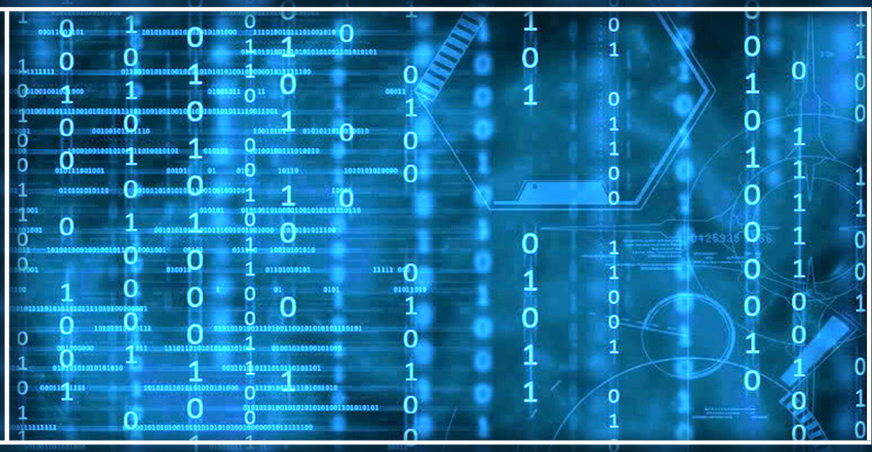
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 9 Issue 2
February 2018



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 9 Issue 2 February 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **Aamir Shaikh**
- **Abbas Al-Ghaili**
  Mendeley
- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
  Université Abdelmalek Essaadi Faculté Polydisciplinaire de Larache Route de Rabat, Km 2 - Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**
  Gomal Unversity
- **Abeer ELkorany**
  Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**
  University of the Western Cape
- **Aderemi A. Atayero**
  Covenant University
- **Adi Maaita**
  ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**
  Department of Mathematics and Informatics, Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**
  University of Belgrade, Faculty of organizational
- **Ahmad Saifan**
  yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**
  Ahlia University
- **Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Ajantha Herath**
  Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**
  University Of California, San Diego
- **Albert S**
  Kongu Engineering College
- **Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Alexane Bouënard**
  Sensopia
- **ALI ALWAN**
  International Islamic University Malaysia
- **Ali Ismail Awad**
  Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**
  Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University
- **Anews Samraj**
  Mahendra Engineering College
- **Anirban Sarkar**
  National Institute of Technology, Durgapur
- **Anthony Isizoh**
  Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**
  University of Naples Federico II
- **Anuj Gupta**
  IKG Punjab Technical University
- **Anuranjan misra**
  Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
  University Technology Malaysia(UTM)
- **Aree Mohammed**
  Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**
  University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**
  Constantine 1 University
- **Ashok Matani**
  Government College of Engg, Amravati
- **Ashraf Owis**
  Cairo University
- **Asoke Nath**

CONTENTS

Dynamic Time Warping and FFT: A Data Preprocessing Method for Electrical Load Forecasting

Juan Huo
School of Electrical and Automation Engineering
Zhengzhou University, Henan Province, China

***Abstract*—For power suppliers, an important task is to accurately predict the short-term load. Thus many papers have introduced different kinds of artificial intelligent models to improve the prediction accuracy. In recent years, Random Forest Regression (RFR) and Support Vector Machine (SVM) are widely used for this purpose. However, they can not perform well when the sample data set is too noisy or with too few pattern feature. It is usually difficult to tell whether a regression algorithm can accurately predict the future load from the historical data set before trials. Here we demonstrate a method which estimates the similarity between time series by Dynamic Time Warping (DTW) combined with Fast Fourier Transform (FFT). Results show this is a simple and fast method to filter the raw large electrical load data set and improve the learning result before looping through all learning processes.**

***Keywords*—Load forecast; Dynamic Time Warping (DTW); Fast Fourier Transform (FFT); random forest; Support Vector Machine (SVM)**

I. Introduction

In electrical engineering, load forecasting speculates and predicts the future power load demand for a certain period of time from historical load data. The accuracy of load forecasts has important effect on power system operations. For power management system, the Day-ahead scheduling process consists of the following principal functions: (1) assemble and update Day-ahead transmission outages; (2) produce Day-ahead zonal load forecast; (3) tabulate and evaluate non-firm transactions; (4) perform automated mitigation of generator offers.

Historical load data is important to set up prediction model and the training features. Most of the research for the prediction methods focus on the forecast methods while did not mention too much about the data preparation process [1], [2]. In most study cases for day ahead load forecast, data of the adjacent days have been selected manually as the training data source. As well known, data pre-processing has significant impact on predictive accuracy, even for some data mining techniques which can balance error in class population of unbalanced datasets [3]. Thus one new method which combines the estimation of DTW and FFT is introduced to act as a reference for raw data pre-processing and feature selection for electrical load data. The electrical load data source is evaluated in both time and frequency domain by Dynamic Time Warping (DTW) and Fast Fourier Transform (FFT) before training. DTW and FFT are supposed to help feature reselection and data re-sampling for data pre-processing purpose. Both DTW and FFT have been widely used to identify the similarity and patterns between two data sets. In the following sections, the function of DTW&FFT for time series similarity and pattern recognition will be tested.

For the purpose of electrical load forecast, we have used Random Forest and Support Vector Machine (SVM) which are popular methods for load forecast in recent years [4]–[6]. In 2001, EUNITE network organized a world wide competition on the daily electrical load prediction problem. In this event, SVM (support vector machine) or SVR (support vector regression) surpassed the other algorithms and claimed the throne for daily electrical load forecast [7], [8]. Some recent papers have found Random Forest Regression (RFR) can also perform well for time series prediction task, sometimes it can even excel SVM for some data sets [4], [5], [9], [10]. But some other papers reserves this opinion and points out they can only be compared when parameters are fixed [6].

The data source used in this paper for test is NYISO (details see Section III-A), which is rich with more than 15 years' historical record for New York Area. For the purpose of electrical load prediction, our initial forecast result is not good. With the analysis from DTW and FFT, the reason is explained. According to the analysis, redundant data sets are filtered out and the new feature is added which results in improvement for the downtown zone (N.Y.C.) of the New York. Another analysis of DTW and FFT for suburb zone also explains why data of suburb zone (North zone) is not suitable for RFR and SVM regression and can be a reference for the other training.

This paper is organized as follows. Section I provides the background knowledge of this work; Section II has the main algorithms demonstrated; Section III introduces the features of the data source; Section IV compares the time efficiency between the traditional and the new algorithm; Section V is the conclusion of this paper.

II. Method and Algorithm

We analyzed the electrical load data by combing DTW and FFT. Besides DTW and FFT. Cross correlation has also been considered once, however it failed to identify the difference between each year's difference for NYISO North Zone data set, thus it is not used as our evaluation reference in this paper. The result of DTW for similarity reference is a distance value $D(U, V)$. For FFT analysis, the resulting common frequency components (frequency with maximum power spectrum amplitudes) are the reference parameters.

A. Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is an algorithm for time series analysis, it has been used for measuring similarity between two temporal sequences which may vary in time of speed. The essence of DTW is to estimate the alignment between two time series. To align two time series, U and V, an n-by-m matrix X is constructed. The (ith,jth) element of the matrix X_{ij} contains the distance $d(u_i, v_j)$ between the two points u_i and v_j. The Euclidean distance is typically used, which corresponds to the alignment between the points u_i and v_j. A warping path, W which is a set of matrix elements that defines a mapping between U and V [11]. Its kth element is defined as

$$w_k = (i_k, j_k) \tag{1}$$

and the warping path W is

$$W = w_1, w_2, ..., w_k, ..., w_K \tag{2}$$

where $max(m, n) \leq K < m + n - 1$

The warping path W is minimized and typically subjected to some constraints such as boundary conditions,continuity and monotonicity. The warping cost can be estimated by different algorithms, the most used one is a recurrence equation that defines the cumulative distance as the distance in the current cell and the minimum of the cumulative distances of the neighbouring elements. Thus the distance between two points is minimized, which can be expressed as:

$$D(U, V) = \min_{W} \left[\sum_{K=1}^{K} d(W_K) \right] \tag{3}$$

where $D(U, V)$ is the estimated distance between two time series U and V. It is an important reference in time domain. DTW has been widely used in different areas to find the matched subsequences between two time series. The speed of DTW has been improved dramatically with different kind of methods and can deal with trillions of time series in a short time [11]–[13].

B. Fast Fourier Transform (FFT)

For the frequency domain, we analyze the amplitude spectrum of FFT (Fast Fourier Transform) to find the common frequency components of two time series. The FFT is a kind of discrete Fourier transform algorithm which reveals periodicities in input data as well as the relative strengths of any periodic components. The input data is decomposed into smaller frequency complex components. By this way, it is more convenient to find the similarity pattern in frequency domain.

C. Training Algorithm and Process

The artificial intelligent model we have used for training purpose is the Random Forest Regression (RFR) and Support Vector Regression (SVR). We have used R package libraries for the implementation of these algorithms [14]–[16]. After several trials of cross validation between different years, we find the default parameters of R package is generally ok for SVM. The main parameter of RFR are the number of trees $ntree$ and the number of variables to partition at each tree node $mtry$, which do not have remarkable impact on the resulting accuracy according to the investigation of previous papers [6]. The tree number we have chosen for RFR is 1000 and variable number is 10 which are good enough to get satisfactory result.

III. Data Profile

The data set is a public electrical load forecasting database, New York Independent System Operator (NYISO)[1]. This data source recorded real-time load demand in every five minutes measured in MW. For comparison, we have studied two zones: N.Y.C. and North. The zonal forecast models use weather information of every day gathered from stations across of New York. Each hour's load was averaged for day ahead load prediction task. Fig. 1 shows the hourly load profile of N.Y.C. while Fig. 5 is the load of North zone. As N.Y.C. is the central part of New York, the load demand is obviously different from the North Zone, where the population density is low (the population for N.Y.C. is about 8 million, while the population for the North zone is only 82 thousand).

A. N.Y.C. Zone

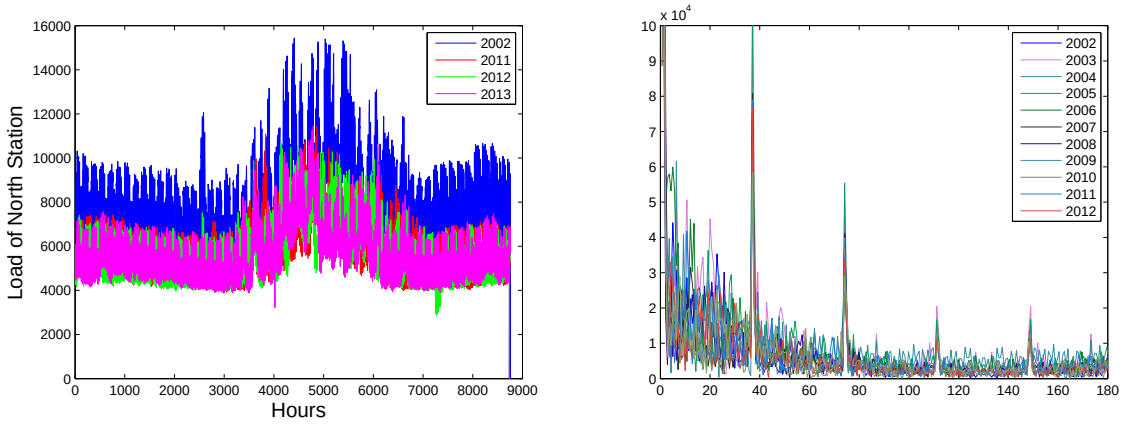
The load demand of a year varies regularly with season in N.Y.C. As in Fig. 1(a), a summer day is obviously superior to the other days each year. Fig. 1(b) is our analysis of all years' load with FFT, the amplitude of frequency spectrum of the recent years are highly overlapped over the main bands.

For load prediction task, we use year 2013 as the test target to be predicted. Before the forecast process, the load of every month from 2002 to 2012 is compared with the corresponding month's load of year 2013. The DTW distance D_{ij} ($i = 1...12$, $j = 2002...2012$), is calculated for every paired month. Thereafter, D_{ij} is divided by the average load of target month's adjacent days, which results in ND_{ij}. Anova analysis of ND_{ij} is shown in the two figures of Fig. 2, which are grouped by dimension month i and year j respectively. Fig. 2 shows the load of year 2002 to 2005 differs from the recent years 2006 to 2012 apparently, the average DTW distance of year 2002 to 2005 is nearly two times of the other years. Fig. 2(b) shows the different distance varies with month. Although the deviation of each month group is high, the summer period ($i = 6...9$) has larger distance than the other months. This is consistent with our normal observation that the load variation of summer period is more uncertain than the other months.

B. RFR and SVM Prediction

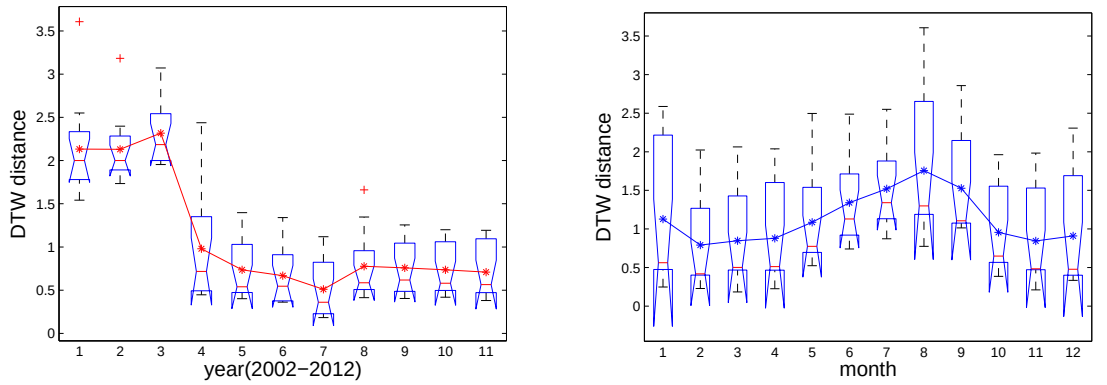
After analysis, the above data were then put into our training system of RFR and SVM to validate our hypothesis that there is noise or outlier values in the data set for regression purpose when month $i = 6...9$ or year $j = 2002...2005$. The features of the input vectors are initially set in Table I. The training result is evaluated as by a most common used parameter for electrical load forecast measurement, which is named as MAPE(mean absolute percentage error) [6]. Normally, MAPE calculates the average error of one day 24h. The formula of MAPE is shown in equation 4,where X_i is the predicted value

[1] http://www.nyiso.com



(a) (b)

Fig. 1: N.Y.C. load profile. (a) The hourly load profile by year. (b) FFT analysis and amplitude spectrum of different years' load.



(a) (b)

Fig. 2: Anova analysis of DTW distance $ND_{ij} = D_{ij}/AVG(L_{2013})$. The time series of each month from year 2001 to 2012 is paired with the corresponding month of test year 2013.

and R_i is the real electrical load data on the i_{th} day or hour of the prediction period (day or hour).

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|X_i - R_i|}{R_i} \tag{4}$$

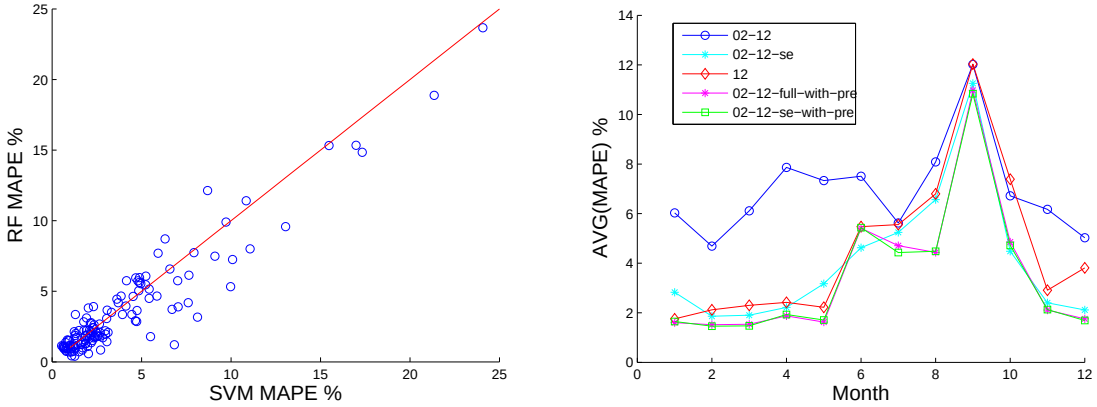
The training process is as follows: for every month of 2013, one day of the first week is randomly selected, as the first day of the test set to be predicted. The hourly load of that day to be predicted is labeled as S_{ik} $(i = 1 \ldots 12, k = 1 \ldots 24)$, in which i represents the month sequence and k represents the hour in the range of 0 to 23. All days before S_i are in the training set. The load 11 days after S_i is the test data to be predicted. The data set is grouped by hour k. Data of each hour group is trained for the corresponding hour to be predicted. For example, if hour 23:00 of March 1st 2013 is to be predicted. The load data at 23:00 is sampled from everyday before March 1st 2013. Features of table I are also collected from that day. Since each train results in 11 days' predicted value, at last we get a MAPE matrix with size 11×12.

