensions of the data the CNN can be of 1D, 2D, 3D. Since our data is not in the image form, it has only one dimension, so we have used the 1D convolutional neural network in our CNN model. The underlying architecture of 1D CNN used in the proposed model in this paper shown in figure 4.

The detailed description of the different layers used in this CNN model is given in the Table 2.

Using the layers mentioned in table 2, we developed our final CNN model, and for optimization, we have used Stochastic gradient descent (sgd) optimizer with 0.01 as the learning rate and decay value of 1e-6.

The summary of the CNN layers and the hyper-parameters of the developed model is shown in figure 5.

After developing the models we have adopted the following research methodology for our research given below.

1) From the complete feature set (U) of 41 features in the data set, we made random seven groups of features with an equal number of features in each group.

$$G_{(i)} \subset U_{fori:(1,2,3,....,7)} \tag{4}$$

2) Perform IG algorithm using ranker search approach on these groups with random features using WEKA 3.6 [22].
3) Choose a threshold value of information gain and select all the features that have the information gain value higher than that value. In this paper, we used 0.4 as a threshold value of information gain.
4) Perform the classification using the features selected in Step 3 using neural networks and CNN models proposed in this paper and note down the test accuracy value for that feature set.
5) Make combination of two groups and perform Steps 2 to 4 on the new group.

$$G_1G_I = G_1 \cap G_{(i)\,fori:(1,2,3,....,7)} \tag{5}$$

Similarly make other random combinations taking three, four, five, six and seven groups together.

6) Select the feature set that gives the highest value of accuracy and run the developed models on this feature set to compare the performance of the models.

V. Experiment and Results

The experiment is carried out on the system with Intel® Core(TM) i7-3770 CPU @ 3.40GHz processor using 8.00 GB RAM. We have used Anaconda idle for writing the codes in Python 3.7. The results are summarized as:

A. Data set selection

On analyzing and studying the feature set and the data in KDD 99 and NSL-KDD data sets, the former seems to be not suitable for the development of any model for intrusion detection. When the proposed models and algorithms are run on the KDD 99 data set, we get extensively high value of accuracy for each model. The very extensively high (99.97%) value of accuracy for KDD 99 data set is due to various redundant and duplicate records present in KDD 99 data set. So NSL-KDD data sets give more informative results as compared to KDD 99.

TABLE II: complete description of the CNN layers and hyper-parameters used

S. NO.	CNN Layer	Description
1	Embedding Layer with 64 embedding dimensions and input data length of 8	This layer is used in CNN with text data to encode the input text data into unique integers
2	Dropout(0.2)	Regularization layer in CNN to prevent over fitting by ignoring the randomly selected neurons during training
3	Conv1D with 20 filters, (6,6) kernel size, 'softmax' as activation function and stride value of 2	This layer produces tensor of output by creating a kernel that is convolved with the input layer
4	Max Pooling Layer	The layer reduces the dimensionality by applying pooling function
5	Dense Layer with 'relu' as activation function and hidden dimension value as 1000	It is a regular CNN layer to connect neurons, where each neuron receives input from all the neurons from the previous layer
6	Dropout(0.2)	It is a regularization layer in CNN to prevent overfitting by ignoring the randomly selected neurons during training
7	Activation Layer with 'relu' activation function	Applies activation function to an output
8	Dense with output array of size (*, 1) and 'relu' activation function.	It is a regular CNN layer to connect neurons, where each neuron receives input from all the neurons from the previous layer
9	Activation Layer with 'relu' activation function	Applies activation function to an output.

B. Feature Selection

On analyzing the features in the original data set, we found that some features are not very informative: some of them are noisy, meaningless or irrelevant for the development of our proposed model. Moreover preparing the data set by taking all the features available in NSL-KDD is an unsustainable task for any real-time intrusion detection system.So, we have reduced the features in NSL-KDD data set to a feature set of 8 sub optimal features using information gain algorithm that gives relevant and correct information to the system, and that can be extracted for any real-time system. Having just 8 features reduces the complexity and dimensional spaces of our model. The impact on accuracy and performance after selecting a smaller subset of features is negligible. Table 3 gives the complete description of the features selected and their corresponding information gain value.

TABLE III: Complete description of the CNN layers and hyper-parameters used

Index	Feature Name	Description	Info gain value
3	Service	This layer is used in CNN with text data to encode the input text data into unique integers	1.1058
4	Flag	Status of connection	0.96423
5	src_bytes	No.of B from source to destination	1.34652
23	Count	No.of connections to the same host as the current connection at a given interval	0.82116
24	Srv_count	No. of connections to the same service as the current connection at a given interval	0.42836
29	Same_srv_rate	percent of connections to the same service	0.87165
30	Diff_srv_rate	percent of connections to different services	0.939
35	Dst_host_diff_srv_rate	percent of connections to different hosts on the same system	0.88612

C. Performance of the models developed using the selected feature set

The performance is evaluated by calculating the accuracy of each model, using $y_t estandy_p redictedvalues$.

Table 4 shows the accuracy minimum, average and maximum accuracy value using different combinations of features when the developed MLP, SVC and CNN models are trained and tested on NSL-KDD data set. The maximum value of accuracy is achieved using the optimal feature set of 8 features that we have selected in the previous section. Also the value of accuracy achieved by using these features is also higher than the accuracy value ever achieved on this data set using MLP Classifier which is 98.19%. Further on comparing the

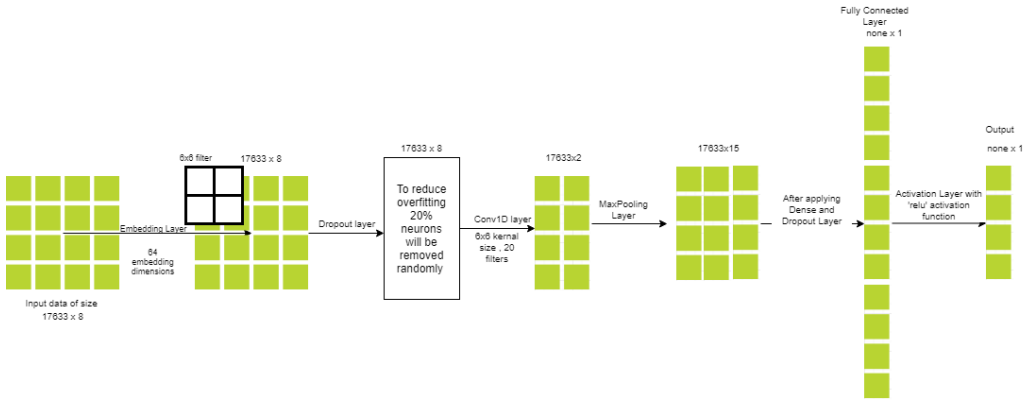


Fig. 4: Basic architecture of the CNN model proposed

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 8, 64)	704000
dropout_5 (Dropout)	(None, 8, 64)	0
conv1d_3 (Conv1D)	(None, 2, 15)	5775
global_max_pooling1d_3 (Glob	(None, 15)	0
dense_5 (Dense)	(None, 1000)	16000
dropout_6 (Dropout)	(None, 1000)	0
activation_5 (Activation)	(None, 1000)	0
dense_6 (Dense)	(None, 1)	1001
activation_6 (Activation)	(None, 1)	0

Total params: 726,776
Trainable params: 726,776
Non-trainable params: 0

Fig. 5: Summary of the CNN model developed

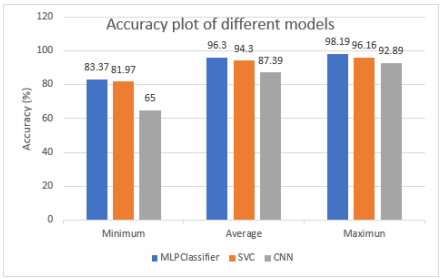


Fig. 6: Comparision of results shown in Table 4

performance of the different models we find that MLP is best for development of a NIDS giving the higher value of accuracy.

TABLE IV: Accuracy value comparison of different models

Trait	MLP	SVC	CNN
Minimum	83.37	81.97	65
Average	96.30	94.35	87.39
Maximum	98.19	96.16	92.89

VI. Conclusion

This paper proposed the feature set consisting of eight features out of 41 features present in Network data. The various machine learning and deep learning models were trained and tested using this feature set out of which MLP Classifier showed the best performance in terms of complexity, performance and training time. The future work is to implement this machine learning model using the selected feature set for the network to develop a real-time intrusion detection system (IDS).

Acknowledgment

The authors would like to express sincere gratitude to Director, CSIR-CSIO for providing infrastructural facilities. This study was supported under the FTT project MLP 0043.

References

[1] Jamuna .A and Vinodh Ewards S.E, "Efficient Flow based Network Traffic Classification using Machine Learning", in International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 2, pp.1324-1328, March -April 2013.

[2] Asmaa Shaker Ashoor and Prof. Sharad Gore, "Importance of Intrusion Detection System (IDS)", in International Journal of Scientific Engineering Research, I061227.

[3] Peyman Kabiri and Ali A. Ghorbani, "Research on Intrusion Detection and Response:A Survey", in International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005.

[4] Dayu Yang, Alexander Usynin, and J. Wesley Hines, "Anomaly-Based Intrusion Detection for SCADA Systems", in International conference for data security, vol 4,page no:1563-1566,2012.

[5] Vinod Kumar and Dr. Om Prakash Sangwan, "Signature Based Intrusion Detection System Using SNORT", in International Journal of Computer Applications Information Technology Vol. I, Issue III, page no 35-41, November 2012.

[6] Rishabh Gupta, Soumya Singh, Shubham Verma and Swasti Singhal, "Intrusion Detection System Using SNORT", in International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 04, page no: 2100-2104, Apr -2017.

[7] Mansour Sheikhan and Amir Ali Sha'bani, "Fast Neural Intrusion Detection System Based on Hidden Weight Optimization Algorithm and Feature Selection", in World Applied Sciences Journal 7 (Special Issue of Computer IT): 45-53, 2009.

[8] Mouhammad Alkasassbeh and Mohammad Almseidin, "Machine Learning Methods for Network Intrusion Detection", in The 20th International Conference on Computing, Communication, 2018

[9] Danijela D. Protić, "Review of KDD CUP'99 , NSL-KDD and KYOTO 2006+ data sets" in Vojnotechnicki Glasnik /Military technical, 2018, Vol. 66, Issue 3 , page no 580-596.

[10] Mahbod Tavallaee, Natalia Stakhanova, and Ali Akbar Ghorbani, Member, IEEE, "Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods" , in IEEE transactions on systems, man and cybernatics—PART C: applications and review, Vol. 40, NO. 5, September 2010.

[11] Mouhammad Alkasassbeh and Mohammad Almseidin, "Machine Learning Methods for Network Intrusion Detection", The 20th International Conference on Computing, Communication, 1 Sep, 2018

[12] Kajal Rai, M. Syamala Devi and Ajay Guleria, "Decision Tree Based Algorithm for Intrusion Detection", in Int. J. Advanced Networking and Applications Volume: 07 Issue: 04 Pages: 2828-2834, 2016

[13] Zhu Xiaoliang, Wang Jian, YanHongcan and Wu Shangzhuo, "Research and Application ofthe improved Algorithm C4.5 on Decision Tree", in 2009 International Conference on Test and Measurement, page no 184-187,2009.

[14] Shahrzad Zargari and Dave Voorhis, "Feature Selection in the Corrected KDD-dataset", in Third International Conference on Emerging Intelligent Data and Web Technologies, page no 174-180, 2012

[15] Saad ALBAWI , Tareq Abed MOHAMMED and Saad AL-ZAWI, "Understanding of a Convolutional Neural Network" , in International Conference on Engineering and Technology (ICET), Antalya, Turkey,08 March 2018

[16] Vinayakumar R, Soman KP and Prabaharan Poornachandran, "Applying Convolutional Neural Network for Network Intrusion Detection", in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, page no: 1222-1228, 2017.

[17] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set" , Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), 2009.

[18] L.Dhanabal1 and Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", in International Journal of Advanced Research in Computer and Communication Engineering,Vol. 4, Issue 6, June 2015.

[19] Bhupendra Ingre and Anamika Yadav, "Performance Analysis of NSL-KDD dataset using ANN", in SPACES-2015 conference, Dept of ECE, K L UNIVERSITY, January 2015.

[20] Mohamed Faisal Elrawy, Ali Ismail Awad and Hesham F. A. Hamed, "Intrusion detection systems for IoT-based smart environments: a survey", in Journal of Cloud Computing: Advances, Systems and Applications, 2019.

[21] Ning Yang , Tianrui Li and Jing Song , "Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation", in International Journal of Computational Intelligence Systems, October 2007.

[22] Dr. Sudhir B. Jagtap and Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA", in International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore.

Generative model chatbot for Human Resource using Deep Learning

Salim Akhtar Sheikh
School of Computing and Information Technology
Manipal University Jaipur
Jaipur, India
salim.179305007@muj.manipal.edu

Vineeta Tiwari
Centre for development of Advance Computing
Pune, India
vineetat@cdac.in

Sunita Singhal
School of Computing and Information Technology
Manipal University Jaipur
Jaipur, India
sunita.singhal@jaipur.manipal.edu

***Abstract*—Human Resource is the working environment inside a business that is in charge of everything master related which unites selecting, checking, picking, verifying, on boarding, preparing, advancing, paying, and terminating delegates and freely utilized substances. Human Resource is besides the working environment that stays over new request controlling how experts should be treated amidst the selecting, working, and consummation process. Here we will focus on the enrolling some bit of Human Resource. A Chatbot is an automated structure expected to begin a dialog with human customers or diverse Chatbots that gives through text. The Chatbots which is being proposed for Human Resource is Artificial Intelligence based Chatbot for major measurement profiling of contenders for the explicit task. The learning strategy utilized for the Chatbot here is assorted neural structure which includes deep learning techniques like recurrent neural network.**

Keywords—Recurrent neural network, chatbot, artificial intelligence, neural network, deep learning. Seq2Seq model.

I. INTRODUCTION

Artificial Intelligence [1] (AI) is the knowledge of machines and the part of PC innovation which targets to make it. AI alludes back to the capacity of a pc or a pc-empowered robot machine to approach certainties and get results a way much like the thought strategy for people in examining, decision making and explaining issues. The goal of AI frameworks is to handle complex inconveniences in strategies much like human rationale and thinking. Principal AI course books characterize the segment as "the analyze and format of down to earth sellers, "wherein a functional specialist is an instrument that sees its environment and takes moves which amplify its risks of accomplishment. John McCarthy, who begat the timeframe in 1956, characterizes it as "the innovation and designing of making smart machines.

Computer based intelligence [6] is a general idea that incorporates various (regularly covering) disciplines. These draw upon learning and systems from arithmetic, insights, software engineering and space explicit mastery to make models, programming projects and devices. These product projects and apparatuses can attempt complex assignments with results that are practically identical, if worse, to customary manual methodologies.

Artificial [1] neural systems are an endeavour at displaying the data preparing capacities of sensory systems. In this way, above all else, we have to consider the fundamental properties of organic neural systems from the perspective of data handling. This will enable us to configuration conceptual models of Artificial neural network (ANN) systems, which would then be able to be re-enacted and examined.

Despite the fact that the models which have been proposed to clarify the structure of the mind and the sensory systems of certain creatures are distinctive in numerous regards, there is a general agreement that the embodiment of the task of neural gatherings is "control through correspondence" Animal sensory systems are made out of thousands or a large number of interconnected cells. Every single one of them is an extremely unpredictable game plan which manages approaching sign from multiple points of view. In any case, neurons are somewhat moderate when contrasted with electronic rationale entryways. These can accomplish exchanging times of a couple of nanoseconds, while neurons need a few milliseconds to respond to a boost. Overall, the cerebrum is equipped for tackling issues, which no computerized PC can yet effectively manage.

Huge and progressive systems administration of the cerebrum is by all accounts the basic precondition for the development of cognizance and complex conduct. Up until this point, be that as it may, researcher and nervous system specialists have focused their exploration on revealing the properties of individual neurons. Today, the components for the creation and transport of sign from one neuron to the next are surely known physiological marvels, however how these individual frameworks coordinate to shape complex and hugely parallel frameworks fit for inconceivable data handling accomplishments has not yet been totally explained. Arithmetic, material science, and software engineering can give priceless assistance in the investigation of these mind boggling frameworks. It isn't astounding that the investigation of the mind has turned out to be a standout amongst the most interdisciplinary regions of logical research as of late.

The primary distinction between neural systems and regular PC frameworks is the enormous parallelism and repetition which they abuse so as to manage the inconsistency of the individual registering units. Additionally, natural neural systems are self-arranging frameworks and every individual neuron is [5] likewise a fragile self-sorting out structure equipped for handling data from multiple points of view.

A chatbot is a conversational programming framework that is intended to imitate correspondence capacities of an individual that connects consequently with a client. It speaks to another, cutting edge type of client help fueled by computerized reasoning by means of a talk interface. Chatbots depend on AI strategies that comprehend normal language, distinguish importance, feeling, and plan for significant reactions. For instance, it makes it simple for clients to get reactions to their questions in a helpful manner without investing their energy holding up in telephone lines or send rehashed messages. Chatbots can lessen normal taking care of time and cost of human asset. In any case, it is difficult to accomplish these functionalities, as it requires different complex communications between frameworks. The generative model chatbot is utilized for the improvement of brilliant bots that are very best in class in nature. This sort of chatbot is all

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

around infrequently utilized, as it requires the usage of complex calculations. Generative models are nearly hard to assemble and create. Preparing of this kind of bot requires contributing a ton of time and exertion by giving a large number of models. This is the means by which the profound learning model can take part in discussion. In any case, still, we can't make certain what reactions the model will produce.

II. LITERATURE REVIEW

A. chatbot overview

A chatbot is a conversational programming system which is intended to impersonate correspondence limits of an individual that discusses thusly with a customer. It speaks to a new, current type of human asset help controlled by artificial intelligence by means of talk interface.

Chatbots depend on AI procedures that get it normal language, distinguish importance, feeling, and structure for significant reactions. For instance, it makes it simple for Human Resource the board to get reactions to their questions in an advantageous route without investing their energy holding on to complete it physically or send rehashed messages. Nonetheless, it is difficult to accomplish these functionalities, as it requires different complex associations between frameworks.

B. Scientific classification of chatbot

Chatbot applications can be gathered into four extraordinary classes, to be specific administration, business, diversion and warning chatbot [17]. Administration chatbots are intended to give offices to clients. For instance, coordination's firm to react to inquiries concerning their representative subtleties and give the important data through texting channel rather than messages or telephone calls.

As indicated by [19] [20], chatbot application can be requested into two social events, for instance, task-masterminded and non-task-arranged. Undertaking centered chatbots hope to help the clients with finishing certain undertakings and have short trades. For example, Siri, Google Now, Alexa talk experts can give travel headings. Then again, Non-task-orchestrated chatbots center around conversing with customers to FAQ's and actuation.

In this paper, we segregated chatbot applications into four gatherings, for example, objective based, information based, administration based what's more, reaction created based as shown in Figure 1. The focal point of this exploration is on generative based chatbot. In this paper, there are diverse produced put together models that depend with respect to four classes to be explicit— Format based Model, Generative Model, Retrieval-based Model and Web document Model as showed up in Figure 2.

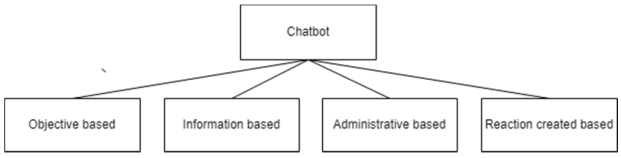


Fig. 1: Scientific classification of chatbot

III. RELATED WORK

There have been two or three models shown by specialists in past years. Ongoing advancement in profound adapting, ANN system models have appeared and guarantee for structure self-learning chatbots. Be that as it may here have been a couple of related undertakings to address the seq2seq model issues with significant learning philosophies, for example, repetitive neural systems (RNN), profound neural systems (DNN) and convolutional neural systems (CNN). Goodfellow [3] has sorted AI into three methodologies:

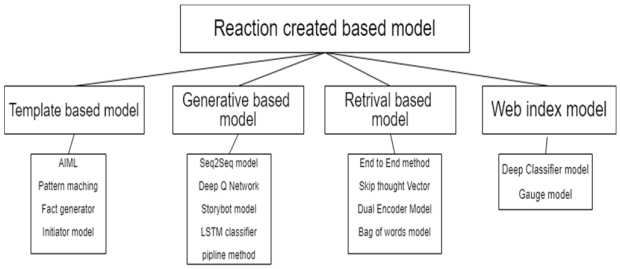


Fig. 2: Classification of generative based model

A. Knowledge Base

Most early work in AI can be referenced in this structure. Learning-based structures have been helping people to oversee issues which are dispassionately dangerous, anyway basic for machines. These issues regularly are effectively spoken to with a lot of formal principles. A case of this could be Mycin [2] which was a device created at Stanford University in 1972 to treat blood diseases [4]. Mycin was based on standards and had the option to recommend an appropriate treatment plan to a patient with blood contaminations. Mycin would request extra data at whatever point required, making it a strong instrument for now is the ideal time. While Mycin was at standard with therapeutic specialists of the day, it fundamentally tackled principles. These checks were required to be made formally, which was an incredible endeavour. Consequently, this methodology confines any AI model to one specific, restricted space, notwithstanding being hard to improve. This might be a reason that none of these endeavours has provoked an important accomplishment [3].

B. Machine Learning

AI attempts to overcome the hindrances of hard-coded benchmarks of the Knowledge Base way to deal with overseeing AI. Artificial intelligence can expel structures from data rather than relying upon principles. Basic Machine Learning systems like straight relapse and credulous Bayes techniques become familiar with the relationship among highlights and the yield class or esteem. They have been utilized to make basic models, for example, lodging value forecast and spam email discovery. AI procedures enabled machines to see some information of this present reality.

The expectations rely upon connection among highlights and yield esteem. These methods, however, the forecasts rely upon relationship among highlights and yield esteem. These strategies, nonetheless, are limited to the highlights, which are planned by the modeler, which again can be a troublesome assignment. This basically implies every element ought to be spoken to as a lot of highlights. Think about the issue of face discovery for instance. The modeler can address a face with numerous highlights, for instance, having a specific shape and structure, be that as it may, this is difficult to appear on a pixel-to-pixel premise.

Another detriment of this procedure is that the depiction of data is basic. Consider a request undertaking of segregating

two components by delineation a line between them. This assignment is vast on the Cartesian portrayal at any rate is direct for polar delineation. To accomplish the best wants, the modeler needs to experience the method of highlight arranging, which joins tending to the information for a model as a pre-processing step. Both database and AI methodologies foresee that we ought to have basic space learning and tendency one reaction to this issue is to utilize AI to find the mapping from delineation to yield, yet moreover the depiction itself [3]. This is the spot delineation learning comes into the image.

C. Representation Learning

The requirement for portrayal taking in originates from the confinements of unbending nature of 24 knowledgebase and AI approaches. We need the model to have the option to get familiar with the portrayal of information itself. Learned portrayals regularly result in much preferred execution over can be gotten with hand-structured portrayals [3]. Think about the case of face ID. As people, we can see a face from various overview centres, unquestionable lighting conditions, arranged facial highlights, for instance, scenes or hairs. This delineation of information is dynamic moreover, can be thought of as a pecking requesting of simple to complex insights which draw in us to recognize various information that we experience. In any case, this information is for all intents and purposes hard to appear by virtue of the haphazardness.

Deep Learning attempts to overcome this test by imparting complex depictions to the extent less troublesome portrayals [3]. Profound Learning is a subset of delineation getting, having different layers of neurons to learn portrayals of information with various pieces of considering [5]. Deep learning models the human character, with cerebrum neurons like managing units and the likelihood of the relationship between the neurons eagerly taking after burdens. Deep Learning[7] design is like an Artificial Neural Network (ANN), however with persistently concealed layers (as such, more neurons) which engages us to show the more multifaceted segments of our brains.

D. Early Approaches

There have been a couple of models displayed by authorities in past years. Ongoing advancement in profound adapting, profound neural system models have appeared and guarantee for structure self-learning chatbots. Be that as it may there have been a couple of related undertakings to address the seq2seq model issues with significant learning strategies, for example, intermittent neural systems (RNN), profound neural systems (DNN) and convolutional neural systems (CNN) [8].

Each assignment comprises of a few setting query-solutions. It arranged and discharged by Facebook. Each errand expected to test an outstanding bit of thinking what's more, towards testing a particular most remote reason for QA learning model. The result shows that this procedure much of the time neglecting to meet desires concerning various approaches, for example, dynamic memory systems, start to finish systems. Be that as it may, it would in general do well on undertakings with true/false inquiries. The Authors proposed that the model may be enhanced on the off chance that it is supplanted with a consideration system that treats sentences freely. Furthermore, sentence choice diagram can be supplanted by an increasingly perceptive sentence determination module with learnable weight.

One investigation [11] presented a consideration instrument that permits DNN to concentrate on various pieces of their information. Looked for after by [12] the makers structure an end that the utilization of a fixed-length setting vector is defective for disentangling long sentences. Their system was multilayered Long Short-Term Memory (LSTM) with an obliged vocabulary. They utilized one LSTM to portray information advancements to a vector of fixed estimation and after that another basic LSTM to extricate up the objective groupings from the vector. When they accessory an information sentence with an objective sentence, each word is a long way from it's looking word. The customary division between relating words in data and the target sentence is unaltered. In the midst of setting up, the producers had the decision to turn the sales the words in the source sentences, regardless not the target sentences. Thusly, they presented some momentary conditions that made the learning issue a lot less difficult. The straightforward trap of turning around the words in the source sentences are the key duty of their work. The yielded delayed consequence of their work got a BLEU score of 34.81 on the WMT'14 dataset to reranks1000 speculations. It was related with four basic LSTM layers utilizing a shaft look decoder and 1000 parameters at each LSTM layer. The outcome bolstered that the viewpoint would clearly well on other seq2seq issues. At long last, they exhibited that a basic, clear and the unoptimized approach could beat a make LSTM structure. the unoptimized approach could beat a make LSTM structure.

In the seq2seq approach, the decoder needs to screen the yield and caused substance to can be changed from the crucial parts in the source. Be that as it may, standard seq2seq model fights with making long reactions since it needs to screen everything. Besides, the decoder has fixed-length verified state vector which prompts cluttered or despite conflicting yields. To fight this, one of the most recent assessment [15] has presented chatbot's reaction age issue through a reasonable methodology, alluded as seq2seq with impression model and stochastic shaft seek unraveling procedure. Creators' characterized information the strategy is the trade history and the yield movement is the reaction. At the fundamental, glimpse model included on the encoder side and after that readied dataset on fixed-length domains of the goal side. It engaged scaling up needing to increasingly unmistakable datasets without running into any memory issues. Second, to make long, unfaltering and different responses using MAP-disentangling of the segment look structure. It is to seclude the reranking over shorter bits and re-rank domain by-section. Consequently, infusing assorted variety prior amid the unraveling procedure. At long last, they shaped target-side-considered framework along with the decoder so it can screen what has passed on. At last, they coordinated target-side-consideration system into the decoder so it can monitor what has produced. At last, creators have arranged on a joined dataset of over 2.3B talk messages from the web. In human, evaluation thinks about showed that their technique passed on longer responses with a higher degree surveyed as commendable and astonishing. In [13], the creators proposed although based seq2seq segment. They redesigned the top checked vector at the decoder side with a weighted ordinary of the encoder secured vectors. The stacks can be settled through a summed up framework where the weight structure is a touch of the parameter to be told By including regard for the immense seq2seq model, the creators had the option to all the more likely adjust data sources and

yields. Another examination [14] proposed Deep Learning which reliant on consideration based seq2seq RNN [9] for inquiry seeing, advancement recommendation and customer cooperation. The writers previously connected the seq2seq model in Deep Learning to handle and change client examination concerning. The rethink is submitted to the recommended structure to recoup an enormous measure of specific answers. What's more, they evaluated the seq2seq model to score and pick a superior mentioning answer. The outcome displayed that with the idea part and LSTM the model could change demand with better quality reviewed by BLUE score. At long last, to make dynamic client interest, they amassed a model of a chatbot, which can show mentioning that fabricates data gain, taking into consideration an increasingly productive client expectation ID process. Toward the completion of their examination, they overviewed by BLEU and human judge examination. Both exhibit noteworthy enhancements contrasted and current best in class frameworks. Be that as it may, it is steady research and required progressively indispensable improvement and examination. This examination should need to recognize this open passage and improve their work with a solid examination, which can check better gainfulness for a client to get the data they need. Another investigation presented a novel advancement for seq2seq learning with a Deep Q-Network (DQN) [10] which interprets the yield storing up iteratively. In every complement, an encoder-decoder LSTM used to thusly make edifying features to address the inward states and detail a snappy review for DQN. The snappy outline contains words probabilities at each time step and DQN makes sense of how to settle on decision on which movement will be browsed the once-over to modify the present progression.

For assessment, a straight forward and powerful technique for disentangling look as proposed by [15] to convey shaft seek calculation. At each timestep, the decoder builds up each isolated sentence in the bar search for with every conceivable word in the vocabulary finally the model was set up to discharge up 10,000 regular sentences. Their starters showed that when showed up contrastingly in connection to a section look LSTM [18] decoder the proposed procedure performed truly well. While deciphering sentences from the preparation set, it fundamentally outflanked, as far as BLUE score acquired. It was settled subject to the closeness between the objective sentence and the decoded yield sentence after DQN makes a move. It was settled subject to the closeness between the objective sentence and the decoded yield sentence after DQN makes a move.