We did a comparison of RFR and SVM by using 132 pairs of MAPE (all days of year 2013). T-test of these pairs proves the hypothesis that the difference between SVM and RFR prediction comes from a normal distribution with mean equal to zero and $p < 0.01$. The scatter plot of SVM vs RFR is shown in Fig. 3(a), which indicates in our load forecast task the difference between SVM and RFR is not significant.

With the initial five features in Table I, the first train has used all years' data $(2002 - 2012)$. The prediction error is very high. The average MAPE is more than 5% every month and is shown as the blue circle line labeled with the set range "02-12" in Fig. 3(a). Then we select the data of the year $2006 \ldots 2012$ for $ND_j <$ AVG(ND) measured by DTW. The average MAPE of this selected data group "02-12-se" is

TABLE I: Input Features

ID	Features	Range	Note
1	**Month number**	$1 \ldots 12$	
2	**Weekday**	$1 \ldots 7$	
3	**Holiday**	binary	1 is holiday and weekend, 0 workday
4	**Minimum temprature**	$15.8 \ldots 102.2^{\circ}F$	
5	**Maximum temprature**	$1.4 \ldots 84.2^{\circ}F$	
6	***Load before 24h**	$0 \ldots 15503.68$ MW	not used in the initial train



(a) (b)

Fig. 3: MAPE comparison for RFR and SVM. (a) MAPE comparison between RFR and SVM. (b) Average MAPE of RFR with different years' data.

shown as the cyan star line in Fig. 3(b). The red diamond line "12", which only has the data of year 2012, has mean MAPE($= 4.5652$). Although, the mean MAPE($= 4.0556$) of this cyan star line "02-12-se" is only slightly lower than "12", it is much better than the mean MAPE($= 6.9319$) of "02-12". This is consistent with our observation of DTW test results in Fig. 2. The time domain difference between the early years' data $j = 2002 \ldots 2005$ and the recent years' data lowered the prediction accuracy for the recent year. On the other hand, in section III-A, we have observed that although there is large deviation between the early years' data and the recent years' data on amplitude, they share the same main frequency bands. This means there is still similar pattern during their variation. Therefore in a new train, a new feature "Load before 24h" shown in Table I was added for training.

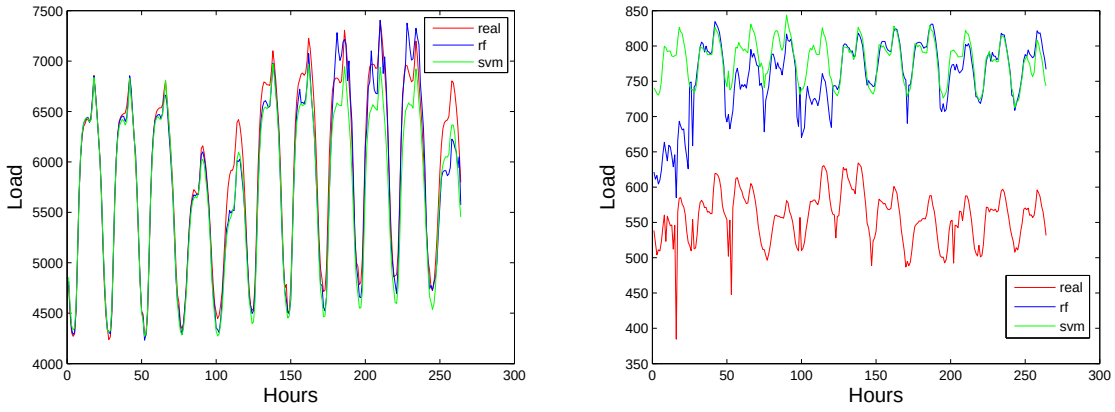
With previous day's load added, the prediction result of RFR is improved, shown as green "02-12-se-with-pre" (DTW selected group) and "02-12-full-with-pre" (all data sets) in Fig. 4(a). There is no significant difference between the results of these two groups, as RFR has classified the early years' data automatically according to the new added feature "Load before 24". The mean MAPE of "02-12-se-with-pre" (=3.4944) and "02-12-full-with-pre" (=3.5336) are both lower than any training without the new feature. The DTW selected group "02-12-se-with-pre" still performs slightly better. For all the training result, summer time is always the worst prediction period $i = 6 \ldots 9$, which is also coherent with our DTW distance hypothesis in Section III-A.

C. North Zone Data

We then use the similar method to test the North Zone's dataset. With fewer population, the load variation is irregular and has more uncertain factors shown in Fig. 5(a). We once tried cross correlation to analyze the difference between different years. However the correlation coefficient value is always above 0.95 between years, just like the N.Y.C. This does not provide much useful information. Whereas the DTW distance analysis show the DTW distance of North Zone is all more than five times of the average load, $ND_{ij} > 5$. This indicates the deviation between the north zone's data is very large. The FFT analysis in Fig. 5 also shows few coherence in frequency domain from year 2011 to 2014. We then have a trial to use the Random Forest and SVM directly to predict the day ahead hourly load of 2014. All the six features with "*Load of last 24h" in Table I were used. Fig. 6 shows the monthly average MAPE is very high and even above 50%. Thus this proves north zone's data is not suitable for prediction with regression methods.

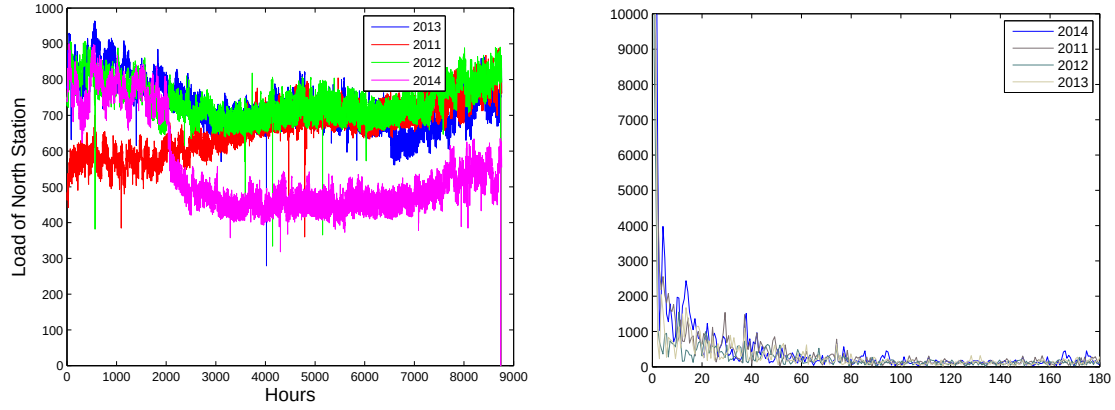
IV. Time Cost

We did test to evaluate the time cost of DTW&FFT by tic-toc function in matlab. The computer has $3.4GHz$ CPU and $8GHz$ RAM. The total time cost of DTW&FFT to compare the year 2011,2012 with 2013 separately for every month is 0.2243 second. When we use data sets of 2011 and 2012 to predict the load trend of 11 days of 2013 in one trial, the average cost for SVM and RFR varies. The time cost for SVM can be as small as 0.046 second. But to have a global



(a) (b)

Fig. 4: Load prediction of one train (December, 2013). The red line is the real load, “RF” and “SVM” represents RFR and SVM respectively. (a) The N.Y.C. zone. (b) The north zone.



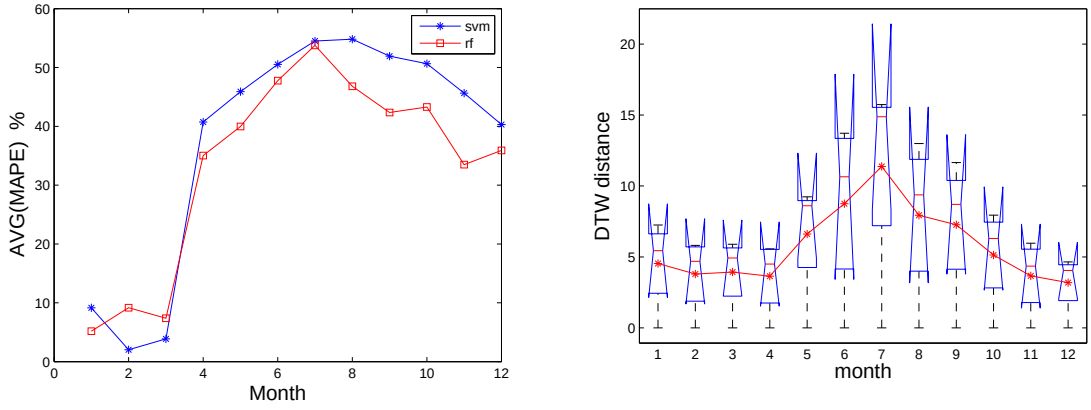
(a) (b)

Fig. 5: North Zone. (a) The load profile. (b) FFT amplitude spectrum of north zone.

estimation of a whole year and find the special period like the summer time, we need to organize new prediction for different period multiple times and make the algorithm much more complex than DTW&FFT. This is the same for RFR, for which the situation can be worse. Because the time cost of RFR strongly depends on ntree and mtry, when $ntree = 10$ and $mtry = 2$, the time cost is less than 0.008 second, however, when ntree increases, such as $ntree = 1000, mtry = 10$, the time cost is 1.1 second on average for one trial. Generally speaking, DTW&FFT is better for global observation of the difference between time series. Some times, when only FFT is implied, it is enough to warn the low performance of regression with time cost only 0.045 second, such as the North zone data.

V. Conclusion

The above results have shown the method which combines DTW and FFT together can help to evaluate the data set for re-sampling and feature selection. Data set group selected by DTW and FFT performs better than group which has not been preprocessed. Especially for electrical load data with too few pattern features, DTW&FFT not only can identify the bad data set for prediction, but also can analyze the reason for potential prediction failure in both time and frequency domain. In addition, the computation of DTW and FFT is simpler than SVM and RFR learning process and thus is time saving compared to loop through all data sets and try different predictors. DTW&FFT algorithm has advantage to view the global features of electrical load forecast time series. Such algorithm composition can help to analyze the quality and property of the electrical load time series and should be treated as an important reference for electrical load data pre-processing.



(a) (b)

Fig. 6: (a) RFR and SVM prediction monthly average MAPE for North Zone. (b) Anova analysis of DTW distance $ND_{ij} = D_{ij}/AVG(L_{2014})$. The time series of each month from year 2011 to 2013 is paired with the corresponding month of 2014.

REFERENCES

[1] H. K. Alfares and M. Nazeeruddin, "Electric load forecasting: literature survey and classification of methods," *International Journal of Systems Science*, vol. 33, no. 1, pp. 23–34, 2002.

[2] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, 2001.

[3] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing," *European Journal of Operational Research*, vol. 173, no. 3, pp. 781–800, 2006.

[4] Y. Chen, S. Guo, H. Chen, L. I. Wanhua, K. Guo, and Q. Huang, "Electricity customers arrears alert based on parallel classification algorithm," 2016.

[5] X. Wu, J. He, P. Zhang, and J. Hu, "Power system short-term load forecasting based on improved random forest with grey relation projection," *Dianli Xitong Zidonghua/automation of Electric Power Systems*, vol. 39, no. 12, pp. 50–55, 2015.

[6] A. Lahouar and J. Ben Hadj Slama, "Day-ahead load forecast using random forest and expert input selection," *Energy Conversion and Management*, vol. 103, pp. 1040–1051, 2015.

[7] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load forecasting using support vector machines: A study on EUNITE competition 2001," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 1821–1830, November 2004.

[8] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] Z. Y.-w. M. B.-h. J. Huan, "Research of medium and long term precipitation forecasting model based on random forest," *Water Resources and Power*, vol. 33, no. 6, pp. 6–10, 2015.

[11] T. C. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011, fu, Tak-chung.

[12] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012, August 12, 2012 - August 16, 2012*, ser. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp. 262–270.

[13] G. T, "Computing and visualizing dynamic time warping alignments in r: The dtw package." *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[15] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015, r package version 1.6-7. [Online]. Available: https://CRAN.R-project.org/package=e1071

[16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

A Serious Game for Improving Inferencing in the Presence of Foreign Language Unknown Words

Pedro Gabriel Fonteles Furtado*, Tsukasa Hirashima and Hayashi Yusuke
Graduate School of Engineering
Hiroshima University
Hiroshima Prefecture, Japan

***Abstract*—This study presents the design of a serious game for improving inferencing for foreign language students. The design of the game is grounded in research on reading theory, motivation and game design. The game contains trial-and-error activities in which students create conversations and then watch these conversations play out. Making mistakes results in students receiving feedback and being requested to try again. An evaluation of the system was also conducted, in which participants used both simple text and the game. Post-test scores for using the game were significantly higher than scores when reading the text. User reception to the system was also positive. These results suggest that serious games can be effective for enhancing inferencing when foreign language students face unknown words. Implications for reading comprehension and for incidental vocabulary learning are also discussed.**

***Keywords*—Serious game; foreign language; contextual inference; unknown words**

I. Introduction

Inferencing is the process of making connections when trying to interpret a text [1]. Inferencing information from the text is a necessary component of reading. To understand a text, readers must use their previous knowledge as base for the inferences of the new information. It also plays a role when disambiguating the meaning of words and clauses [2]. The inferencing necessary is not extensive, unless the information is too unfamiliar, or the language proficiency of the reader is not sufficient for that specific text [3]. One case of lack of proficiency is lack of vocabulary. Unknown words will often increase the amount of inferencing necessary for understanding the passage. In such cases, the reader will have to rely on the words he does know, the current information he collected from the text and his own background knowledge. In some cases, the reader may also infer the meaning of the unknown word. Children learn thousands of words per year and those words are mostly acquired from context [4], [5]. This also happens for second language (L2) learners [6]–[9].

As such, inference from context in the presence of unknown words plays two roles in reading comprehension:

- Understanding the passage by using the remaining information;
- Inferring the meaning of the unknown word to aid in understanding the passage.

Learning words from context is a form of incidental learning. Incidental learning is the accidental learning of information without intention of remembering that information [10], [11]. The amount of words the students learn from incidental learning might make one believe that inferring the meaning of a word from context is an easy task. The problem is that students often fail to pick up this contextual information in the presence of unfamiliar vocabulary. Learners may ignore the word and give up on understanding the given passage. There may not be enough information in the context to infer the meaning of the word. It can also be the case that students infer the wrong meaning of a word [10], [12], [13]. Even when using dictionaries, students don't look up the meaning of all the words [14], specially when there are too many new items [15]. Also, new words usually need to be encountered multiple times [4] in order to be learned.

One factor that has been shown to affect incidental learning is task involvement load. Vocabulary enhancing techniques have been used to increase the effectiveness of incidental vocabulary learning [12], [16]–[19]. Past studies also agree that the higher the involvement load, the higher the effectiveness of incidental vocabulary learning. A meta-review on this matter can be seen in [20]. These tasks, however, have limitations. First, time spent in the classroom is limited. This limits the amount of time users will spend interacting with these activities. Users cannot be expected to engage for long periods of time in these tasks on their free time. Yet, as mentioned before, students are able to learn a large amount of words in their school years which cannot be attributed to explicit learning. Voluntary reading, often done outside the classroom, have been associated with better language acquisition [21] and better vocabulary test scores [22]. Voluntary reading works because of the large volume of reading done. Students read in such large volumes that they have multiple encounters with the unknown words. However, the amount of reading necessary and the time it takes to for measurable progress to be made can make extensive reading hard to implement [23].

Current research agrees that reading comprehension is positively affected by motivation [24]–[26]. Students with poor reading skills show more correlation between their motivation and their reading performance [24]. Motivation also affects the total amount of reading done [27]. Interventions to increase reading comprehension have been shown to increase reading performance [28]. Motivation is also a predictor of success in language learning in general. It was shown that motivation had the stronger correlation with language grades and self-evaluation [29]. In that study, motivation outperformed both the attitude towards the learning situation, integrativeness and orientations.

One approach to handling motivation is Digital Game

Based Learning (DBGL). One definition for DBGL is "the innovative learning approach derived from the use of computer games that possess educational value or different kinds of applications that use games for learning and education purposes such as learning support, teaching enhancement, assessment and evaluation of learners" [30]. DBGL includes the use of both commercial games and serious games. Serious games are digital games made for more than entertainment [31]. One core element of games that affects motivation is challenge. Appropriate challenge that matches the skill of the user will greatly affect the experience [32]. If it's too easy, the player will be bored. If it's too hard, the player will be discouraged. This fits with the conditions to achieve flow state, a popular construct in entertainment research [33]. It also fits with the need for competence from the Self Determination Theory explained in [34], [35]. As such, proper challenge is one of the factors that influences engagement and motivation in game design. However, game design is not about arbitrarily creating challenge. A game must be both accessible and easy to use while still providing a hard experience for the player [36]. This means that a game's challenge should not be born from usability issues. It's necessary to focus on usability in game design. Also cited as an important element is for the player to have freedom to fail and try again, as much as the user needs [32].

On the limitation of using commercial games, [37] conducted a research comparing vocabulary recall in players and watchers of a music game. Players actually interacted with the game and watchers were asked to simply watch the game. Watchers had a much higher score in vocabulary recall. The research reported that players were divided between listening to the words or doing well in the game. This shows that extraneous cognitive load can get in the way of reading, depending on game genre. This shows that while using commercial games is cost-effective, there might be a loss in learning gains compared to a well-designed educational game.

One popular game genre is visual novels. It involves reading a narrative through long periods of time [38]. Visual novels, unlike books, only show one snippet of text at a time. After the player does some sort of interaction with the game (pressing a button, for instance) the game advances to the next snippet of text. This means that players are presented with a limited amount of text at a time, meaning that players don't have to keep track of their progress in a book's page, for example. Visual novels also have graphical elements like backgrounds and character artwork. This gives the player a vision of what is going on inside the story. This facilitates reading comprehension [39]. Many visual novels also have some game-play elements between story-line sections, such as [40]. Alternation between story and game-play is a recurring element in game design, and is said to have beneficial elements, such as rewarding the player and improving pacing [41].

The serious game used in this study contains an activity designed to induce students to infer information from context. It locks them into a trial-and-error feedback loop while they attempt to construct a conversation. It combines this activity with a story, similar to a visual novel. In the sections below, the design of the game will be further explained, with a focus put on how it improves inferencing and on its motivational elements. Then, the experiment will be described, and the results analyzed. The game used in this study has been explored before in the context of extensive reading [42], [43].

This study aims to answer weather or not inferencing improved by using the game when compared to simply reading text. It also aims to present how the design of the game interacts with inferencing.

Section II presents related work in the field, exploring other works that used games for language teaching with a focus on reading. Section III presents the design of the game and details of the experiment. The game's design portion starts by presenting the challenges of designing a game for improving inferencing, based on the discussion presented in Section I. It then presents the various aspects of the game and ends by summarizing how the aspects of the game answer the design challenges. Then it introduces various aspects of the experiment done. Section IV presents the results obtained through the experiment and discusses their meaning. Section V concludes the study. It presents implications of the results on the field, the shortcomings of the study and possible future research.

II. Related Work

DBGL has been used successfully for language teaching on various fields, ranging from situated vocabulary learning [44], conversational visual novels [45], commercial games in the classroom [46], [47], relating language gains to gaming habits [48] and so on. The synthesis done in [49] about video games and second language learning concluded that games have a positive impact on learning, specially for vocabulary, with the experimental group surpassing the traditional study control group in some cases [50]–[53]. This shows that it is possible to have gains when reading content in games. The gains from vocabulary measured are due to incidental learning while playing. This shows that inferencing while reading also happens when playing games.

Despite this, there has not been much research that focused specifically on designing a software focused on supporting reading as the main activity in the context of DBGL and foreign languages. The work of [54] attempted to use a augmented reality game to enhance reading comprehension but it failed to show gains in reading comprehension. It did, however, show motivational gains. Some works [55]–[57] focus on first language primary reading skills (among other fields) for young children and showed positive results, but does not focus on reading long texts, focusing instead on more basic skills, such as individual word reading. The study of [58] focuses on first language reading comprehension, but does not evaluate the actual learning gains and does not expose how the design of the game relates to actually attaining those skills. Other works, like [59], address L2 reading but do not go in depth in designing the application to integrate with the reading process.

As far as inferencing during foreign language reading goes, none of the work reviewed addressed the process directly.

III. Methodology

A. Design of the Game

1) Design Challenges: Summarizing the material presented in Section I, the challenges in designing an activity for

Fig. 1. Screen-shot of a story sequence in the game.

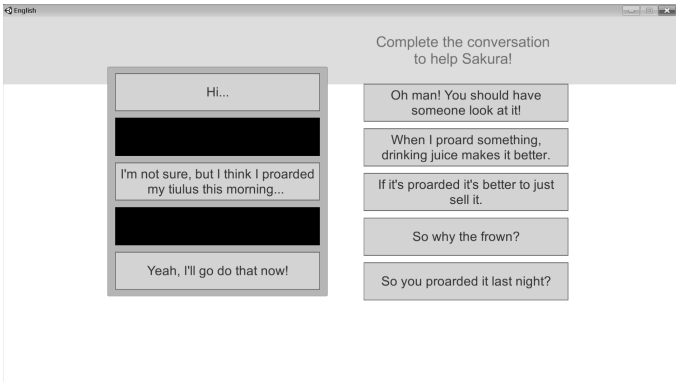


Fig. 2. Screen-shot of the conversation assembling activity.

improving inference from text in the presence of unknown words are:

- Students may ignore passages which contain unknown words;
- Students may infer the wrong meaning of a word;
- Activities would benefit from being intrinsically motivating, which implies:
  - Better performance in reading comprehension and language learning in general;
  - Possibility for the activity to be used in students' free time, thus avoiding time limitations of the classroom;
  - Compatibility with extensive reading.

If the activity is a game, it would also present the following challenges:

- Game elements should not detract from reading or from inferencing;
- The game needs be challenging but not too challenging;
- The activity should allow for students to fail as much as they need in order to progress.

2) Game Introduction: The game features a combination of story segments and activity segments. Screen-shots of the segments can be seen in Fig. 1 and 2, respectively.

In the story segment, we can see the characters present in the scene, a box displaying a piece of dialog and a background depicting the current location. Those elements can all be seen in Fig. 1. Upon user input, the story advances. This makes the next line of dialog or narration appear. This dynamic continues until the scene ends. While reading is prevalent, no elements are present to induce or improve contextual inferencing in this segment.

The inducement of contextual inference happens in the activity segments. During activity segments, users attempt to construct a conversation that solves a certain in-story goal. The conversation is constructed by inserting the pieces of the conversation into the empty slots, as shown in Fig. 2. The design of activity segments will be further discussed below.

3) The Conversation Construction Activity's Design: This activity consists of constructing a conversation and watching it play out. If the constructed conversation is inappropriate, a new conversation will be formed that will give the user insight into why that conversation is wrong and into how to create the appropriate conversation. From now on we'll refer to the phase of constructing a conversation as the assembling phase and the phase of watching the conversation play out as the result phase. Those two phases will be further developed in the subsections below.

The ideal behavior of the user for this activity can be seen in Fig. 3.

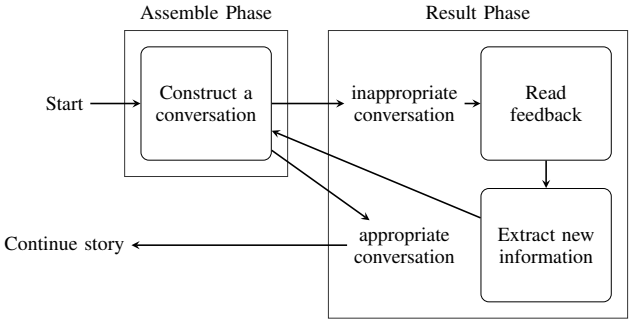


Fig. 3. Ideal user behavior flow for the conversation construction activity.

4) Conversation Construction's Assembling Phase: This phase consists of forming a sequential dialog by inserting conversation pieces into a grid, like in Fig. 2. However, the user can only insert the pieces related to what one person says. What the other person says is already fixed on the grid and cannot be moved. This was a deliberate decision to reduce ludo-narrative dissonance. Ludo-narrative dissonance is when game-play and story have a mismatch [60]. If players ask themselves "if I am the main character in the narrative, how come I can control what the other person will say?" immersion would break. Whenever a student fills up all vacant spaces with conversation pieces a button will appear in the interface. Pressing that button will take the student back to the screen of the story segments and the resulting conversation will play out.

Regarding Fig. 3, this refers to the "Construct a conversation" node.

5) Conversation Construction's Result Phase: First, the system must check if the conversation is appropriated or

not, by comparing it to the answer. If the conversation is appropriate, it will be shown to the player as it is, and the story will go on. This refers to the "appropriate conversation case" in Fig. 3.

However, if it's incorrect, the system must logically assemble a new conversation based on the player's constructed conversation. this is done by using the following steps:

- Find the player's first mistaken conversation piece in the conversation by comparing the correct conversation with the assembled conversation from top to bottom;
- Discard all conversation pieces below the player's first mistaken conversation piece;
- Insert the text that has been previously prepared as a reaction to the mistaken conversation piece. This text will show up after the mistaken conversation piece;
- Insert the text that has been previously prepared as a clue for the correct conversation piece that would fit in the position the player made his first mistake. This text will appear after the text of the previous step.

In Fig. 3, this would be the inappropriate conversation case. This new generated conversation is then shown to the player. After the generated conversation ends, the player will go back to the conversation construction screen. This process can be better understood in Fig. 4.

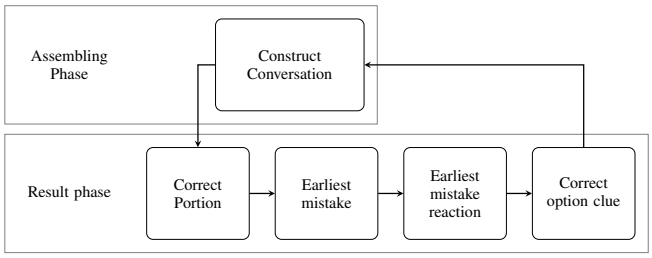


Fig. 4. Flow chart for mistakes during the conversation construction activity.

In the "First mistake reaction", when the conversation goes in to an unexpected flow, the user feedback begins, where users can acquire information on why the card related to the "first mistake" is unappropriated and insight in to what conversation piece would be appropriate for that particular slot. Users who are reading attentively will also be able to clearly point out which conversation piece has been considered inappropriate, since the feedback (the change in the conversation flow) begins at that moment. This feedback is effective because it uses the player's inappropriate input to generate a conversation, similarly to an error-based simulation (such as the one used in the work of [61]). Instead of simply stating "this conversation is wrong, the correct one is this one", it allows players to reflect on their input in a more effective way. This refers to the "Read feedback" and "Extract new information" nodes in Fig. 3.

6) "Look for Clues" Functionality: In Section III-A5, it has been stated that feedback starts at the "First mistake reaction". To further enhance this fact to the player and to support him in relating this feedback to the "First mistake", a screen was added, called the "Look for Clues" message, which will appear right before the "First mistake". The message says the following "This conversation will not go as expected! Read it to find clues!!!". This happens in the middle of the conversation and it tells the user that:

- The created conversation has a problem.
- Until that point the conversation did not have a problem.
- There is something wrong with the conversation piece that appears right before that message.
- Looking for clues in whatever is coming up next is what the game expects them to do.

This was designed to further induce the ideal behavior in Fig. 3.

7) Game Design Elements: This session will describe the game design elements that have been incorporated into the design of the game. Their effects and importance will also be discussed. References for these elements can be seen in Section I.

Challenge and Freedom to Fail: Our approach for challenge has been through natural, emergent difficulty. As we've previously shown, reading comprehension for L2 learners can be a fairly difficult task. On extensive reading there is a focus on choosing texts with appropriate difficulty to mitigate this difficulty. The conversation construction task involves extracting information from the text and using that information. As such, it should have a difficulty similar to the reading comprehension process. The difference is that feedback is provided. In our feedback loop, progress will make it simpler for him to solve the activity. This way, every time the user tries to solve the task, he should have more information and the task should become easier.

About freedom to fail, the user is free to fail in our design. Furthermore, he is rewarded with feedback from his failure.

Visual Novel: The game is very similar to a visual novel and could be classified as such. This was not an arbitrary design decision. As discussed before, visual novels have a number of elements that make them an effective reading application.

User Interaction: In our conversation construction activities, drag-and-drop is the main form of interaction used. In [62], drag-and-drop is encouraged and described as an intuitive way to move content through the system. Our conversation construction activity has been designed with this in mind for its intuitiveness and for providing a fast way to construct the conversation. This approach has been used in applications like Monsakun to achieve similar effects and they have been well received [63].

8) Addressing the Design Challenges: Students may ignore passages which contain unknown words: A user that displays this behavior is not performing according to the ideal behavior displayed in 3. If the user ignores a passage, he would have trouble building the conversation. Because of this, the chances of the user making a mistake would rise. Upon making a mistake, the user would then be presented with feedback. At

that moment, the "look for clues" functionality described above further points the user to reading the feedback.

Furthermore, the chances of the user solving the activity by luck is 5%, given the default setup of five conversation pieces and two empty slots. It is low enough to make reading the feedback a more suitable strategy than trying to make the correct conversation by luck.

Students may infer the wrong meaning of a word: This would also imply in students making a mistake in the conversation construction activity. The expectancy is that the user will be able to correct his misunderstanding from reading the feedback.

Activities would benefit from being intrinsically motivating: The designing focuses on intrinsic motivation by balancing challenge, offering freedom to fail and by using drag-and-drop for ease of use. As discussed before, these are the elements related to the intrinsic motivation in games.

Game elements should not detract from reading or from inferencing: As seen before, the game has two types of segments, story segments and conversation construction segments. Both segments include reading. There are no actions to be done in-game that don't involve reading in some way. Conversation construction segments would result in multiple readings of the conversation pieces. Students are also expected to be carefully reading the feedback. As such, instead of detracting, the design has a focus on improving inferencing.

The game needs be challenging but not too challenging: One challenge of L2 reading is matching the difficulty of the text to the skills of the user. As such, difficulties in reading are highly content based. The trial-and-error with feedback design mitigates this issue by making the activity progressively easier as the user keeps on reading the feedback. This has been further explored in the subsection Game Design Elements.

The activity should allow for students to fail as much as they need in order to progress: This is a natural part of the trial-and-error design. Further details above in the subsection Game Design Elements.

B. Game Development

The game has been developed using the C# language and the game engine Unity [64]. A scripting language was created to describe the story sequences and the conversation construction activities. Since the story sequences are similar to visual novels, the commands used are similar in format to the ones found in Ren'py, a visual novel engine [65]. A script interpreter was written to render the scene for the players, while also handling input.