ELIZA is one of the first ever chatbot programs formed. It uses sharp composed by hand formats to make answers that take after the customer's data articulations. Starting now and into the foreseeable future, limitless hand-coded, rule-based chatbots have been made. In addition, different programming frameworks unequivocally planned to support building talk administrators have been made.

These chatbot programs are fundamentally the same as in their centre, in particular that they all utilization manually written standards to produce answers. Generally, straightforward example coordinating or watchword recovery procedures are utilized to deal with the client's info expressions. At that point, rules are utilized to change a coordinating example or a watchword into a predefined answer. A simple example is shown below in AIML.

<category>
<pattern>What is your name?</pattern>
<>template>My name is Alice<template>
</category>

Here if the info sentence coordinates the sentence composed between the sections the answer composed between the sections is yielded. Another model is appeared beneath where the star image is utilized for supplanting words. For this situation whatever word pursues the word like it will be available in the reaction at the position determined by the token:

IV. METHODOLOGY

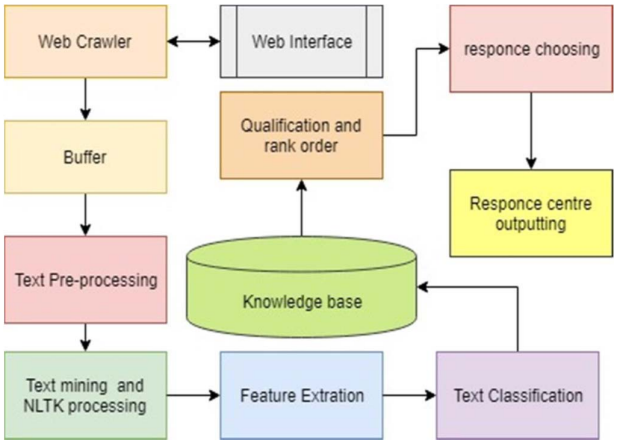


Fig. 3: Architecture for the proposed system

The proposed framework figure 3 starts with a web crawler with the capacity to get plain substance from the web where they move the profiles for applying for the action. The web crawler used here is web spider for extracting the concerned keywords from the uploaded profiles. So as to keep away from capacity limit issues, buffering has been utilized. The assistance enables the web crawler to keep the quantity of pages inside the memory obstruction by controlling the season of new pages. The plain substance of various profiles is pre-dealt with to crash unwanted pictures, for instance, emphases, stop words, or non-English letters and words. After pre-setting up, the substance is mined to give split sentences for the entire substance. By using python library the sentences are part into individual words and at some point later pos-named into talk parts. It goes through the following stages namely lower case, punctuation removal, and stop words removal, spelling correction, tokenization, stemming and lemmatization.

Then, different feature extracted as if personal details, qualification, work experience, etc what's more, place them into the database. The sentences are rankly coordinated after appraisal as per the highlights removed. Intimating the rank sales, the best responses for the perfect objective can be picked and send to the HR division. In order to create a chatbot, or really do any machine learning task, of course, the first job we have is to acquire training data, then we need to structure and prepare it to be formatted in an "input" and "output" manner that a machine learning algorithm can digest. Arguably, this is where all the real work is when doing just about any machine learning. The building of a model and training/testing steps are the easy parts.

A. Jacard's Coefficient

Jaccard's coefficient appraises the closeness between two instructive records by separating the quantity of typical properties between the researched sets by all smaller person of highlights [23]. For model, if X and Y are two sets, by then Jaccard's coefficient between them is:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

The methodology used in this framework is n-grams which are accessible inside the Python library. N-grams is a module that consolidates the quantity of words in a sentence or a substance and considers each word a gram. There is a section in the N-grams module which awards finding the intersectional words between two strategies of words.

B. Mean Reciprocal Rank

The outcomes are assessed utilizing Mean Reciprocal Rank (MRR) technique which is progressively appropriate to quantify the execution of the framework actualized. MRR is determined identifying with the accompanying connection:

$$MRR = \frac{1}{n}\sum_{1=1}^{n}\frac{1}{r_i} \quad (2)$$

Where,
MRR is Mean Reciprocal Rank.
n is number of queries.
i is individual query number.
r_i the reciprocal rank of the right solution.

V. Experiments and Results

The python libraries used for implementing the chatbot are absolute import, division, print function, Unicode literals, pytorch, csv, randon, re, os, unicodedata, codecs, itertools, math. After importing the libraries we have load the data and preprocess the data. To make the bot learn we have taken the HR Corpus dataset from Kaggle, which is rich obvious dataset of employee information. This dataset is huge and differing, and there is an incredible variety of language custom, timeframes, supposition, and so forth. Our expectation is that this assorted variety makes our model hearty to numerous types of information sources and questions. We have made a pleasantly organized information document in which each line contains a tab secluded request sentence and a response sentence pair.

The figure 4 flowchart explains the implementation of the proposed system. Next we change the Unicode strings to ASCII and all letters to lowercase and lastly trim all non-letter characters, with the exception of essential accentuation. At last, to help in preparing assembly, we split the text into sentences and moved through sentences with length more imperative than max length limit. Diminishing the component space will be in like manner loosen up the inconvenience of the limit that the model must make sense of how to deduced which is accomplished by procedure to trim words and filter out pairs with words and trim the symbol from the dataset.

Now we prepare data for models one approach to set up the handled information for the models is the seq2seq interpretation we utilize a bunch size of 1, implying that we ought to just change over the words in our sentence sets to their looking at documents from the vocabulary and feed this to the models. To accommodate sentences of different sizes in a comparative cluster, we made our bundled data tensor of shape (max_length, batch_size), where sentences shorter than the max_length are zero padded after a POS tag_token. Subsequently, we transpose our information group shape to (max_length, batch_size), with the goal that ordering over the principal measurement restores a period venture over all sentences in the clump. We handle this transpose verifiably in the zero Padding capacity. It furthermore reestablishes a tensor of lengths for all of the courses of action in the bundle which will be passed to our decoder later. Train Data basically takes a lot of sets and returns the info and target tensors utilizing the previously mentioned capacities. The target of a seq2seq model is to recognize a variable-length plan as data and return a variable-length gathering as a yield using a fixed-sized model.

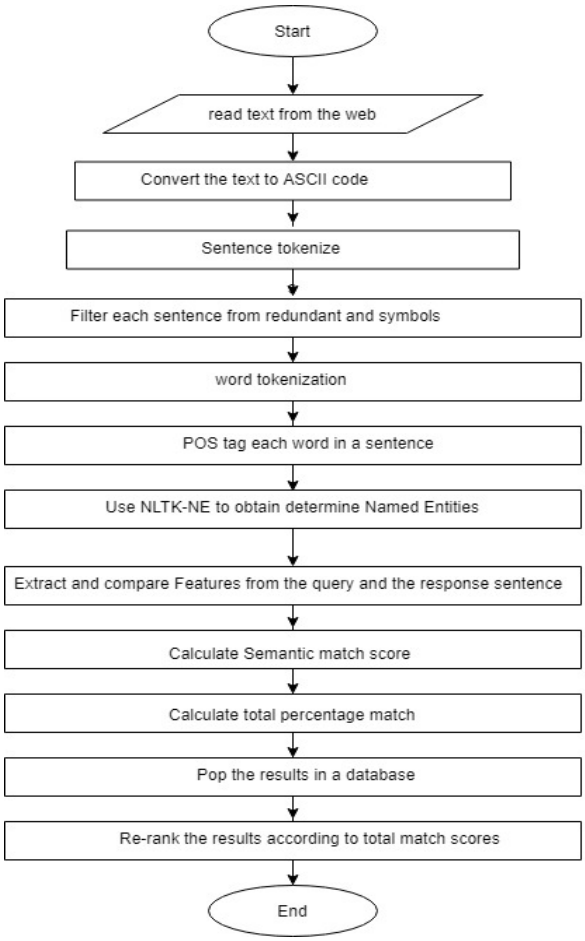


Fig. 4: Flowchart of implemented system

By using two separate abundance neural nets together, we can accomplish this endeavor. One RNN goes about as an encoder, which encodes a variable length data social event to a fixed-length setting vector. On a fundamental level, this setting vector will contain semantic information about the sales sentence that is a promise to the bot. The second RNN is a decoder, which takes a data word and the setting vector, and returns a hypothesis for the going with the word in the technique and a peddled state to use in the going with cycle. The architecture of basic encoder and decoder model is depicted in figure 5.

The encoder RNN [16] goes over through the data sentence one token promptly, at each time step yielding a "yield" vector and a "shrouded state" vector. The verified state vector is then sat back headway, while the yield vector is recorded. The

encoder changes the setting it saw at each point in the get-together into a tremendous measure of centers in a high-dimensional space, which the decoder will use to pass on a basic yield for the given endeavor. At the point of convergence of our encoder is a multi-layered Gated Recurrent Unit. We will use a bidirectional variety of the GRU, suggesting that there are essentially two self-directing RNNs one that is reinforced the data development in normal sequential mentioning and one that is animated the data strategy in switch request. The yields of each system are summed at each time step. Utilizing a bidirectional GRU will give us the upside of encoding both past and future setting.

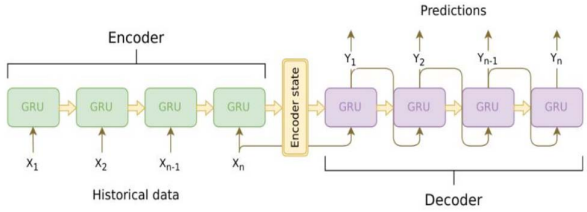
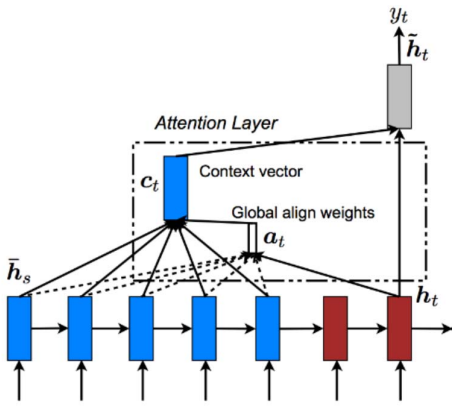


Fig. 5: Architecture model of encoder and decoder

The decoder RNN makes the response sentence in a token-by-token course of action. It uses the encoder's setting vectors, and inside covered states to make the going with the word in the party. It keeps 49 making words until it yields an EOS_token, talking very a long way from the sentence. A typical issue with a vanilla seq2seq decoder is that in the occasion that we depend totally on the setting vector to encode the whole information progression's noteworthiness, resolved using the decoder's current disguised state and the encoder's yields. The yield thought burdens have a similar shape as the data movement, enabling us to duplicate them by the encoder yields, giving us a weighted absolute which demonstrates the bits of encoder regard revolve around figure 6 depicts this unimaginable. Since we have depicted our thought submodule, we can execute the genuine decoder model. For the decoder, we have physically attracted our get-together one-time inclusion at the soonest opportunity. This construes our presented word tensor and GRU yield will both have shape (1, batch_size, hidden_size).

Presently we the encoder yields, giving us a weighted absolute which demonstrates the bits of encoder regard revolve around figure 5 depicts this unimaginable. Since we have depicted our thought submodule, we can execute the genuine decoder model. For the decoder, we have physically attracted our get-Ftogether one-time inclusion at the soonest opportunity. This construes our presented word tensor and GRU yield will both have shape (1, batch_size, hidden_size).



Fig, 6: Block diagram of Global attention

Generally speaking, the Global consideration system can be condensed by the accompanying figure 7. We have executed the "Consideration Layer" as an alternate nn.Module called Attn. The yield of this module is a softmax institutionalized burdens tensor of shape (batch_size, 1, max_length).

TABLE-I: QUERY COMPARISON

No.	Queries	Nearest matching Sentence	Combination Including Jaccard's Coefficient combination Matching Score %	Data order in Jaccard's combination	Combination Including cos similarity combination Matching Score %	Data order in cos similarity combination n	No. of data
1	Where are you from ?	I am from San francisco	32.6	2	43.57	1	4270
2	What is your experience ?	I have 3 years expereince	30.7	3	38.33	1	4167
3	What is your qualification ?	M Tech	31.8	3	42.45	2	4273

For cases of the questions, the nearest match sentences, moreover, the nearest match scores for the two blends in which is the mix including Jaccard's coefficient and the mix including cosine equivalence, are created in Table-I. The fundamental results have showed up in Table-I give the most surprising scored and the closest matches out of in excess of 4000 records.

The assessment results utilizing MRR are addressed in Table-II. The assessment of Table-II shows that the presentation of the proposed structure increments fundamentally by utilizing the Cosine closeness mix. The MRR of Cosine closeness blend (52.38) gives improvement by around 10 from the ordinary MRR of bot connection (42.25).

TABLE II. ASSESSMENT RESULT

No.	Combination	MMR Score
1	Combination using Jaccard's coefficient	36.12
2	Combination using cosine similarity	52.38
3	A. Moschitti, and S. Quarteroni [24].	42.25

```
> hello?
Bot: hello .
> where am I?
Bot: you re in a hospital .
> who are you?
Bot: i m a lawyer .
> how are you doing?
Bot: i m fine .
> are you my friend?
Bot: no .
> you're under arrest
Bot: i m trying to help you !
> i'm just kidding
Bot: i m sorry .
> where are you from?
Bot: san francisco .
> it's time for me to leave
Bot: i know .
> goodbye
Bot: goodbye .
```

Fig. 7: Interaction with the bot

After training we interact we the bot and above is the sample conversation that we recorded.

VI. CONCLUSIONS

With the consideration of the proposed system and other comparative study on chatbot. We have implemented seq2seq model of deep learning to have conversation and adapt self-learning. The chatbot will learn using bidirectional RNN, one as encoder and the other as decoder. The test results demonstrate that the most noteworthy scored sentences are nearest to an inquiry. Assessment results demonstrate that the framework execution ascends outright by utilizing cosine likeness metric for the lexical match. The MRR of Cosine closeness blend (52.38) gives improvement by around 10 points from the typical MRR of bot connection (42.25) The test was conducted using Cornell movie dialog dataset. In future the model can be trained with increasing no. of hidden layers to make it more accurate and no. of iterations in model training will be taken 8000.

REFERENCES

[1] J. Vanian, "Google Adds More Brainpower to Artificial Intelligence Research Unitin Canada," Fortune, 2016.

[2] B. Copeland, "MYCIN," in Encyclopedia Britannica, Inc., 2017.

[3] Y. Lecun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436, 2015.

[4] I. N. Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni and S. F. .R. Alves, Artificial Neural Networks A Practical Course, Springer International Publishing, 2017.

[5] R. Collobert and J. Weston, "A unified architecture for natural language processing:deep neural networks with multitask learning," 25th international conference on machine learning, pp. 160-167, 2008.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, April 2017.

[7] R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval based human-computer conversation system," 39th International ACM SIGIR conference on research and development in information retrieval, pp. 55-64, 2016.

[8] T. N. Sainath, A. R. Mohamad, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," IEEE international conference on acoustics, speech and signal processing, pp. 8614–8618, 2013.

[9] Y. You, A. Buluç, and J. Demmel, "Scaling deep learning on GPU and knights landing clusters", The international conference for high performance computing, networking, storage and analysis, pp. 1-12, 2017.

[10] R. Socher, Y. Bengio, and D. M. Christopher. "Deep Learning for NLP," Tutorial Abstracts of ACL 2012, pp. 5-5, 2012.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," Conference on empirical methods in natural language processing, pp. 1532-1543, 2014.

[12] J. Epstein and W. Klinkenberg, "From eliza to internet: a brief history of computerized assessment," Computers in Human Behavior, vol. 17 (3), pp. 295-314, 2001.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 3rd International conference on learning representations, pp. 1-15, 2015.

[14] I. Sutskever, O. Vinyals, and Q.V. Le, "Sequence to sequence learning with neural networks," 27th International conference on neural information processing systems, vol. 2, pp. 3104-3112, 2014.

[15] J. Weizenbaum, "ELIZA: a computer program for the study of natural language communication between man and machine," Magazine: Commincations of the ACM, vol. 9 (1), pp. 36-45, 1966.

[16] Z. Yin, K.-h. Chang, and R. Zhang, "DeepProbe: Information directed sequence understanding and chatbot design via recurrent neural networks," 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2131-2139, 2017.

[17] S. Barker, "How chatbots help," MHD Supply Chain Solutions, vol. 47 (3), pp. 30, 2017.

[18] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems: Recent Advances and New Frontiers," ACM SIGKDD Explorations Newsletter, vol. 19 (2), pp. 25-35, 2017.

[19] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu,"Using of jaccard coefficient for keywords similarity." International multi conference of engineers and computer scienctist, vol. 1, pp. 1-5, 2013.

[20] A. Moschitti and S. Quarteroni, "Linguistic kernels for answer reranking in question answering systems," Information Processing & Management, vol. 47 (6), pp. 825-842, 2011.

KeySED:An Efficient Keyword based Search over Encrypted Data in Cloud Environment

Kasturi Dhal
Dept. of Comp. Sc. & Engg.
Silicon Institute of Technology
Bhubaneswar, India
kdhal@silicon.ac.in

Satyananda Champati Rai
Dept. of Information Technology
Silicon Institute of Technology
Bhubaneswar, India
satya@silicon.ac.in

Prasant Kumar Pattnaik
School of Computer Engineering
KIIT University
Bhubaneswar, India
patnaikprasant@gmail.com

Somanath Tripathy
Dept. of Comp. Sc. & Engg.
Indian Institute of Technology, Patna
Bihta, India
som@iitp.ac.in

***Abstract*—To maintain confidentiality of certain specific data, it is required to encrypt the data by using Ciphertext-Policy Attribute-Based Encryption (CPABE) mechanism, before its upload onto a untrusted cloud server. Moreover either to search an item in a given encrypted data or to carry out required computation is not feasible without compromising the confidentiality. To ease the process of search operation over outsourced encrypted data, it is delegated to the cloud server. The CSP performs the outsourced search only when the user is authorized as per the data owner. Further, the searched result returned by the cloud server should be verified by the user. To address the above issue we propose a fine grained attribute based keyword search model. The authenticity of the user is verified by the cloud server using CPABE technique. Further only authorized users can generate valid trapdoor messages which are used later to perform search operation by the cloud server. The proposed model takes comparatively less computation time in comparison to the considered related existing models. The empirical results and mathematical proof of the proposed model shows that it would be suitable for integration into resource constraint devices.**

***Index Terms*—KeySED, Keyword search, Cloud Server, Encryption**

I. Introduction

Due to the advancement of technologies, the huge amount of data generated per second. However, to store and process such a huge amount of data, the user prefer to outsource the data to the cloud server. Confidentiality and privacy are priority for health and military data. To maintain confidentiality, the data owner outsources the data in encrypted form. However, encrypted data restricts the cloud server to execute some operations like search. However, as the cloud server is untrusted the user should verify whether the returned result is correct or not.

We use CPABE technique [1] to achieve access control in keyword based search which allows the data owner to decide flexible access control policy to search. Further based on the access control a user is allowed to perform search operation. Further we try to slightly enhance the performance of the keyword based search model to make it accessible using resource constraint devices.

978-1-7281-2087-4/19/$31.00 © 2019 IEEE

A. Related Work

To facilitate keyword based search operation, the data owner generates tokens which subsequently used by the cloud server to search on behalf of the user. Further many models of keyword based search have been proposed with the aim to improve efficiency along with additional functionalities. The existing keyword based search models are classified into following two categories *Keyword based search on symmetric key encryption* [2]–[6], [8]–[10] and *Keyword based search on asymmetric key encryption* [11]–[15].

Yang et al. [17] first introduced a hybrid model including the features of attribute based encryption and keyword based search to achieve fine grained access control on search operation so that the search operation is outsourced to the remote server without compromising confidentiality and privacy. Further the proposed model also supports user revocation.

Yang et al. [16] introduced a keyword based search model where a user can delegate its partial access rights to other users to perform keyword based search on the encrypted data using proxy re-encryption mechanism. Miao et al. [18] introduced attribute based keyword search over encrypted data to support multiple owners along with multiple keywords used for search operation.

Though the existing keyword based search models [19]–[21] achieve verifiability with fine grained access control but with high computational cost. Still there is scope not only to improve performance of keyword based search model with additional functionality but also the search result must be verified by the user.

B. Our contribution

In this paper we investigate various directions in which keyword based search model for encrypted data can be refined.

The proposed model tries to upgrade the keyword based search model. The contributions of our model are outlined as follows:

- We proposed a comparatively efficient keyword search model. Further, attribute based encryption technique is used to achieve fine grained access control. The user must generate valid trap message if the user is authorized by the data owner to preform search operation.
- The user delegates the search operation to the cloud server. The search operation is performed without compromising confidentiality and privacy of the encrypted file. The cloud server should not find the correlation between files and keywords that is which files belong to a particular keyword. To reduce the search time of the cloud server the bloom filter is used.
- We implement the proposed model and compare its computational complexity with existing models. The experimental results depicts the performance of our model is comparatively computationally efficient than existing models.

C. Organization

The organization of the rest of the paper outlined as follows. In Section 2 the mathematical backgrounds are briefly discussed. We introduced the proposed scheme's system model in section 3. We investigate the computational cost of the suggested model and compare it with other existing models in section 4. Section 5 finally conclude the model with future scope.

II. Preliminaries

The section briefly outlines mathematical backgrounds required to develop and analyze cryptography and keyword based search schemes.

A. Bilinear Map

Let G and G_T are two cyclic groups having prime order p, and g is a generator of G. A map $e : G \times G \rightarrow G_T$ is said to be bilinear if the following properties are satisfied as per [1]:

1) Bilinearity: $\forall\ a, b \in Z_P$ and $u, v \in G, e(u^a, v^b) = e(u, v)^{ab}$.
2) Non-degeneracy: $e(g, g) \neq 1$. The generator of the group G must not map to identity element of the group G_T.
3) Computability: $\forall \in u, v \in G_1$ an computationally efficient algorithm exists to compute $e(u, v)$.

The Bilinear Map is a mathematical model to map two elements of one group to an element of another group. This mapping is mostly used in some cryptographic model to increase complexity of security algorithm so that it can not be broken by the intruder.

B. Access Structure

Let $\{P_1, P_2, \ldots, P_n\}$ be set of parties which represents an attributes set. A collection $\mathbf{A} \subseteq 2^{\{P_1,\ldots,P_n\}}$ is said to be monotone if $\forall B, C$ if $B \in \mathbf{A}$ and $B \subseteq C$ then $C \in \mathbf{A}$. An access structure is a set $\mathbf{A}$ which is non empty subsets of $\{P_1, P_2, \ldots, P_n\}$. The elements in $\mathbf{A}$ are authorized and the elements not belongs to set $\mathbf{A}$ are unauthorized [1]. Any access structure represented as a boolean function. An access tree is used to represent a boolean function where leaf nodes represent attributes and internal nodes represents logical operators like AND, OR.

C. Linear Secret Sharing Scheme(LSSS)

The objective of LSSS is to share a secret among the users so that the secret can only be extracted by the authorized users. Every monotonic access structure has an equivalent linear secret sharing scheme. The following properties makes a Secret Sharing Scheme π [23] over a set of parties P to be linear secret sharing scheme.

1) Each party computes the share of s in the form a vector over Z_P.
2) Let us consider share-generating matrix $\mathbf{M}$ of order $(n \times l)$. The share generating matrix $\mathbf{M}$ is used to generate share for the parties. The row $i \in [1, \ldots, n]$ of share generating matrix labeled by a function $\rho(i)\colon [1, \ldots, n] \rightarrow P$. Let $s \in Z_P$ be the secret to be shared among authorized parties. The data owner choose a random vector $\vec{v} = [s, r_2, \ldots, r_n]$, where $r_2, \ldots, r_n \in Z_P$ are random values from Z_P. The data owner computes $\vec{\lambda} = M\vec{v}$, is a vector containing n shares of secret s for each authorized party according to linear secret sharing scheme. The party $\rho(i)$ share computed as $M_i\vec{v}$.

Further, the following linear reconstruction property is attained by every linear secret sharing scheme.

1) Let an access structure $\mathbf{A}$ is represented by $(\mathbf{M}, \rho)$ and π is a LSSS for $\mathbf{A}$. Let $I = \{i : \rho(i) \in S\}$ be the set of rows in $\mathbf{M}$ which are labeled by attributes in the authorized set S. Then there exists constants $\{w_i \in Z_P\}_{i \in I}$ such that $\sum_{i \in I} w_i \lambda_i = s$. As a result of which only authorized set can able to derive secret s but unauthorized sets can not able to derive secret s using their shares.

III. System Model

This section gives a brief description of the working principle of the proposed model. The role of each entity involved in the proposed model also presented in this section.The flow diagram of the proposed model shown in Fig. 1. The following entities participated in the model: *Data Owner*, *User*, *Cloud service Provider*.

DataOwner: To manage and process the data efficiently the data owner encrypts the data and outsources it to the cloud server. Data owner shares data only with the authorized users. The user facilitated by the data owner to perform a keyword-based search. To enhance the search operation index of the file is generated using keywords. Though the search operation outsourced, the data owner specifies a policy to select the authorized users to perform a search operation using access control policy.

User: With a objective to reduce computation time of user, the keyword-based search is delegated to the cloud server. But without reducing confidentiality. The user computes the trapdoor message. Subsequently used by the cloud server to

perform keyword-based search. Only authorized users decrypt the result.

Cloud Service Provider: The cloud server is responsible for a search operation on the instruction of the user as pay use basis. If the trapdoor message is valid, then the cloud server sent the files containing the keyword. Though the cloud service provider is untrusted, it tries to behave as per the service level agreement (SLA). The cloud server tries to gain as much information for the files associated with a keyword. To minimize the computation cost the cloud server may send the previously computed search result.

IV. Proposed Model

The working principle of the proposed keyword based search scheme with fine gained authorization control as decided by the data owner is presented in this section. Let total n attributes $Att = \{a_1, a_2, \ldots, a_n\}$ are in the system.

Let G is group with bilinear property having order P and a generator g. Let $H : \{0,1\}^* \leftarrow Z_p$ be a collision resistance hash function. The working principle segregates the proposed model into following phases 1) *System Initialization* 2) *Secret Key Generation* 3) *Encryption* 4) *Search* 5) *Authorization Validation*.

In detail description of each phase of the proposed model are given as follows.

Phase 1:System Initialization

Setup$(k, Att) \rightarrow (PK, MSK)$. The attribute authority begins the system using setup() process. It takes security parameter k and attribute universe Att as input to produce public key(PK) and master key(MSK). The algorithm first chose a bilinear group G having prime order P based on security parameter. Further, the process chose a random element $v_j \in Z_P$ for each attribute $j \in Att$. Then the setup process computes PK as per equation (1)and MSK as given in equation(2). Finally, setup process published the PK but the MSK is kept as secret.

$$PK = \left\{G, g, \{pkatt_j = g^{v_j} \forall j \in Att\}\right\} \quad (1)$$

$$MSK = \left\{\{skatt_j = v_j \forall j \in Att\}\right\} \quad (2)$$

Phase 2:Secret Key Generation

Keygen$(PK, MSK, S) \rightarrow SK$. The attribute authority generates the secret key for a user based on the attributes $S \subseteq Att$. The Keygen process compute the user's secret key SK by taking PK, MSK and user's attribute set S as input to compute the user's secret key SK. The Keygen process chose a random number $r \in Z_P$, unique for each user in the system. Then the secret key is computed as per equation(3). Finally, the secret key SK sent to the user. The SK used in the process of Trapdoor generation process and authorization validation process.

$$SK = \left\{D = g^r, \forall j \in S : D_j = g^{r/v_j}\right\} \quad (3)$$

Phase 3: Encryption

The confidentiality of the outsourced data is maintained by the data owner using the encryption techniques. In the proposed model, the authorized user able to perform a keyword-based search. The data owner assigns each file F with an unique identifier F_{id}. Then keyword set KW identified from each file F to perform the search. The data owner encrypts each file F using $AES.Encrypt(F, SymKey)$ with a symmetric key $SymKey$ denoted as F_{ENC}. Further, to facilitate the data sharing process, the data owner encrypts the $SymKey$ using the CPABE technique and attach as the header of the file.

The data owner executes IndGen process to compute the index for each file used in susequent search process. The index is related to policy to facilitate only authorized user to search.

IndGen$(PK, (M, \rho), KW) \rightarrow CI$:

The data owner runs index generation process to produce the index ciphertext CI as output by taking PK, access control policy $\mathbf{A}$ and list of keyword W as input. The data owner first converts access control policy to an equivalent monotone access control policy (M, ρ). The data owner first chose a random number $s \in Z_p$. Let Y be set of attributes associated with the each row of M. To share the secret s IndGen process chose random numbers $s_2, \ldots, s_r$ where r denotes number of rows in the matrix M. Then it generates a random vector $\vec{v} = (s, s_2 \ldots, s_r)$. The share of s for each row M_j computed as $\forall j \in Y \lambda_j = M_j v_j$ where M_j denotes j^{th} row of M.