C. Design of the Experiment

This study used a within-subject design with two conditions for counterbalancing: text-game and game-text. Afterwards, participants took a post-test and some of the users took a user perception survey. Text-game read a text and then played the game. Game-text played the game and then read the text. All measurements were done in the end so that measuring would not affect the behavior of the users. This flow can be seen in Fig. 5. The post-test had three sections:

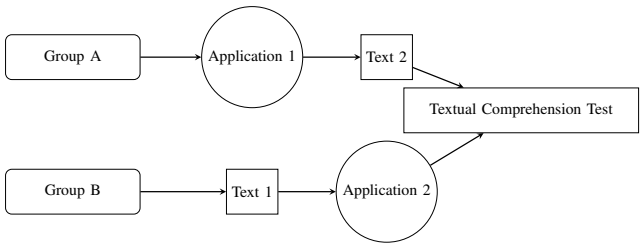


Fig. 5. Experiment flow diagram

remembering section, textual comprehension section and word comprehension questions. Both game and text included dummy words to create a situation where users are reading a material with unknown words.

13 Japanese University students participated in the study and were randomly assigned to each condition.

Two textual contents were used in this study, A and B. Both contents have a game form and a textual form. Thus, we have game A, game B, text A and text B. The text-game group used text A and game B. The game-text group used game A and text B. Content A and B were found to be appropriate or below the difficulty of Grade 2 in the Common Core State Standards [66]. As such, both contents are considered to be accessible and equivalent in difficulty. This was measured using the TextEvaluator tool [67]. Scores given by the tool were found to have high correlation with judgment presented by human experts [68]. Content A has 164 words while content B has 226 words.

The post-test was divided into three sections:

- Remembering section: users were asked to write as much as they could remember with as much detail as possible.
- Textual interpretation section: users were asked questions such as "Did Brian ever get angry in the story? If yes, why did he get angry?"
- Word comprehension section: users were asked to explain the meaning of the dummy words and to translate phrases that used the dummy words.

The textual interpretation questions were designed around the passages that contained dummy words. This means that inferencing information in the presence of unknown words is necessary to correctly answer the questions. Only the last two sections are used to calculate the scores. The first section was included for the possibility of an exploratory analysis, but it is not addressed in this study.

The user perception survey had four questions. Three of them comparing text and game on ease of content understanding, motivation to read and on suitability for studying English. The last question was about the usability of the game.

The experiment was performed in participants individually. They interacted with the game on a computer. The text was read through a PDF file. While interacting with the game, participants were taught that clicking would advance the story. They were also taught how to drag-and-drop to build the conversations. The post-test and the user perception survey were both administered through an online form.

TABLE I. AVERAGE SCORES AND STANDARD DEVIATION FOR THE TWO CONDITIONS AND FOR ALL PARTICIPANTS

	Application M (SD)	Text M (SD)	N
Game-text	0.76 (0.09)	0.58(0.24)	7
Text-game	0.77 (0.18)	0.44(0.24)	6
All participants	0.77 (0.13)	0.52(0.24)	13

TABLE II. SURVEY RESULTS. QUESTION ONE TO QUESTION THREE IS ABOUT COMPARING GAME AND TEXT. QUESTION FOUR IS ABOUT USABILITY

Which one is easier to understand?	
The game is better	28.6%
The game is a bit better	28.6%
They are the same	42.9%
The text is a bit better	0.00%
The text is better	0.00%
Which one makes you want to read it more?	
The game is better	57.1%
The game is a bit better	28.6%
They are the same	0.00%
The text is a bit better	0.00%
The text is better	14.3%
Which one do you think is better for studying English?	
The game is better	14.3%
The game is a bit better	42.9%
They are the same	42.9%
The text is a bit better	0.00%
The text is better	0.0%
Was the application easy to use?	
Easy to use	71.4%
A little bit easy to use	14.3%
Not easy, not hard	0.0%
A little bit hard to use	14.3%
Hard to use	0.0%

IV. RESULTS AND DISCUSSION

Table I shows scores obtained by the two conditions and for all participants. Scores have a minimum value of 0 and a maximum value of 1. Game scores had a lower standard deviation than text scores.

A two-way analysis of variance was conducted on the influence of two independent variables (medium, order of use) on the post test scores. Medium includes two levels (game, text) and order of use consisted of two levels (first, second). The only significant effect at the .05 significance level was for the medium factor. The main effect for medium yielded an F ratio of $F(1, 22) = 11.16$, $p < .01$, indicating a significant difference between using the game ($M = 0.77$, $SD = 0.13$) and reading the text ($M = 0.52$, $SD = 0.24$). The main effect for order yielded an F ratio of $F(1, 22) = 1.06$, $p > .05$, indicating that the effect for order was not significant, first ($M = 0.61$, $SD = 0.24$) and second ($M = 0.67$, $SD = 0.23$). The interaction effect was not significant, $F(1, 22) = 0.86$, $p > .05$.

The above results suggest that using the game results in more information absorbed than using the text. It also suggests that order of use (which one is used first, and which one is used second) does not affect the amount of information absorbed.

As for the user perception survey results, found in Table II, the following trends were found:

- In the area of interest, all users except for one had a positive opinion towards the game, with over half of the users completely favoring the game.
- On perceived comprehensibility and perceived learning, half of the users had a positive opinion while the other half had a neutral opinion.
- On usability, one user found the game a little bit hard to use, while the clear majority thought the game was easy to use.
- The user who felt the game is a little bit hard to use is the only one user that was unfavorable towards the game in any of the areas. He also favored printed text in the area of interest.

Those trends show that the hypothesis was true. About the one user that was unfavorable towards the game, his scores were checked in order to see if his unfavourability affected his scores. Surprisingly, he was the only user to get a perfect grade related to the content in the game version he used, suggesting that the comprehensibility scores are not affected by dislike of the game. These results fit well with past findings suggesting good affective reception from learners in relation to DBGL, as reported in [69] and in other works ([55], [70], etc.).

Users higher comprehensibility when using the game can be attributed to being able to read the feedback information to solve the conversation construction problems. This suggests that users were performing according to the ideal behavior previously defined, indicating that our efforts to create an activity that can only be practically solved by displaying the needed behavior have been successful. When using the text, users may have been more likely to ignore passages or to make mistakes during inferencing. This gain in performance is reflected not only in reading comprehension but also in incidental vocabulary learning, since the experiment included dummy words. Thus, results suggest that users are able to infer partial meaning of the words better when using the game.

V. CONCLUSION

Results suggest that users are able to infer information from context better by using the game. This implies that activity designs based on creating a trial-and-error task with automatic feedback can be useful for improving reading comprehension and improving incidental vocabulary learning. Qualitative results have also been positive.

As for the perception of the game as an English studying tool in comparison to the paper version, around half of the users pointed to them being equally effective. And yet the comprehension scores for the game version have been much higher. This contradiction between user's perceived learning effectiveness and the actual effectiveness has also been reported by [59]. Low perceived learning is also one of the challenges of extensive reading, so making DBGL tools have a higher perceived learning by students should positively impact their performance and studies in that direction are necessary, such as measuring differences in flow and motivation between incidental and explicit learning.

Remaining issues would be performing additional experiments to show more compelling evidence of the increases in reading comprehension, since the number of participants was small. Also, the low perceived learning and the fact that the design relies on the presence of conversations are also limitations. Expansions to this research could focus on making learning more explicit by mixing the narratives with explicit vocabulary teaching, thus making the learning process more

obvious to the student. Another problem is that, currently, producing content for the game is a complex task. Creating a tool to assist this process would allow content to be created by teachers and other content creators.

Another possible next step is adapting the design to be generated based on natural language processing techniques without human input. This would allow for a large amount of game content to be created. This would have implications for improving the performance of extensive reading programs.

REFERENCES

[1] G. Brown and G. Yule, *Discourse analysis*. Cambridge university press, 1983.

[2] W. Grabe, *Reading in a second language: Moving from theory to practice*. Ernst Klett Sprachen, 2009.

[3] W. P. Grabe and F. L. Stoller, *Teaching and researching: Reading*. Routledge, 2013.

[4] W. E. Nagy and P. A. Herman, "Incidental vs. instructional approaches to increasing reading vocabulary." *Educational perspectives*, vol. 23, no. 1, pp. 16–21, 1985.

[5] W. E. Nagy and R. C. Anderson, "How many words are there in printed school english?" *Reading research quarterly*, pp. 304–330, 1984.

[6] W. Grabe and F. Stoller, "Reading and vocabulary development in a second language: A case study," *Second language vocabulary acquisition: A rationale for pedagogy*, pp. 98–122, 1997.

[7] M. Horst, "Learning l2 vocabulary through extensive reading: A measurement study," *Canadian Modern Language Review*, vol. 61, no. 3, pp. 355–382, 2005.

[8] M. Pigada and N. Schmitt, "Vocabulary acquisition from extensive reading: A case study," *Reading in a foreign language*, vol. 18, no. 1, p. 1, 2006.

[9] B. Dupuy and S. D. Krashen, "Incidental vocabulary acquisition in french as a foreign language." *Applied Language Learning*, vol. 4, pp. 55–63, 1993.

[10] J. H. Hulstijn, "Retention of inferred and given word meanings: Experiments in incidental vocabulary learning," in *Vocabulary and applied linguistics*. Springer, 1992, pp. 113–125.

[11] R. Schmidt, "Deconstructing consciousness in search of useful definitions for applied linguistics," *Consciousness in second language learning*, vol. 11, pp. 237–326, 1994.

[12] J. H. Hulstijn, M. Hollander, and T. Greidanus, "Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words," *The modern language journal*, vol. 80, no. 3, pp. 327–339, 1996.

[13] B. Laufer and D. D. Sim, "Measuring and explaining the reading threshold needed for english for academic purposes texts," *Foreign language annals*, vol. 18, no. 5, pp. 405–411, 1985.

[14] J. H. Hulstijn, "When do foreign-language readers look up the meaning of unfamiliar words? the influence of task and learner variables," *The Modern Language Journal*, vol. 77, no. 2, pp. 139–147, 1993.

[15] F. R. Jones, "Learning an alien lexicon: A teach-yourself case study," *Second Language Research*, vol. 11, no. 2, pp. 95–111, 1995.

[16] M. Hill and B. Laufer, "Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition," *International Review of Applied Linguistics*, vol. 41, no. 2, pp. 87–106, 2003.

[17] M. Horst, T. Cobb, T. Cobb, and P. Meara, "Beyond a clockwork orange: Acquiring second language vocabulary through reading." *Reading in a foreign language*, vol. 11, no. 2, pp. 207–223, 1998.

[18] S. Knight, "Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities," *The Modern Language Journal*, vol. 78, no. 3, pp. 285–299, 1994.

[19] Y. Kim and K. McDonough, "The effect of interlocutor proficiency on the collaborative dialogue between korean as a second language learners," *Language Teaching Research*, vol. 12, no. 2, pp. 211–234, 2008.

[20] S. Huang, V. Willson, and Z. Eslami, "The effects of task involvement load on l2 incidental vocabulary learning: A meta-analytic study," *The Modern Language Journal*, vol. 96, no. 4, pp. 544–557, 2012.

[21] S. Krashen, "Free voluntary reading: New research, applications, and controversies," *Anthology series-Seameo regional language centre*, vol. 46, no. 1, 2005.

[22] R. C. Anderson, P. T. Wilson, and L. G. Fielding, "Growth in reading and how children spend their time outside of school," *Reading research quarterly*, pp. 285–303, 1988.

[23] C. Davis, "Extensive reading: an expensive extravagance?" *ELT journal*, vol. 49, no. 4, pp. 329–336, 1995.

[24] S. Logan, E. Medford, and N. Hughes, "The importance of intrinsic motivation for high and low ability readers' reading comprehension performance," *Learning and Individual Differences*, vol. 21, no. 1, pp. 124–128, 2011.

[25] L. Baker and A. Wigfield, "Dimensions of children's motivation for reading and their relations to reading activity and reading achievement," *Reading research quarterly*, vol. 34, no. 4, pp. 452–477, 1999.

[26] P. L. Morgan and D. Fuchs, "Is there a bidirectional relationship between children's reading skills and reading motivation?" *Exceptional children*, vol. 73, no. 2, pp. 165–183, 2007.

[27] A. Wigfield and J. T. Guthrie, "Relations of children's motivation for reading to the amount and breadth or their reading." *Journal of educational psychology*, vol. 89, no. 3, p. 420, 1997.

[28] J. T. Guthrie, A. Wigfield, P. Barbosa, K. C. Perencevich, A. Taboada, M. H. Davis, N. T. Scafiddi, and S. Tonks, "Increasing reading comprehension and engagement through concept-oriented reading instruction." *Journal of educational psychology*, vol. 96, no. 3, p. 403, 2004.

[29] A.-M. Masgoret and R. C. Gardner, "Attitudes, motivation, and second language learning: a meta–analysis of studies conducted by gardner and associates," *Language learning*, vol. 53, no. 1, pp. 123–163, 2003.

[30] S. Tang, M. Hanneghan, and A. El Rhalibi, "Introduction to games-based learning," *Games Based Learning Advancements for Multi-Sensory Human Computer Interfaces. New York: IGI Global*, 2009.

[31] T. Susi, M. Johannesson, and P. Backlund, "Serious games: An overview," 2007.

[32] S. Deterding, "The lens of intrinsic skill atoms: A method for gameful design," *Human–Computer Interaction*, vol. 30, no. 3-4, pp. 294–335, 2015.

[33] N. D. Bowman, "A pat on the back: Media flow theory revis (it) ed." *Rocky Mountain Communication Review*, vol. 4, no. 1, 2008.

[34] E. L. Deci and R. M. Ryan, "Motivation, personality, and development within embedded social contexts: An overview of self-determination theory," *The Oxford handbook of human motivation*, pp. 85–107, 2012.

[35] A. K. Przybylski, C. S. Rigby, and R. M. Ryan, "A motivational model of video game engagement." *Review of general psychology*, vol. 14, no. 2, p. 154, 2010.

[36] J. Juul and M. Norton, "Easy to use and incredibly difficult: on the mythical border between interface and gameplay," in *Proceedings of the 4th international conference on foundations of digital Games*. ACM, 2009, pp. 107–112.

[37] K. Kuwada, "The effect of interactivity with a music video game on second language vocabulary recall," *About Language Learning & Technology*, vol. 74, 2010.

[38] D. Cavallaro, *Anime and the visual novel: narrative structure, design and play at the crossroads of animation and computer games*. McFarland, 2009.

[39] J. Liu, "Effects of comic strips on l2 learners' reading comprehension," *TESOL quarterly*, vol. 38, no. 2, pp. 225–243, 2004.

[40] S. Chunsoft, "Danganronpa: Trigger happy havoc," UMD, 2010.

[41] H. Hancock, "Better game design through cutscenes," *Gamasutra.[Online]. Available: http://www. gamasutra. com/features/20020401/hancock*, vol. 1, 2002.

[42] P. Furtado, T. Hirashima, and Y. Hayashi, "Transforming foreign language narratives into interactive reading applications designed for comprehensibility and interest," in *International Conference on Artificial Intelligence in Education*. Springer, 2017, pp. 510–513.

[43] P. G. F. Furtado, T. Hirashima, and Y. Hayashi, "Development and experimental evaluation of an interactive reading application designed for comprehensibility and interest," *Workshop Proc. of ICCE2017*, vol. 25, 2017.

[44] M. E. C. Santos, T. Taketomi, G. Yamamoto, M. M. T. Rodrigo, C. Sandor, H. Kato *et al.*, "Augmented reality as multimedia: the case for situated vocabulary learning," *Research and Practice in Technology Enhanced Learning*, vol. 11, no. 1, p. 1, 2016.

[45] I. D. Agusalim, "Developing visual novel game of english conversation for dep eepis." *Journal of Education and Practice*, vol. 6, no. 33, pp. 113–124, 2015.

[46] H.-J. H. Chen and T.-Y. C. Yang, "The impact of adventure video games on foreign language learning and the perceptions of learners," *Interactive Learning Environments*, vol. 21, no. 2, pp. 129–141, 2013.

[47] Y. A. Rankin and M. W. Shute, "Re-purposing a recreational video game as a serious game for second language acquisition," *Serious game design and development: Technologies for training and learning*, pp. 178–195, 2010.

[48] P. Sundqvist and L. K. Sylvén, "World of voccraft: Computer games and swedish learners l2 english vocabulary," in *Digital games in language learning and teaching*. Springer, 2012, pp. 189–208.