Algorithm 1 IndexGen$(PK, (M, \rho), KW)$

```
for Each File F do
    for Each row M_j ∈ M do
        C_j = g^{V_j λ_j}
    end for
    Chose a random number s
    Chose m bit bloom filter and n psudo random function
    f_1(), ..., f_n()
    for Each Kewword KW_j do
        C'_j = g^{sH(KW_j)}
        for i ← 1, n do
            x_{KW,i} = f_i(KW, key)
            Insert x_{KW,i} to the F_id bloom filter
        end for
    end for
    ACC = {(M, ρ), C_j}, KC = {C'_j}
    CI = {ACC, KC}, (bloomfilter, F_id)}
end for
```

The algorithm(1) gives brief details of steps to compute index ciphertext(CI). Further, key is the key for the bloom filter shared with the user. The process choses n different hash function $f_1, \ldots, f_n$ for the insert operation of the bloom filter. The index ciphertext consists of two part. One part of the index ciphertext is associated with the access control policy known as Authorization Control Ciphertext (ACC). The other one is related to keywords known as Keyword Ciphertext (KC). Lastly, the encrypted file alongwith Index Ciphertext stored in

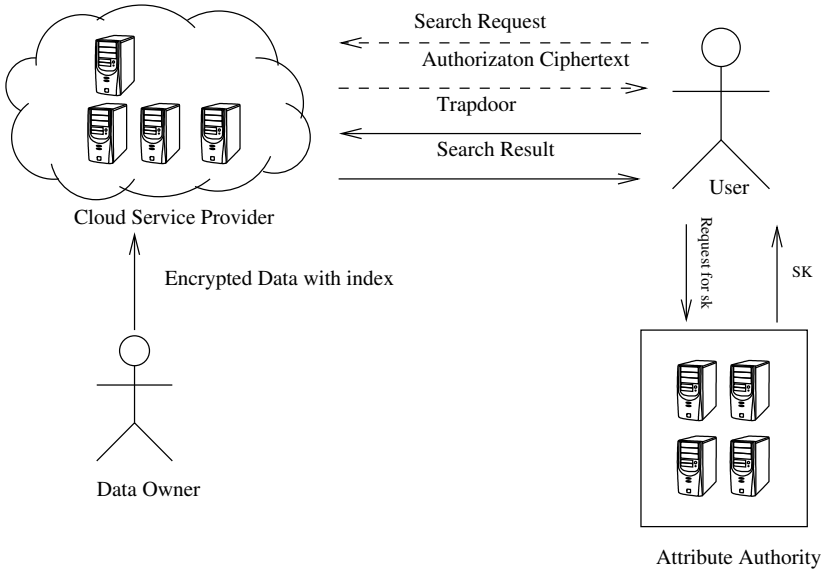


Fig. 1. System Model

the cloud server. The model achieves both data sharing and keyword-based search.

To computation time of search operation in the proposed model is reduced due to the usage of the bloom filter [25] [26] [27]. Each keyword of a file mapped to the bloom filter of that file. Before the process starts, the data owner generates search key key with the authorized user.

When the user requests for the search for the first time, the cloud server sent authorization control ciphertext(ACC) if the user is authorized to search.

Phase 4: Trapdoor Generation

The user generates a trapdoor message TD using the process described next. The trapdoor message TD is used later in the search process.

Trapdoorgen$(SK, KW_j, key) \rightarrow TD$:

The process takes theSK, keyword $K\tilde{W}_j$ and keyword based search key key as input to results the trapdoor message (TD). The codeword for each keyword computed. The details steps to generate the trapdoor message given in the algorithm(2). Finally, the user sends it to the cloud server subsequently.

Algorithm 2 Trapdoorgen(SK, AVR, KW_j, key)

1: **for** Each Psudorandom Function $f_i()$ **do**
2: $x_{KW_j,i} = f_i(KW_j, key)$
3: **end for**
4: $TD = \{x_{KW_j}\}$

Phase 5: Search

The keyword has not disclosed by the user to the cloud server during the process of outsourcsed search process. The encrypted file with the keyword $K\tilde{W}_j$ is searched by the cloud server using the following process after receiving Trapdoor from the user.

Search$(CI, TD) \rightarrow Searchresult$. The cloud server initiates the search process to return the search result to the user. The process takes index ciphertext CI along with Trapdoor message TD as input to search the file in encrypted form. Bloom filter is searched to reduce the search time rather than entire cloud storage. The details steps given in algorithm 3. Finally, the search result is returned to the user. The user with decryption permission decrypts the file to check if it contains keyword KW_j or not.

Algorithm 3 search(CI, TD)

1: **if** All x_{KW_j} position of the bloom filter is set **then**
2: Return the searched file
3: **else**
4: File Not Present
5: **end if**

Phase 6: Authorization Validation The user executes the decrypt process to check whether the correct file sent by the cloud server or not without decryption. The decrypt scheme described next to produce the authorization validation result(AVR). If the result is true, then the user decrypts the file.

Decrypt$(PK, ACC, SK) \rightarrow F$ or $\perp$:

The Decrypt process produces the File F or no result by taking the public key, ciphertext and secret key of the user as input. The process verifies the correctness of the sent file as given in equation(5). If the equation(5) holds the correct encrypted file with keyword $K\tilde{W}_j$ is sent by the cloud server. Let I denotes number of attributes in the access control policy embedded in the ACC. The decrypt process chose set of constants $\sigma_j \in Z_P \forall j \in I$ to reconstruct secret as $s = \sum_{j \in I} \lambda_j \sigma_j$ using linear secret sharing scheme. The detail steps of the decryption given in algorithm 4. Equation(4)gives

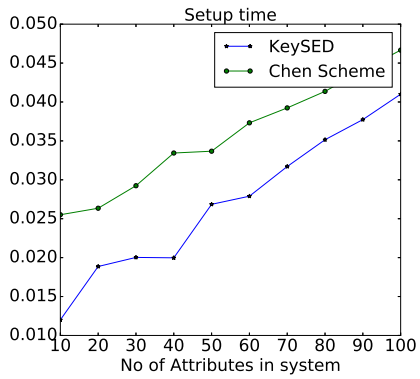


Fig. 2. Setup

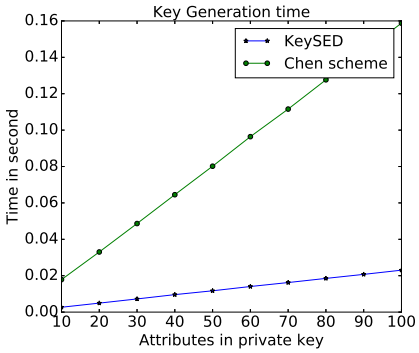


Fig. 3. KeyGen

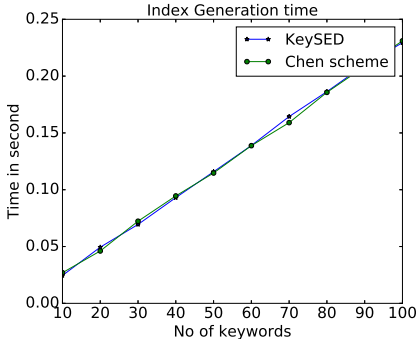


Fig. 4. index

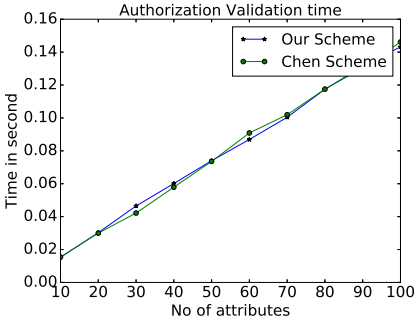


Fig. 5. Authorization

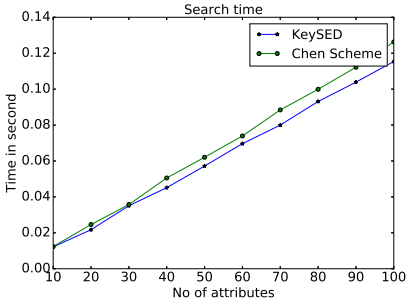


Fig. 6. Searchtime

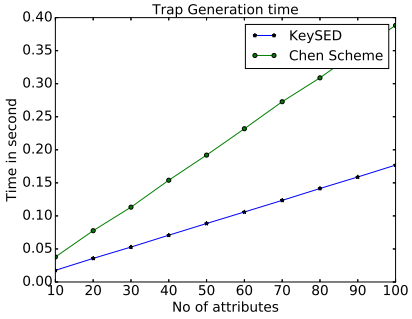


Fig. 7. TrapGeneration

steps to derive T.

Algorithm 4 Decrypt(PK, ACC, SK)

Let $I = \{\forall att \in SK\}$
$T = 1$
for Each row $M_j \in M$ **do**
  if $\rho(M_j) \in I$ **then**
    Chose a random number σ_j
    $T = \prod e(D_j, C_j)^{\sigma_j} = \prod e(g^{r/v_j}, g^{v_j \lambda_j})^{\sigma_j}$
  end if
end for
$AVR = \{T = e(g,g)^{rs}\}$
Chose a random number u
$D' = AVR^u = e(g,g)^{rsu}$
$D'' = D^{u/H(KW_j)} = g^{ru/H(KW_j)}$
if $e(C'_{KW_j}, D'') == D'$ holds **then**
  Sent file is correct
  $SymKey = CPABE.Decrypt()$
  $F = AES.Decrypt(F_{ENC}, SymKey)$
end if

$$T = \prod_{j \in I} e(D_j, C_j)^{\sigma_j} = \prod_{j \in I} e(g^{r/v_j}, g^{v_j \lambda_j})^{\sigma_j} = \prod_{j \in I} e(g,g)^{r \lambda_j \sigma_j} = e(g,g)^{r \sum_{j \in I} \lambda_j \sigma_j} = e(g,g)^{rs} \quad (4)$$

$$e(C'_j, D'') = e(g^{sH(w_j)}, g^{ru/H(\tilde{w})}) = e(g,g)^{rsu} = D' \quad (5)$$

V. Performance Analysis

In this section the performance of our proposed model is being investigated by conducting several experiments. The simulation results shows the computational time of our proposed model takes comparatively less time than the existing models.

A. Experimental Analysis

We compare our proposed model with the Chen [21] by simulating these two schemes. We execute these schemes using python Charm crypto library [22] on a Ubuntu 16.04 system with 2.40 GHz and 4 GB RAM. While simulating, we take 160-bit elliptic curve group over a 512-bit finite field to decide the security parameters of the system. The average of 30 trials are taken into consideration while comparing the simulation results. We vary the number of attributes from 10 to 100 with step value 10. We take 100 distinct keywords extracted from 500 pdf files.

Figure 2 shows the execution time taken by the setup processes.We vary the number of attributes between 10 and 100 while comparing execution time. Figure 2 depicts the setup process of our model takes comparatively less time.

It has observed from Figure 3 that the keygen algorithm of our model takes significantly less computation time than that of the chen model [21].

Figure 4 shows the comparison of computational time in Indgen process for the two schemes. Both schemes Indgen process takes almost equal time.

Figure 5 shows the comparison authorization validation result(AVR) process. The equal time required by both the schemes.

Figure 6 demonstrates the search process of the proposed model takes slightly less time in comparison with the chen model [21].

Figure 7 depicts the comparison of Trapdoor generation process computational time for the two schemes. To generate a trapdoor message our model requires only two exponent operations. The proposed model takes less time to generate trapdoor.

The simulation results show that our scheme is comparatively takes less computational time than the scheme in [21]. The proposed model is more suitable for real life keyword based search application.

VI. Conclusion

A Keyword search model for resource constraint devices with control proposed. Whether the user has the right to perform the search operation or not is verified using the authorization control method. The computational cost of the proposed model different schemes compared with the scheme in [21]. The simulation results ensures that the proposed model is appropriate for real life application to search over the encrypted data using resource-limited devices. In this work, a keyword-based search model presented, which is suitable in the untrusted server like cloud servers. As future work, the proposed model extended to verifiable keyword-based search model. Further, the proposed model may be improved to support flexible authorization control with attribute revocation.

References

[1] J. Bethencourt, A. Sahai and B. Waters, "Ciphertext-policy attribute-based encryption," in IEEE symposium on security and privacy(SP'07), pp. 321–334, May 2007.

[2] S. D. Xiaoding, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data," in Proceeding 2000 IEEE Symposium on Security and Privacy, pp. 44–55, 2000.

[3] C. Y . Cheng and M. Mitzenmacher "Privacy preserving keyword searches on remote encrypted data ," in International Conference on Applied Cryptography and Network Security, pp. 442–455, June 2005.

[4] R. Curtmola, J. Garay, S. Kamara and R. Ostrovsky,"Searchable symmetric encryption: improved definitions and efficient constructions," in Journal of Computer Security, vol. 19(5), pp. 895–934, 2011

[5] M. Chase and K. Seny, "Structured encryption and controlled disclosure, " in International Conference on the Theory and Application of Cryptology and Information Security, pp. 577–594, 2010, Springer, Berlin, Heidelberg.

[6] K. Kaoru and Y. Ohtaki, "UC-secure searchable symmetric encryption," in International Conference on Financial Cryptography and Data Security, pp. 285–298, 2012, Springer, Berlin, Heidelberg.

[7] K. Seny, and K. Lauter, "Cryptographic cloud storage," in International Conference on Financial Cryptography and Data Security, pp. 136–149, 2010, Springer, Berlin, Heidelberg.

[8] K. Seny, C. Papamanthou and T. Roeder, "Cs2: A searchable cryptographic cloud storage system," Microsoft Research, TechReport MSR-TR-2011-58, May 2011.

[9] K. Seny, C. Papamanthou and T. Roeder, "Dynamic searchable symmetric encryption, " In Proceedings of the 2012 ACM conference on Computer and communications security, pp. 965–976. ACM, 2012.

[10] Q. Chai and G. Gong, "Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers," in IEEE International Conference on Communications (ICC), pp. 917-922. IEEE, 2012.

[11] J. Camenisch, K. Markulf , R. Alfredo and S. Caroline, "Blind and anonymous identity-based encryption and authorised private searches on public key encrypted data," in International Workshop on Public Key Cryptography, pp. 196–214, 2009, Springer, Berlin, Heidelberg.

[12] D. Boneh, G. Di. Crescenzo, R. Ostrovsky and G. Persiano, "Public key encryption with keyword search," in International conference on the theory and applications of cryptographic techniques, pp. 506–522, 2004, Springer, Berlin, Heidelberg.

[13] B . Waters, D. Balfanz, G. Durfee and D. K. Smetters, "Building an Encrypted and Searchable Audit Log, " in NDSS, vol. 4, pp. 5–6. 2004.

[14] M. Bellare, A. Boldyreva and A. ONeill, "Deterministic and efficiently searchable encryption, " in Annual International Cryptology Conference, pp. 535–552, 2007, Springer, Berlin, Heidelberg.

[15] J. Baek, R. Safavi-Naini and W. Susilo, "Public key encryption with keyword search revisited," In International conference on Computational Science and Its Applications, pp. 1249–1259, 2008, Springer, Berlin, Heidelberg.

[16] Y. Yang and M. Ma, "Conjunctive keyword search with designated tester and timing enabled proxy re-encryption function for e-health clouds, " IEEE Transactions on Information Forensics and Security 11, vol. 4 ,pp. 746–759, 2016.

[17] Y. Yang, "Attribute-based data retrieval with semantic keyword search for e-health cloud," Journal of Cloud Computing 4. 1, 2015: 10.

[18] Y. Miao, J. Ma, X. Liu, F. Wei, Z. Liu and X. A. Wang, "m 2-ABKS: Attribute-based multi-keyword search over encrypted personal health records in multi-owner setting," Journal of medical systems, vol. 40, no. 11, pp. 246, 2016.

[19] Q. Zheng, S. Xu and G. Ateniese, "VABKS: verifiable attribute-based keyword search over outsourced encrypted data," In IEEE INFOCOM 2014-IEEE Conference on Computer Communications, pp. 522–530, Apr 2014.

[20] W. Sun, S. Yu, W. Lou, Y.T. Hou and H. Li, "Protecting your right: Verifiable attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 27(4), pp.1187–1198, 2016.

[21] Z. Chen, F. Zhang,P. Zhang, J. K .Liu, J. Huang, H. Zhao and J .Shen, "Verifiable keyword search for secure big data-based mobile healthcare networks with fine-grained authorization control," Future Generation Computer Systems, vol. 87, pp.712–724, 2018.

[22] J.A Akinyele, C. Garman, I. Miers, M.W. Pagano, M. Rushanan, M. Green and A.D Rubin, "Charm: a framework for rapidly prototyping cryptosystems," Journal of Cryptographic Engineering, vol. 3, pp.111–128, 2013.

[23] A. Lewko and B. Waters, "Decentralizing attribute-based encryption. In Annual international conference on the theory and applications of cryptographic techniques," pp. 568–588, 2011.

[24] M.T Goodrich and M. Mitzenmacher, "Invertible bloom lookup tables," In 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 792–799, IEEE.

[25] B. Wang, B., S. Yu, W .Lou and Y. T. Hou, "Privacy-preserving multi-keyword fuzzy search over encrypted data in the cloud," In IEEE INFOCOM 2014-IEEE Conference on Computer Communications, pp. 2112–2120, 2014,April.

[26] E.J.,Goh, "Secure indexes," IACR Cryptology ePrint Archive,Vol.2003, pp.216,2003.

[27] S.K., Nayak and S. Tripathy," SEMFS: Secure and Efficient Multi-keyword Fuzzy Search for Cloud Storage," In International Conference on Information Systems Security, pp. 50–67, Springer, Cham, 2017, December.

Merged LSTM Model for emotion classification using EEG signals

Anumit Garg
Department of Electronics and Communication Engineering
Dr. B.R. Ambedkar National Institute of Technology
Jalandhar, India
anumitgarg@gmail.com

Ashna Kapoor
Department of Electronics and Communication Engineering
Dr. B.R. Ambedkar National Institute of Technology
Jalandhar, India
ashnakapoor07@gmail.com

Anterpreet Kaur Bedi
Department of Electronics and Communication Engineering
Dr. B.R. Ambedkar National Institute of Technology
Jalandhar, India
anterpreetbedi27@gmail.com

Ramesh K. Sunkaria
Department of Electronics and Communication Engineering
Dr. B.R. Ambedkar National Institute of Technology
Jalandhar, India
sunkariark@gmail.com

***Abstract*— The applicability of contemporary deep learning techniques have seen considerable improvements in the field of biomedical signal analysis. Emotion analysis using EEG signals is one such problem that has been studied and worked upon extensively in recent times. In this paper we have proposed a novel methodology to classify emotions using signal processing techniques such as wavelet transform and statistical measures for feature extraction and dimensionality reduction followed by developing state of the art neural architecture for the classification task. A merged LSTM model has been proposed for binary classification of emotions. The model's applicability and accuracy has been validated using DEAP dataset which is the benchmark dataset for emotion recognition.**

Keywords—deep recurrent neural networks (RNN), long short term memory (LSTM), discrete wavelet transform, statistics, dimensionality reduction, feature extraction, merged LSTM, DEAP dataset.

I. INTRODUCTION

Emotions are non verbal cues that define the mental state of an individual associated with his/ her thoughts, feelings, external stimuli etc. According to Charles Darwin emotions are adaptations gained through evolution that allow both humans and animals to maximize their chances of survival and reproduction [1]. Emotions play an instrumental role in decision making capability of an individual [2]. With maturation of technology, availability of huge amount of dataset and deeper understanding of human psychology, the field of cognitive analytics [3] has seen many breakthroughs in recent years.

Study of emotions based on text, speech, facial expressions, body pose, gesture have contributed a lot in this field [4] [5] [6] [7] [8] but analysis based on physiological signals is an emerging and more promising domain of research as these signals are spontaneous and highly involuntary in nature. Physiological signals are generated by the body in response to the functioning of various physiological systems. Hence these signals hold information which can be extracted to find out the state of these physiological systems. Most commonly used physiological signals for the purpose of emotion analysis include [9] [10] skin temperature variation, Electro Dermal Activity (EDA), Electrocardiogram (ECG), Electroencephalogram (EEG), etc.

In the present work, we use Electroencephalogram (EEG) signals for the purpose of emotion recognition. Here we study electroencephalogram (EEG) signals, their utility for the application of emotion recognition and propose a Merged Recursive Neural Network (LSTM) architecture framework for emotion classification using EEG data. We have validated our procedure and model on DEAP dataset which is a benchmark dataset for emotion recognition using EEG signals [11].

II. ELECTROENCEPHALOGRAM (EEG) SIGNAL BASICS

EEG signals are the recordings of spontaneous electrical activity of brain via metal electrodes and conductive media from the scalp. The voltage fluctuations are mainly caused due to the flow of ionic currents within neurons and reflect cortical electrical activity.

Despite limited spatial resolution, EEG data proves to be a valuable asset for research and diagnostic purposes. The disorders like various fluctuations in the EEG signals are of utmost importance for the diagnosis of epilepsy, sleep disorders, coma and cases of brain death [12]. Unlike CT, MRI or PET, it offers millisecond range temporal resolution. These signals are non-invasive in nature and are acquired without much effort from the body. They greatly reflect the influence of emotions on automatic nervous system.

Though the spectrum of acquired EEG signal is continuous, ranging from 0 Hz to half of the sampling frequency, the brain state of an individual may make certain frequencies that are dominant over other frequencies. On this basis EEG signals have been classified into four major groups:

1) *Beta: frequency greater than 13 Hz.*
2) *Alpha: frequency range from 8 to 13 Hz.*
3) *Theta: frequency ranges from 4 to 8 Hz.*
4) *Delta: frequency ranges from 0.5 to 4 Hz.*

Based on experiments the highest electrical activity is observed from cerebral cortex due to its proximity to the surface and most of the meaningful information is acquired from this particular region.

III. RELATED WORKS

Emotions are psycho- physiological process tantamount of a person's mental health and state of mind. Emotions

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

represent the perception of a being about the environment and world around him. As this field gains more impetus, EEG technology has become more affordable and less intrusive, thus making it suitable for effective deployment in healthcare industry.

The public release of DEAP dataset has motivated medical practitioners and researchers to dive deeper into the field of emotion analysis and has laid a firm foundation for them to build, test and deploy their novel ideas in this domain. Since the release of DEAP dataset many researchers have made a huge contribution and built effective algorithms to be validated on this dataset.

The preprocessing and dimensionality reduction of DEAP dataset is based upon the work of Jahankhani et al. [13]. This preprocessing methodology provides a starting baseline for feature extraction without much loss of information. It includes application of discrete wavelet transform to extract features from time series EEG signals [14]. This methodology provides compact representation of EEG signals in time- frequency domain. Further, to decrease the dimensionality statistical features such as maximum, minimum, mean and standard deviation were calculated.

Wavelet representation is better than Fourier representation because Fourier representation localizes a function only to its frequency domain and the signal cannot be significantly represented in spatial domain while wavelets are used to localize the function in frequency domain as well as spatial domain. Similarly the work of Kolte et al. [15] analyses the effects of various mother wavelet function on EEG signal analysis and concludes Daubachies (db4) wavelet transform for denoising and preprocessing the time series data [16].

Further the work of Samarth Tripathi et al. [17] explores the effectiveness of Neural Networks to classify user emotions based on DEAP dataset. The paper suggests two different neural models, a simple Deep Neural Network and a Convolution Neural Network for emotion classification. Their work is the testament of the fact that neural networks form a robust set of classifiers for analyzing physiological signals particularly EEG signals.

Hussein et al. [18] proposed the use of deep recurrent neural networks, particularly the long short-term memory (LSTM) model [22], for epileptic seizure diagnosis. Initially, the raw EEG signals were divided into non overlapping segments which were fed into deep recurrent neural layer (LSTM), in order to obtain a high level representation of EEG signals. Further the output of LSTM was fed into a dense layer followed by a softmax unit to create label predictions.

Compared to other works that are quite sensitive to noise, the LSTM model maintains its high detection performance even in the presence of common EEG artifacts and white noise as well. This model is thus robust in noisy as well as real life conditions.

IV. METHODOLOGY

Deep learning architectures have led us to achieve some of the most remarkable solutions to solve many contemporary problems. The power of deep learning lies in its supremacy over other statistical models due to high accuracy when trained on large dataset, which could only be achieved with huge amount of data, computation power and its layered architecture. Unlike traditional machine learning algorithms, deep learning algorithms learn about the data in hierarchical fashion which eliminates the need of domain expertise and reduces the need of extensive feature engineering.

However the EEG signal acquisition gives us a huge amount of data for processing. For instance, DEAP dataset consists of 40 trials for each volunteer and each trial includes readings from 40 EEG channels. Each channel consists of 8064 data points, thus accounting for total of 322560 readings for a single trial. With limited hardware resources, this data is very difficult to process and might contain some data that does not contribute much to the accuracy of the proposed model. Therefore we begin by cleaning our dataset and reducing its dimensionality by suitable preprocessing methodology.

Feature Extraction from EEG Signal is achieved using discrete wavelet transforms. The choice of optimal wavelet algorithm depends entirely on the application desired. In EEG Signal Feature Extraction, Db of order 2 is proved to be the optimal mother wavelet function [15] as compared to other mother wavelet functions. In this work there is overlap between iterations to pick up each detail that may be missed by other wavelet algorithms. At each step of iterations wavelet function is applied to input data. The looping variable is incremented by two at each iteration and thus if original data has N values, the computation will calculate N/2 differences.

In the present work, we propose a neural model for our research, a merged recurrent neural network followed by fully connected dense layers. The model uses contemporary deep learning techniques such as dropouts [19] to reduce over fitting of the data, LSTM cells to remove the problem of vanishing gradient and short term memory; and ReLu activation function is also used to include non-linearity in our model. The model architecture is implemented using Keras and Tensor flow libraries in python.

A. Recurrent Neural Network:

Neural networks form the basis of deep learning algorithms. In these networks, we assume all inputs and outputs to be independent of each other; i.e. the parameters of particular neuron are not affected by the input- output sequence of any other neuron. But for sequential data in which inputs and outputs are dependent on each other, the assumptions made by neural networks does not allow us to capture the temporal information contained in the data.

There are countless learning tasks that deals with sequential data such as image captioning, video analysis, handwriting analysis, study of genomes and numerical time series data obtained through sensors. In these tasks neural networks fails miserably due to their incapability to capture temporal information [20]. Unlike neural networks, recurrent neural networks (RNNs) have internal memory which enables it to retain past information as well. Thus, RNNs have a deeper understanding of sequential data and its context as compared to other algorithms.

A single layer of recurrent neural network has large number of processing units commonly called RNN cells. From equation (1) and (2) we see that each of these cells takes two inputs: activation from previous cell ($a^{<t-1>}$) and current input from dataset ($x^{<t>}$). These input variables are

multiplied with their respective weight matrices and a bias term (b_a) is added to them. Further they are passed through corresponding activation functions to generate an input for the next cell. Output ($\hat{y}^{<t>}$) is calculated by multiplying ($a^{<t>}$) with the weight matrix (w_{ya}) and adding a bias term (b_y) followed by a softmax function. The structure on an RNN cell and flow of data through it is shown in Fig. 1.

$$a^{<t>} = tanh(\, w_{aa}\, a^{<t-1>} + w_{ax}\, x^{<t>} + b_a) \quad (1)$$

$$\hat{y}^{<t>} = softmax(\, w_{ya}\, a^{<t>} + b_y) \quad (2)$$

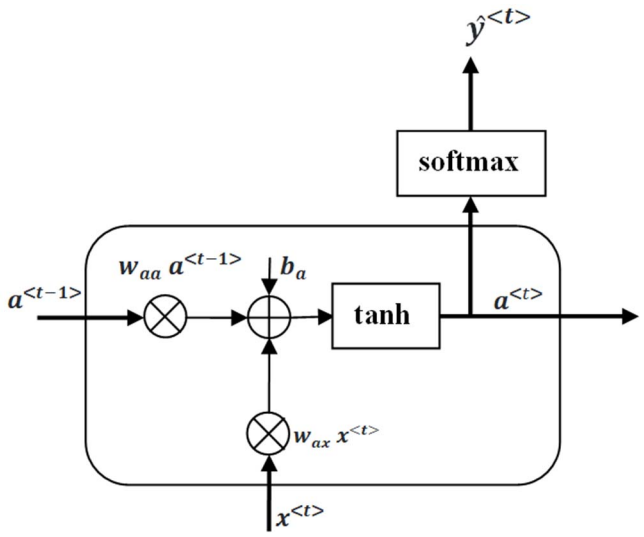


Fig. 1. Recurrent Neural Network cell

B. LSTM basics:

RNN effectively enables us to remember short sequences and model time- dependent data. However, they do suffer from one of the major problems of vanishing gradient [21]. During back propagation, the gradient reduces to such a small value that no parameters are significantly updated, thus restricting RNN to learn effectively from the data.

There are number of measures to overcome this problem. In the present work, LSTMs [22] have been used to rectify the problem of vanishing gradient and also enable our model to learn long sequences. The distinguishing feature of LSTMs that makes them more appropriate choice over RNNs is its internal architecture shown in Fig. 2. There are three gates which provide LSTM its functionality.

Here, $c^{<t>}$ represents the cell state, $a^{<t>}$represents output from the block while $c'^{<t>}$ represents candidate for cell state at timestamp (t). Now, to obtain the memory vector for current time stamp the candidate is calculated as given in equation (3).

$$c'^{<t>} = tanh(\, w_c\, [a^{<t-1>}, x^{<t>}] + b_c) \quad (3)$$

1) Forget gate: it is responsible for the removal of information from the cell state that is no longer needed for the LSTM to understand and learn from the data. As shown in equation (4), the gate takes two inputs, input ($x^{<t>}$) at time stamp t and the activation from previous state ($a^{<t-1>}$). These set of input states are multiplied to the weight matrix (w_f) and a bias term (b_f) is added to it. The resultant is passed through sigmoid function. The output of sigmoid function further decides which value to keep and which value to discard. If the output is '0', the forget gate no longer remembers the past value while for output equals '1', the past value is retained.

$$\Gamma_f = \sigma(w_f\, [a^{<t-1>}, x^{<t>}] + b_f) \quad (4)$$

2) Update gate: it is primarily responsible for the addition of information to the cell state. As shown in equation (5), the gate takes two input, input ($x^{<t>}$) at time stamp t and the activation from previous state ($a^{<t-1>}$). These set of input states are multiplied to the weight matrix (w_u) and a bias term (b_u) is added to it, then the resultant is passed through sigmoid function.

Based on the output of sigmoid function, the gate ensures that information of significant importance is added.

$$\Gamma_u = \sigma(w_u\, [a^{<t-1>}, x^{<t>}] + b_u) \quad (5)$$

3) Output gate: Output gate filters the information that is not required in the output by the LSTM cell. As shown in equation (6), the gate takes two input, input ($x^{<t>}$) at time stamp t and the activation from previous state ($a^{<t-1>}$). These set of input states are multiplied to the weight matrix (w_o) and a bias term (b_o) is added to it. The resultant is passed through sigmoid function. The output gate selects only useful information from current cell and propagate it as an input to the next cell.

$$\Gamma_o = \sigma(w_o\, [a^{<t-1>}, x^{<t>}] + b_o) \quad (6)$$

Once the value of all the gates are calculated final cell state and activation for the next cell are predicted using these values as shown in equation (7) and (8).

$$c^{<t>} = \Gamma_u * c'^{<t>} + \Gamma_f * c^{<t-1>} \quad (7)$$

$$a^{<t>} = \Gamma_o * tan(\, c^{<t>}) \quad (8)$$

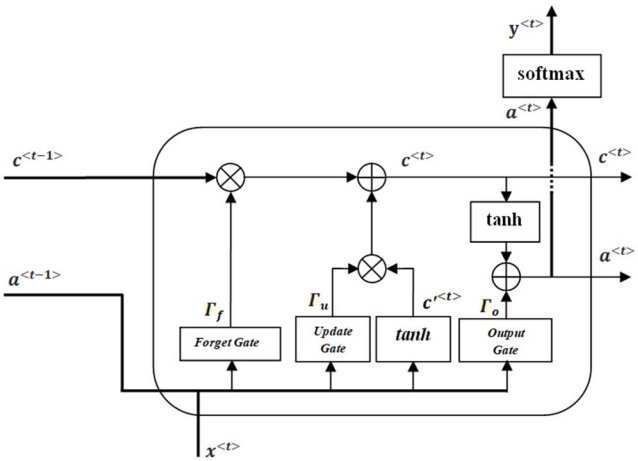


Fig. 2. LSTM cell

V. Proposed Algorithm

Biomedical signals are highly non stationary in nature and their statistical characteristics changes abruptly over time. Therefore in our proposed algorithm we begin by dimensionality reduction and feature extraction from our dataset using discrete wavelet transform followed by use of statistics to capture the nature and trend in variations of the dataset. Validation of the algorithm is done using DEAP Dataset. Mean, median, maximum, minimum, standard deviation, range, skewness and kurtosis are statistical measures used over subsets of our dataset.

In our research work we have used a merged LSTM model for the classification problem. The features extracted from the dataset are split into training and testing data and fed to the model. The model is trained and validated on the test set; a detailed description of the flow of algorithm is shown in Fig. 4. Our merged LSTM model uses LSTM layers and dense layers as the first half of the network. The preprocessed data of each of the 40 channels is fed simultaneously to this network, and this is done to provide more detailed learning to each network specific to a particular channel. Later the second half of this network consists of two fully connected layers followed by an output layer. The outputs from each of the first half of the network are concatenated and fed as input to the second half of the network as shown in Fig. 3.

It is evident from numerous researches that not all channels or regions of the brain contribute equally towards predicting the occurrence of each emotion and some regions show more activity and display a particular emotion more prominently [23]. For example disgust is found to be associated with right-sided activation in the frontal and anterior temporal regions as compared to happy condition. Thus intuitively the merged LSTM model captures the very contribution of each channel and helps us in predicting the emotional state of a person more effectively.

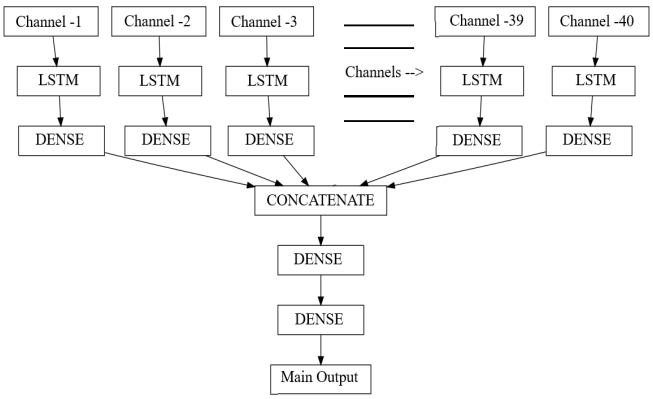


Fig. 3. Merged LSTM model

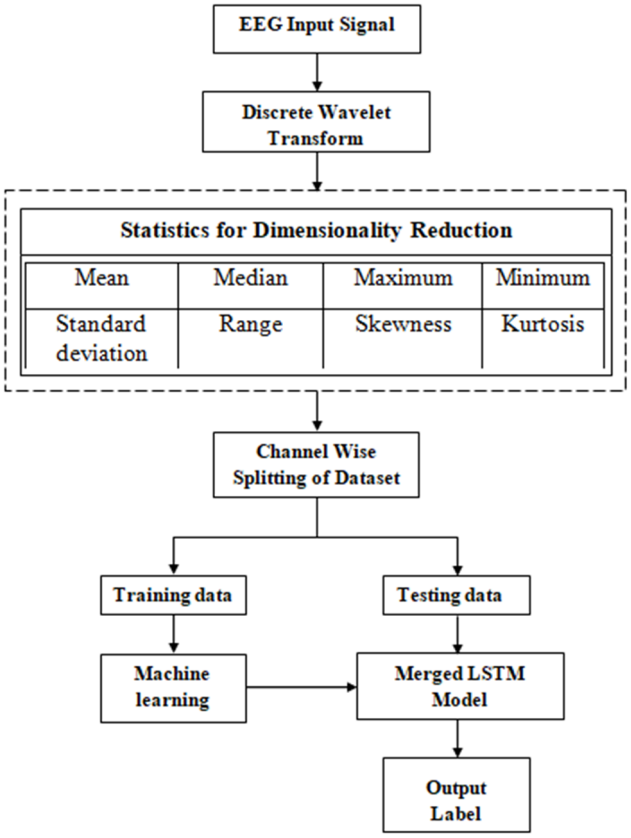


Fig. 4. Flow of algorithm

VI. Results

We have validated our algorithm on DEAP dataset which is a multimodal dataset for emotion analysis using EEG, physiological and video signals. The dataset consists of two parts. First part comprises of the ratings from an online self-assessment where 120 one-minute extracts of music videos were each rated by 14-16 volunteers based on arousal, valence and dominance. While second part comprise of the participant ratings, physiological recordings and face video of an experiment where 32 volunteers (16 men and 16 women, aged between 19 and 37) watched a subset of 40 above mentioned music videos. The proposed music videos induced various emotions in different users who then rated these videos on the valence- arousal scale [25], the same videos had an on-line evaluation that could be used for comparison.

However in present work we have used preprocessed and segmented version of data that has been made available in MATLAB and pickled Python/ Numpy formats. The data is segmented into 60 sec trial and a 3 sec pre-trial baseline removed. This version of data is well suited for testing a classifier or regression technique without the hassle of processing all data first. The entire dataset contains files of 32 participants and each file contains two arrays as illustrated in Table I.

TABLE I. DEAP DATASET DESCRIPTION

Array name	Array shape	Array contents
Data	40x40x 8064	video/trial x channel x data
Labels	40x4	video/trial x label (valence, arousal, dominance, liking)

Comparing our results with the existing state of the art methods clearly depicts supremacy of the proposed algorithm over existing methods. Samarth Tripathi et al. [17] Deep Neural Model achieves accuracy of 75.78% and 73.125% on valence and arousal two class emotion classification. Li et al. [26] propose a method of Deep Belief Networks for Emotion Identification and achieve an accuracy of 58.4% on valence, 64.2% on arousal, 65.8% on dominance and 66.9% on liking. The proposed algorithm gave us staggering accuracy of 84.89% on two class valence classification and 83.85% on two class arousal classification using a batch size of 128 when the model is run on 60 epochs. The same model was trained on all four emotions and the results are shown in TABLE II.

TABLE II. ACCURACY ON DIFFERENT EMOTIONS

Emotion	Accuracy
Valence	84.89%
Arousal	83.85%
Dominance	84.37%
Liking	80.72%

VII. CONCLUSION

In this paper, we built upon the prior research work in the field of EEG data analysis and Emotion recognition. Further we explored the possibility of using Recurrent Neural networks to classify emotions using EEG signals. Our study once again proved the efficiency of neural networks in solving some of the biggest biomedical problems in a profound way. The proposed merged LSTM model provides state of the art classification accuracy; we achieved an accuracy of 84.89%, 83.85%, 84.37% and 80.72% on binary classification of valence, arousal, dominance and liking respectively. This also proves that neural networks form a robust set of classifiers for EEG signals surpassing traditional learning techniques.

The present model also lays a firm foundation for future work on emotion analysis. Our future work would focus more on development and improvement of existing model for multiclass classification of a each emotion.

VIII. REFERENCES

[1] Darwin, Charles, and Phillip Prodger. The expression of the emotions in man and animals. Oxford University Press, USA, 1998.

[2] Loewenstein, George, and Jennifer S. Lerner. “The role of affect in decision making.” Handbook of affective science619.642 (2003): 3.

[3] Gudivada, Venkat N., et al. “Cognitive analytics: Going beyond big data analytics and machine learning.” Handbook of statistics. Vol. 35. Elsevier, 2016. 169-205.

[4] Lang, Peter J. "A bio‐informational theory of emotional imagery." Psychophysiology 16.6 (1979): 495-512.

[5] Fasel, Beat, and Juergen Luettin. "Automatic facial expression analysis: a survey." Pattern recognition 36.1 (2003): 259-275.

[6] Strapparava, Carlo, and Rada Mihalcea. "Learning to identify emotions in text." Proceedings of the 2008 ACM symposium on Applied computing. ACM, 2008.

[7] Gunes, Hatice, and Massimo Piccardi. "Bi-modal emotion recognition from expressive face and body gestures." Journal of Network and Computer Applications 30.4 (2007): 1334-1345.

[8] Busso, Carlos, et al. "Analysis of emotion recognition using facial expressions, speech and multimodal information." Proceedings of the 6th international conference on Multimodal interfaces. ACM, 2004.

[9] Kim, Kyung Hwan, Seok Won Bang, and Sang Ryong Kim. "Emotion recognition system using short-term monitoring of physiological signals." Medical and biological engineering and computing 42.3 (2004): 419-427.

[10] Bong, Siao Zheng, M. Murugappan, and Sazali Yaacob. "Analysis of electrocardiogram (ecg) signals for human emotional stress classification." International Conference on Intelligent Robotics, Automation, and Manufacturing. Springer, Berlin, Heidelberg, 2012.

[11] Koelstra, Sander, et al. "Deap: A database for emotion analysis; using physiological signals." IEEE transactions on affective computing 3.1 (2012): 18-31.

[12] Teplan, Michal. "Fundamentals of EEG measurement." Measurement science review 2.2 (2002): 1-11.

[13] Jahankhani, Pari, Kenneth Revett, and Vassilis Kodogiannis. "Data mining an EEG dataset with an emphasis on dimensionality reduction." 2007 IEEE Symposium on Computational Intelligence and Data Mining. IEEE, 2007.

[14] Adeli, Hojjat, Ziqin Zhou, and Nahid Dadmehr. "Analysis of EEG records in an epileptic patient using wavelet transform." *Journal of neuroscience methods* 123.1 (2003): 69-87.

[15] Chavan, A., and M. Kolte. "Optimal Mother Wavelet for EEG Signal Processing." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2.12 (2013): 5959-5963.

[16] Raghuveer R. Bapordikar A , „ Wavelet Transforms – Introduction to theory and applications.‟ Addis on – Wesley , 2000

[17] Tripathi, Samarth, et al. "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset." Twenty-Ninth IAAI Conference. 2017.

[18] Hussein, Ramy, et al. "Epileptic seizure detection: a deep learning approach." arXiv preprint arXiv:1803.09848 (2018).

[19] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[20] Sundermeyer, Martin, et al. "Comparison of feedforward and recurrent neural network language models." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.

[21] Hochreiter, Sepp. "The vanishing gradient problem during learning recurrent neural nets and problem solutions." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 6.02 (1998): 107-116.

[22] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[23] Davidson, Richard J., et al. "Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology: I." Journal of personality and social psychology58.2 (1990): 330.

[24] Tieleman, Tijmen, and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude." COURSERA: Neural networks for machine learning 4.2 (2012): 26-31.

[25] Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.

[26] Li, Xiang, et al. "EEG based emotion identification using unsupervised deep feature learning." (2015).

A Novel Chaotic based Privacy preserving machine learning model on large distributed client applications

Ch. Nanda Krishna
Research Scholar
Department Of Computer Science and Engineering
JNTUA
Anantapuramu
chnk1789@gmail.com

Dr.K.F.Bharati
Assistant Professor
Department Of Computer Science and Engineering
JNTUA
Anantapuramu
kfbharathi@gmail.com

***Abstract*— With the rapid growth of data size, data storage and computational memory, it is essential to implement an advanced privacy preserving model on large datasets. Machine learning framework is used to extract essential hidden patterns which are plain text format for decision making in distributed applications. Privacy preserving data mining (PPDM) has emerged as an essential area for data confidential in terms of data exchange, decision making and data publication. Privacy preserving is a popular data privacy model for securing individual decision patterns from unauthorized access. As the decision-making patterns of the data owner are stored and distributed publicly, it leads to the misuse of information in distributed applications. Some privacy information about business organizations, industries and individuals has to be encoded before it is publicly shared or published. In this paper, a novel chaotic privacy preserving model is designed and implemented on the large distributed data for privacy preserving. Here, different traditional privacy preserving models are used to compare the proposed model to the traditional models**

Keywords*— *Chaotic hash, privacy preserving, distributed applications.

I. INTRODUCTION

Recently, data is accumulating tremendously in almost all fields such as industrial organizations, scientific research, educational institutions, medical science, business sectors and government organizations, etc. These organizations started looking for privacy preserving from these databases for future prospects with less computational memory and time. Machine learning is a cutting edge technology derived from the field of Artificial Intelligence to extract hidden knowledge or rules from large databases. In machine learning, Classification algorithms, association rule mining algorithms, clustering algorithms and outlier mining algorithms are used in a wide range of applications for privacy preserving. Privacy-Preserving Data Mining (PPDM) integrates the concepts of data mining functions along with probable privacy preservation [1]. PPDM completely depends upon Secure Multi-party Computation (SMC). This method has wide implementation in the area of privacy preservation. PPDM is responsible for adding extended security to the sensitive information in data mining [1].

Privacy Preserving Data Publishing techniques hide the data using some form of transformation-based approaches or cryptographic approaches before sharing the data to third party. The way in which the data is going to be used is unknown at the first step. Hence the privacy preserving techniques transform the data into a form suitable for any computation. The published data can be used for any purpose not specifically for data mining. Data randomization is a process of performing transformation operation such as scaling, rotation, noise addition which would change the original value. The mapping or correlation among the data is completely lost during the randomization process. Cryptographic approaches use various encryption algorithms to preserve the data against intruders. But once the data is decrypted the data becomes plain and a semi-honest adversary can cause attacks.

With a classification scheme, this work can locate a variety of techniques that may be just suitable for a given scenario. Some techniques have been proposed. Given the varied situations in which PPDM is used, it appears unlikely that one, single privacy preserving technique may outperform all other techniques in every way. Every technique has its pluses and minuses. Therefore, it becomes very important to conduct a wide evaluation of all privacy preserving techniques. One of the Traditional techniques for Privacy preserving data mining is the Randomization algorithm and it was first introduced by Agrawal and Srikanth (2000). Randomization allows many users to provide data that may be sensitive to be effectively used in a centralized data mining system, all the while keeping a check on whether sensitive values get disclosed in the process.

II. RELATED WORK

The imbalanced data problem is relaxed in unsupervised self-organizing learning with support vector ranking as mentioned in [2]. In this method variables are selected by the model adopted by support vector machines to deal with this problem. ESOM also known as Emergent Self Organizing Map is used to cluster the ranker features to provide for unsupervised cluster classification. A Kolomogrov Smirnov statistic based on decision tree method(KS tree) [3] is the latest method in which complex problem is divided into several easier sub problems, in that case imbalanced distribution becomes less daunting. This method is also used for feature selection removing the redundant ones. After division, a two-way resampling is employed to determine optimal sampling criteria and rebalanced data is used to incorporate into logistic regression models. If all the posterior probabilities are considered, simpleMIL will work impressively well. This approach gives emphasis on intensity and texture distributions. But, these types of classifiers are used where scans are gathered from single domain. In case of cross domains just like multiple scanners, the above mentioned classifiers are not much effective. In order to overcome the said issue, the classification algorithm is

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

implemented in a multi-class dataset. The Gaussian texture features algorithm shows better performance as compared to intensity features. An efficient weighting technique which depends upon classifiers can distinguish among database scans from various domains. This technique can definitely enhance the results of the traditional classification approaches.

In case of distributed data mining approach, security of communications and encryption are two major areas of concern. Hence, privacy preserving along with encryption schemes are often implemented on distributed applications. All distributed application are capable of storing data through two models, those are:- vertically partitioned data model and horizontally partitioned data model. Vertically partitioned data can be defined as data located in various sites and the stored data are separated from each other in such a manner that no data overlapping occurs. Horizontally partitioned data is stated as data located in numbers of sites based on the records. The sites do not have information about records of other sites. Many useful and valid association rules are generated by considering horizontally partitioned data [6, 7]. There have been developed tremendous amount of research works in the field of cryptography to resolve the privacy preservation issues. Let us consider an example of Secure Multi-Party Computation (SMC). SMC is always implemented in multi-user networks where each user has the responsibility to work together with other users in order to complete computing tasks and maintaining privacy as well [8][9].

Current research strategies involve different privacy preservation techniques and process them with the data mining models to discover decision patterns. The main goal of privacy protection scheme is to develop a data mining model in order to retrieve useful knowledge present in sensitive information. There are two major issues in the above mention PPDM approach, those are:-

1. The most common issue is to provide protection to sensitive information such as name, ID number, and address and income level.

2. The second problem involves in the process of preserving sensitive information by implementing knowledge expression in database which is also known as knowledge database (KD hidden).

[7] Introduced a basic algorithm in order to construct an efficient decision tree classifier which is termed as ID3 algorithm. In the proposed privacy preservation data mining framework, the database administrator preserve patterns to restrict values in class label. The information are downgraded from high (secure environment) to low (public environment) and the presence of inference patterns are also considered in this mining model. In the above proposed framework, the potential downgraded details are integrated with cost measure which is not initialized to low. It also enhances the confidentiality measure of the data mining framework. In case of horizontally partitioned data [8], all the pre-existing classification techniques need attribute(s) exchange between the parties. In case of vertically partitioned data, traditional approaches need exchange of each and every individual instance and attribute information from the communicating parties. Many cryptographic approaches are introduced in the field of Secure Multiparty Computation to achieve accurate privacy and security. This approach is an extension of Lindell's research [9] with association rule mining techniques, classification schemes [10] and clustering approaches. Few approaches are efficient in achieving privacy, whereas most algorithms are incapable of maintaining privacy and restricting their information disclosure.

The main applications of Decision Tree Classifiers based privacy preserving models includes: radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, speech recognition, and so on. Many extended models have been carried out to improve a feature selection measures for the existing decision tree models.

Generally, decision trees are represented as directed graphs having nodes and edges. The root node and the intermediate nodes always represent tests, whereas the results of tests are denoted by the outgoing edges. Some researchers termed these intermediate nodes as inner nodes. Besides these, the leaf nodes represent different class labels which are responsible for determining the privacy patterns in large databases.

In [11], they analyzed the problems of traditional privacy preserving data mining approach and proposed an effective solution on small datasets. They implemented the conventional privacy preserving ID3 algorithm containing horizontally partitioned databases. [12] Introduced a new method for privacy preservation of ID3 algorithm by considering horizontally partitioned data. They also considered multiple parties, instead of considering two parties like vertically partitioned data approaches. Here, Gini feature selection measure is implemented in place of entropy for privacy-preserving ID3 algorithms in order to construct decision tree patterns. All the parties are able to evaluate the gain value of attributes mutually. This approach operates perfectly when database is partitioned into two or multiple parties. Entropy generally helps in the process of constructing balanced trees [10], studied and analyzed various privacy preservation data mining approaches. Implementation of certain approaches may restrict some basic functions of data mining. They developed a new method to achieve privacy preservation and named it as Privacy Preserving Record Linkage (PPRL). This technique permits different database linkage to organizations along with privacy on sensitive data

III. CHAOTIC MAPS: PRELIMINARIES

Chaotic system has several important characteristics, including pseudo-randomness and sensitivity to the original initial conditions. These features have excellent diffusion and confusion properties, which are essential for encryption. For data integrity and message authentication process, chaotic hash functions are used to improve the security and randomization. Chaos has often been used to build key hash functions in recent years. Chaotic maps are functions that are highly sensitive to their original conditions, and after a while develop into very different consequences. An infinite change in its initial conditions leads to chaotic functions producing completely different output patterns. It's a very useful feature for encryption. In encryption algorithms, so far many different chaotic maps have been used. Due to their interesting characteristics such as sensitivity to small changes in initial conditions and parameters, mixing property, ergodicity, and so on, chaos has been widely used in the last

decade. Most chaotic hash functionality can only be exercised sequentially [1–5], however. In particular, until the previous one is processed, processing of the current message unit cannot start. This inevitably reduces efficiency significantly.

IV. CHAOTIC ENCRYPTION MODEL

The skew tent map is defined as follows:

$$f_\varphi(m) = \begin{cases} \dfrac{m}{\varphi}, & 0 < m \le \varphi \\ \dfrac{m-1}{\varphi - 1}, & \varphi < m \le 1 \end{cases} \quad \text{--(1)}$$

The inverse function of the skew tent map is given by

$$f_\varphi^{-1}(m) = \varphi x \text{ or } f_\varphi^{-1}(m) = 1 + (\varphi - 1)m \quad \text{-- (2)}$$

In the nonlinear dynamic system, chaos is a ubiquitous phenomenon. The Lyapunov exponent can show most non-linear systems and the numerical number of them can reflect the degree of radiation of the adjacent locus.

$$\text{Hqper} - \text{chaoti c Lü sqstem} = \begin{cases} \dot{p} = a(q-p) \\ \dot{q} = -pr + cq \\ \dot{r} = pq - br \end{cases} \quad \text{--(3)}$$

$$\text{Hqper} - \text{chaoti c Chen sqstem} = \begin{cases} \dot{p} = a(q-p) \\ \dot{q} = -pr + dp + cq - q \\ \dot{r} = pq - br \\ \dot{q} = p + k \end{cases} \quad \text{--(4)}$$

$$\text{Hqper} - \text{chaoti c Rossler sqstem} = \begin{cases} \dot{p} = -q - r \\ \dot{q} = p + aq + w \\ \dot{r} = b + pr \\ \dot{w} = -cr + dw \end{cases} \quad \text{--(5)}$$

The chaotic tent map used here is a one-dimensional calculation and a piece-wise linear map.

Effective chaotic function is based on inherent merits of chaos such as the sensitivity to small changes to the original conditions and parameters, mixed properties, ergodicity, unstable, periodic orbits with long periods of time and one-way iteration. A linear chaotic map (PWLCM) and a four-dimensional chaotic map (CATCM) are used for the hash generation and encryption algorithm for strong data security and integrity verification.

In this algorithm, each node's data is taken as input to compute the high randomized hash value. Client identification M_ID and cloud node data M are the input parameters to the integrity algorithm. Message M is partitioned into blocks and sub-blocks of size 4 bytes. Each byte in the sub-block partition is processed using a series of polynomial transformations. In the polynomial transformations, QR decomposition and Cauchy randomization are used to improve the sensitivity level of the round key.

New keys are generated using the Q and R matrices of QR decomposition measure. These matrices are used to find the Cauchy polynomial computation and rank. This process is repeated until the number of rounds. Generated integrity value is used to validate the cloud node integrity against the attacks in WMNs. In this algorithm, the size of the input hash value is not fixed to constant value. Different sizes of hash values such as 512, 1024, 2048, 4096 are generated using the proposed algorithm. The computational efficiency of the proposed algorithm is better than the existing integrity techniques in terms of time and sensitivity. Finally the generated hash integrity value is used in phase 2 to verify the signature of the cloud node in the data exchange or communication process.

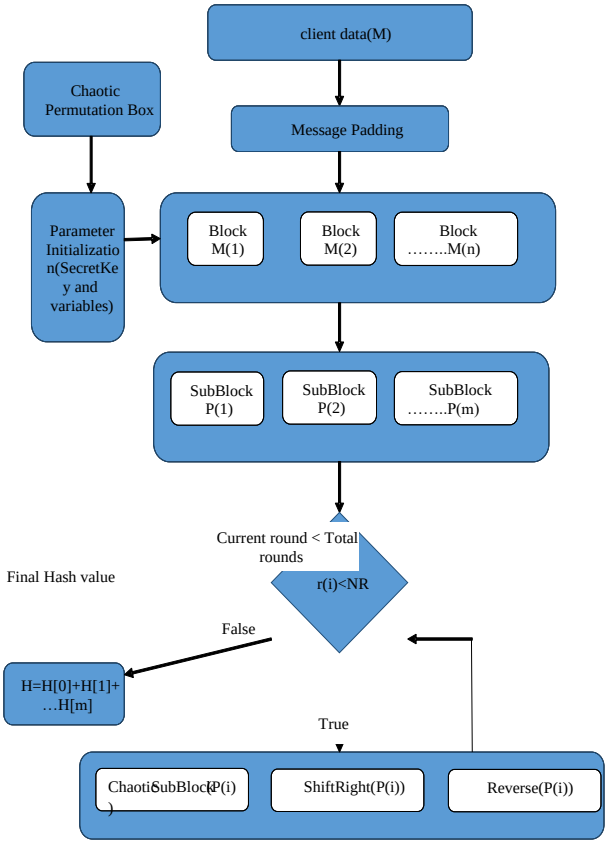


Fig. 1. Chaotic hash based Privacy Preserving Model

Model is better than the traditional methods in terms of privacy and accuracy are concerned.

TABLE I. COMPARATIVE ANALYSIS OF PRESENT HOMOMORPHIC MODEL TO THE EXISTING MODELS IN TERMS OF PRIVACY PRESERVING ON ONE PARTITIONING ATTRIBUTE.(MAX DEPTH=1)

Instances	*RSA*	*ECC*	*Homomorphic Encryption*
#50	0.919	0.942	0.951
#100	0.922	0.938	0.936
#150	0.924	0.929	0.932

Instances	*RSA*	*ECC*	*Homomorphic Encryption*
#200	0.924	0.942	0.946
#250	0.917	0.924	0.941
#300	0.924	0.925	0.954
#350	0.922	0.932	0.952
#400	0.915	0.924	0.924
#450	0.914	0.931	0.928
#500	0.918	0.925	0.951

Table 1, describes the performance of the proposed model to the traditional models for privacy preserving patterns. In the present system, the privacy preserving model performs better than the traditional models on one partitioning attribute.(Max depth=1).

TABLE II. COMPARISON OF THE PRESENT HOMOMORPHIC MODEL TO THE TRADITIONAL MODELS ON TWO PARTITIONING ATTRIBUTE.(MAX DEPTH=2).

Instances	*RSA*	*ECC*	*Homomorphic Encryption*
#50	0.917	0.935	0.948
#100	0.913	0.922	0.944
#150	0.924	0.941	0.95
#200	0.919	0.933	0.944
#250	0.922	0.923	0.936
#300	0.914	0.922	0.936
#350	0.919	0.941	0.93
#400	0.921	0.94	0.927
#450	0.922	0.942	0.954
#500	0.914	0.944	0.954

Table 2 describes the performance of the proposed model to the traditional models for privacy preserving patterns. In the present system, the privacy preserving model performance is better than the preserving patterns. In the present system, the privacy preserving model performs better than the traditional models on two partitioning attribute.(Max depth=2).

TABLE III. COMPARISON OF THE PRESENT HOMOMORPHIC MODEL TO THE TRADITIONAL MODELS ON THREE PARTITIONING ATTRIBUTE.(MAX DEPTH=3).

Instances	*RSA*	*ECC*	*Homomorphic Encryption*
#50	0.913	0.923	0.924
#100	0.922	0.941	0.951
#150	0.917	0.939	0.953
#200	0.915	0.943	0.936
#250	0.914	0.941	0.947
#300	0.913	0.941	0.926
#350	0.922	0.943	0.948
#400	0.921	0.936	0.948
#450	0.918	0.923	0.929
#500	0.92	0.926	0.926
#550	0.92	0.923	0.946
#600	0.915	0.937	0.945
#650	0.914	0.934	0.95
#700	0.915	0.937	0.953
#750	0.919	0.938	0.925
#800	0.92	0.931	0.942
#850	0.921	0.931	0.951
#900	0.913	0.93	0.923
#950	0.913	0.944	0.951
#1000	0.924	0.923	0.953

Table 3, describes the performance of the proposed model to the traditional models for privacy preserving patterns. In the present system, the privacy preserving model performs better than the traditional models on three partitioning attribute.(Max depth=3).

TABLE IV. RUNTIME COMPUTATION OF AVERAGE PARTITIONING BASED PRIVACY PRESERVING MODEL TO THE TRADITIONAL METHODS.