[49] A. Yudintseva, "Synthesis of research on video games for the four second language skills and vocabulary practice," *Open Journal of Social Sciences*, vol. 3, no. 11, p. 81, 2015.

[50] S. Vahdat and A. R. Behbahani, "The effect of video games on iranian efl learners vocabulary learning," *Reading*, vol. 13, no. 1, 2013.

[51] C. I. Hitosugi, M. Schmidt, and K. Hayashi, "Digital game-based learning (dgbl) in the l2 classroom: The impact of the un's off-the-shelf videogame, food force, on learner affect and vocabulary retention," *CALICO Journal*, vol. 31, no. 1, p. 19, 2014.

[52] L. Aghlara and N. H. Tamjid, "The effect of digital games on iranian children's vocabulary retention in foreign language acquisition," *Procedia-Social and Behavioral Sciences*, vol. 29, pp. 552–560, 2011.

[53] S. Suh, S. W. Kim, and N. J. Kim, "Effectiveness of mmorpg-based instruction in elementary english education in korea," *Journal of Computer Assisted Learning*, vol. 26, no. 5, pp. 370–378, 2010.

[54] H. Tobar-Muñoz, S. Baldiris, and R. Fabregat, "Augmented reality game-based learning: Enriching students experience during reading comprehension activities," *Journal of Educational Computing Research*, vol. 55, no. 7, pp. 901–936, 2017.

[55] R. Rosas, M. Nussbaum, P. Cumsille, V. Marianov, M. Correa, P. Flores, V. Grau, F. Lagos, X. López, V. López *et al.*, "Beyond nintendo: design and assessment of educational video games for first and second grade students," *Computers & Education*, vol. 40, no. 1, pp. 71–94, 2003.

[56] M. Ven, L. Leeuw, M. Weerdenburg, and E. Steenbeek-Planting, "Early reading intervention by means of a multicomponent reading game," *Journal of Computer Assisted Learning*, vol. 33, no. 4, pp. 320–333, 2017.

[57] D. Hooshyar, M. Yousefi, and H. Lim, "A procedural content generation-based framework for educational games: Toward a tailored data-driven game for developing early english reading skills," *Journal of Educational Computing Research*, p. 0735633117706909, 2017.

[58] L. S. Gaytán-Lugo and S. C. Hernandez-Gallardo, "Towards improving reading comprehension skills in third graders with a serious game," *ICCE 2012*, p. 25, 2012.

[59] B. E. Shelton, D. Neville, and B. McInnis, "Cybertext redux: using interactive fiction to teach german vocabulary, reading, and culture," in *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 3*. International Society of the Learning Sciences, 2008, pp. 128–129.

[60] C. Hocking, "Ludonarrative dissonance in bioshock: The problem of what the game is about," in *Well Played 1.0*. ETC Press, 2009, pp. 114–117.

[61] T. Horiguchi, I. Imai, T. Toumoto, and T. Hirashima, "Error-based simulation for error-awareness in learning mechanics: An evaluation." *Educational Technology & Society*, vol. 17, no. 3, pp. 1–13, 2014.

[62] I. Apple. (2017) Drag and drop - user interaction - human interface guidelines for macos apps. Apple Inc. [Online]. Available: https://developer.apple.com/macos/human-interface-guidelines/user-interaction/drag-and-drop/

[63] T. Hirashima, T. Yokoyama, M. Okamoto, and A. Takeuchi, "Long-term use of learning environment for problem-posing in arithmetical word problems," in *Proceedings of ICCE*, vol. 2008, 2008, pp. 817–824.

[64] U. G. Engine, "Unity game engine-official site," *Online][Cited: February 17, 2018.] http://unity3d. com*, 2018.

[65] T. Rothamel, "The ren'py visual novel engine," https://www.renpy.org/, 2006, accessed: 2017-02-28.

[66] C. C. S. S. Initiative *et al.*, *Common Core State Standards for English Language Arts & Literacy In History/Social Studies, Science, and Technical Subjects*. Washington, DC: Council of Chief State School Officers & National Governors Association., 2010.

[67] K. M. Sheehan, I. Kostin, D. Napolitano, and M. Flor, "The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment," *The Elementary School Journal*, vol. 115, no. 2, pp. 184–209, 2014.

[68] K. M. Sheehan, "A review of evidence presented in support of three key claims in the validity argument for the textevaluator® text analysis tool," *ETS Research Report Series*, vol. 2016, no. 1, pp. 1–15, 2016.

[69] T. Hainey, T. M. Connolly, E. A. Boyle, A. Wilson, and A. Razak, "A systematic literature review of games-based learning empirical evidence in primary education," *Computers & Education*, vol. 102, pp. 202–223, 2016.

[70] E. Ojanen, M. Ronimus, T. Ahonen, T. Chansa-Kabali, P. February, J. Jere-Folotiya, K.-P. Kauppinen, R. Ketonen, D. Ngorosho, M. Pitkänen *et al.*, "Graphogame–a catalyst for multi-level promotion of literacy in diverse contexts," *Frontiers in psychology*, vol. 6, 2015.

Real Time Computation for Robotic Arm Motion upon a Linear or Circular Trajectory

Liliana Marilena Matica, Cornelia Győrödi, Helga Maria Silaghi, Simona Veronica Abrudan Cacioara
Faculty of Electrical Engineering and Information Technology, University of Oradea
Oradea, Romania

***Abstract* —The computation method proposed in this paper, named ADNIA (Analysis Differential Numeric Interpolate Algorithms), computes waypoints Cartesian coordinates for TCP (tool centre point) of a robotic arm, for a motion on an linear or circular imposed trajectories. At every sampling period of time, considering real-time software implementation of ADNIA, the location matrix of a robotic arm is computed. This computation method works with a well-defined value of motion speed; it results a maximum computation precision (for those motions).**

Keywords—Waypoints; location matrix; position vector; orientation of a robotic arm; orientation versors; Analysis Difference Numeric Interpolate Algorithm (ADNIA); linear or circular ADNIA

I. INTRODUCTION

The paper describes the real-time computation method for robotic arm motions, upon imposed trajectories, named ADNIA (**A**nalysis **D**ifferential **N**umeric **I**nterpolate **A**lgorithms).

Some considerations regarding the importance of industrial robots, in manufacturing process, are described in this first paragraph.

Industrial production (manufacturing) is a complex process, the decisions result of combination and use of involved factors. About manufacturing process, the important factors are: environmental and natural resources, scientific and technological resources, quality of management. The multitudes of factors that determine industrial production are identified by the so called socio-political environment generously given. About manufacturing, an important method of qualitative growth is manufacturing automation. Automation, not only refers to the development of the means of production, but also to its integrative aspects, in relation of all system levels of industrial structure, to the human factor.

Many automation facilities have as purpose the replacement of human activity, in the manufacturing process; the mainly functions of them is to help about the automatization of the machining tasks, concerning the technologic processes; tasks requiring the intervention of the hand human operator, under the supervision of the eye (the whole action is coordinated by the human brain). Possibly, such operations are: start-stop equipment, loading and unloading, assembly-disassembly work items (pieces, parts), change or handling tools, testing, processing, inspection, comparison, repair or maintenance.

Regarding industrial production (manufacturing) systems, it is needful to emphasize the means of work; one particularly powerful is: industrial robots [1]-[8]. About industrial production efficiencies and necessarily required organizational activities: leadership, management, correlation in time and space, the industrial robots are very useful, in order to achieve objectives. So, the concept of interest allocation (as a function of effective industrial production plan and management) is implemented by industrial robots, in the manufacturing process. Among the important objectives of manufacturing (industrial production) systems, (regarding the concept of allocation), it is the growth adaptability of those systems to the changing social demand or to the environment, while reducing resource consumption. An important role, particularly in this respect, has the industrial robots [1], [5].

In purpose to define a location and command a motion of an industrial robot, more exactly, of a robotic arm (a specific industrial robot that is similar with human arm), it must be defined the location matrix [2], [4]:

$$G = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

The location matrix contains the axles components of orientation versors: $\vec{n};\vec{o};\vec{a}$, (three vectors with module value equal with 1 value, $|\vec{n}|=1,|\vec{o}|=1,|\vec{a}|=1$) and the position vector: $\vec{p}$, Fig. 1, (definition: a versor is a vector having module value equal with 1 value).

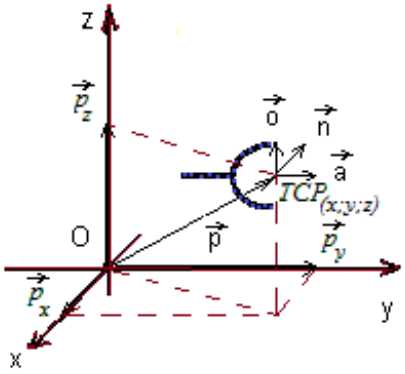


Fig. 1. Orientation versors and position vector.

The motion of a robotic arm may be imposed, upon a linear or a circular trajectory; an imposed trajectory of motion is necessary during an industrial process executed by the robotic arm, as painting or welding industrial process. The imposed trajectory for a robotic arm may be defined by computation of every location matrix for every waypoint, upon the trajectory. This computation, working with *ADNIA* algorithms, will be described in this paper. Those algorithms were implemented about metal cutting industrial process, on CNC machines. [7] Those algorithms were not been described about robotic arms motion; the paper describes the adaptation of this algorithm for robotic arm motion.

The paper is structured as follows: Section 2 presents the computation method about robotic arm motion, on a linear or circular imposed trajectory and Section 3 describes the computations about a variable orientation of a robotic arm, during the motion on a linear or circular imposed trajectory. Based on the results of the tests performed, several conclusions are presented in the last section.

II. ADNIA for Linear or Circular Trajectory of Robotic Arms

This paragraph describes the computation method about robotic arm motion, on a linear or circular imposed trajectory; the robotic arm has a constant orientation, during the motion.

About a robotic arm, the motion trajectory of TCP (***T**ool **C**entre **P**oint*) is defined by variation of $\vec{p}$ vector, Fig. 1. The axle components of $\vec{p}$ vector: $\vec{p}_x; \vec{p}_y; \vec{p}_z$ defines the Cartesian coordinates of TCP. ADNIA algorithm computes the (indexed *k*) Cartesian coordinates of waypoints, $P_k(x_k;y_k;z_k)$ upon the linear trajectory. The input data of linear ADNIA algorithm are: Cartesian coordinates of first point (index *0*) and end point (index *F*) of the linear trajectory and the motion speed: *v*.

The ADNIA computation starts with the number of sampling periods of time, necessarily for traverse the linear trajectory, named *N* [2], [4] (the value of sampling periods of time is Δt):

$$N = round\left(\frac{\sqrt{(x_F - x_O)^2 + (y_F - y_O)^2 + (z_F - z_O)^2}}{v \cdot \Delta t}\right) \qquad (2)$$

The *N* value must be an integer value; about (2), the rounding computation find the next integer value of the value computed (for example, if the result value is 22.1 the round value must be 23).

Then, the linear ADNIA computes the values of linear space steps, necessary to be performed at every sampling period of time, for each axle (executed motion at every sampling period of time), named axle steps: $\delta_x, \delta_y, \delta_z$. The interpolating process of the linear trajectory run and computes the waypoints, $P_k(x_k;y_k;z_k)$, considering [2]:

$$x_k = x_O + k \cdot \delta_x$$

$$y_k = y_O + k \cdot \delta_y \qquad (3)$$

$$z_k = z_O + k \cdot \delta_z$$

Let consider the interpolation process of a linear trajectory; defined by Table I:

TABLE I. VALUES OF CARTESIAN COORDINATES

Step No. (index)	**Cartesian coordinate of intermediary points**		
	X_I	Y_I	Z_I
0 (start ADNIA)	10	20	30
1	10.1	22	30.03
2	10.2	24	30.06
3	10.3	26	30.09
4	10.4	28	30.12
5	10.5	30	30.15
6	10.6	32	30.18
7	*10.7*	*34*	*30.21*
8	10.8	36	30.24
9	10.9	38	30.27
10 (stop ADNIA)	11	40	30,3

About this interpolation process, the axle steps have the values: $\delta_x = 0.1$; $\delta_y = 2$; $\delta_z = 0.03$; and the number of necessarily steps for traverse the linear is: *N =10,* (for every axle).

Let consider this constant orientation of the robotic arm, during the motion on the linear trajectory: $\vec{n}_x = 1$; $\vec{o}_y = 1$; $\vec{a}_z = 1$; the location matrix about a waypoints upon the trajectory, named G_k , have the axle components values of position vector as listed in Table I. For example, on the seventh step upon the linear trajectory, the location matrix is:

$$G_k = G_7 = \begin{bmatrix} 1 & 0 & 0 & x_{k=7} \\ 0 & 1 & 0 & y_{k=7} \\ 0 & 0 & 1 & z_{k=7} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 10.7 \\ 0 & 1 & 0 & 34 \\ 0 & 0 & 1 & 30.21 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (4)$$

The big advantage of ADNIA algorithms is the constant value of motion speed.

Next considerations explain the situation of a variable orientation of robotic arm, (the ADNIA algorithms, explained in this paper, work about a linear or circular trajectory, with a constant or variable orientation of the robotic arm).

The linear ADNIA (previously explained), works with Cartesian coordinates. The circular ADNIA works with spherical coordinates: (R, φ, ϕ) , Fig. 2 [7].

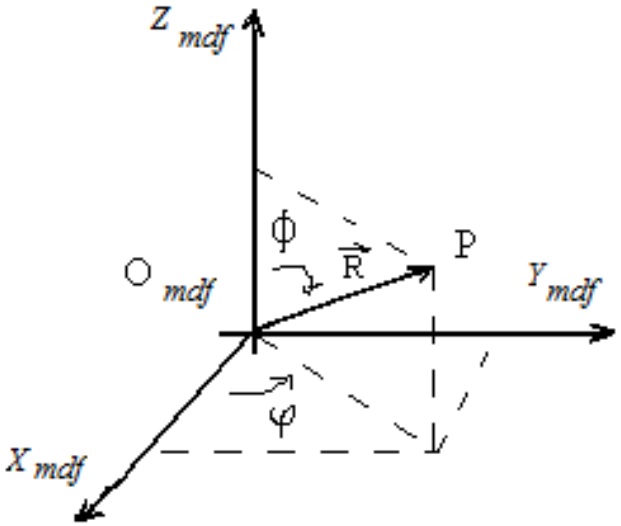


Fig. 2. The spherical coordinates.

In Fig. 2, the Cartesian coordinates system, named $OXYZ_{mdf}$; is properly chosen, in purpose to have a constant value of R, for spherical coordinates of waypoints upon the circular trajectory.