Instances	*RSA*	*ECC*	*Homomorphic Encryption*
#50	295	259	170
#100	601	500	339
#150	890	786	522
#200	1188	1042	682
#250	1529	1194	841
#300	1795	1541	1020
#350	2118	1840	1212
#400	2385	2073	1387
#450	2725	2278	1558
#500	2996	2488	1706
#550	3371	3018	1903
#600	3672	2925	2083
#650	3954	3079	2243
#700	4246	3565	2433
#750	4602	3664	2540
#800	4770	4353	2715
#850	5174	3989	2938
#900	5393	4409	3132
#950	5664	5025	3197
#1000	5863	5264	3415

Table 4 Describes the computational time of the present partitioning based privacy preserving model to the traditional models on different partitions. From the table, it is clearly observed that the present model has less computational runtime compared to the existing models.

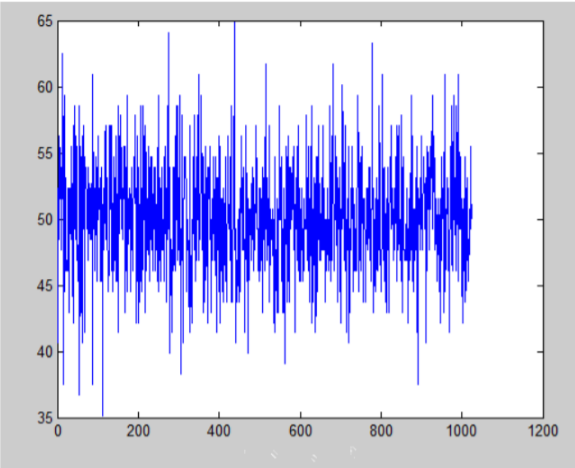


Fig. 2. Proposed Chaotic hash

Figure 2, describes the chaotic behaviour of the input data fields using the hash function. As the size of the input instances increases, the randomization of the hash functions also increase.

V. CONCLUSION

To improve the privacy of the original dataset, while retaining the patterns, this work implemented a novel framework using the partitioning based homomorphic model with less computational time and high accuracy. Machine learning models such as classification, clustering or feature selection models are applied on the multiple datasets for pattern analysis. The disclosures of personal data or patterns have become a major problem in large datasets.

REFERENCES

[1] Mahesh, R., &Meyyappan, T. (2013, February). Anonymization technique through record elimination to preserve privacy of published data. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on (pp. 328-332). IEEE.

[2] Usha, P., Shriram, R., &Sathishkumar, S. (2014, February). Sensitive attribute based non-homogeneous anonymization for privacy preserving data mining. In Information Communication and Embedded Systems (ICICES), 2014 International Conference on (pp. 1-5). IEEE.

[3] Prakash, M., &Singaravel, G. (2015). An approach for prevention of privacy breach and information leakage in sensitive data mining. Computers & Electrical Engineering, 45, 134-140.

[4] J. Le Ny and G. Pappas, “Differentially private filtering,” Automatic Control, IEEE Transactions on, vol. 59, no. 2, pp. 341–354, Feb 2014.

[5] Z. Huang, S. Mitra, and G. Dullerud, “Differentially private iterative synchronous consensus,” in Proceedings of the 2012 ACM .

[6] LiorRokach and Oded Maimon “TopDown Induction of Decision Trees Classifiers – A Survey”, IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002

[7] Zhou Shui-Geng, Li Feng, Tao Yu-Fei, Xiao-Kui. Privacy Preservation in Database Applications: A Survey.Chinesejournerl of computer,2009

[8] Yan Zhao1 Ming Du2 Jiajin, Le1 Yongcheng Luo1, A Survey on Privacy Preserving Approaches in Data Publishing. First International Workshop on Database Technology and Applications, 2009

[9] Agrawal, Shashank, et al. ”Function Private Functional Encryption and Property Preserving Encryption: New Definitions and Positive Results.” IACR Cryptology ePrint Archive 2013 (2013): 744

[10] Attrapadung, Nuttapong, and BenotLibert. ”Functional encryption for public-attribute inner products: Achieving constant-size ciphertexts with adaptive security or support for negation.” J. Mathematical Cryptology 5.2 (2012): 115-158.

[11] Stefano Guarino,"Provable Storage Medium for Data Storage Outsourcing",IEEE TRANSACTIONS ON SERVICES COMPUTING,2014.

[12] JinguangHan,"Improving Privacy and Security in Decentralized Ciphertext-Policy Attribute-Based Encryption",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 3, MARCH 2015.

[13] JiantingNing,"White-Box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting Flexible Attributes",IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 6, JUNE 2015.

[14] ChangjiWang,"An Efficient Key-Policy Attribute-Based Encryption Scheme with Constant Ciphertext Length",Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 810969.

Prospect of Stein's Unbiased Risk Estimate as Objective Function for Parameter Optimization in Image Denoising Algorithms – A Case Study on Gaussian Smoothing Kernel

Simi V.R[1]., Damodar Reddy Edla[1], Justin Joseph[2] and Venkatanareshbabu Kuppili[1]

[1]Department of Computer Science & Engg.,
National Institute of Technology, Goa – 403401, India.

[2]Department of Electrical Engineering,
Indian Institute of Technology, Gandhinagar-382355, India

***Abstract*- Stein's Unbiased Risk Estimate (SURE) is considered as an indirect method for predicting Mean Squared Error (MSE) in the absence of ground-truth, as its computation requires only noisy observation and denoised image. SURE is usually used as an objective function for optimizing the operational parameters of denoising algorithms, adequate for real-time images. Hence, a close analysis of the performance of SURE on standard test images is worthy of investigation. Pearson's Correlation (r) of SURE with Mean Absolute Error (MAE) between denoised images and ground-truth is analyzed in this paper, on Shepp-Logan Phantom and simulated Magnetic Resonance (MR) images, at different noise levels. Denoised images which differ in terms of MAE against ground-truth are produced by varying the standard deviation of a Gaussian smoothing kernel ($0.01 \leq \sigma \leq 0.04$) of fixed dimension, 9×9. Values of correlation between SURE and MAE on Shepp-Logan and simulated MR images are $r = -0.99 \pm 0.02$ and $r = 0.48 \pm 0.36$, respectively. Concordance of SURE with MAE is observed to be poor on simulated MR images, especially at higher noise levels. SURE is suitable for optimizing the parameters of denoising kernels only when the underlying function used to compute the kernel is fully differentiable by the noisy observation.**

***Keywords*- Gaussian Smoothing Kernel; Image Denoising; Optimization; Stein's Unbiased Risk Estimate.**

I. INTRODUCTION

Parameter optimization of denoising algorithms in real-time applications is troublesome due to the lack of noise-free ground-truth. Hence, there is a practice of optimizing the parameters of denoising algorithms on standard test images, provided prior information about noise statistics in it and their ideal noise-free estimate or ground-truth is known. In this practice, Mean Squared Error (MSE) or Peak Signal to Noise Ratio (PSNR) between the restored and ground-truth images is used as objective functions for selecting optimum value of operational parameters of denoising algorithms. The parameters learned like-wise will be used on real-time images with unknown ground-truth. Perhaps, the parameters fixed according to minimum MSE criterion on standard images may be sub-optimal for real-time images. Stein's Unbiased Risk Estimate (SURE) is considered as an indirect method for of predicting MSE, in the absence of ground-truth, from the denoised output and noisy observation of the image. Hence, usage of SURE as objective function for selecting the operational parameters of denoising algorithms on real-time images is common in image processing.

Fundamentally, SURE is based on the assumption that noise which corrupts the image follows Gaussian distribution. There were attempts in literature to modify SURE to make it suitable for other distributions. In this direction, Montagner *et al*. [1] introduced an extension of SURE suitable for noise with mixed Poisson–Gaussian distribution. In line with this, a generalization of SURE for exponential families was derived by Y. C. Eldar [2]. SURE depends on the kernel or the function used to compute noise-free estimate of the image. Derivation of SURE corresponding to complex denoising algorithms is tedious. To alleviate the complexity of deriving SURE for arbitrary denoising kernels or filter functions, Ramani *et al*. [3] proposed a Monte-Carlo technique which does not require any functional form of the denoising operator. In this technique, SURE is derived from the response of the denoising function or operator to a noisy input of known characteristics.

As already pointed out, the practice of using minimum SURE as target criterion for selecting the operational parameters of denoising algorithms on real-time images is common in image computing. For example, SURE was used by D.V.D. Ville and M. Kocher [4-5] as target function for selecting optimum values of operational parameters of Non-Local Means (NLM) filter. Similarly, minimum SURE criterion was followed by Ramani *et al*. [6] to optimize regularization parameter of Total Variation (TV) minimization denoising algorithm. For the selection of the threshold in wavelet-based image denoising [7-9] also SURE is used as objective function. SURE was employed by Qiu *et al*. [10] for determining the optimal affine transform coefficients in the local linear denoising algorithm.

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

In spite of the extensive application of SURE, a close analysis of its performance on standard test images is missing in literature. Pearson's Correlation of SURE with MAE between denoised images and ground-truth is analyzed in this paper, on Shepp-Logan Phantom and simulated Magnetic Resonance (MR) images, at different noise levels. Such an objective analysis of performance of SURE is first one of its kind. One of the major highlights of this paper is the demonstration of a systematic way of formulating SURE functions for denoising kernels, on a standard Gaussian smoothing kernel. Pitfalls associated with derivation of SURE for arbitrary denoising kernels and its application are discussed in this paper. Inferences drawn out of the objective analysis is versatile as a road-map comprising guidelines for practical usage of SURE.

Formulation of SURE for a Gaussian smoothing kernel and specifications of Shepp-Logan Phantom as well as simulated MR images are discussed in section 2 of this paper. In section 3, correlation of SURE with MAE among denoised and ground-truth images is analyzed, on Shepp-Logan Phantom and simulated MR images, at various noise levels.

II. METHODOLOGY

SURE is computed from the noisy image, denoised output and the noise estimate. The generic expression of SURE is [4-11],

$$SURE(\hat{y}) = \frac{1}{MN}\,\|\hat{y} - x\|^2 + \frac{2\sigma_n^2}{MN}\,div_x(\hat{y}) - \sigma_n^2 \qquad (1)$$

In (1),

$$x = y + n \qquad (2)$$

where, 'x' is the noisy observation and 'y' is the noise-free image. '$\hat{y}$' represent approximate estimate of the noise-free image, 'y', computed using an arbitrary filter function or kernel. 'n' is the noise vector. Basic assumption behind SURE is that the noise which corrupts the image is 'additive white Gaussian'. 'σ_n' is standard deviation of additive white Gaussian noise. $||\hat{y} - x||^2$ stands for L_2 norm or Euclidean distance between denoised estimate and noisy observation. The normalization factor in (1) is the total number of pixels in the image whose dimension is M×N. The divergence function in (1) is given by [4-11],

$$div_x(\hat{y}) = \sum_{\hat{y}} \frac{\partial \hat{y}_u}{\partial x_u} \qquad (3)$$

For an ordinary Gaussian smoothing kernel, the intensity of an arbitrary pixel in the denoised estimate, '$\hat{y}_u$' can be expressed as a weighted sum of its neighborhood pixels. It is expressed as,

$$\hat{y}_u = \sum_{x_v \in \Theta_u} G_{uv} x_v \qquad (4)$$

In (4), 'Θ_u' represent the neighborhood of 'x_u'. Weights in the Gaussian kernel, 'G' can be computed as,

$$G_{uv} = \frac{e^{-\left(\frac{(u-v)^2}{2\sigma^2}\right)}}{\sum_{\Theta_u} e^{-\left(\frac{(u-v)^2}{2\sigma^2}\right)}} \qquad (5)$$

In (5), 'σ' is the SD of the Gaussian function. The differential term in the divergence function in (3) when the generic expression of Gaussian smoothing operation given in (4) is substituted in it,

$$\frac{\partial \hat{y}_u}{\partial x_u} = \frac{\partial \sum_{x_v \in \Theta_u} G_{uv} x_v}{\partial x_u} \qquad (6)$$

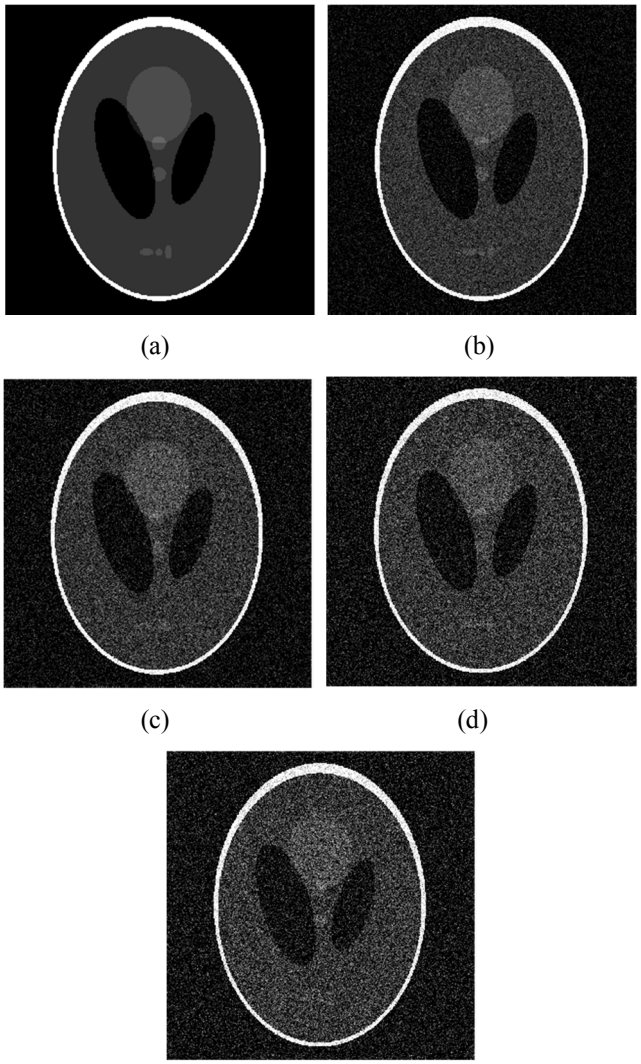
It can be noted in (5) that underlying function of Gaussian smoothing kernel is simply the function of the spatial distance, $(u\text{-}v)^2$, between the contextual pixel, 'x_u' and one of its arbitrary neighbors 'x_v', $x_v \in \Theta_u$. The function is differentiable only when u=v. When this condition is satisfied, the Gaussian function becomes equal to one. As the whole differential term in (3) reduces to one, the divergence function in (3) becomes equal to MN. Eventually, for an ordinary Gaussian smoothing operation, SURE reduces to,

$$SURE(\hat{y}) = \frac{1}{MN}\,\|\hat{y} - x\|^2 + \sigma_n^2 \qquad (7)$$

As the estimate of noise, 'σ_n' is a constant value, SURE depends only on the L_2 norm between denoised estimate and noisy observation, as evident in (7). In effect, dynamic variability of SURE and its concordance with MAE shall be compromised. In the case of any arbitrary denoising function, accuracy of SURE depends on the accuracy of the method employed for estimating SD of noise in the noisy observation. Highly reliable noise estimation techniques are rare. This fact makes the paradigm more complicated. In the present study, SD of noise is computed using a simple model, $\sigma = 1.4826 \times MAD\{x\}$ [12]. MAD stands for Median Absolute Deviation and {x} represent lexicographically ordered pixel intensities within the noisy observation 'x'.

Shepp-Logan Phantom [13] used in the experiment is generated in Matlab®. Noisy phantom images (forty distinct noise levels, Standard Deviation (SD) of noise = $0.01 \le \sigma \le 0.04$) are produced by adding white Gaussian noise to the ground-truth. Simulated MR images are produced by an MRI simulator available online named as BrainWeb [14-15]. The MR slices in this database belong to three pulse sequences, which are Proton Density (PD), T1 and T2. The inter-slice thickness of the simulated images employed in this experiment is 5 mm. Noise-free and noisy-images with noise levels, 3%, 5%, 7% and 9% are used. A few examples of images used in the experiment are furnished in figure 1and figure 2 for the reference of readers.

Denoised images which differ in terms of MAE against ground-truth are produced by varying the standard deviation of a Gaussian smoothing kernel ($0.01 \le \sigma \le 0.04$) of fixed dimension, 9×9. The degree of smoothing increases with SD of Gaussian mask. All experimental analysis is performed in Matlab®.

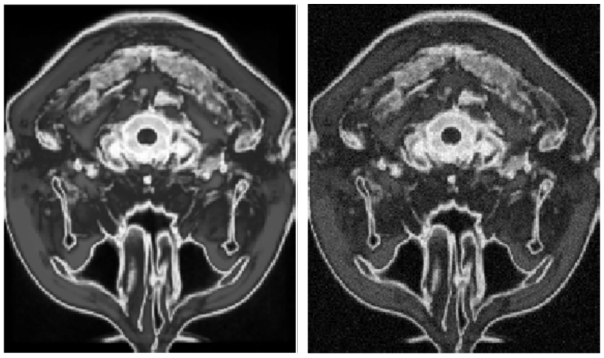


(a) (b)

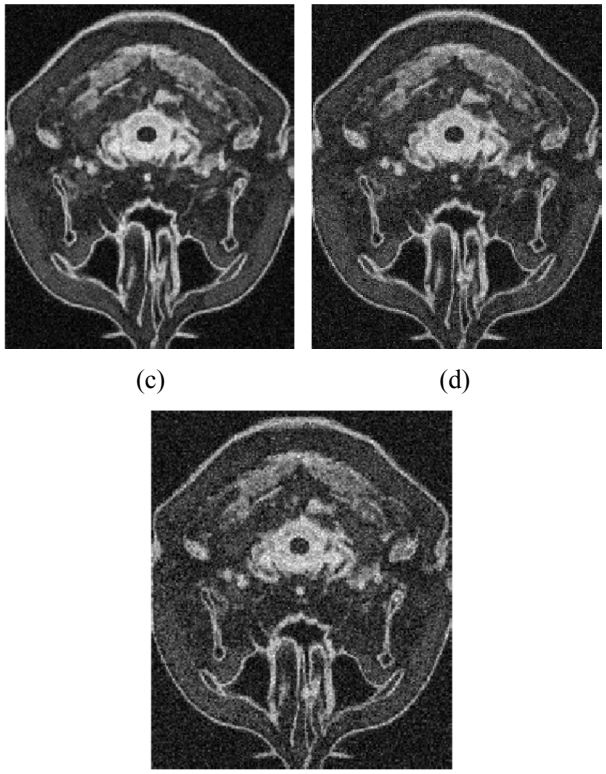
(c) (d)

(e)

Fig 1: Shepp-Logan Phantom (a) Noise-free observation (b) Noisy observation 1 ($\sigma = 0.01$) (c) Noisy observation 2 ($\sigma = 0.02$) (d) Noisy observation 3 ($\sigma = 0.03$) (e) Noisy observation 4 ($\sigma = 0.04$)



(a) (b)



(c) (d)

(e)

Fig 2: Simulated MR (a) Noise-free observation (b) Noisy observation 1 ($\sigma = 3\%$) (c) Noisy observation 2 ($\sigma = 5\%$) (d) Noisy observation 3 ($\sigma = 7\%$) (e) Noisy observation 4 ($\sigma = 9\%$)

III. RESULTS & DISCUSSIONS

Variation of MAE and SURE against SD of Gaussian mask on Shepp-Logan for various noise levels are shown in figure 3 and figure 4, respectively. Variation of MAE and SURE against SD of Gaussian mask on simulated MR images for various noise levels are shown in figure 5 and figure 6, respectively. MAE comes down as the SD of Gaussian kernel increases and starts increasing slowly after certain value of SD, as apparent in figure 3. As SD of the kernel increases, noise gets eliminated more effectively and denoised image comes closer to the ground-truth. As SD is increased further, MAE starts increasing due to the blurring effect. MAE is comparatively lower at low noise levels than high noise levels. MAE versus SD of Gaussian kernel shifts up as the noise level increases. In figure 5, the variation of MAE with respect to the variation in SD of Gaussian kernel is more ideal and curvy. SURE do not follow the pattern of MAE, faithfully. Perhaps it could be due to the poor dynamic variability of SURE. The reason for poor dynamic variability of SURE has been discussed in the previous section. Further inferences about the concordance of SURE with MAE is drawn out from the Pearson's correlation between SURE and MAE, in the discussions below.

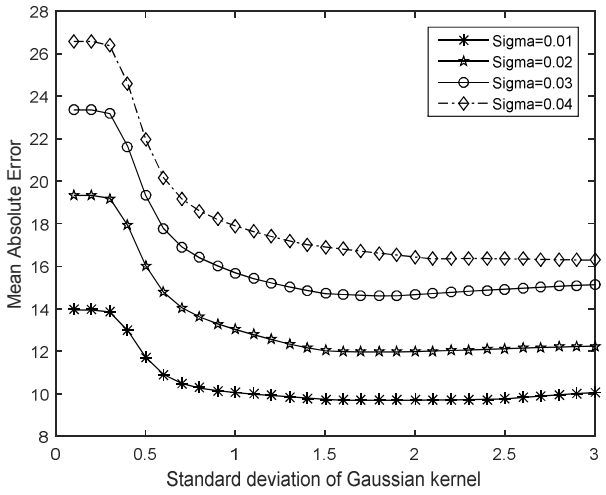


Fig 3: Variation of MAE against SD of Gaussian mask on Shepp-Logan, for various noise levels

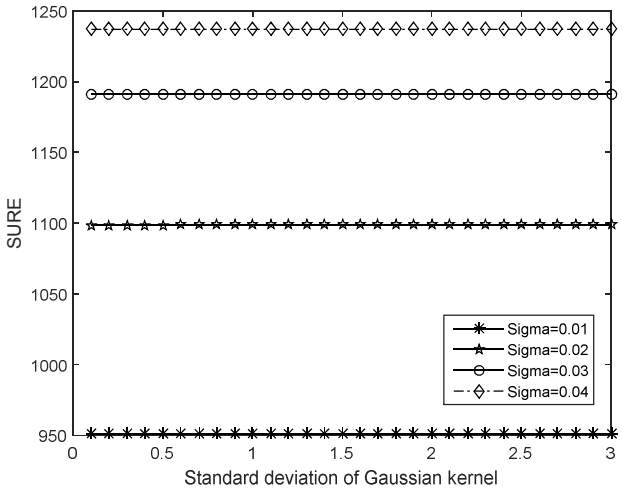


Fig 4: Variation of SURE against SD of Gaussian mask on Shepp-Logan, for various noise levels

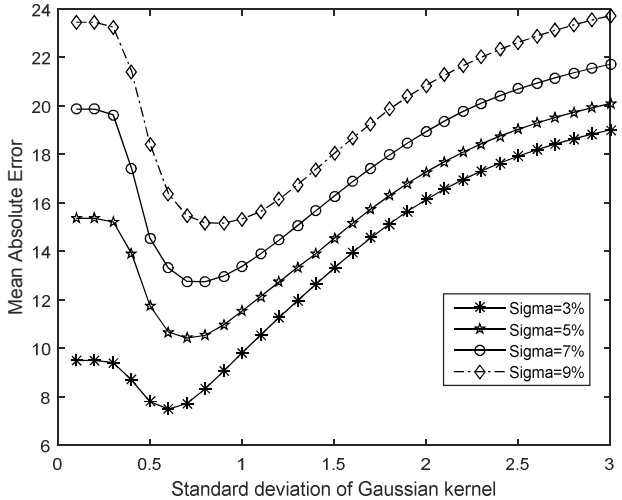


Fig 5: Variation of MAE against SD of Gaussian mask on simulated MR, for various noise levels

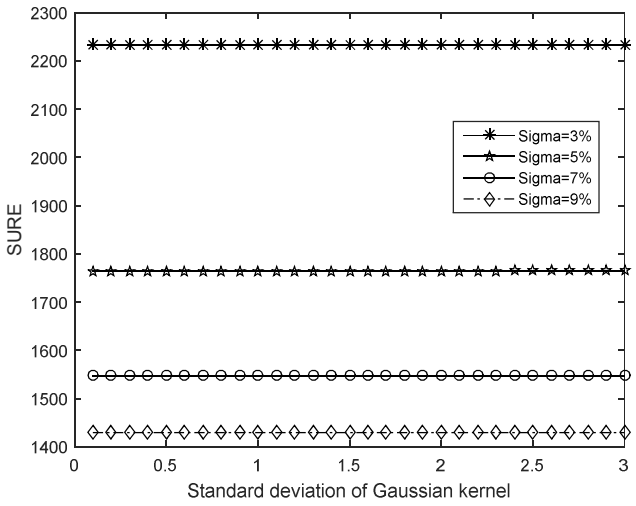


Fig 6: Variation of SURE against SD of Gaussian mask on simulated MR, for various noise levels

Values of Pearson's correlation coefficient (r) between SURE and MAE on Shepp-Logan Phantom and simulated MR images at different noise levels are shown in table I. On Shepp-Logan Phantom, values of correlation coefficient are close to -1 at all four noise levels. SURE is observed to be well-correlated with MAE, on Shepp-Logan, regardless of the noise level. In fact, the value of 'r' is expected to be positive as SURE is supposed to be directly proportional to MAE. On simulated MR images, values of 'r' are positive at all noise levels. However, concordance of SURE with MAE is observed to be poor on simulated MR images, especially at higher noise levels (σ = 7% & σ = 9%). In summary, values of correlation between SURE and MAE on Shepp-Logan and simulated MR images are r = -0.99 ± 0.02 (for 40 various noise levels, $0.01 \leq \sigma \leq 0.04$) and r = 0.48 ± 0.36 (for 4 various noise levels, σ = 3%, σ = 5%, σ = 7% & σ = 9%), respectively. The reason for discrepancy regarding proportionality of SURE to MAE between Shepp-Logan and simulated MR is unknown. This observation questions the reliability of the test images or SURE itself. Both Shepp-Logan and simulated MR images are used in medical image computing, over the last decade.

TABLE I: PEARSON'S CORRELATION BETWEEN MAE AND SURE ON SHEPP-LOGAN & PHANTOM AT VARIOUS NOISE LEVELS

Test Image	Noise Level	Pearson's Correlation	Summary
Shepp–Logan	σ = 0.01	-0.9608	-0.99 ± 0.02
	σ = 0.02	-0.9946	
	σ = 0.03	-0.9957	
	σ = 0.04	-0.9990	
Simulated MR	σ = 3%	0.8980	0.48 ± 0.36
	σ = 5%	0.6103	
	σ = 7%	0.3252	
	σ = 9%	0.0725	

IV. CONCLUSION

Pearson's Correlation of SURE with MAE between denoised images and ground-truth was analyzed in this paper, on Shepp-Logan Phantom and simulated MR images, at various noise levels. SURE did not exhibit appreciable

concordance with the benchmark metric, MAE, on simulated MR images, especially at higher noise levels, even though the correlation was satisfactory on Shepp-Logan. SURE is observed to be suitable for denoising kernels whose elements or weights are computed from the pixel intensities in the noisy observation. For example, elements in the Gaussian smoothing kernel depend only on the spatial distance between contextual and its neighborhood pixels and not on the intensity of pixels in the noisy observation. In such situations, the divergence function in SURE reduces to a constant. As SD of noise in the noisy observation is a constant value, SURE will depend only on the Euclidean distance between denoised image and noisy observation.

Unlike a Gaussian smoothing kernel, in the kernel of a bilateral filter, weights in the radiometric kernels are computed from the difference between intensities of the contextual pixel and its neighborhood pixels, within the noisy observation. As the Gaussian function corresponding to the radiometric kernel in bilateral filter comprises pixel intensity of noisy observation, the function is fully differentiable by the noisy observation. Whereas, the Gaussian function corresponding to an ordinary Gaussian smoothing kernel is only minimally differentiable. There is an important take-home message that before putting SURE for optimizing parameters of an arbitrary filter kernel or function, prospect of SURE on the specific kernel need to be investigated analytically.

REFERENCES

[1] Y. Le Montagner, E. D. Angelini and J. Olivo-Marin, "An Unbiased Risk Estimator for Image Denoising in the Presence of Mixed Poisson–Gaussian Noise," in IEEE Transactions on Image Processing, vol. 23, no. 3, pp. 1255-1268, March 2014.

[2] Y. C. Eldar, "Generalized SURE for Exponential Families: Applications to Regularization," in IEEE Transactions on Signal Processing, vol. 57, no. 2, pp. 471-481, Feb. 2009.

[3] S. Ramani, T. Blu and M. Unser, "Monte-Carlo Sure: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms," in IEEE Transactions on Image Processing, vol. 17, no. 9, pp. 1540-1554, Sept. 2008.

[4] D. Van De Ville and M. Kocher, "Nonlocal Means With Dimensionality Reduction and SURE-Based Parameter Selection," in IEEE Transactions on Image Processing, vol. 20, no. 9, pp. 2683-2690, Sept. 2011.

[5] D. Van De Ville and M. Kocher, "SURE-Based Non-Local Means," in IEEE Signal Processing Letters, vol. 16, no. 11, pp. 973-976, Nov. 2009.

[6] S. Ramani, Z. Liu, J. Rosen, J. Nielsen and J. A. Fessler, "Regularization Parameter Selection for Nonlinear Iterative Image Restoration and MRI Reconstruction Using GCV and SURE-Based Methods," in IEEE Transactions on Image Processing, vol. 21, no. 8, pp. 3659-3672, Aug. 2012.

[7] F. Luisier and T. Blu, "SURE-LET Multichannel Image Denoising: Interscale Orthonormal Wavelet Thresholding," in IEEE Transactions on Image Processing, vol. 17, no. 4, pp. 482-492, April 2008.

[8] T. Blu and F. Luisier, "The SURE-LET Approach to Image Denoising," in IEEE Transactions on Image Processing, vol. 16, no. 11, pp. 2778-2786, Nov. 2007.

[9] F. Luisier, T. Blu and M. Unser, "A New SURE Approach to Image Denoising: Interscale Orthonormal Wavelet Thresholding," in IEEE Transactions on Image Processing, vol. 16, no. 3, pp. 593-606, March 2007.

[10] T. Qiu, A. Wang, N. Yu and A. Song, "LLSURE: Local Linear SURE-Based Edge-Preserving Image Filtering," in IEEE Transactions on Image Processing, vol. 22, no. 1, pp. 80-90, Jan. 2013.

[11] H. Kishan and C. S. Seelamantula, "Sure-fast bilateral filters," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 1129-1132.

[12] A. Foi, M. Trimeche, V. Katkovnik and K. Egiazarian, "Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data," in IEEE Transactions on Image Processing, vol. 17, no. 10, pp. 1737-1754, Oct. 2008.

[13] A.K. Jain, Fundamentals of Digital Image Processing, Englewood Cliffs, NJ, Prentice Hall, pp. 439, 1989.

[14] C.A. Cocosco, V. Kollokian, R.K.S. Kwan and A.C. Evans, “BrainWeb: Online Interface to a 3D MRI Simulated Brain Database,” NeuroImage, vol. 5, issue 4, 1997.

[15] J. Yang, J. Fan, D. Ai, S. Zhou, S. Tang and Y. Wang, “Brain MR image denoising for Rician noise using pre-smooth non-local means filter,” BioMedical Engineering On Line, vol. 14, issue 2, pp. 1-20, 2015.

NPRank: Nexus based Predicate Ranking of Linked Data

Sakthi Murugan R., Ananthanarayana V.S.
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, Mangalore - 575025, Karnataka, India.
sakthimuruga@gmail.com

***Abstract*—In the typical use case of browsing Linked Data in DBpedia, the user would find an average of 180 facts attached to each entity. These facts are ordered alphabetically based on predicates, but a logical ordering of these facts is a better option. In this article, we present a Nexus based predicate ranking of Linked Data facts named NPRank. The key idea of NPRank is, the importance of a predicate is directly proportional to its familiarity among its group called Nexus. NPRank is a language and endpoint independent model allowing seamless integration and querying of data from multiple endpoints. Nexus score generated to rank predicates also assists in fragmentation of large data and bring in more hidden data from the SPARQL endpoints. Our experiments, conducted with the ranking of the Linked Data facts, corresponding to most visited pages of Wikipedia; from 275 active SPARQL endpoints, achieves better performance than the state-of-the-art methods.**

***Index Terms*—Linked data, Predicate Ranking, Semantic Relationship.**

I. Introduction

Linked Data refers to the collection of interrelated structured data on the web, in machine-readable format [1]. Linked Data lies at the heart of the Semantic Web [2]. Linked Data provides a common framework to publish and exchange of structured data on the web. The Linked Data is represented using RDF data model [3]. In RDF, the unit data is called facts, which are expressed in the form of subject-predicate-object called triples. The subject is also termed as the concept, the predicate is called as property or relationship, and the object may be specified as a concept or a value.

Ontologies are used to represent the relationship; some of the good ontologies include Schema.org [4], FOAF [5] and Dublin Core [6]. SPARQL is the standard to query RDF data [7]. Many domains publish their Linked Data via SPARQL endpoints which are accessible via SPARQL queries. DBpedia is the nucleus of Linked Lata, containing data extracted from the Wikipedia webpage and has links connecting other endpoints [8].

DBpedia[1] has 1628825380 triples with 9079963 distinct subjects, wherein typical use case of browsing in DBpedia, the user would find an average of 180 facts attached to each entity. In case, if the user would like to browse an entity from all the SPARQL endpoints available over the internet, then the number of triples and its predicates will be huge.

For instance, a user looking for the country United States from all the SPARQL endpoints[2] will be in front of 87650 triples containing 806 distinct predicates with the same data being repeated across the endpoints. Supposing the user is interested in knowing the President of United States, he will be finding it difficult to reach the President predicate, since the list also contains many other unuseful predicates.

Machines can track the entities as the user browse through, but not the predicates; since a single entity may have a list of predicates, and it is tough to get user interested predicate unless the user explicitly queries for the predicate. In DBpedia, the predicates are alphabetically ordered, but a logical ordering of these predicates is much better than the alphabetical ordering.

In this article, we have presented a Nexus based ranking of predicates. The term 'Nexus' is used in this article in the context '*relationships connecting to the core*'. For instance, in case of Nexus of 'Book' its 'Title', 'Author', 'Language' and 'Category' are some of its attributes needs to be ranked. Similarly, in case of concept Nexus of 'Country' its 'President', 'capital' and 'population' are some of its attribute which needs to be ranked. The ranking of these attributes is based on Nexus score. NPRank also makes the facts human-readable by clustering the predicates based on Nexus. In Linked Data browsing, the users need to visit different SPARQL endpoints to get the information, but the NPRank will act as a one-point frontend to access all the SPARQL endpoint.

The structure of the article is as follows. Preliminaries considered for the work are described in section II. Section III presents the related works. Section IV provides a detailed description NPRank ranking technique. Section V presents the experimental results. Finally, section VI presents the conclusion and future work.

[1]http://dbpedia.org/sparql, Accessed on 9th March 2018.

[2]We have considered 275 endpoints, which is discussed in section V

978-1-7281-2087- 4/19/$31.00 ©

II. Preliminaries

Table I contains the list of prefix used this article to shorten the URI, i.e., the URI http://www.w3.org/1999/02/22-rdf-syntax-ns#Property is shortened as rdf:Property. The prefix example is used for illustration of examples.

TABLE I
Prefix

Prefix	URL
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
owl	http://www.w3.org/2002/07/owl#
dc	http://purl.org/dc/elements/1.1/
gold	http://purl.org/linguistics/gold/
foaf	http://xmlns.com/foaf/0.1/
skos	http://www.w3.org/2004/02/skos/core#
dbr	http://dbpedia.org/resource/
dbo	http://dbpedia.org/ontology/
dbp	http://dbpedia.org/property/
ogp	https://ogp.me/ns#
example	http://example.org/

III. Related Works

A. Predicate Ranking

Very few works are there in ranking the predicates.

Dessi and Atzori [9], [10] proposed a Machine Learned Ranking (MLR) approach to rank the RDF predicates among entities. MRL uses the machine learning approach with nine different features. In particular, we are interested in the main feature, i.e., ranking based on the frequency of the predicate. This model considers the frequency of the predicate across the data, for example, the count of dbo:birthDate is almost used across all the domain will have a higher frequency. But in some domain like Actor predicates with the overall low score like dbp:famousMovie and dbo:award are more important than the dbo:birthDate. The other setback of this model is the feature IsEnglish where the results are biased towards English predicates. The training methods adopted by MRL are supervised by semantic web expert and students. It will be more complex to train the semantic web to rank based on the user, as not only the data is enormous, but also the ontology is huge, and diverse across various domains. With DBpedia alone containing 483605 number of distinct rdf:types and we require domain experts in each field to rank. For example, the bioinformatics ontology and medical ontology requires the expert with the domain knowledge to order the predicates.

Lee et al. [11] proposed Semantic Search Ranker (SSR) with three measures to rank the semantic search results Number of meaningful semantic paths, Coverage of keywords and Discriminating power of keywords. The semantic path weight is computed using the number of ways the entity can be reached in the RDF graph. Example, for any book entity, the predicate example:hasTitle only one title value, but the predicate example:writtenBy can have many author names assigned to it. So the author computes higher rank for example:hasTitle and lower rank for example:writtenBy. But SSR didnt address the added case such as a book has many single-valued predicated like example:title, example:publisher, example:publishingDate, example:volume, example:issue and so on. SSR didnt address the method to rank among those with same semantic path weight.

Ruback et al. [12] proposed SELEcTor to find the similarity between two entities in Linked Data by ranking and comparing their features. SELEcTor generates the ranked features with the help of SPARQL query. The SPARQL query matches the path pattern to get all the features, and rank based on the count. But some features like rdf:type, dc:subject, owl:sameAs, rdf:label, rdf:comment, etc., are standard across entities which are of an entirely different domain which may be ranked similarly by SELEcTor. Feature identification of SELEcTor is aided by a domain expert.

Cyrille et al. [13] proposed Holistic Ranking for RDF Entities (HARE) to rank resources, predicates, literals and triples. HARE ranks the predicates considering it as a resource which is general and common across the domain. But in our approach, the rank for the predicates are specific to the domain (class of ontology).

B. Ranking RDF

Arnaout and Elbassuoni [14] proposed a framework to rank the triple pattern in relevance to keywords of RDF knowledge graph. They also proposed a query relaxation technique if the exact keyword match didnt match any triple pattern.

Marx et al. [15] proposed RDF ranking using the real user query logs. But the actual user query is mostly related to domain and range, and minimal for predicates. This ranking method is perfect in ranking classes, and when it comes to predicates, the user query logs alone is not sufficient.

Sahar et al. [16] proposed Dual Walk based Ranking (DWRank) for ranking ontologies. DWRank uses training-based approach to compute the rank based on the connectivity of the ontologies within and with other ontologies. This approach is most preferred by knowledge engineers to find the ontologies, and not best suited for ranking the RDF data.

Noia et al. [17] proposed a Semantic Path-based Ranking (SPrank) for ranking concepts. SPrank uses machine learning approach on user interactions like clicks, purchases, video watching to understand the likes and dislikes of the user and recommend the concepts that user may be interested in based on it.

Motivated by these techniques, we focused on a different approach to rank the RDF predicates.

IV. NPRank

NPRank is a language and endpoint independent, predicate-based ranking model, used to rank the triples of a concept. To

implement NPRank, we need to get all the types of Nexus and get the list of predicates associated with each Nexus. Once we got the predicates, we need to find the Nexus score, which will indicate the rank of the predicate for a given Nexus. NPRank will also bring all the triples related to the concept from all the available SPARQL endpoints and present to the user in a readable format.

A. Nexus Generation

The publicly available SPARQL endpoints can be found at W3C[3] and SPARQLES[4] [18]. SPARQLES is a public SPARQL endpoint monitoring website which provides the availability of the SPARQL endpoints. Get all the SPARQL endpoints and remove the duplicate SPARQL endpoints, where the same data is being hosted in multiple URLs.

In DBpedia, the Nexus can be identified using the gold:hypernym property. The rdf:type is not suitable for determining Nexus as a single subject may be associated with more than one rdf:type. The SPARQL query 1 is used to extract all the available gold:hypernym. The distinct values of the gold:hypernym obtained from different SPARQL endpoints make the Nexus.

SPARQL Query 1: Select Hypernyms and Count

```
SELECT
 DISTINCT(?hypernym), (COUNT(?hypernym)
          AS ?hcount)
 WHERE {
  ?subject gold:hypernym ?hypernym.
 } ORDER BY DESC(COUNT(?hypernym))
```

Once the Nexus is generated, the properties associated with the Nexus is obtained by passing the gold:hypernym URI in place of example:Country in SPARQL query 2.

SPARQL Query 2: Select Predicates and Count

```
SELECT
 DISTINCT(?predicate), (COUNT(?predicate)
          AS ?pcount)
 WHERE {
  ?subject gold:hypernym example:Country;
           ?predicate ?object.
 } ORDER BY DESC(COUNT(?predicate))
```

The SPARQL query 2 is repeated with all the SPARQL endpoints to obtain the predicate and its respective count across the Nexus. The predicate obtained from different SPARQL endpoints with the same Nexus are then clustered together.

[3]https://www.w3.org/wiki/SparqlEndpoints, Accessed on March 10th, 2018.
[4]http://sparqles.ai.wu.ac.at/api/endpoint/list

B. Nexus Score

The predicates are first split into two categories,

1) Common predicates
2) Core predicates

1) Common Predicates: The common predicates are those like rdf:type, owl:sameAs, dc:subject, rdfs:label, dbo:wikiPageID, foaf:primaryTopic, rdfs:comment, dbo:abstract and foaf:name which appears across different Nexus. The common predicates are by default aligned as in figure 1 and are customisable as per user preference.

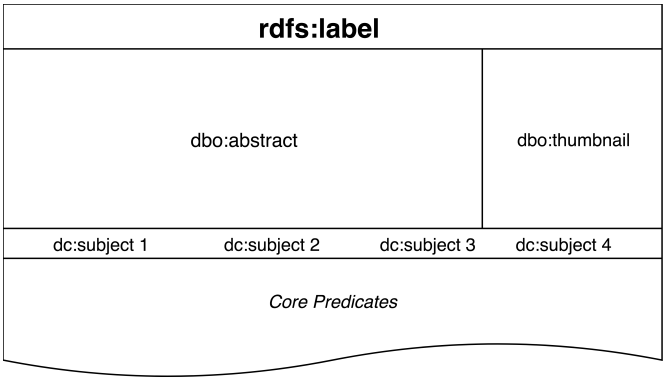


Fig. 1. Display of Common Predicates in NPRank

2) Core Predicates: Core predicates are those other than the common predicates which define the Nexus. Nexus scores are generated for core predicates with the motive '*The importance of predicate is directly proportional to its familiarity among its group*'. Let '*p*' be the predicate belonging to Nexus '*n*'. Let $Count_{p,n}$ be the count of the predicate '*p*' in Nexus '*n*'. Let $Count_{a,n}$ be the count of all the predicates in Nexus '*n*'. The Nexus score for the predicate '*p*' in Nexus '*n*' is computed using the formula 1.

$$NexusScore_{p,n} = \frac{Count_{p,n}}{Count_{a,n}} \tag{1}$$

a) Note:: The same predicate will have different Nexus score in different Nexus.

C. Retrieving

To retrieve all the data corresponding to a subject, we must first get all the owl:sameAs URI corresponding to the given subject. The triples corresponding to the given URI and all the owl:sameAs URIs are crawled using the SPARQL query 3. False owl:sameAs can be identified with the Nexus match of the two concepts. The SPARQL query 3 also retrieves the rdfs:label corresponding to predicate and object whose application is discussed in section IV-D. We also need to bring the indirect triples which can be obtained by from owl:symmetricProperty and owl:transitiveProperty.

a) Cleaning the Triples:: Once the triples are retrieved from different SPARQL endpoints, it is merged, and duplicates are removed. Indirect duplicate values such as rdfs:subPropertyOf and owl:equivalentProperty are merged.

SPARQL Query 3: Select Labels of Predicates and Objects

```
SELECT
 ?pred ?obj ?predLabel ?objLabel
WHERE{
 example:UnitedStates ?pred ?obj.
  OPTIONAL{ ?pred rdfs:label ?predLabel.
   FILTER langMatches(
              lang(?predLabel),"en").}
  OPTIONAL{ ?obj rdfs:label ?objLabel.
   FILTER langMatches(
              lang(?objLabel),"en"). }
}
```

D. Representation

In case of triple, for any given $< subject >$, the $< predicate >$ is always displayed as URI, and $< object >$ are displayed as URI in most of the cases. There are many works concentrate on mapping the natural language query to URI, but displaying natural language text to the user was not the main concern. We display the human-readable text from skos:prefLabel, rdfs:label and rdfs:comment filtered with the language of user preference to the replaceable URI. In SPARQL query 3, we pass the user prefered language, and bring the rdfs:label with the user prefered language and display it to the user. We also display image and video URI embedded directly as image and videos and not as links, thereby increasing human readability.

The readability of the triples can be improved by grouping the relevant predicates on rdf:subPropertyOf. In case of actor displaying personal information like example:name, example:dob, example:birthPlace and predicates related to work like example:movies, example:numberOfFilms, example:awards separately improves the readability. Grouping allows the user to look for the information quickly by navigating to the group. The ordering of the groups is based on the uniqueness of the group towards the Nexus.

V. Experimental Results

This benchmark suite for NPRank includes a collection of top resources from the most visited pages of Wikipedia [19]. We have crawled 776 SPARQL endpoints from W3C[3] and SPARQLES[4] [18], out of which only 275 endpoints where available online for queriying, and the rest were offline[5]. Table II shows the amount of data available in the SPARQL endpoints.

We generated the about 40958 Nexus from DBpedia. Since some of them are similar we merged those with predicate match greater than 70%. We set the threshold to find the common predicate as 70%, i.e., those perdicate occur in more than 70% of the Nexus are filtered as common predicate. We then generated the Nexus score for the core predicates.

[5]Queried on March 10th, 2018.

TABLE II
Measures from Available SPARQL Endpoints

No. of Available SPARQL endpoints	275
Sum of Triples	26552150101
Sum of Distinct Subjects	471971762
Sum of Distinct Predicates	1769459597
Sum of Distinct Types	3167816
Sum of rdfs:lables	286533038
Sum of Object as Values	12621058948

We got the top viewed pages of Wikipedia across fourteen different categories and conducted our evaluation. Table III shows the amount of data crawled in each category. The data corresponding to the resources is obtained from all the available SPARQL endpoints. We have crawled about 26318000 triples from all the available SPARQL endpoints and cleaned it to 153404 triples. There is a 4390% increase in the number of triples and 618.18% increase in the number of predicates when compared to DBpedia. Predicates range from a minimum of 49 to 1176 with an average of 395 predicates per concept. If the user is looking for triples from any single endpoint, he would get an average of 645 triples per endpoint and still face the challenge of viewing the same data being repeated across the endpoints. There are some semantic search engines which address the issue of bringing the data from multiple endpoints, but integrating and ranking at the level of predicates is still not adopted by most of it.

Evaluating precision, recall and F-Measure will have no impact, as we have used URI to retrieve all the relevant triples from all the available SPARQL endpoints. We have developed NPRank using Java with the support of Apache Jena[6] package.

We have also extracted microdata embedded in webpages for the same top resources from the most visited pages of Wikipedia. We have extracted 328086 microdata from the web search result of Google[7] and DuckDuckGo[8] for the keyword in rdfs:label for the respective resources. The predicates obtained are mostly with the prefix from the Open Graph protocol[9] like ogp:site_name, ogp:url, ogp:title and ogp:description with an average of five predicates per website. We couldnt get enough predicate variety to generate NPRank for microdata.

VI. Conclusion

We have presented NPRank, an alternate method to rank the predicates of a given resource in Linked Data. The Nexus of Linked Data makes it easier to cluster and rank the triples. Ranking the relationships on the Nexus can better sort the Linked Data with more relevant triples on top. With NPRank,

[6]https://jena.apache.org
[7]https://www.google.co.in
[8]https://duckduckgo.com
[9]http://ogp.me

TABLE III
RETRIEVED TRIPLES AND PREDICATES

Category	Title	No. of Endpoints	DBpedia Predicates	Predicates from all Endpoints	Distinct Predicates from all Endpoints	DBpedia Triples	Triples from all Endpoints	Distinct Triples from all Endpoints
Country	United States	241	141	153750	806	321	13947933	87726
City	New York City	243	112	220660	1176	311	4085522	20678
People	Donald Trump	241	53	61600	332	318	846961	4275
Singers	Michael Jackson	244	62	71377	389	385	993614	5409
Actors	Kim Kardashian	234	32	35011	196	186	275644	1495
Sportsmen	Cristiano Ronaldo	236	45	79760	431	265	990065	5320
Pre-modern people	Jesus	241	45	36027	202	254	726232	3944
3rd-millennium people	Willow Smith	234	29	35139	198	164	170766	944
Music bands	The Beatles	241	43	70833	363	257	1146209	6232
Sport teams	Manchester United F.C.	241	46	94004	499	237	839524	4489
Films and TV series	The Big Bang Theory	236	54	59035	311	236	448409	2374
Albums	Thriller	226	57	62583	336	246	449743	2463
Books and book series	Harry Potter	241	41	46447	242	150	1109830	6518
Pre-modern books	Inferno(Dante)	233	15	9022	49	89	287548	1537
	Average	238	55	73946	395	244	1879857	10957

we got an average of 618.18% increase in the number of predicates when compared to DBpedia. We also got an average of 4390% increase in the number of triples when compared to DBpedia. NPRank is suitable for Linked Data from SPARQL endpoints and not for microdata embedded in webpages.

REFERENCES

[1] T. Berners-Lee, "Linked data. design issues for the world wide web," *World Wide Web Consortium. http://www. w3. org/DesignIssues/LinkedData. html*, 2006.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[3] G. Klyne and J. J. Carroll, "Resource description framework (rdf): Concepts and abstract syntax," 2006.

[4] R. V. Guha, D. Brickley, and S. Macbeth, "Schema. org: evolution of structured data on the web," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, 2016.

[5] D. Brickley and L. Miller, "Foaf vocabulary specification 0.91," 2007.

[6] D. C. M. Initiative *et al.*, "Dcmi home: dublin core metadata initiative (dcmi)," 2016.

[7] S. Harris, A. Seaborne, and E. Prudhommeaux, "Sparql 1.1 query language," *W3C recommendation*, vol. 21, no. 10, 2013.

[8] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[9] M. Atzori and A. Dessi, "Ranking dbpedia properties," in *WETICE Conference (WETICE), 2014 IEEE 23rd International*. IEEE, 2014, pp. 441–446.

[10] A. Dessi and M. Atzori, "A machine-learning approach to ranking rdf properties," *Future Generation Computer Systems*, vol. 54, pp. 366–377, 2016.

[11] J. Lee, J.-K. Min, A. Oh, and C.-W. Chung, "Effective ranking and search techniques for web resources considering semantic relationships," *Information Processing & Management*, vol. 50, no. 1, pp. 132–155, 2014.

[12] L. Ruback, M. A. Casanova, C. Renso, and C. Lucchese, "Selector: discovering similar entities on linked data by ranking their features," in *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. IEEE, 2017, pp. 117–124.

[13] Axel-Cyrille, N. Ngomo, M. Hoffmann, R. Usbeck, and K. Jha, "Holistic and scalable ranking of rdf data," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec 2017, pp. 746–755.

[14] H. Arnaout and S. Elbassuoni, "Effective searching of rdf knowledge graphs," *Journal of Web Semantics*, vol. 48, pp. 66 – 84, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570826817300677

[15] E. Marx, A. Zaveri, D. Moussallem, and S. Rautenberg, "Dbtrends: Exploring query logs for ranking rdf data," in *Proceedings of the 12th International Conference on Semantic Systems*. ACM, 2016, pp. 9–16.

[16] A. S. Butt, A. Haller, and L. Xie, "Dwrank: Learning concept ranking for ontology search," *Semantic Web*, vol. 7, no. 4, pp. 447–461, 2016.

[17] T. D. Noia, V. C. Ostuni, P. Tomeo, and E. D. Sciascio, "Sprank: Semantic path-based ranking for top-n recommendations using linked open data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 9, 2016.

[18] P.-Y. Vandenbussche, J. Umbrich, L. Matteis, A. Hogan, and C. Buil-Aranda, "Sparqles: Monitoring public sparql endpoints," *Semantic Web*, vol. 8, no. 6, pp. 1049–1065, 2017.

[19] "Wikipedia: Multiyear ranking of most viewed pages," https://en.wikipedia.org/wiki/Wikipedia:Multiyear_ranking_of_most_viewed_pages, (Accessed on 03/14/2018).

Coefficient of Correlation for Spherical Fuzzy Sets in Computational Application

Abhishek Guleria
Department of Mathematics
Jaypee University of Information Technology
Waknaghat, Solan, INDIA
abhishekguleriahappy@gmail.com

Rakesh Kumar Bajaj
Department of Mathematics
Jaypee University of Information Technology
Waknaghat, Solan, INDIA
rakesh.bajaj@gmail.com

***Abstract*—Correlation and correlation coefficients are the most utilized statistical tools and important measures in the field of engineering, intelligence sciences, data analysis, decision making, biological sciences, etc. In the present communication, we have proposed a new measure of correlation coefficient of two T-spherical fuzzy sets based on the newly defined information energy measure under the perception of the four parameters of impreciseness - degree of membership, degree of indeterminacy (neutral), degree of non-membership and the degree of refusal. Further, by implementing the principle of maximum correlation coefficient over the proposed correlation coefficient, the methodology for solving the computational problem of pattern recognition has been provided with the help of an example. A comparative analysis in contrast with the existing methodologies has been presented with comparative remarks and additional advantages.**

***Index Terms*—Spherical fuzzy set, T-spherical fuzzy set, Information energy, Correlation coefficient, Pattern recognition.**

I. Introduction

The researchers in the field of fuzzy sets and information are well aware that various generalizations of the notion of fuzzy sets [20] and Intuitionistic Fuzzy Sets (IFSs) [1] have taken place to model the uncertainties and the hesitancy inherent in many practical circumstances for a wider coverage of flexibility. Yager [19] revealed that the existing structures of fuzzy set and intuitionistic fuzzy set are not capable enough to depict the human opinion in more practical/broader sense and introduced the notion of Pythagorean fuzzy sets which effectively enlarged the span of information by introducing the new conditional constraint.

The concept of membership/belongingness (yes), non-membership/non-belongingness (no) and indeterminacy/neutral (abstain) have been well described in the definition of intuitionistic fuzzy set as well as in the definition of Pythagorean fuzzy set. As an extension, if we consider an example of voting system where voters can be categorized into four different classes–one who votes for (yes), one who votes against (no), one who neither vote for nor against (abstain), one who refused for voting (refusal). It may be noted that the concept of 'refusal' is not being taken into account by any of the sets stated above.

978-1-7281-2087- 4/19/$31.00 © 2019 IEEE

In order to deal with such circumstances and to develop a concept which would be sufficiently close to humans nature of flexibility, Cuong [4] introduced the concept of picture fuzzy set in which all the four parameters, i.e., degree of membership, degree of indeterminacy (neutral), degree of non-membership and the degree of refusal have been taken into consideration. Recently, Mahmood et al. [13] introduced the notion of Spherical Fuzzy Set (SFS) and T-spherical Fuzzy Set (TSFS) which give additional strength to the idea of picture fuzzy set by broadening/enlarging the space for the grades of all the four parameters.

Next, Kifayat et al. [12] studied the geometrical comparison of fuzzy sets, intuitionistic fuzzy sets, Pythagorean fuzzy sets, picture fuzzy sets along with spherical and T-spherical fuzzy sets in detail. Also, they studied various existing similarity measures for intuitionistic fuzzy sets and picture fuzzy sets with their limitations that they could not be applied in the broader setup as of the spherical fuzzy set. Further, they proposed various types of similarity measures for TSFS with their usefulness in various fields. Garg et al. [6] presented a new improved interactive aggregation operators for TSFSs with application in decision-making. Also, Guleria and Bajaj [9] introduced the notion of eigen spherical fuzzy sets and devised an algorithm to find the greatest and the least eigen spherical fuzzy sets to solve some of the decision-making problems. Next, Guleria and Bajaj [8] successfully proposed the notion of T-spherical fuzzy soft set and studied some new aggregation operators along with some applications in the field of decision-making.

In the recent past, various researchers have extensively studied different types of information measures in connection with the correlation coefficients, similarity measures, entropy, distance measures, discriminant measures which are available in the literature. These measures have found many applications in the different areas of decision making, pattern recognition, econometrics, medical analysis, etc.

The notion of correlation and correlation coefficient are important statistical tools which play a vital role in various engineering applications viz. intelligence sciences, biological sciences, data analysis, pattern recognition, etc. The correlation coefficient for intuitionistic fuzzy set was first provided by Gerstenkorn and Manko [7]. For probability spaces, the

coefficient of correlation was studied by Hong and Hwang [10]. Next, Mitchell [14] proposed the correlation coefficient between IFSs by interpreting an IFS as an ensemble of fuzzy set. Various researchers developed different correlation coefficients for IFSs and interval-valued intuitionistic fuzzy sets with their viewpoint and applied them in the different fields ([15], [16]). Huang and Guo [11] proposed a revised and improved correlation coefficient of the intuitionistic fuzzy sets and generalized these correlation coefficients over the interval-valued intuitionistic fuzzy sets. They also discussed some properties of the proposed correlation coefficients and proposed methodology for solving the problem of medical diagnosis & clustering analysis. Chen [3] introduced a correlation based closeness index for interval-valued Pythagorean fuzzy set and discussed its properties. By utilizing the correlation-based closeness index, an algorithm for solving multi-criteria decision making problem under the interval-valued Pythagorean fuzzy environment has also been provided. The paper also demonstrated the feasibility and effectiveness of the proposed methodology through a comparative analysis in contrast with the well known existing methods. Singh [17] introduced the concept of correlation coefficient for picture fuzzy sets as an extension of the correlation coefficient for IFSs and also proposed the weighted correlation coefficients for the picture fuzzy sets with their application in clustering analysis.

In case of picture fuzzy set, the restriction corresponding to the degree of membership, the degree of neutral membership (abstain) and the degree of non-membership is that their sum is less than or equal to 1, where the condition in case of SFSs is that their squared sum is less than or equal to 1. This difference in constraint conditions gives an additional advantage for a wider coverage of information span in dealing with the correlation coefficients for the spherical fuzzy sets. Hence, in order to have wider utility of the information measures, we are extending and proposing new correlation coefficients for TSFSs and SFSs.

The outline of the present work is as follows. In section 2, we study some basic preliminaries in reference with the spherical fuzzy set, T-spherical fuzzy set and correlation coefficient. Considering the fact that T-spherical fuzzy sets have the immense capability to model the imprecise, vague, uncertain or incomplete information inherent in the real-world applications, we propose correlation coefficient of two T-spherical fuzzy sets with respect to the proposed information energy and correlation function in section 3. Particular cases have also been dealt. Further, the proposed correlation coefficient has been utilized for proposing a new methodology for solving the computational application problem of pattern recognition in section 4. An illustrative example has also been provided for the application under consideration. In section 5, a comparative analysis depicting the advantages and listing some important remarks has been carried out. Finally, the paper is concluded in section 6 by stating the scope for the future work.