Every waypoint upon the circular trajectory is defined by its spherical coordinates: $P_m(R,\varphi_m,\phi_m)$. Considering the polar angle (indexed *m)* : φ_m and azimuthally angle (indexed *m)* : ϕ_m , the Cartesian coordinates of waypoints, indexed *m* , (the point that must be reach after *m* angle steps) are computed with relations [2]:

$$\varphi_m = \varphi_O + m\cdot\delta_\varphi = \varphi_O + m\cdot\left(\frac{\varphi_F - \varphi_O}{N}\right) \quad (5)$$

$$\phi_m = \phi_O + m\cdot\delta_\phi = \phi_O + m\cdot\left(\frac{\phi_F - \phi_O}{N}\right) \quad (6)$$

The notations have the similar meanings as previously explained: index *0* point upon the first point of the circular trajectory, $P_0(R,\varphi_0,\phi_0)$; index *F* point upon the last point of the circular trajectory, $P_F(R,\varphi_F,\phi_F)$; *N* is the number of angle steps, necessary for traverse the circular trajectory, its value may be computed from the speed value of motion. For example, considering: $N=15$; $\varphi_O = 0^0, \phi_O = 92^0$; $\varphi_F = 90^0, \phi_F = 47^0$; $R=22$, so the spherical coordinates of initial and final point are: $P_0(22;\ 0^0;\ 92^0)$; $P_F(12;\ 90^0;\ 47^0)$; it results the value of angle steps that must be performed, on every sampling time period:

$$\delta_\varphi = \frac{90^0 - 0^0}{15} = 6^0$$
$$\delta_\phi = \frac{47^0 - 92^0}{15} = -3^0 \quad (7)$$

The Cartesian coordinates of waypoints: $P_m(x_m, y_m, z_m)$, upon the circular trajectory, may be computed:

$$\begin{aligned} x_m &= R\cdot\sin(\phi_0 + m\cdot\delta_\phi)\cdot\cos(\varphi_0 + m\cdot\delta_\varphi) = \\ &= 22\cdot\sin(90^0 - m\cdot 3^0)\cdot\cos(0^0 + m\cdot 6^0) \\ y_m &= R\cdot\sin(\phi_0 + \cdot\delta_\phi)\cdot\sin(\varphi_0 + m\cdot\delta_\varphi) = \\ &= 22\cdot\sin(90^0 - m\cdot 3^0)\cdot\sin(0^0 + m\cdot 9^0) \\ z_m &= R\cdot\cos(\phi_0 + m\cdot\delta_\phi) = 22\cdot\cos(90^0 - m\cdot 3^0) \end{aligned} \quad (8)$$

For example, after three steps, the Cartesian coordinates of waypoint $P_3(x_3, y_3, z_3)$ are:

$$\begin{aligned} x_3 &= 22\cdot\sin(90^0 - 3\cdot 3^0)\cdot\cos(0^0 + 3\cdot 6^0) \\ y_3 &= 22\cdot\sin(81^0)\cdot\sin(18^0) \\ z_3 &= 22\cdot\cos(81^0) \end{aligned} \quad (9)$$

Because $90^0 - 3\cdot 3^0 = 81^0; 0^0 + 3\cdot 6^0 = 18^0$. About robotic arm, after three steps upon the circular trajectory, the location matrix is (considering the defined constant orientation of the robotic arm):

$$G_3 = \begin{bmatrix} 1 & 0 & 0 & 22\cdot\sin(81^0)\cdot\cos(18^0) \\ 0 & 1 & 0 & 22\cdot\sin(81^0)\cdot\sin(18^0) \\ 0 & 0 & 1 & 22\cdot\cos(81^0) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

The computation method was implemented about motion command for a welding process, in a mining equipment manufacturing company. Computation precision was: $1\cdot 10^{-5}$ and imposed precision was: $1\cdot 10^{-2}$. The sampling period of time was 18 *ms* ($18\cdot 10^{-3}$ seconds). Welding execution precision was the one imposed by the beneficiary.

The execution precision is limited by the motion precision of the mechanical part of the equipment. Another limitation of the motion command process was the chosen place for mounting the measuring transducer, it was on axis of electrical engines (it was not on the TCP); possible malfunctions of the motion transmission chain can't be detected.

The software implementation worked with Assembly language for Intel microprocessor. The challenge for the future is to implement the same computation with Assembly language for Intel microcontroller.

III. ADNIA FOR VARIABLE ORIENTATION OF ROBOTIC ARM DURING THE MOTION

This paragraph describes the computations about a variable orientation of a robotic arm, during the motion on a linear or circular imposed trajectory.

Let considers the versors: $\vec{i}, \vec{j}, \vec{k}$; those versors define the three axles: OX, OY and OZ, Fig. 3:

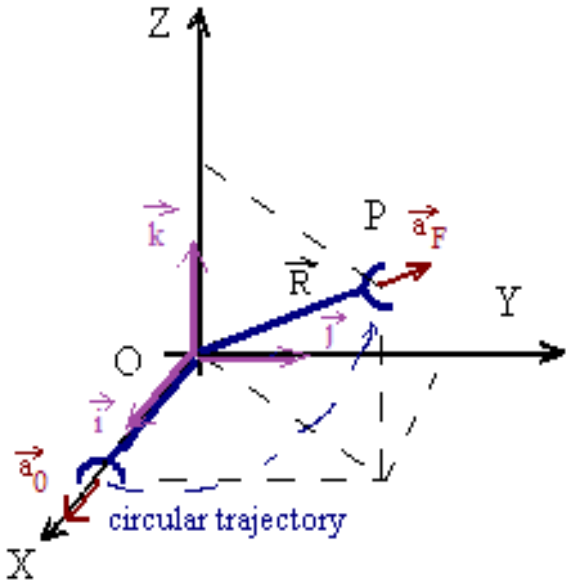


Fig. 3. Circular ADNIA about versor $\vec{a}$.

Considering those versors, the orientation versor $\vec{a}_m$, on every *m* sampling time period (indexed *m*), from the first position (indexed *0*) to the last position (indexed *F*) is:

$$\vec{a}_m = a_{m,x} \cdot \vec{i} + a_{m,y} \cdot \vec{j} + a_{m,z} \cdot \vec{k} \qquad (11)$$

Let considers the previous variation of polar and azimuthal angle, from example described about circular trajectory interpolation; after 3 steps, the axle components of versor $\vec{a}$ computed with ADNIA algorithm are (in (9), the radium *R* have the value equal with 1 value) :

$$\begin{aligned} a_{3;x} &= 1 \cdot \sin(81^0) \cdot \cos(18^0) \\ a_{3;y} &= 1 \cdot \sin(81^0) \cdot \sin(18^0) \\ a_{3;z} &= 1 \cdot \cos(81^0) \end{aligned} \qquad (12)$$

Let considers a linear trajectory for robotic arm motion as described in Table I (Step No. 3) and the interpolation process of versor $\vec{a}$ as described before; it may compute the location matrix, after three steps of motion process, with next relation:

$$G_3 = \begin{bmatrix} 1 & 0 & 1 \cdot \sin(81^0) \cdot \cos(18^0) & 10.3 \\ 0 & 1 & 1 \cdot \sin(81^0) \cdot \sin(18^0) & 26 \\ 0 & 0 & 1 \cdot \cos(81^0) & 30.09 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (13)$$

The acceleration and the deceleration of the movement are not included in those computations; it may be another part of the computations [4].

The described algorithm (about $\vec{a}_m$) may be applied for interpolate the versor $\vec{o}_m$, (14) and the versor $\vec{n}_m$ (15):

$$\vec{o}_m = o_{m,x} \cdot \vec{i} + o_{m,y} \cdot \vec{j} + o_{m,z} \cdot \vec{k} \qquad (14)$$

$$\vec{n}_m = n_{m,x} \cdot \vec{i} + n_{m,y} \cdot \vec{j} + n_{m,z} \cdot \vec{k} \qquad (15)$$

IV. CONCLUSIONS

The ADNIA algorithms may be successfully applied for computing the location matrix for every waypoint upon an imposed linear or circular trajectory, in the case of the robotic arm motion. The imposed trajectory enforces the axle components values of position vector: $\vec{p}$; its components upon the three axles must be computed with (3) for a linear trajectory and with (7) for a circular trajectory; the motion speed is imposed and it has a constant value.

At every sampling period of time, during the real-time running of ADNIA algorithm, the Cartesian coordinates of another waypoint is computed, thus it results the maximum computation precision, (for motion command process, upon a linear or circular imposed trajectory).

ADNIA algorithms work with an imposed constant value for motion speed, upon the trajectory, this being an advantage.

During the trajectory traverse with a variable orientation of robotic arm, about computation of orientation versors (vector with module equal to 1), the number of angle steps is equal with steps number computed for trajectory traverse (it is imposed by interpolation process of the trajectory). The axle components of orientation versors must be computed according with circular ADNIA.

REFERENCES

[1] E. Ciupan, F. Lungu, C. Ciupan, "ANN Method for Control of Robots to Avoid Obstacles", International Journal of Computers, Communications & Control, Vol. 9, Nr. 5, 2014, ISSN 1841-9836, Available: http://univagora.ro/jour/index.php/ijccc/article/view/813/pdf_141 , accessed January 2018.

[2] L.M. Matica, "About Adapting Traversing Trajectories ADN Interpolating Algorithms for Industrial Robots", Journal of Computer Science & Control Systems; 2009, Vol. 2, Issue 2, Available: https://www.researchgate.net/publication/40422467_About_adapting_traversing_trajectories_ADN_interpolating_algorithms_for_industrial_robot , accessed January 2018.

[3] Z. Kovendi, I.C. Rada, L. Magdoiu, A. Corha, C. Bondici, "Checking Algorithms on Differential Equations with Known Analytical Solution" Journal of Computer Science & Control Systems, 2015/5/1, Available: https://scholar.google.ro/citations?user=8ouqKx0AAAAJ&hl=ro, accessed January 2018.

[4] L.M.Matica, H. Oros, "Speed Computation for Industrial Robots Motion Followed by Accurate Positioning. International Journal of Computers", Communications & Control, February 2017, http://univagora.ro/jour/index.php/ijccc/article/download/2785/1061 , accessed January 2018.

[5] T. Kunz, M. Stilman, "Optimal Trajectory Generation for Path with Bounded Acceleration and Velocity", Robotics: Science and Systems 8th Conference, 2012, Sydney, ISBN 978-0-262-51968-7, Available: https://books.google.ro/books?hl=en&lr=&id=NOrxCwAAQBAJ&oi=fnd&pg=PA209&dq=kunz+time-Optimal&ots=dMImeillUZ&sig=0N0K2_OJa6H17-4jqwCMrQOz8dk&redir_esc=y#v=onepage&q=kunz%20time-Optimal&f=false accessed February 2018.

[6] B. J. Evans, C. D. Cook, "Evaluation of Real-Time Robot Control Systems", Conference proceedings: Field and Service Robotics, 1998, Springer, London, Available: https://link.springer.com/chapter/10.1007/978-1-4471-1273-0_57 accessed February 2018.

[7] L.M. Matica, Conducerea robotilor industriali. Editura Universitatii din Oradea, ISBN 978-973-759-481-5, 2008

[8] K. Krishnaswamy, J. Sleeman, T. Oates, "Real-Time Path Planning for Robotic Arm", Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments , May 2011, University of Maryland, Baltimore, Available: https://ebiquity.umbc.edu/_file_directory_/papers/575.pdf accessed February 2018.

A Game Theoretic Approach to Demand Side Management in Smart Grid with Multiple Energy Sources and Storage

Aritra Kumar Lahiri, Ashwin Vasani, Sumanth Kulkarni, Nishant Rawat
School of Computing, Informatics,
Decision Systems Engineering,
Arizona State University
Tempe, USA

***Abstract*—A smart grid is an advancement in electrical grid which includes a variety of operational and energy measures. To utilize energy distribution in an efficient manner, demand side management has become the fore-runner. In our research paper, we use game theory as a tool to model our system as a Stackelberg game. We make use of different energy sources like solar energy, energy from battery and energy from the provider to run appliances of a subscriber. We consider the scenario where the subscriber can give excess energy that is generated, back to the grid, thereby reducing the load on the grid during peak hours. We design a pricing scheme to calculate the utilities of the subscriber and the provider and show how our model maximizes the utility of the entire system, thereby showing the existence of a Nash equilibrium.**

***Keywords*—Demand side management; utility; smart grid; solar; battery; energy provider; fairness; proportional division; utilitarian division; Nash equilibrium**

I. Introduction

Traditional grids have been around for more than a 100 years. They were developed to meet the energy needs of the 19th century. Demands of the 21st century are ever changing and despite reasonable patchwork, traditional grids are not capable to adapt to that change any more. This brings smart grid into the picture. Smart grid is an electrical grid which allows two-way communication between providers and subscribers. It makes use of information technology to quickly adapt to the changing needs of the subscribers. Smart grid provides digitization of communications by aligning the providers and subscribers in a much more effective way. One of the main emerging areas in the smart grid research is Demand Side Management (DSM) [1], [2].

A. Demand Side Management as a Framework

Demand side management [1] can be characterized as a set of interconnected and flexible programs that facilitates efficient usage of electric power. It helps in controlling electric power consumption by shifting load accordingly and at the same time making a fair allocation of resources among subscribers. This also ensures that the subscribers are provided incentive when they reduce their consumption during peak hours to the hours when the smart grid is less loaded. Demand side management techniques provides smart pricing schemes that can ensure load shifting by implementing various scheduling schemes for appliances and also managing interactions at social (service level agreement between energy provider and subscribers) as well as technical (between smart meters and energy providers) interaction [3], [4]. The nature of demand side management and smart grid suggests the use of Game Theory as an analytical tool.

B. Game Theory as a Tool

Game theory is a framework [1] which makes use of mathematical rules and relations to study complex interactions between players. This is the primary reason why game theory can be used in developing an optimal solution for cases of cooperation or conflict with players acting in their best interest. Due to involvement of fair allocation of power and pricing schemes in demand side management, game theory can provide a natural mechanism for developing them.

C. Multiple Sources of Energy

We have also explored the possibility of multiple energy sources in this paper and considered solar energy to be an energy source. Solar energy is a form of renewable energy resource, which is cheap and easily available and can be utilized as an alternate way to obtain energy by the subscribers. We present our research in demand side management using storage devices where we are considering low cost efficient energy storing device like battery for charging during off peak hours and discharging during peak hours to reduce the load on the grid as well as give incentives to the subscribers. The consideration of battery is in tune with the fact that weather plays an effective part in obtaining energy from solar resource, hence it ensures that subscribers have sufficient amount of energy to run their basic appliances during hours of high payload. Finally we propose a model where the subscribers are strategically choosing their way of using the storage devices and when to buy energy from the providers.

II. Related Work and Motivation

One of the important papers [1] in this direction is authored by E. Salfati and R. Rabinovici. The research direction of that paper focuses primarily on the single energy provider system with multiple subscribers where the subscribers compete for resources from the energy provider.

Usually, if all the subscribers are given access to get energy for all their appliances, there is a possibility that the energy demand would exceed the total energy which can be supplied by the grid. Hence we utilize game theory in demand side management to manage the shared resource in a fair manner such that it does not disturb the subscriber interest, plus ensuring optimal usage of the grid.

Our paper focuses on the usage of multiple energy sources. The rise of alternate energy sources, specifically solar energy which an independent subscriber can harness on his own has made subscribers self-sufficient to an extent. We aim to consider a flexible system where subscribers can generate their own power and utilize it, while at the same time having the option to consume power provided by external energy provider. We believe this system increases the subscriber satisfaction since it decreases the cost of energy that a subscriber has to pay for, in case of a single energy provider. But with solar energy, we run into a few roadblocks. Changes in weather make solar energy inconsistent. It cannot be relied on a hundred percent. An effective strategy here would be to use a low cost efficient energy storage device which could be used to harness the solar energy and provide energy to the subscribers when there is little to no solar energy availability. Having multiple sources of energy would be beneficial for the subscriber as well as the energy provider. From the subscribers' perspective, the cost which they have to pay to the energy provider would be lessened and from the energy providers' perspective, the overall load on the grid would be reduced while satisfying the subscribers demand.

III. Problem Statement

Traditional smart grids face an immense amount of load during peak hours because of the high energy demand of the subscribers. This is detrimental to both the players because the subscribers need to pay a lot more money during peak hours and the energy provider is not able to meet the demands of all the subscribers, thereby reducing the utility of both players, which in turn reduces the utility of the entire system.

We aim to present a demand side management system that enables subscribers to be self-sufficient. We have to come up with a solution which enables us to use solar energy to help the subscribers increase their utility by generating more energy. But since solar energy is an unreliable source of energy, we also have to harness the excess energy from solar and store it in a low cost energy efficient device - battery, which can be used when solar energy is inconsistent. This would make the subscriber self-sufficient to a large extent. This demand side management system will work to keep the utilization of energy from the grid to as less as possible during peak hours. We aim to do this by suggesting a scheduling scheme for the different energy sources that makes a utility-efficient decision while choosing the correct energy source to employ for the subscriber.

Another important aspect which needs to be addressed is the giving back of excess energy generated by the subscriber to the grid during the peak hours. To effectively do this, we have to come up with a pricing scheme which we will develop in order for users to increase their utility. The pricing scheme has to take into consideration, the amount of energy being taken by the subscribers and the amount of energy being given back to the provider. We have to devise a method to calculate how much the subscribers will get paid based on their peak time assistance. So the challenge is to ensure a synergistic relationship among the provider and subscriber, such that the combined utility of our system is maximized.

IV. Our Research Goal

Our path of research focuses on a demand side management approach that is concerned with the cumulative load of all the subscribers. Hence we formulate a static non-cooperative game with N subscribers that devise a scheduling algorithm for direct load shifting of appliances based on the Asynchronous Consumption Mode (ACM) mentioned by authors in [1]. ACM is characterized by Quality of Service Metrics (QoS) classified according to subscriber characterization of appliances. Following is a brief overview of these metrics:

1) ***Constant Consumption Rate (CCR):*** This level of service includes appliances that must be operated whenever their demand is generated, for e.g. refrigerators and room lights.