II. Preliminaries

In this section, we study some important notions in connection with spherical fuzzy set, T-spherical fuzzy set and correlation coefficient, which are available in the basic literature.

Definition 1: [13] A T-spherical fuzzy set S in X is given by

$$S = \{< x, \mu_S(x), \eta_S(x), \nu_S(x) >| \; x \in X\};$$

where $\mu_S : X \to [0,1]$, $\eta_S : X \to [0,1]$ and $\nu_S : X \to [0,1]$ denotes the degree of membership, degree of neutral membership (abstain) and degree of non-membership respectively and satisfy the condition

$$\mu_S^n(x) + \eta_S^n(x) + \nu_S^n(x) \leq 1; \; \forall x \in X.$$

The degree of refusal for any T-spherical fuzzy set S and $x \in X$ is given by

$$r_S(x) = \sqrt[n]{1 - (\mu_S^n(x) + \eta_S^n(x) + \nu_S^n(x))}.$$

The correlation coefficient establish a relation between two variables, i.e., how they move together in a correlated way. It is one of the most used index in statistics which reflect a linear relationship between two variables. Next, we outline the basic preliminaries related to the correlation coefficients in context with intuitionistic fuzzy sets.

Some of the important correlation coefficients between two intuitionistic fuzzy sets I_1 and I_2 over $X = \{x_1, x_2, x_3, \ldots, x_m\}$ (universe of discourse) proposed by various researchers are given below.

- **Gerstenkorn and Manko [7]**:

$$K(I_1, I_2) = \frac{\sum_{i=1}^{m} \left(\mu_{I_1}(x_i)\mu_{I_2}(x_i) + \nu_{I_1}(x_i)\nu_{I_2}(x_i)\right)}{\sqrt{\left[\sum_{i=1}^{m} \left(\mu_{I_1}(x_i) + \nu_{I_1}(x_i)\right)\right]\left[\sum_{i=1}^{m} \left(\mu_{I_2}(x_i) + \nu_{I_2}(x_i)\right)\right]}}.$$

- **Szmidt and Kacprzyk [16]**:

$$r_{IFS}(I_1, I_2) = \frac{1}{3}\left[r_1(I_1, I_2) + r_2(I_1, I_2) + r_3(I_1, I_2)\right];$$

where,

$$r_1(I_1, I_2) = \frac{\sum_{i=1}^{m} \left(\mu_{I_1}(x_i) - \overline{\mu_{I_1}}\right)\left(\mu_{I_2}(x_i) - \overline{\mu_{I_2}}\right)}{\left(\sum_{i=1}^{m} (\mu_{I_1}(x_i) - \overline{\mu_{I_1}})^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{m} (\mu_{I_2}(x_i) - \overline{\mu_{I_2}})^2\right)^{\frac{1}{2}}},$$

$$r_2(I_1, I_2) = \frac{\sum_{i=1}^{m} \left(\nu_{I_1}(x_i) - \overline{\nu_{I_1}}\right)\left(\nu_{I_2}(x_i) - \overline{\nu_{I_2}}\right)}{\left(\sum_{i=1}^{m} (\nu_{I_1}(x_i) - \overline{\nu_{I_1}})^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{m} (\nu_{I_2}(x_i) - \overline{\nu_{I_2}})^2\right)^{\frac{1}{2}}},$$

and

$$r_3(I_1, I_2) = \frac{\sum_{i=1}^{m} \left(\pi_{I_1}(x_i) - \overline{\pi_{I_1}}\right)\left(\pi_{I_2}(x_i) - \overline{\pi_{I_2}}\right)}{\left(\sum_{i=1}^{m} (\pi_{I_1}(x_i) - \overline{\pi_{I_1}})^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{m} (\pi_{I_2}(x_i) - \overline{\pi_{I_2}})^2\right)^{\frac{1}{2}}}.$$

It may be noted that the correlation coefficients between two intuitionistic fuzzy sets, I_1 and I_2, express the feature of relative strength. In addition to this, it also represents a positive (negative) relationship between the sets.

III. Correlation Coefficients of T-spherical Fuzzy Sets

Let $R = \{ < x_i, \mu_R(x_i), \eta_R(x_i), \nu_R(x_i) >| \; x_i \in X; \; i = 1, 2, 3 \ldots, m\}$ be a T-spherical fuzzy set over X (universe of discourse). In order to propose new correlation coefficients for T-spherical fuzzy sets, we first propose the information energy for a T-spherical fuzzy set as

$$I(R) = \sum_{i=1}^{m} \left((\mu_R^n)^2(x_i) + (\eta_R^n)^2(x_i) + (\nu_R^n)^2(x_i) + (r_R^n)^2(x_i) \right). \quad (1)$$

Consider two T-spherical fuzzy sets given by

$$R = \{< x_i, \mu_R(x_i), \eta_R(x_i), \nu_R(x_i) >| \; x_i \in X; \; i = 1, 2, 3 \ldots, m\}$$

and

$$S = \{< x_i, \mu_S(x_i), \eta_S(x_i), \nu_S(x_i) >| \; x_i \in X; \; i = 1, 2, 3 \ldots, m\}$$

over the given domain of discourse X.

We propose the following definition for the correlation function between two T-spherical fuzzy sets R and S as

$$C(R,S) = \sum_{i=1}^{m} \Big[\mu_R^n(x_i)\mu_S^n(x_i) + \eta_R^n(x_i)\eta_S^n(x_i) + \nu_R^n(x_i)\nu_S^n(x_i) \quad + r_R^n(x_i)r_S^n(x_i) \Big]. \quad (2)$$

It may be easily verified that the correlation function given by equation (2) satisfies the following necessary and sufficient conditions:

- $C(R,R) = I(R)$;
- $C(R,S) = C(S,R)$.

Next, we propose the following definitions of correlation coefficients for TSFSs:

Definition 2: Let R and S be two T-spherical fuzzy sets over the domain of discourse X given by

$$R = \{< x_i, \mu_R(x_i), \eta_R(x_i), \nu_R(x_i) >| \; x_i \in X\}$$

and

$$S = \{< x_i, \mu_S(x_i), \eta_S(x_i), \nu_S(x_i) >| \; x_i \in X\}.$$

The correlation coefficient between R and S is defined as

$$K_1(R,S) = \frac{C(R,S)}{[I(R) \cdot I(S)]^{1/2}} \quad (3)$$

Theorem 1: The correlation coefficient of two T-spherical fuzzy sets $K_1(R,S)$, given by equation (3), is a valid statistical measure.

Proof: It may be noted that a proposed correlation coefficient measure must fulfill the following well established axioms [16]:

- **Axiom 1:** $K_1(R,S) = K_1(S,R)$.
- **Axiom 2:** $0 \leq K_1(R,S) \leq 1$.
- **Axiom 3:** $K_1(R,S) = 1$ if and only if $R = S$.

Since $C(R,S) = C(S,R)$, therefore in view of the equation (3) in the definition 2, Axiom 1 is quite obvious, i.e.,

$$K_1(R,S) = K_1(S,R).$$

Next, we prove that Axiom 2 is satisfied. By definition, $K_1(R,S)$ is clearly a non-negative quantity. It is now sufficient to show that $K_1(R,S) \leq 1$.

We consider

$$\begin{aligned} C(R,S) &= \sum_{i=1}^{m} \Big[\mu_R^n(x_i)\mu_S^n(x_i) + \eta_R^n(x_i)\eta_S^n(x_i) + \nu_R^n(x_i)\nu_S^n(x_i) \\ &\quad + r_R^n(x_i)r_S^n(x_i) \Big] \\ &= \Big[\mu_R^n(x_1)\mu_S^n(x_1) + \eta_R^n(x_1)\eta_S^n(x_1) + \nu_R^n(x_1)\nu_S^n(x_1) \\ &\quad + r_R^n(x_1)r_S^n(x_1) \Big] + \\ &\quad + \Big[\mu_R^n(x_2)\mu_S^n(x_2) + \eta_R^n(x_2)\eta_S^n(x_2) + \nu_R^n(x_2)\nu_S^n(x_2) \\ &\quad + r_R^n(x_2)r_S^n(x_2) \Big] + \\ &\quad + \ldots + \\ &\quad + \Big[\mu_R^n(x_m)\mu_S^n(x_m) + \eta_R^n(x_m)\eta_S^n(x_m) + \nu_R^n(x_m)\nu_S^n(x_m) \\ &\quad + r_R^n(x_m)r_S^n(x_m) \Big]. \end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \Big[C(R,S)\Big]^2 &\leq \{ \Big((\mu_R^n)^2(x_1) + (\eta_R^n)^2(x_1) + (\nu_R^n)^2(x_1) + (r_R^n)^2(x_1) \Big) + \\ &\quad + \Big((\mu_R^n)^2(x_2) + (\eta_R^n)^2(x_2) + (\nu_R^n)^2(x_2) + (r_R^n)^2(x_2) \Big) + \\ &\quad + \ldots + \\ &\quad + \Big((\mu_R^n)^2(x_m) + (\eta_R^n)^2(x_m) + (\nu_R^n)^2(x_m) + (r_R^n)^2(x_m) \Big) \} \\ &\quad \times \{ \Big((\mu_S^n)^2(x_1) + (\eta_S^n)^2(x_1) + (\nu_S^n)^2(x_1) + (r_S^n)^2(x_1) \Big) + \\ &\quad + \Big((\mu_S^n)^2(x_2) + (\eta_S^n)^2(x_2) + (\nu_S^n)^2(x_2) + (r_S^n)^2(x_2) \Big) + \\ &\quad + \ldots + \\ &\quad + \Big((\mu_S^n)^2(x_m) + (\eta_S^n)^2(x_m) + (\nu_S^n)^2(x_m) + (r_S^n)^2(x_m) \Big) \} \\ &\quad = \sum_{i=1}^{m} \Big[(\mu_R^n)^2(x_i) + (\eta_R^n)^2(x_i) + (\nu_R^n)^2(x_i) + (r_R^n)^2(x_i) \Big] \\ &\quad \times \sum_{i=1}^{m} \Big[(\mu_S^n)^2(x_i) + (\eta_S^n)^2(x_i) + (\nu_S^n)^2(x_i) + (r_S^n)^2(x_i) \Big]. \end{aligned}$$

$$\Rightarrow [C(R,S)]^2 \leq I(R) \cdot I(S).$$

Therefore, in view of the definition 2, it is clear that $K_1(R,S) \leq 1$ which proves the Axiom 2.

Further, if $R = S$, i.e., $\mu_R = \mu_S$, $\eta_R = \eta_S$ and $\nu_R = \nu_S \; \forall \; x_i \in X$, then from equation (3), we get $K_1(R,S) = 1$. It is easy to note that the converse is also true which proves the Axiom 3. This completes the proof of the theorem.

Particular Cases:

1) For $n = 2$, the correlation coefficient between T-spherical fuzzy sets (3), becomes the correlation coefficient between the spherical fuzzy sets.
2) For $n = 1$, the correlation coefficient between T-spherical fuzzy sets (3), becomes the correlation coefficient between the picture fuzzy sets [17].
3) For $n = 2$ and $r_R = 0$ & $r_S = 0$ (*absentia of degree of refusals*), the correlation coefficient between T-spherical fuzzy sets (3), becomes the correlation coefficient between the Pythagorean fuzzy sets [5].
4) For $n = 1$ and $r_R = 0$ & $r_S = 0$ (*absentia of degree of refusals*), the correlation coefficient between T-spherical

fuzzy sets (3), becomes the correlation coefficient between the intuitionistic fuzzy sets [7].

IV. Computational Applications of the Proposed Correlation coefficients

In this section, computational application of solving the problem of pattern recognition has been demonstrated by applying the proposed correlation coefficients. We first present the "*Principle of Maximum Correlation Coefficient [5]*" in the light of spherical fuzzy sets as follows:

Principle of Maximum Correlation Coefficient: Let us consider 'm' different classes of patterns in which there are many members in each class. Here, we represent each member of each class by a T-spherical fuzzy set, say, A_α where, $\alpha = 1, 2, \ldots, m$ in X (universe discourse). If we have an unknown pattern Q (as another T-spherical fuzzy set) with us which is to be recognized in terms of its possible belongingness to one of the 'm' classes, then the degree of closeness index (correspondence) between A_α and Q is given by

$$\alpha^* = arg \max_{\alpha} \{K(A_\alpha, Q)\}; \tag{4}$$

where $K(A_\alpha, Q)$ will be computed based on the proposed correlation coefficients. More the value of α^*, more will be the belongingness of the pattern Q in the α^{th} class.

For simplicity of the calculation in the following illustrative example of pattern recognition, we take the value $n = 2$ and $m = 3$:

Let us take three representative patterns A_1, A_2 and A_3 from three different classes C_1, C_2 and C_3 under consideration respectively, which are being described by T-spherical fuzzy sets over the domain of discourse $X = \{x_1, x_2, x_3\}$:

$$A_1 = \{(x_1, 0.5, 0.4, 0.2), (x_2, 0.4, 0.3, 0.4), (x_3, 0.4, 0.5, 0.1)\};$$
$$A_2 = \{(x_1, 0.6, 0.5, 0.1), (x_2, 0.5, 0.1, 0.3), (x_3, 0.5, 0.5, 0.1)\};$$
$$A_3 = \{(x_1, 0.4, 0.4, 0.2), (x_2, 0.4, 0.5, 0.2), (x_3, 0.3, 0.3, 0.4)\}.$$

Consider an unknown sample pattern Q which is given by

$$Q = \{(x_1, 0.4, 0.4, 0.2), (x_2, 0.5, 0.6, 0.1), (x_3, 0.3, 0.4, 0.4)\}.$$

Now, the main objective of the problem is to find out the class to which Q belongs. As per the principle of maximum correlation coefficient stated above, the computed values of the correlation coefficients with respect to the proposed one, i.e, equations (3) are tabulated as follows:

TABLE I
Computed Values of Correlation Coefficients

Classes	$K_1(A_\alpha, Q)$
A_1	0.9239
A_2	0.8411
A_3	**0.9692**

From Table I, it may be easily observed that the unknown pattern Q belongs to the class C_3.

V. Comparative Analysis and Advantages of the Proposed Work

In this section, we carry out a comparative analysis to validate the performance of the proposed correlation coefficient of TSFSs with some of the existing approaches under IFS and picture fuzzy set environment. The detailed analysis and advantages of using the proposed approach along with illustrative examples are presented below:

A. Correlation Coefficients in Pattern Recognition

In order to validate the performance of the proposed correlation coefficients in pattern recognition, we consider an example where there are three representative centrally located patterns A_1, A_2 and A_3 from three different classes C_1, C_2 and C_3 under consideration respectively. It may be noted that the patterns described by the spherical fuzzy sets have significantly wide coverage than intuitionistic fuzzy set or picture fuzzy set, i.e., a decision maker is not strictly bounded in forwarding its opinion. We present the following computations for the patterns under consideration with the domain of discourse as $X = \{x_1, x_2, x_3\}$:

- **Correlation Coefficient by [7]**: Suppose the representative patterns are given in the form of intuitionistic fuzzy set as follows:

  $$\text{“}A_1 = \{(x_1, 0.4, 0.5), (x_2, 0.7, 0.1), (x_3, 0.3, 0.3)\}\text{”};$$
  $$\text{“}A_2 = \{(x_1, 0.5, 0.4), (x_2, 0.7, 0.2), (x_3, 0.4, 0.3)\}\text{”};$$
  $$\text{“}A_3 = \{(x_1, 0.4, 0.5), (x_2, 0.7, 0.1), (x_3, 0.4, 0.3)\}\text{”}.$$

  Consider an unknown sample pattern Q in the form of intuitionistic fuzzy set which is given by

  $$\text{“}Q = \{(x_1, 0.1, 0.1), (x_2, 1.0, 0.0), (x_3, 0.0, 1.0)\}''.$$

  Now, the main objective of the problem is to find out the class to which Q belongs. Based on the correlation coefficient proposed by [7], we obtain the following values:

  $$K_{IFS}(A_1, Q) = K_{IFS}(A_2, Q) = K_{IFS}(A_3, Q) = 0.6292.$$

- **Correlation Coefficient by [17]**: Suppose the representative patterns are given in the form of picture fuzzy set as follows:

  $$\text{“}A_1 = \{(x_1, 0.4, 0.5, 0.1), (x_2, 0.7, 0.1, 0.1), (x_3, 0.3, 0.3, 0.2)\}\text{”};$$
  $$\text{“}A_2 = \{(x_1, 0.5, 0.4, 0.1), (x_2, 0.7, 0.2, 0.1), (x_3, 0.4, 0.3, 0.1)\}\text{”};$$
  $$\text{“}A_3 = \{(x_1, 0.4, 0.5, 0.1), (x_2, 0.7, 0.1, 0.1), (x_3, 0.4, 0.3, 0.2)\}\text{”}.$$

  Consider an unknown sample pattern Q in the form of picture fuzzy set which is given by

  $$\text{“}Q = \{(x_1, 0.1, 0.1, 0.4), (x_2, 1.0, 0.0, 0.0), (x_3, 0.0, 1.0, 0.0)\}\text{”}.$$

  Based on the correlation coefficient proposed by [17], we obtain

  $$K_{P_1}(A_1, Q) = 0.7937, K_{P_1}(A_2, Q) = 0.7746, K_{P_1}(A_3, Q) = 0.7721.$$

- **Proposed Correlation Coefficients**: Suppose the representative patterns are given in further translated form of spherical fuzzy set ($n = 2$) as follows:

  $$\text{“}A_1 = \{(x_1, 0.4, 0.5, 0.2), (x_2, 0.7, 0.1, 0.3), (x_3, 0.3, 0.3, 0.5)\}\text{”};$$
  $$\text{“}A_2 = \{(x_1, 0.5, 0.4, 0.3), (x_2, 0.7, 0.2, 0.2), (x_3, 0.4, 0.3, 0.3)\}\text{”};$$
  $$\text{“}A_3 = \{(x_1, 0.4, 0.5, 0.4), (x_2, 0.7, 0.1, 0.4), (x_3, 0.4, 0.3, 0.4)\}\text{”}.$$

Similarly, if we consider an unknown sample pattern Q in the form of spherical fuzzy set which is given by

$$\text{“}Q = \{(x_1, 0.1, 0.1, 0.9), (x_2, 1.0, 0.0, 0.0), (x_3, 0.0, 1.0, 0.0)\}\text{”};$$

then using the proposed correlation coefficient, we obtain the following values:

$$K_1(A_1, Q) = 0.393535, K_1(A_2, Q) = 0.4047, K_1(A_3, Q) = 0.460449.$$

Comparative Remarks: Based on the above calculations and analysis, the following are the important comparative remarks:

- All the values the correlation coefficients $K_{IFS}(A_i, Q);\ \forall i$ obtained by [7] are found to be identical which shows a kind of limitation/drawback as an unknown pattern could not belong to all the classes simultaneously.
- However, on the other hand, the values the correlation coefficients $K_{P_1}(A_i, Q);\ \forall i$ obtained by [17] are different but the difference is not that much prominent as desired.
- Based on the results obtained using the proposed correlation coefficients, we clearly assert that the values for classification are significantly differentiable in both the cases. Certainly, the proposed correlation coefficients of spherical fuzzy sets have an added advantage of allowing the decision maker to give their opinion freely without any restriction.

B. Advantages of the Proposed Work

In view of the above detailed analysis, the proposed correlation coefficient of T-spherical fuzzy sets and spherical fuzzy sets are found to be worthy enough in contrast with the existing related literatures. The following are the major advantages of the proposed work:

- As mentioned earlier in the introduction, the incorporation of intuitionistic fuzzy sets and picture fuzzy sets has some limitations and not able to capture the full information specification of the situation. Therefore, the additional exponents of the degrees of membership, neutral membership, non-membership and degree of refusal in case of the spherical fuzzy sets certainly provide a wider coverage and wider geometrical span.
- Therefore, the proposed correlation coefficients have capabilities to address the related dependability on the imprecise information which has a degree of refusal in a more reliable manner.
- The discussion on the result obtained in case of pattern recognition shows that the proposed work handled the generalized framework in an effective and consistent way.

VI. Conclusions and Scope for Future Work

The correlation coefficient for T-spherical fuzzy sets has been well introduced along with its validity proof. The proposed correlation coefficients are able to incorporate the four parameters, i.e., the degree of membership, neutral membership, non-membership, and the refusal degree. Hence, the correlation coefficient cover an add-on reliability and is able to capture more flexibility of the imprecise information. In view of the revised principle of maximum correlation coefficients, the proposed correlation coefficient has been implemented in solving the problem of pattern recognition under T-spherical fuzzy environment. Some important comparative remarks and advantages of the proposed methodology have been listed. It has also been concluded that the results obtained are found to be consistent and methodologies outlined above can be extended for larger dimensional problems in future. It is well known that the patterns within a cluster are more correlated to each other than related to the different clusters. In literature, various researchers utilized the concept of correlation coefficients in the clustering of the patterns [18]. In future, the proposed correlation coefficient can also be comprehensively used in the cluster analysis when the information data is to be taken in the form of SFSs. In addition to this, the application may further be projected in the field of bidirectional approximate reasoning [2] [17].

References

[1] Atanassov K. T., Intuitionistic fuzzy sets, Fuzzy Sets and Systems, vol. 20, pp. 87–96, 1986.

[2] Chen S. M., Hsiao W. H., Jong W. T., Bidirectional approximate reasoning based on interval-valued fuzzy sets, Fuzzy Sets and Systems, vol. 91, pp. 339–353, 1997.

[3] Chen T. Y., An interval-valued Pythagorean fuzzy compromise approach with correlation-based closeness indices for multiple-criteria decision analysis of bridge construction methods, Complexity, 2018.

[4] Cuong B. C., Picture fuzzy sets first results. Part 1, in preprint of seminar on neuro-fuzzy systems with applications, Institute of Mathematics, Hanoi, May, 2013.

[5] Garg H., A Novel Correlation Coefficients between Pythagorean Fuzzy Sets and Its Applications to Decision-Making Processes, International Journal of Intelligent Systems, vol. 31(12), pp. 1234–1252, 2016.

[6] Garg, H., Munir, M., Ullah, K., Mahmood, T., Jan, N., (2018) Algorithm for T-Spherical Fuzzy Multi-Attribute Decision Making Based on Improved Interactive Aggregation Operators, Symmetry, 10, pp. 670.

[7] Gerstenkorn T., Manko J., Correlation of intuitionistic fuzzy sets, Fuzzy Set and Systems, vol. 44, pp. 39–43, 1991.

[8] Guleria, A. and Bajaj, R.K., T-spherical Fuzzy Soft Sets and its Aggregation Operators with Application in Decision Making, Scientia Iranica, 2019. [doi: $10.24200/sci.2019.53027.3018$]

[9] Guleria, A. and Bajaj, R.K., Eigen spherical Fuzzy Sets and its Application in Decision Making, Scientia Iranica, 2019.

[10] Hong D. H., Hwang S. Y., Correlation of intuitionistic fuzzy sets in probability spaces, Fuzzy Sets and Systems, vol. 75 pp. 77–81, 1995.

[11] Huang H. L., Guo Y., An Improved Correlation Coefficient of Intuitionistic Fuzzy Sets, Journal of Intelligent Systems, 2017.

[12] Kifayat U., Khan Q., Jan N., Similarity Measures for T–Spherical Fuzzy Sets with Applications in Pattern Recognition, Symmetry, vol. 10, 193, 2018.

[13] Mahmood T., Kifayat U., Khan Q., Jan N., An approach toward decision making and medical diagnosis problems using the concept of spherical fuzzy sets, Neural Computing and Applications, 2018.

[14] Mitchell H. B., A correlation coefficient for intuitionistic fuzzy sets, International Journal of Intelligent Systems, vol. 19, pp. 483–490, 2004.

[15] Park J. H., Lim K. M., Park J. S., Kwun Y. C., Correlation coefficient between intuitionistic fuzzy sets, Fuzzy Information and Engineering, vol. 2, pp. 601–610, 2009.

[16] Szmidt E. and Kacprzyk J., Correlation of intuitionistic fuzzy sets, Lectuer Notes in Computer Science, 6178, 169–177, 2010.

[17] Singh P., Correlation coefficients for picture fuzzy sets, Journal of Intelligent and Fuzzy Systems, vol. 28, pp. 591–604, 2015.

[18] Xu Z. S., Chen J., Wu JJ., Clustering algorithm for intuitionistic fuzzy sets, Information Sciences, vol. 178, pp. 3775–3790, 2008.

[19] Yager R. R., Pythagorean fuzzy subsets, In, Proceedings of Joint IFSA World Congress and NAFIPS Annual Meeting, Edmonton, Canada, pp. 57–61, 2013.

[20] Zadeh L. A., Fuzzy sets, Information and Control, vol. 8, pp. 338–353, 1965.

An Innovative Query Tuning Scheme for Large Databases

Chaman Wijesiriwardana
Dept. of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
chaman@uom.lk

M.F.M. Firdhous
Dept. of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
firdhous@uom.lk

***Abstract*— In today's competitive world businesses place more emphasis on information considering them as the more valuable asset of the organization. Hence data is given more attention than that is given to hardware or software and even sometimes to the human resources. The utility of data relies on the fact that how soon it can be turned into useful information and to strategic knowledge. Data is generally stored in a database that powers the company's information system. Relational databases store data in a series of related tables as rows and columns. When a query is issued seeking data from the database, it extracts data from multiple tables depending on the criteria set in the query and presents it as a single set of rows and columns. Hence, the response time of SQL queries becomes a visible factor for the performance of the entire information system. Therefore, query optimization becomes a vital function for preventing performance degradation. In this paper, the authors present two SQL tuning techniques that improve the performance of data retrieval from relational databases over non tuned SQL queries.**

Keywords*— RDBMS, query tuning, query optimization, SQL, large databases.

I. INTRODUCTION

In today's information era, data has been considered as the most important and strategic asset for any organization. Hence more and more organizations are moving towards implementing their own information systems for aiding with efficient decision making [1], [2]. A database management system (DBMS) is the main component of any information system irrespective of their size and nature of operation. The DBMS provides function for organizing, controlling and cataloging of data for efficient and secure storage and retrieval. Since the DBMS plays a vital role in the background of any information system, the performance of the information system directly depends on that of the DBMS. Thus the optimization of database systems plays an important role in improving the efficiency and responsiveness of information systems to user requests.

The required data is retrieved from a database by issuing instructions to the system commonly known as queries [3]. These queries are generally written in a high level language called Structured Query Language (SQL). When a query is issued it undergoes a series of operation known as query processing for retrieving the required data from the database. These operations include the translation of queries in high-level database languages into low level machine instructions, a range of query-optimizing and the actual evaluation of queries [4]. Figure 1 shows the main steps involved in query processing. These steps can be grouped into three main parts as parsing and translation, optimization and evaluation.

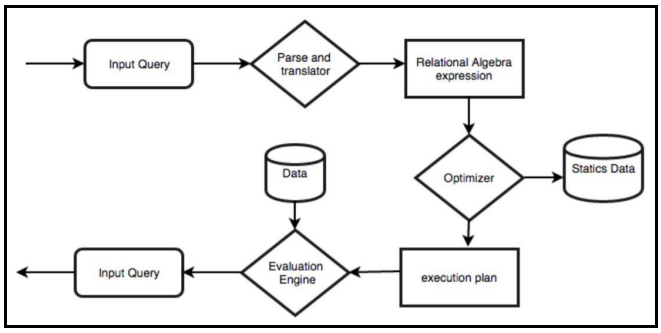


Fig. 1. Steps in Query Processing [4]

In the query processing shown in Figure 1, the query optimizer module plays an important role. The query optimizer determines the best strategy for performing a query. It is the query optimizer's job to decide, for example, whether to use indexes for a given query or not, and which join techniques are more appropriate when joining multiple tables. These decisions have a tremendous effect on SQL performance. Hence, query optimization is a key technology for every application, from operational systems to data warehouse and analysis systems to content-management systems. Query optimizers generally employ a cost-based strategy to identify and select the best technique to use in a given situation. In a cost-based optimization strategy, multiple execution plans are generated for a given query, and then an estimated cost is computed for each plan and then the query optimizer chooses the plan with the lowest estimated cost [5].

Query optimizer is not directly accessible to users. Once a query is submitted to the server and parsed, they are then passed to the query optimizer for optimization [6]. However, some database engines allow guiding the query optimizer with hints. Optimization time increases rapidly for queries joining many tables. Though, randomized or heuristic search algorithms can reduce query optimization time for large join queries by considering fewer plans, they sacrifice plan optimality. Though the commercial DBMS generally execute query plans in parallel, the optimization of such plans still is carried out in series [7], [8]. Hence, writing SQL queries efficiently (SQL query tuning) plays a major role in optimizing the performance of DBMSs. Research has shown that proper SQL query tuning can improve the performance of DBMSs by more than 50% [9]. In this paper the authors present two query tuning techniques that can greatly improve the response time of query processing in relational database management systems.

978-1-7281-2087-4/19/$31.00 ©2019 IEEE

This paper is organized into four main sections as follows: Section I gives an overall introduction to the paper with a brief analysis of the problem area explored. Sections II provides a general description of query processing along with a discussion on similar work carried out by other researchers. Section III presents the techniques proposed in this research along the results of tests carried out for evaluating the techniques. Finally Section IV concludes the paper with a discussion on the results.