2) ***Available Consumption Rate (ACR):*** Their priority level of service is just next to CCR and can endure a certain period of delay before they need to be operated after their demand, for e.g. Electric kettle and boiler.

3) ***Unspecified Consumption Rate (UCR):*** They have the least priority level and can be operated when there is an excess energy capacity that can be used without hampering the working of appliances with CCR, for e.g. under-floor heating and outdoor lighting.

The objective is to minimize the overall energy consumption from the energy provider and effectively reduce the charges incurred by the subscribers. The non-cooperative game is experimentally shown to have existence of Nash equilibrium at a global optimal point. As mentioned earlier, we have considered further, that along with non-cooperative game our model will be based on multiple energy sources. We have considered energy storage as an essential criterion, since it has a huge potential in impacting the demand side management. For example, it can be assumed that a user might opt to store energy during hours when energy load is very low, so that it can utilize them during peak hours rather than obtaining them from power stations thereby maximizing their own utility and the utility of the energy providers [3]. Traditional grid systems faced issues related to voltage fluctuations and harmonic distortions while integrating small scale renewable energy sources due to lack of synchronization. But introduction of smart grids prevents these outages and provides multiple options of renewable resources to supply energy to the grid through distributed power generation and storage. Solar energy is one such form of renewable source that we are considering in our research in this paper. Usage of solar energy requires an initial installation cost, but nevertheless is a cheap and easily available source of energy for the subscribers. However, since weather plays an important factor in the generation of solar power, hence it is practical for the subscribers to use low cost efficient devices

like battery to store energy. Using battery will ensure that subscribers can store energy for much longer periods of time and discharge energy from them during peak periods and charging them during low energy demand periods and also from solar energy resource whenever it is available. However, without a proper scheduling algorithm for charging and discharging of storage devices, there might be scenario where all storage devices wanting to charge energy simultaneously leading to a higher peak demand [4], [5]. Thus our primary focus is to devise a non-cooperative game model to address the issue. Hence we have proposed a distributed algorithm that supports all the fairness criteria [1] and helps in maximizing social utility of the system. We consider the four main types of fairness criteria while proposing the algorithm such as:

1) ***Proportional Division:*** This criteria ensures every player gets his due share according to his own valuation of the resource.

2) ***Envy-free Division:*** This criteria ensures that each player gets a valuation which is at least equal to all other players share of resource.

3) ***Equitable Division:*** This criteria ensures each player obtains the exact same valuation of the resource.

4) ***4. Utilitarian Division:*** This fairness criteria ensures that the sum of the individual utilities of all the players is maximized.

Here each subscriber can independently strategize for his energy consumption scheduling based on the QoS metrics and also schedule his charging and discharging of storage device like battery. We have experimentally calculated and verified that this model can lower the total energy cost of the entire system by reducing the peak-to-average ratio and maximization of utilities for each of the subscribers.

Along with this, we have also formulated a Stackelberg game for our model as well as implementing a pricing scheme between the energy providers and subscribers. The subscribers and providers are both players of this game and pricing scheme is determined by the providers. The subscribers will aim to buy extra energy before the typical peak hours and store them in their storage devices while trying to sell it back to the providers during peak hours to minimize their cost. The energy providers on the other hand adjust the pricing scheme accordingly once it sees many subscribers are trying to sell back energy during peak hours and it exceeds the consumption rate at that point of time. This scenario can be effectively modelled as a Stackelberg game [4], keeping in mind that the energy providers will try to maximize their profit and the subscribers will aim to reduce their cost of electricity, thereby maximizing the total utility of the system.

V. Demand Side Management Model

A. System Model

Our system model includes one energy provider providing energy to N subscribers. Fig. 1 gives a complete overview on how energy providers and subscribers are connected. We assume that every subscriber has an energy consumption controller device (smart meter) which can efficiently allocate energy to the subscriber as well as keep a history on the energy consumption of an individual subscriber. The uniqueness of our model is we utilize multiple energy sources, in specific, two other energy sources - solar energy and a battery. The subscribers can use the energy generated by the solar panel and use it for themselves (during peak and off-peak hours). A battery is an energy storage device and energy providing device in our model since solar is not always reliable and the charge from battery can be utilized to power the appliances for a subscriber. Another uniqueness of our model is that the subscribers can give back excess energy to the energy provider whenever available, thus generating incentives for the subscriber as well as the energy provider.

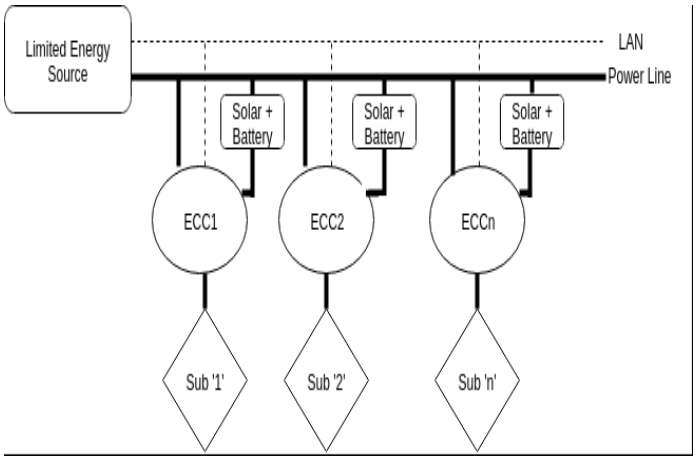


Fig. 1. System Model.

All in all, our model supports energy from three different sources - energy provider, solar energy and battery and the possibility of giving back energy to the grid.

Let N denote the number of subscribers in our model. For every subscriber $n \in N$, we define a parameter called consumption level demand $C_n^d[kW]$ which is the energy consumed by a particular subscriber. We define our time slot vector $t \in T$ to contain different time slots. T can be any number of divisions for a 24 hour period. The reason for this division is to effectively calculate the subscriber demand for a particular time slot and to calculate the utilities of the subscribers based on whether a particular time slot falls under a peak hour time slot or an off peak hour time slot.

B. Control Loop Description

The control loop for our model as seen in Fig. 2 is unique since it uses the fair division allocation as seen in [1] as well as considers the usage of solar, battery and energy from the provider. The energy obtained from solar panels is used by the subscriber until it lasts then battery is used. Only if both the solar and battery charge fail to give a certain amount of power which is decided by the subscriber demand algorithm, then energy from the provider is used. Fig. 3 represents the energy supply design. This approach effectively calculates the utilities of the provider and the subscribers by maximizing potential for both the players in our game. The subscribers only pay for the energy they consume depending on the pricing model set by the leader of the game, the energy provider. As discussed in the previous section, the subscribers can also give back

excess amount of energy generated by them to the energy provider thus receiving incentives.

C. Game Formulation

The primary objective of our research is to maximize the utility of the subscribers and the provider, thereby maximizing the utility of the entire system. We model our idea as a Stackelberg game, where the provider is the leader and the subscribers are the followers. The leader has the authority on the pricing model and the constant factor δ which we will discuss in the upcoming sections. The overall incentives for the subscriber is calculated based on the parameters of amount of energy they are generating, amount of energy they are utilizing, amount of energy they are giving back to the grid.

Provider utility is calculated based on the amount of energy being saved during peak hours. This is discussed in detail in the subsequent sections.

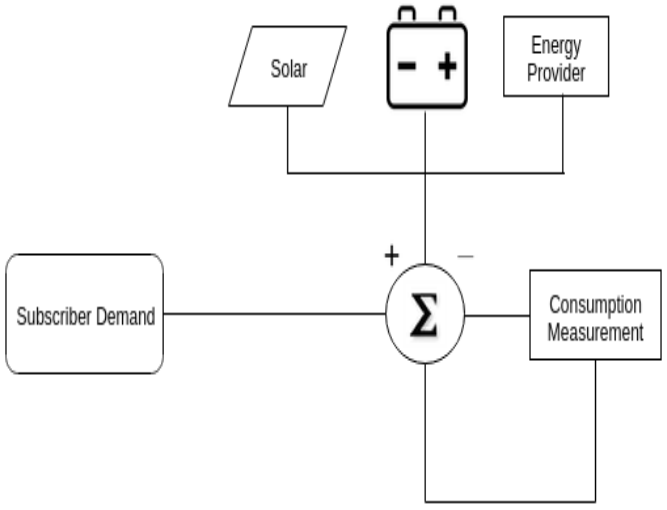


Fig. 2. Control loop description.

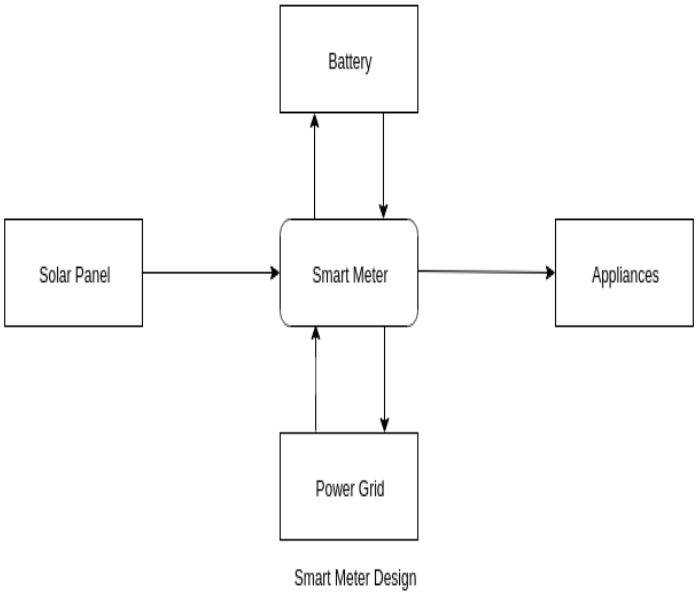


Fig. 3. Energy supply design.

D. Subscriber Demand Algorithm

The subscriber demand algorithm determines the amount of energy needed for every subscriber during the peak and off-peak hour time slots. Our analysis has shown that the peak hours are between *12pm - 7pm* and off-peak hours are between *7pm - 12pm* with a majority of energy providers like APS, SRP etc. From [1], we know about the asynchronous consumption mode, which is categorized into three quality of service metrics $Q = 1, 2, 3$ discussed in *section 4*, where $j \in Q$, we calculate the energy demand for all the available metrics for every subscriber $k \in N$ depending on the available capacity C of the grid and consumption level demand $C_n^d[kW]$ of every subscribers in the time slot T. Fig. 4(a) to (d) describes the algorithm functionality diagrammatically.

Algorithm 1 Subscriber Demand

Require: Time slots $t \in T$
Ensure: Allocation vector (x_i)

```
for each time slot i ∈ T do
    for each QoS metric j ∈ Q do
        if Σ_{k=1}^{n} C_{i,j}^k < C then
            x_{i,j}^k ← C_{i,j}^k
        else
            x_{i,j}^k ← (C_{i,j}^d / Σ_{i=1}^{n} C_{-i,j}^d) × C
        end if
    end for
end for
return (x_1, x_2, ..., x_n)
```

E. Energy Usage Algorithm

In this section, we describe the algorithm which is used to calculate the total energy usage of a particular subscriber $j \in N$ in time slots $i \in T$ as discussed before. This algorithm considers the amount of solar energy generated along with the battery capacity. If both the primary and secondary energy sources do not satisfy the subscriber demand for different time slots obtained from the subscriber demand algorithm, then energy is obtained from the provider and a cost is paid for it. We define an energy usage vector as A_j^i , solar energy as s_j^i, , battery capacity as B, battery charge as bj , the charging rate of battery as c_r, discharging rate of battery as d_r and the number of hours the battery charges as h_r. On the other hand if solar energy is abundant, then we charge the battery as well as give back to the provider for an incentive. Our algorithm takes care of all possible scenarios from the subscriber perspective. The algorithm is as follows:

Algorithm 2 Energy Usage Algorithm

Require: Allocation vector (x_i)
Ensure: Energy usage vector A_j^i

```
for every i ∈ T do
    for every j ∈ N do
        e ← s_j^i − x_j^i
        if e > 0&(b_j + e) ≤ B then
            b_j ← b_j + e
            A_j^i ← A_j^i ∪ +e
        end if
        if e < 0 then
            if b_j ≥ e then
                b_j ← h_r × d_r − e
            else
                e ← b_j − e
                A_j^i ← A_j^i ∪ −e
            end if
        else
            if off − peak then
                b_j ← b_j + h_r × c_r
                A_j^i ← A_j^i ∪ −b_j
            end if
        end if
    end for
end for
return A_j^i
```

F. Utilitarian Algorithm

The last algorithm in our model, which we have come up with takes into account the excess energy produced by the subscriber which is given back to the provider and also the energy obtained by the subscriber from the provider. It gets the energy usage vector A^i_j from Algorithm 2. This algorithm calculates the utilities of the subscribers and the providers. Here, we also introduce a variable δ, which is the basis of our novel pricing model. The utilities of subscribers u_j is calculated by taking the difference of the energy supplied to the provider obtained from the energy usage vector A^i_j and the energy obtained from the provider.

Algorithm 3 Utilitarian Algorithm

Require: Energy usage vector A_j^i
Ensure: Utilities u_j

```
for i:=1,2 do
    if e < 0 then
        u_j ← e × P_i
    else
        u'_j ← δ × e × P_i
    end if
end for
return Δu_j
```

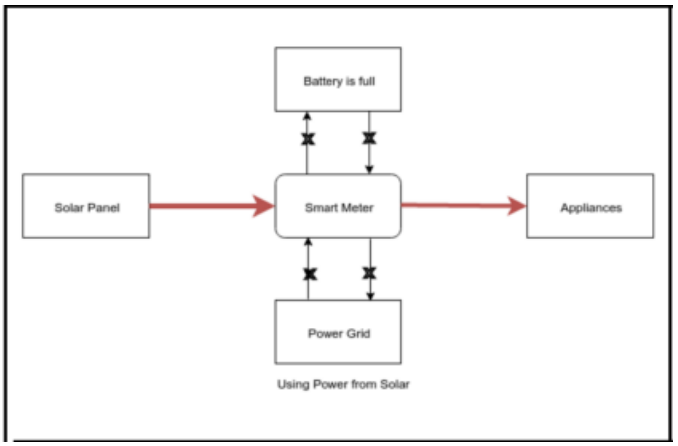
1) Pricing Scheme

In our approach, we surveyed the pricing schemes of multiple energy providers and we reached a consensus that effective pricing schemes are categorized based on two major grouping of times - peak and off-peak hours. δ is a variable with value in [0, 1]. As mentioned before, the leader of the game (provider) sets this value. This value is utilized when the subscribers give energy to the energy provider. If the energy provider needs the energy, he will set the value of δ to be >= 0 or <= 1. Otherwise if the provider does not need the energy from the subscribers, then he could set the value of δ to be 0.

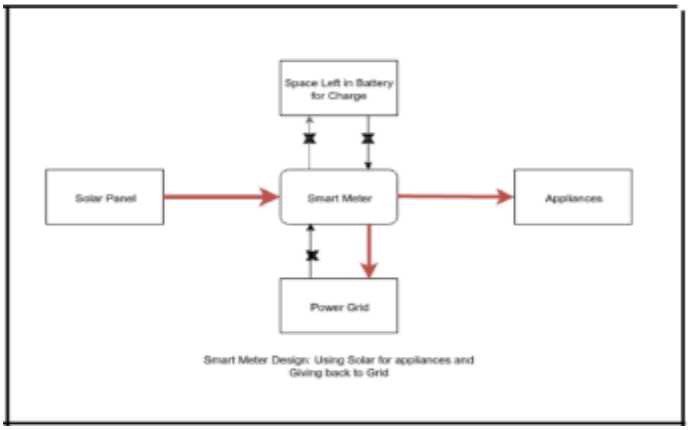
The utility of subscribers can be obtained by $u_j = u_j - u^0_j$. This is the amount of money that the subscribers save or get from the energy provider.