II. Query Tuning in Relational Database Management Systems

Relational Database Management Systems (RDBMS) have been adopted by many organizations making it the de facto standard for database management due to its ease of use and querying capabilities [10]. Information systems installed by companies for improving the efficiency and effectiveness of their operations collect data at exponential rates. These data need to be analyzed, processed and results to be produced to suit the requirements of the users. Thus, the performance of DBMSs become critical in meeting user requirements as delay in response time could affect company performance considerably. Hence, database performance tuning or database performance optimization needs to be carried out to make the database systems run faster. SQL optimization or tuning is the process of improving SQL queries at the application level that typically offers the biggest potential for database performance optimization [11].

A. Query Tuning Process

SQL statements written in a high level language ere used to retrieve data from relational databases. It is possible to get the same results using different SQL queries [12]. But, for better performance it is necessary to use the best, faster and efficient queries that provides the best response time in retrieving, preparing and presenting the data based on user requirements. Hence query tuning becomes an important step in the implementation of any database application.

The main objectives of query tuning are as follows [13]:

- Reduce response time for SQL processing
- To find more efficient ways to process workloads
- Improve search time using indexes
- Join data efficiently between multiple tables

Query tuning is essentially an iterative process as when an issue is fixed, the performance bottleneck may to shift to another part of the system that need to be fixed next [14]. Hence, several cycles of tuning will need to be carried out until the expected performance levels are achieved. Also, manual tuning of SQL statements requires a thorough knowledge of how these statements are executed in the background as well as the experience to understand suitable access paths to produce better response times. Further, there may be many SQL statements to tune in an application making manual SQL tuning more cumbersome. In order to help database programmers, there are many automatic SQL tuning tools available in the market that help them to identify problematic SQL statements and provide suggestions to fix them [15]. Though, automatic tools may help reduce the workload of database designers, still SQL tuning requires a lot of human input [16].

SQL query tuning is an essential skill for database developers and programmers. For arriving at the best SQL queries for a given requirement, developers must possess the thorough knowledge on the query optimizer of the chosen DBMS and the techniques such as cost-based optimizing, heuristic-based optimizing or hybrid (of cost-based and heuristic) optimizing, it employs to identify and select an access path and preparing the query execution plan. The best way to arrive at the set of queries that provide the optimum performance is to write the queries in a number of different ways and compare their reads and execution plans commonly known as the trial-and-error method [17]. Researchers have proposed many techniques and tips that can be used to prepare and test the optimization of query performance [5], [12], [13], [17], [18], [19]. There are many different ways to determine the best way to write queries. Two of the most commonly used methods are looking at the number of logical reads produced by the query and looking at graphical execution plans provided by the SQL Server Management Studio [12]. Next subsection provides an in-depth analysis of related work carried out in SQL tuning by other researchers.

B. Related Work

Karthic et al., in [13] discuss in detail about the importance of SQL query tuning for the performance of RDBMSs. Before proposing any SQL query tuning techniques, they take a detail look at the best way to tuning multiple SQL statements simultaneously. For enabling parallel tuning of SQL statements possible, they propose an approach called broad-brush approach that can save thousands of hours of SQL tuning at a later stage. But, they have stopped short of proposing any SQL tuning techniques but claims that using the DISTINCT keyword in SQL queries has no meaning, when the search is based on a primary key attribute.

Bhajipale et al., have in [12] carried out an in depth study on the need for query tuning in large databases for improving performance. They have put forward several query tuning techniques (tips) as part of their proposal. In their own words, they have accepted query optimization is a deep subject and they cover only the most important points. Hence, the tips forward by them in [12] is not an exhaustive list of tips covering all the techniques or all types of complex queries. Hence, there is a lot more left for others to work on.

Several SQL query tuning/optimization techniques have been presented in [20] by Srinivas et al. Though, several important techniques have been presented, the list is not sufficiently exhaustive to cover all the available techniques. Hence, there is enough opportunity for researchers to forward their proposals for SQL tuning in the future. The other shortcoming of this paper is non evaluation of any of the technique proposed. Hence, there is no assurance that the proposed techniques will work when applied to a real situation.

According to Batra et al., Entity Attribute Value (EAV) storage model has been extensively used to manage healthcare data [21]. They further state that EAV suffers from lack of search efficiency. In order to overcome the lack of search efficiency in EAV, a new Two Dimensional EAV (2D EAV) model has been proposed. The main focus of the

2D EAV is on how to handle template-centric and other health data related query scenarios. The 2D EAV handles sparseness, frequent schema evolution, and efficient query support altogether for standardized Electronic Health Records (EHRs). The main shortcoming of this scheme is the tight focus on handling health related databases. Thus, it lacks the generalizability to other types of databases and queries.

Sharma and Sharma have carried out a survey on query optimization techniques proposed in literature [22]. This work is very shallow as apart from discussing the need and the basics of query processing and optimization, it evaluates only a few query optimization techniques proposed in the literature.

Patel and Patel in [23] have proposed a query optimization technique using Schema Object Base View for large databases. The proposed scheme uses a different parameter for the optimization of queries compared to other techniques like Query Graph, Tableaus or Optimization of Queries with Aggregates. They show that their scheme results in reduced query execution cost and query space yielding better query processing times. But, this article does not present any optimization techniques for improving SQL syntax for achieving better performance.

Patil et al., have in [9] presented several query response time enhancement methods including SQL tuning, indexing and table partitioning with a discussion on their advantages and disadvantages.

Gathering statistics, index management, table reorganization, prediction, data mart, materialized view, partition table, query rewriting, monitoring performance and optimization have been presented as 10 different methods to be used for SQL query tuning. The information presented is only a high level overview of these techniques without any concrete proposals or tips on how to write efficient SQL statements.

Colley and Stanier have studies the new techniques that have been proposed in recent literature for improving the performance of databases [24]. They have looked at three key areas that contribute to query performance. They are namely; database design considerations, query execution optimization and SQL query design. They have only broadly looked at SQL query design by looking at the effect of JOIN clause on the performance. Even this discussion is not comprehensive as it has not covered the different ways how the JOIN clause can be used in an SQL statement. Not any concrete proposals on how to write efficient queries has been put forward.

Query optimization using SQL transformations proposed in [5] looks at one of the four steps in Oracle's query optimizer. The other three steps are Execution plan selection, Cost model and statistics and Dynamic runtime optimization. The SQL transformation is carried out by the database SQL optimizers to convert a given SQL statement into another SQL statement that can be executed more efficiently, but returning the same results. In this paper, the author looks at different SQL transformation techniques Oracle has implemented for converting SQL statements. This paper only presents the different SQL transformation techniques implemented in Oracle and how they are executed. It does not provide any information or tips for writing efficient SQL queries to suit a given requirement or situation.

The optimization problem of dynamically generated SQL queries for tiny-huge, huge-tiny problem has been investigated in [25]. The author of this paper shows that tiny-huge or huge-tiny problem creates an execution dilemma for Cost-based Optimizers (CBO) in the DBMSs resulting in increased response times. In this paper, the author proposes a new approach by combining denormalized columns and re-writing of the security predicates' sub-queries at run-time by leveling the outer and security sub-queries. The test results show that the proposed scheme results in more stable execution plans improving the performance of SQL executions. Though, this approach results in better performance in run time generated SQL queries, this technique is more complex to implement and requires very extensive designs at the beginning making it unsuitable for general SQL query implementations.

Corlatan et al., have in [26] studied query optimization techniques with special emphasis on Microsoft SQL server. They have identified many factors affecting the performance of DBMSs including missing indexes, inexact statistics, badly written queries, deadlocks, using cursors in Transact-SQL operations, fragmentation of indexes and frequent recompilation of queries. They have suggested several tips for optimizing SELECT queries. But, they have not tested the performance of the tips proposed using any world experiments. Hence, these tips can be considered as general suggestions only.

Taniar et al., [27] and Lokhande and Shete [28] have looked at the use of hints in SQL-nested query optimizations. As the query optimizer of a DBMS cannot be directly accessed by users, comments inserted into an SQL statement in the form of hints can instruct the optimizer to perform certain specific operations as per the user requires bypassing the automatic optimization process. They show that providing hints can improve the performance of query execution. They also claim that this technique is easier to understand and use as hints are written in human readable high level language and it follows a simple process. However, the main shortcoming of this techniques is that syntax hints are DBMS implementation specific and cannot be generalized as hints are not integral part of SQL standards.

Habimana has in [17] proposed several tips for writing efficient and faster SQL queries. The proposed techniques have been tested by comparing the performance of tuned queries against that of non-tuned queries. The results show that they are is definite advantages in tuning queries. The main shortcoming of this work is that all the tips are for queries that extract data from a single table. Also, the tips do not cover nested queries. The other shortcoming of the proposal is the non evaluation of the proposed tips. Hence, the effectiveness of the tips cannot be determined. Also, the author has failed to indicate any performance attributes, that can be used by other re4searchers to test the effectiveness of the proposals.

III. PROPOSED QUERY TUNING TECHNIQUES

Best way to write queries can be arrived at using many different ways. Checking the number of logical reads or disk seeks produced by the query and examining the query execution plan using the interface provided by the DBMS vendor or a third party are two commonly used methods for scrutinizing the performance of the query in fetching the

information from the storage device and presenting it to the user or application that called it [12]. For example, in Microsoft SQL Server, the number of logical reads generated by a query can be found by turning the STATISTICS IO option ON using the command *SET STATISTICS IO ON*. Turning STATISTICS IO option on will return many data items related to IO operations in the Messages window of the SQL Server Management Studio. Out of them, logical reads portion is the one that indicates about the number of pages read from the data cache during this query execution. The number of logical reads is the most helpful parameter in estimating the performance of the query as it remains unchanged every time the same query is run irrespective of the changes to the environment. Other parameters like response time may be affected by external factors such as locking by other queries. The lowest value resulting for number of logical reads is the indication of best tuned queries as fewer logical reads typically lead to faster execution times.

The following are the Techniques proposed in this research:

Technique 1:

Avoid *IN* Operator in *WHERE* clause

Solution:

Create the temp table and insert needed data and JOIN to the main query

Steps to be Followed:

If complex query includes an *IN* Operator then it must be removed and the following steps are to be followed to create a temporary table and include it in the query.

- Remove the *IN* operator and create a *#temp* table and insert the necessary data into that table.
- Create an index to the *#temp* table.
- Join the *#temp* with the base table.

Technique 2:

Avoid *temp* Tables as much as possible.

Solution:

If it is really necessary to have a *temp* Table, create it as a *#temp* Table explicitly using the command "Create Table *#temp*".

Steps to be Followed:

Run the DDL command "Create Table #temp".

A. Evaluation

The proposed query tuning techniques were evaluated by running optimized as well as non-optimized queries on a large database. The database contained more than 250,000 records stored in multiple tables. The database was initially optimized by creating an index. The optimization of the database was checked using the following complex query.

```
SELECT DISTINCT 'Doctor fees' AS TrnTypeCode, DFRH.ReceiptNo, DFRH.BHTNo,
    DFRH.ReferenceNo, DFRH.ReceiptAmount, 0.00 AS PaidAmount, DFRH.MachineCode,
    DFRH.MachineBillNo, 'D' AS AdvanceReceiptType, DFRH.CreateUser, DFRH.CreateDate,
    DFRH.ModifiedUser, DFRH.IsVoid,DFRH.SessionID, DFRP.PaymentType,
    PT.[Description] AS [PaymentTypeName], DFRP.PaymentNo, DFRP.CardType,
    DFRP.BankCode, DFRP.ChequeDate, DFRP.CommonReferanceDetails,
    DFRP.SettledAmount, CEL.dLogDate, CEL.dLogOutDate, (TL.Description + ' ' +
    PA.FirstName + ' ' + PA.LastName)  AS PatientName ," as DoctorCode ,
    DFRP.SettledAmount as DocAmount, '' AS ProfessionalName
FROM [HMS].[BILL_TRN_DoctorFeeReceiptHeader] AS DFRH JOIN
    [HMS].[BILL_TRN_DoctorFeeReceiptPayment] AS DFRP ON DFRH.ReceiptNo =
    DFRP.ReceiptNo JOIN [HMS].[Sys_Audit_TRN_CashierEventLog] AS CEL
ON  DFRH.SessionID = CEL.nLogRecId JOIN [HMS].[BILL_Comm_MST_PaymentType] AS
    PT
ON LTRIM(RTRIM(DFRP.PaymentType)) = LTRIM(RTRIM(PT.PaymentCode))
JOIN [HMS].[BILL_Comm_MST_PatientAdmissionHeader] AS PA
ON PA.BHTNo = DFRH.BHTNo
JOIN [HMS].[BILL_TRN_DoctorFeeHeader] AS DFH
ON DFRH.BHTNo = DFH.BHTNo AND DFH.DocReceiptNo = DFRH.ReceiptNo
LEFT OUTER JOIN (SELECT * FROM [HMS].[BILL_Comm_MST_ReferenceData]
    WHERE Modulecode ='BILL_COMM_MST_TITLE') AS TL
ON LTRIM(RTRIM(PA.Title)) = LTRIM(RTRIM(TL.ReferenceCode))
```

Fig. 2. Complex Query Used for Testing Database Performance

Figure 3 shows the performance of the query before and after creating the index. It can be seen that before creating the index, it has taken 3 seconds to process the query, whereas after the index creation it has become a very negligible amount.

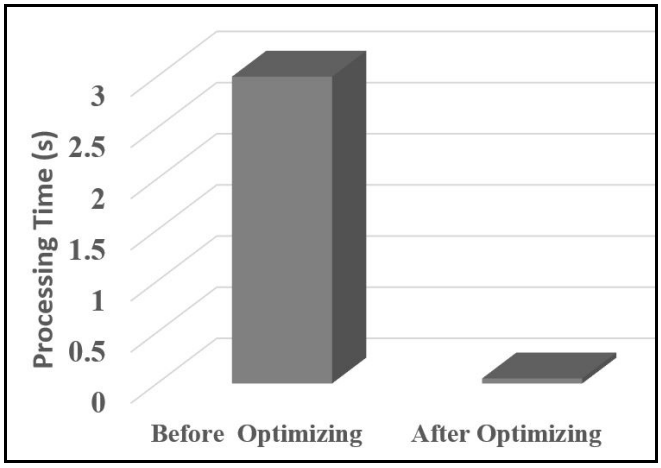


Fig. 3. Performance of Database Before and After Optimization

B. Evaluating Technique 1

Technique 1 was evaluated by creating multiple SQL queries with the IN operator and using Temp Tables in place of the IN operator as proposed in this research. Figures 4 and 5 show two queries with IN operator and their counterparts that use temp Tables.

```
SELECT a.BHTNo, a.FirstName, a.LastName,
FROM hms.INV_TRN_BillEntryDetails AS s
    INNER JOIN hms.INV_TRN_BillEntryHeader AS d ON s.EntryNo = d.EntryNo
    INNER JOIN hms.BILL_Comm_MST_PatientAdmissionHeader AS a ON a.BHTNo =
        d.BHTNo
WHERE ItemCode IN (SELECT itemcode FROM hms.BILL_Comm_MST_Item) AND
    costcentercode
        IN (SELECT costcentercode FROM hms.BILL_Comm_MST_CostCenterHeader) AND
        a.roomno IN (SELECT roomno FROM hms.BILL_Comm_MST_Room)
```

Original Query

```
CREATE TABLE #TempTable(ID varchar(50))
INSERT INTO #TempTable (ID)

SELECT DISTINCT itemcode FROM hms.BILL_Comm_MST_Item
SELECT a.BHTNo, a.FirstName, a.LastName FROM hms.INV_TRN_BillEntryDetails AS s
INNER JOIN #TempTable AS t ON t.ID = s.ItemCode
INNER JOIN hms.INV_TRN_BillEntryHeader AS d ON s.EntryNo = d.EntryNo
INNER JOIN hms.BILL_Comm_MST_PatientAdmissionHeader AS a ON a.BHTNo =
d.BHTNo
WHERE ItemCode IN (select itemcode from hms.BILL_Comm_MST_Item) AND
        costcentercode IN (SELECT costcentercode FROM
        hms.BILL_Comm_MST_CostCenterHeader) AND a.roomno IN (SELECT roomno
        FROM hms.BILL_Comm_MST_Room)

DROP TABLE #TempTable
```

Tuned Query

Fig. 4. 1st Set of Queries for Testing Technique 1

```
SELECT  S.*
FROM hms.INV_TRN_BillEntryDetails AS s
INNER JOIN hms.INV_TRN_BillEntryHeader AS d  ON s.EntryNo = d.EntryNo
INNER JOIN hms.BILL_Comm_MST_PatientAdmissionHeader AS a  ON a.BHTNo =
        d.BHTNo
WHERE ItemCode IN (SELECT DISTINCT itemcode
FROM hms.BILL_Comm_MST_Item) AND costcentercode
IN (SELECT  costcentercode FROM hms.BILL_Comm_MST_CostCenterHeader) AND
        a.roomno
IN (SELECT  roomno FROM hms.BILL_Comm_MST_Room)
```

Original Query

```
SET STATISTICS TIME ON
CREATE TABLE #TempTable (ID varchar(50))

INSERT INTO #TempTable (ID) SELECT DISTINCT itemcode FROM
    hms.BILL_Comm_MST_Item

CREATE NONCLUSTERED INDEX IX_Itemcode on #TempTable(ID)

SELECT S.* FROM hms.INV_TRN_BillEntryDetails AS s
INNER JOIN hms.INV_TRN_BillEntryHeader AS d ON s.EntryNo = d.EntryNo
INNER JOIN hms.BILL_Comm_MST_PatientAdmissionHeader AS a ON a.BHTNo =
    d.BHTNo
INNER JOIN #TempTable AS t ON t.ID = s.ItemCode WHERE costcentercode
IN (SELECT costcentercode FROM hms.BILL_Comm_MST_CostCenterHeader) AND
    a.roomno
IN (SELECT roomno FROM hms.BILL_Comm_MST_Room)

DROP TABLE #TempTable
SET STATISTICS TIME OFF
```

Tuned Query

Fig. 5. 2nd Set of Queries for Testing Technique 1

Figure 6 shows the query response times for executing both original and tuned queries for extracting more than 250,000 records from the database. From, 6, it can be seen that in both cases the query response time has improved by 23.77% and 90.43% respectively. Both are significant improvements over the original response times. The difference in the improvement is a clear indication that the performance enhancement achieved by the proposed technique depends on the complexity of the query too.

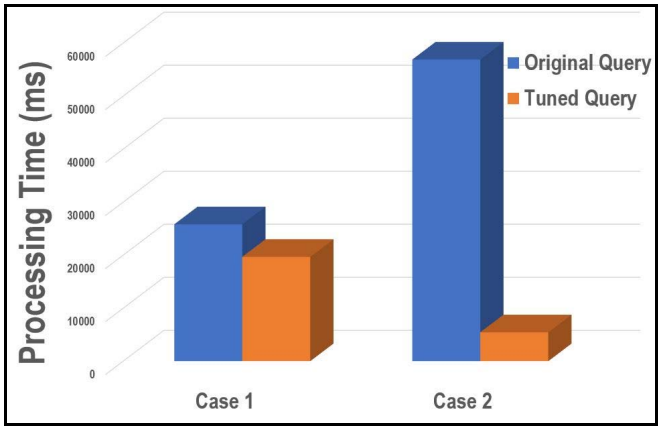


Fig. 6. Query Performance Times - Technique 1

Queries shown in Figure 5 (Case 2) were further analyzed using SQL Sentry Plan Explorer. At the end of the analysis, the total processing times reported by the Sentry Plan Explorer also shows that the Tuned query has better overall response time than the original query.

C. *Evaluating Technique 2*

Technique 2 was also evaluated by running different SQL queries including implicit temp tables and explicitly created #temp tables. Figures 7 shows the queries with implicit temp table and explicitly created #temp table that were executed on the same optimized database used to evaluate Technique 1.

```
SET STATISTICS time ON

WITH BASE AS (SELECT ProductID,  YEAR(TransactionDate) AS TransCurrYear,
        COUNT(1) AS NoTrans
FROM Production.TransactionHistory GROUP BY ProductID,        YEAR(TransactionDate))

SELECT CurrYear.ProductID, CurrYear.NoTrans AS CurrTransCnt, PrevYear.NoTrans AS
    PrevTransCnt, Prev2Year.NoTrans AS Prev2YearCnt
FROM BASE AS CurrYear CROSS APPLY (SELECT * FROM BASE PrevYear WHERE
        CurrYear.ProductID = PrevYear.ProductID AND CurrYear.TransCurrYear =
        PrevYear.TransCurrYear - 1) AS PrevYear
OUTER APPLY (SELECT * FROM BASE Prev2Year WHERE CurrYear.ProductID =
        Prev2Year.ProductID AND CurrYear.TransCurrYear = Prev2Year.TransCurrYear - 2) AS
        Prev2Year

SET STATISTICS time OFF
```

Original Query

```
SET STATISTICS time ON

CREATE TABLE #T1 (ProductID int, TransCurrYear int, NoTrans int);

CREATE CLUSTERED INDEX CI_#T1 ON #T1 (TransCurrYear)

INSERT INTO #T1 SELECT ProductID, YEAR(TransactionDate) AS TransCurrYear,
        COUNT(1) AS NoTrans  FROM Production.TransactionHistory  GROUP BY ProductID,
        YEAR(TransactionDate)  ORDER BY YEAR(TransactionDate);With BASE AS (SELECT
        * FROM #T1)

SELECT CurrYear.ProductID, CurrYear.NoTrans AS CurrTransCnt, PrevYear.NoTrans AS
        PrevTransCnt, Prev2Year.NoTrans AS Prev2YearCnt FROM BASE AS CurrYear
        CROSS APPLY (SELECT * FROM BASE PrevYear  WHERE CurrYear.ProductID =
        PrevYear.ProductID AND CurrYear.TransCurrYear = PrevYear.TransCurrYear - 1 ) AS
        PrevYear OUTER APPLY (SELECT * FROM BASE Prev2Year WHERE
        CurrYear.ProductID = Prev2Year.ProductID  AND CurrYear.TransCurrYear =
        Prev2Year.TransCurrYear - 2 ) AS Prev2Year

 DROP TABLE #T1
 SET STATISTICS time OFF
```

Tuned Query

Fig. 7. Set of Queries for Testing Technique 2

Figure 8 shows the total number of disk scans and logical reads for both original query and the tuned query respectively. The number of disk scans has come down from 27 to 1, which is a 96.3% improvement. On the other hand,

the number of logical reads has also seen a drastic reduction from 1998 to 251, which is a 87.44% improvement over the performance of the original non-tuned query. The disk scans represents the number of times the physical hard drive is accessed for reading data while, the logical reads is the number of times the memory is accessed for data or "cache hits". Since both disk scans and logical reads directly impact the total processing times, these reductions in both disk scans and logical reads will reduce the processing times greatly.

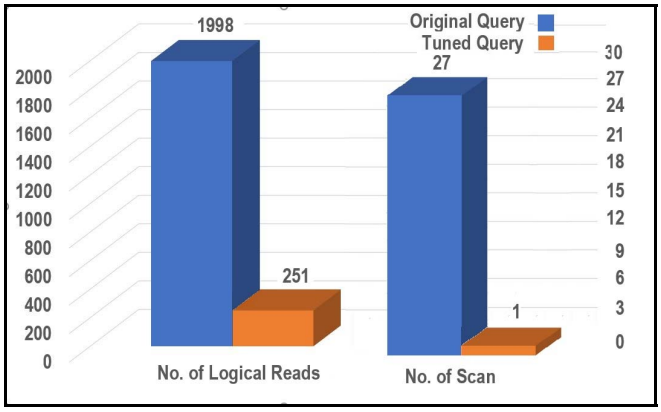


Fig. 8. Query Performance Indicators - Technique 2

IV. CONCLUSIONS

Database performance relies greatly on how efficiently data can be extracted from the database using queries. Research has shown that making small changes to SQL queries can improve the performance of the queries and in turn that of the database substantially. Hence, query tuning has become one of the important functions of database developers and programmers. In this paper, the authors have proposed two query tuning techniques. The proposed techniques have been evaluated extensively in laboratory (development) environments and shown that they improve the performance of queries over their non tuned counterparts to a very large extent.

REFERENCES

[1] A. J. Karim, "The significance of management information systems for enhancing strategic and tactical planning," *Journal of Information Systems and Technology Management*, vol. 8, no. 2, pp. 459–470, 2011.

[2] J. Shao, X. Liu, Y. Li, and J. Liu, "Database performance optimization for SQL server based on hierarchical queuing network model," *International Journal of Database Theory and Application*, vol. 8, no. 1, pp. 187–196, 2015.

[3] J. Yang, P. Subramaniam, S. Lu, C. Yan, and A. Cheung, "How not to structure your database-backed web applications: A study of performance bugs in the wild," in *40th International Conference on Software Engineering*, Gothenburg, Sweden, 2018, pp. 1–11.

[4] A. Wagh and V. Nemade, "Query optimization using multiple techniques," *International Journal of Computer Applications*, vol. 163, no. 3, pp. 30–32, 2017.

[5] M. Sharma, "Query optimization using SQL transformations," *International Journal of IT, Engineering and Applied Sciences Research*, vol. 1, no. 1, pp. 100–104, 2012.

[6] N. Satyanarayana, S. Sharfuddin, and S. J. Bhasha, "New dynamic query optimization technique in relational database management systems," *International Journal of Communication Network Security*, vol. 2, no. 2, pp. 65–69, 2013.

[7] W. S. Han, W. Kwak, J. Lee, G. M. Lohman, and V. Markl, "Parallelizing query optimization," in *34th International Conference on Very Large Data Bases*, Auckland, New Zealand, 2008, pp. 188–200.

[8] I. Trummer and C. Koch, "Parallelizing query optimization on sharednothing architectures," in *42nd International Conference on Very Large Databases*, New Delhi, India, 2016, pp. 660–671.

[9] S. Patil, P. Damare, J. Sonawane, and N. Maitre, "Study of performance tuning techniques," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 2, no. 3, pp. 499–502, 2015.

[10] G. Feuerlicht, "Database trends and directions: Current challenges and opportunities," in *10th Annual International Workshop on DAtabases, TExts, Specifications and Objects*, Stedronin Plazy, Czech Republic, 2010, pp. 163–174.

[11] M. Khan and M. N. A. Khan, "Exploring query optimization techniques in relational databases," *International Journal of Database Theory and Application*, vol. 6, no. 3, pp. 11–20, 2013.

[12] R. Bhajipale, P. Bisen, A. Meshram, and S. S. Thakur, "SQL tuner," *International Journal of Computer Trends and Technology*, vol. 33, no. 1, pp. 29–32, 2016.

[13] P. Karthik, G. T. Reddy, and E. K. Vanan, "Tuning the SQL query in order to reduce time consumption," *International Journal of Computer Science Issues*, vol. 9, no. 4/3, pp. 418–423, 2012.

[14] P. S. Bhadoria, A. Chhabra, and S. Ojha, "SQL performance tuning in oracle 10g and 11g," *International Journal of Software and Web Sciences (IJSWS)*, vol. 7, no. 1, pp. 13–16, 2014.

[15] B. Dageville, D. Das, K. Dias, K. Yagoub, M. Zait, and M. Ziauddin, "Automatic SQL tuning in Oracle 10g," in *13th International Conference on Very Large Databases*, Toronto, Canada, 2004, pp. 1098–1109.

[16] H. Herodotou and S. Babu, "Automated SQL tuning through trial and (sometimes) error," in *2nd International Workshop on Testing Database Systems*, Providence, RI, USA, 2009, pp. 1–6.

[17] J. Habimana, "Query optimization techniques - tips for writing efficient and faster SQL queries," *International Journal of Scientific & Technology Research*, vol. 4, no. 10, pp. 22–26, 2015.

[18] M. L. Rupley, "Introduction to query processing and optimization," Indiana University, South Bend, South Bend, IN, USA, techreport TR- 20080105-1, 2008.

[19] R. Sahal, M. Nihad, M. H. Khafagy, and F. A. Omara, "iHOME: Indexbased JOIN query optimization for limited big data storage," *Journal of Grid Computing*, vol. 16, no. 2, pp. 345–380, 2018.

[20] S. S. Srinivas, B. V. Naik, and J. S. A. Kumar, "Query minimization methods," *International Journal of Scientific & Engineering Research*, vol. 8, no. 5, pp. 30–33, 2017.

[21] S. Batra, S. Sachdeva, and S. Bhalla, "Entity attribute value style modeling approach for archetype based data," *Information*, vol. 9, no. 2, pp. 1–30, 2018.

[22] V. Sharma and L. Sharma, "A survey: Query optimization," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 4, pp. 4094–4095, 2015,.

[23] D. Patel and P. Patel, "An approach for query optimization by using schema object base view," *International Journal of Computer Applications*, vol. 119, no. 16, pp. 21–24, 2015.

[24] D. Colley and C. Stanier, "Identifying new directions in database performance tuning," *Procedia Computer Science*, vol. 121, p. 260-265, 2017.

[25] A. K. Sirohi, "Optimization of dynamically generated SQL queries for tiny-huge, huge-tiny problem," *International Journal of Database Management Systems*, vol. 5, no. 1, pp. 53–68, 2013.

[26] C. G. Corlatan, M. M. Lazar, V. Luca, and O. T. Petricica, "Query optimization techniques in Microsoft SQL server," *Database Systems Journal*, vol. V, no. 2, pp. 33–48, 2014.

[27] D. Taniar, H. Y. Khaw, H. C. Tjioe, and E. Pardede, "The use of hints in SQL-Nested query optimization," *Information Sciences*, vol. 177, pp. 2493–2521, 2007.

[28] A. D. Lokhande and R. M. Shete, "The use of hints in SQL-Nested query optimization," *Journal of Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 54–57, 2012.

.

AUTHOR INDEX