On the other hand, the provider utility u_p can be obtained by the energy being used *by* the subscribers and the energy received from the subscribers.

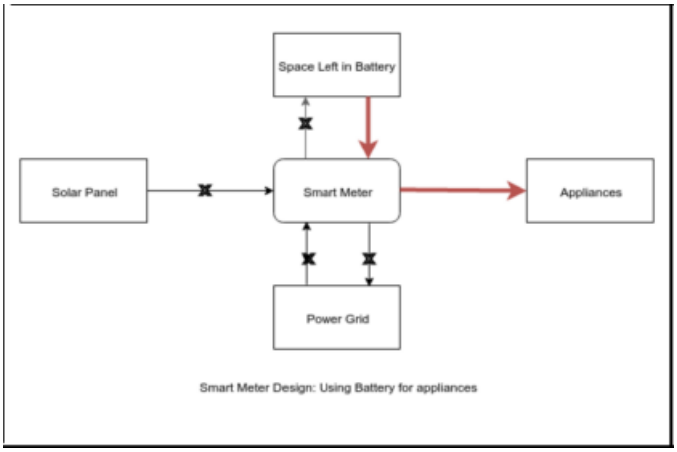
The utility of provider can be obtained by $u_p = u^0_j$ since this corresponds to the energy received from the subscribers and he also make a monetary gain with the amount of power consumed by the subscribers specified in u_j. Owing to these conditions, we see that the utilities of the subscribers and the provider are maximized from our algorithm.



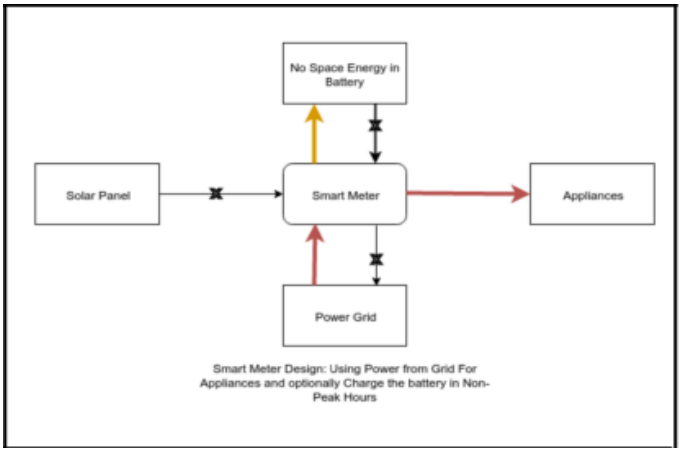
(a) Smart meter actions using power from solar.



(b) Smart meter action using solar for appliances and getting back to grid.



(c) Smart meter actions using battery for appliances.



(d) Smart meter actions using power from grid and optionally charging battery in non-peak hours.

Fig. 4. (a)-(d) Smart meter actions.

G. Existence of Nash Equilibrium

From the utilitarian algorithm, we see that the subscribers make a monetary profit when they give energy back to the provider, thus enhancing their utility. They save up money by using multiple sources of energy as well as receive incentive for giving energy to the grid. The provider profit is also maximized because the subscribers manage the peak hour load as show in Algorithm 2 alongside having a monetary gain from the power distributed during the off peak hours. Since the utilities of both the provider and the subscribers are at maximum, we can effectively say that utilitarian algorithm gives us a *Nash Equilibrium*. Our algorithm also satisfies the fairness criteria mentioned in *Section 4.*

VI. Performance Evaluation

For simulation, we have taken values considering a real life scenario. Our designed algorithm will work on smart meter which has a capacity to control source and destination using digital relay circuits. In our system, we assume that the battery is a two way source, which means the battery can be charged or discharged. In discharging phase, the energy will not flow back to the grid. The grid also has two way capacity. Energy can be given to the grid or it can take back. Solar is a single point source which can generate the energy and appliance can use the energy from solar, battery or the grid. We have taken into consideration a lithium ion battery for this experiment which has a range of 4.4kWh to 7.5kWh [6]. These are the high end expensive battery. For the experiment, we assume that the maximum capacity of the battery will be 4.5kWh and it can charge up to 2.5kWh in one hour. From [7] and [8], we calculate the amount of power needed by different appliances and their general observed schedule. Based on these numbers, we define allocation vector x*i* for each subscriber. We have used a tool to calculate solar energy [10] to get the energy generated by the solar panels. As we can see in Fig.5, there is sufficient amount of power is generated by 15 panels which is sufficient to give electricity to a median home throughout the year. This calculation is done for the Arizona State University location. In our experiment, we have taken readings for the month of January.

A. δ Calculation

In Utilitarian algorithm, the value of the delta is set by the energy provider. Energy providers can maximize their profit by changing the value of δ as discussed in the above sections. During off-peak hours, there is no profit to energy providers, if electricity is given back to the grid. But, in peak hours, when the electricity rates are high along with the demand, user cannot be paid by same electricity price. Hence, energy provider purchase the electricity with the reduced price of δ. Fig. 6 depicts APS charges for subscriber with \$0.20960 per kWh during peak hours (12noon-7pm) and \$0.02601 during off-peak hours [9]. For δ = 0.01, APS can buy back the electricity at the price of \$0.02096 kWh.

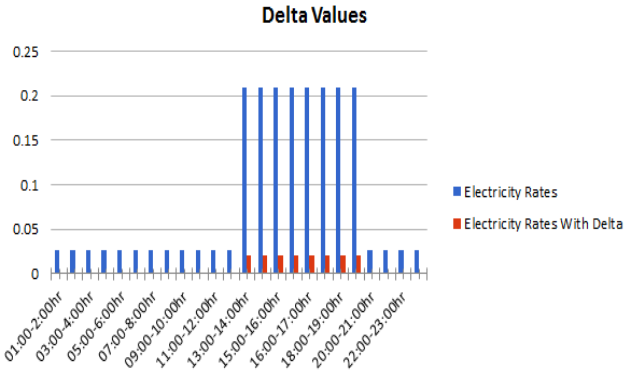


Fig. 5. Delta values.

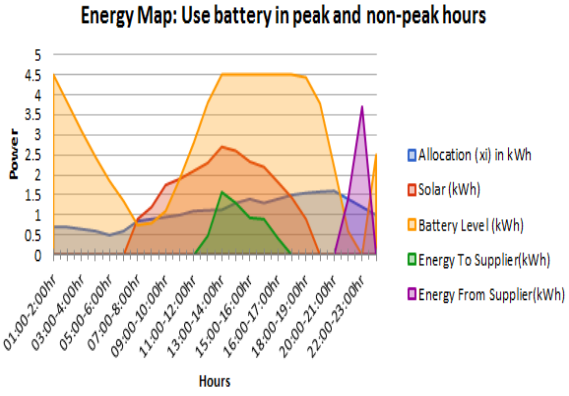


Fig. 6. Energy map with battery in peak and non peak hours.

1) Total Energy Map

Till now, we have calculated the allocation vector, solar power generated per hour, battery with finite capacity (4.5kWh), finite charging rate (2.5kWh) and δ factor for peak (=0.01) and off-peak hours (=0). Based on these values, we plot the below graph as seen in Fig. 7 for the whole day.

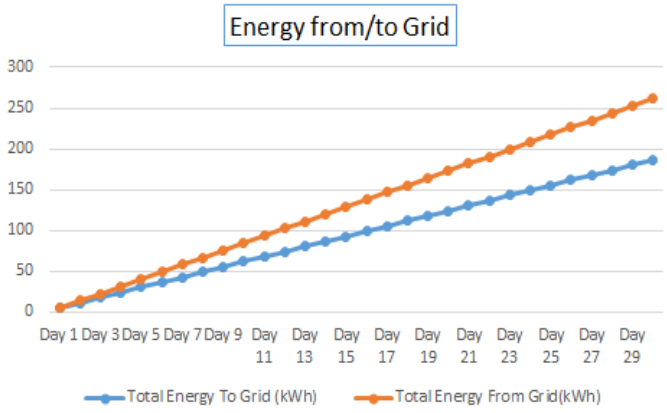


Fig. 7. Subscriber usage of supplier energy.

Initially the battery level is full and hence smart meter based on our algorithm will use battery to fulfil the allocation request. Now, from 6:00 am onwards, the solar energy starts generating the power and can use to fulfil the allocation requests. As the day goes on, more solar energy is generated and this can be used to charge the battery. After 11:00 pm, as the battery is fully charged, the electricity is given back to the power grid. Now, from 4:00 pm onwards, as the sunlight fades, the solar power generation reduces and it can no longer serve the allocations requests. The smart meter now makes a decision to use the power from the battery. By 9:00 pm, the battery is totally drained and hence, smart meter now switches to grid to fulfil the allocation requests. But, as the power is cheap in off-peak hours, smart meter also charges the battery so that the user in next day can send more energy to the grid during peak hours instead of charging the battery.

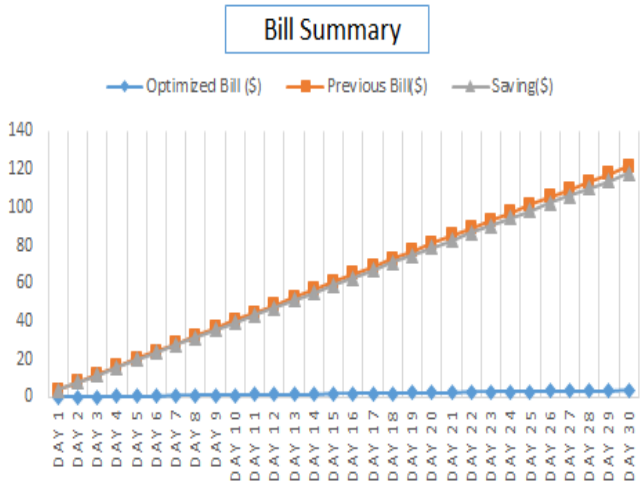


Fig. 8. Saving.

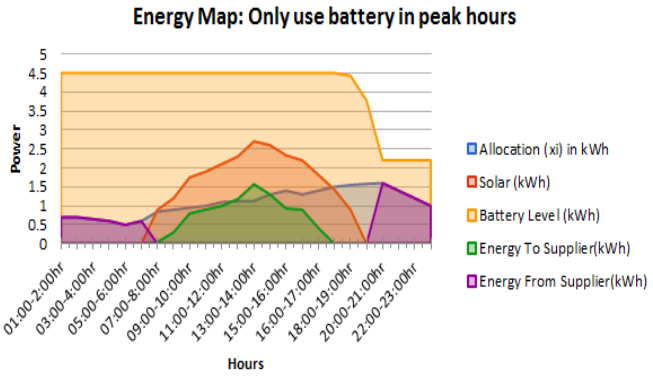


Fig. 9. Energy map: only use battery in peak hours.

After analyzing the whole month usage, we found that the subscriber gives back 186.85kWh of energy to the grid and took 261.75kWh of energy from the grid. The subscriber's monthly expected bill was $121.64469 before our smart meter addition and it came down to $3.58 with total saving of $118 in month of January. With this saving, as plotted in Fig. 8, user can buy back the whole solar panel setup cost ($12K) [10] in 8 years. The above graph in Fig. 9 shows the Energy Map when the battery is used in peak hours or else power is used from the power grid in non-peak hours. As we can see, more energy is given back to the grid as the battery is charged during non-peak hour period. Here, the battery usage is less, but desirable as it is inexpensive and have limited charging - but desirable as it is inexpensive and have limited charging-discharging cycle.

VII. Conclusion

The re-engineering of Asynchronous Consumption Model [1] is the one of the foundations of our research. We categorize appliances based on the three qualities of service metrics to get a subscriber demand. In this paper, we have formulated a novel way to use demand side management by utilizing multiple energy sources. We have come up with the energy usage algorithm which proves that our approach gives better results to the subscribers based on our analysis. We modelled our system as a Stackelberg game where the energy provider was the leader and the subscribers were the followers. Our pricing model controlled by the provider based on the time slots is another effective strategy which is proven in the results. Overall, the subscriber and the provider utility is maximized which paves way to the maximization of utility of our entire system. We also showed the existence of a Nash equilibrium which is also the optimal case in our design.

A possible future direction would be to dynamically calculate prices for a particular time slot in our pricing scheme.

VIII. Contributions

Multiple papers have been published in the aspect of demand side management in smart grid. The paper [2] researched on utilizing a storage device during peak hours, but it fails to optimize the subscriber utility in general as it focuses

only on peak hours. The same issue was in the paper [4] which discusses about storage devices for smart grid. While conducting our research, we took a step back and found the fundamental problem was with energy generation, which is why we focused on solar energy. But since solar energy is inconsistent, we focused on the usage of solar energy to charge a storage device, battery, in our case. Little research [4] in smart grid is in the direction where subscribers can give energy back to the grid. We considered this case and focused our research in this direction and proved that with an effective pricing scheme, subscribers can gain incentives to participate in our plan.

ACKNOWLEDGMENT

We thank Professor Xue, the teaching assistants, Ruozhou and Xiang for their continued encouragement and support. We have learnt a lot about demand side management and its vast applications in smart grid while independently researching for this paper. To conclude, this course has provided us with solid foundation and knowledge that will definitely help us in our future endeavours.

REFERENCES

[1] Salfati, E.; Rabinovici, R., "Demand-side management in smart grid using game theory" in Electrical & Electronics Engineers in Israel (IEEEI), 2014 IEEE 28th Convention of , vol., no., pp.1-5, 3-5 Dec. 2014 doi: 10.1109/EEEI.2014.7005769

[2] Nguyen, Hung Khanh, Ju Bin Song, and Zhu Han. "Demand side management to reduce peak-to-average ratio using game theory in smart grid." Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on. IEEE, 2012.

[3] Saad, Walid et al. 'Game-Theoretic Methods For The Smart Grid: An Overview Of Microgrid Systems, Demand-Side Management, And Smart Grid Communications'. IEEE Signal Process. Mag. 29.5 (2012): 86-105.

[4] Soliman, Hazem M., and Alberto Leon-Garcia. 'Game-Theoretic Demand-Side Management With Storage Devices For The Future Smart Grid'. IEEE Trans. Smart Grid 5.3 (2014): 1475-1485.

[5] Wang, Kun et al. 'A Game Theory-Based Energy Management System Using Price Elasticity For Smart Grids'. IEEE Transactions on Industrial Informatics 11.6 (2015): 1607- 1616.

[6] http://bosch-solar-storage.com/the-battery/lithium-battery/

[7] http://www.energysavings.com/energy-consumption.html

[8] http://www.nrc-cnrc.gc.ca/eng/dimensions/issue6/smart_technology.html

[9] https://www.aps.com/library/rates/et-2.pdf

[10] http://www.wunderground.com/calculators/solar.html

Design of Mobile Application for Travelers to Transport Baggage and Handle Check-in Process

Sara Y. Ahmed
Department of Information Technology,
Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia
Department of Scientific Computing,
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
PO Box 42808, 21551, Jeddah, Saudi Arabia

***Abstract*—In this paper, an Android based application called 'Baggage Check-in Handling System' is developed for helping travelers/passengers transport their baggage to the airport and handle the check-in process. It is merging the idea of online baggage check-in, and tracking technology together. The application is stimulated from the rapid growth of on-demand ride services, such as UberX and Lyft and the wide spread adoption of smart-phones. The proposed system enables travelers to make an appointment before the flight's take-off by requesting a driver to pick up the traveler's baggage to transport to the airport. Then, travelers can track the driver's location using Geographical Position System (GPS). Eventually after the check-in process, the driver will send a unique barcode provided for the baggage to travelers through the application. As a result, the traveler will have the choice of directly proceeding to the flight gate. The application is created for Android platform operating system, and developed in Java programming language using the Android software development kit (SDK). Additionally, data between database and server have been exchanged using phpMyAdmin. The application uses an authentication technique called Secure Hash Algorithm (SHA). This technique is designed to improve the scalability of authentication and reduce the overhead of access control.**

***Keywords*—Baggage handling system; tracking technology; baggage barcode; android platform; Android software development kit (SDK); phpMyAdmin, Secure Hash Algorithm (SHA)**

I. INTRODUCTION

In recent years, airport departure processes have drastically changed, as traveling became a necessity to many individuals. Most travelers/passengers face a lot of difficulties and complications during traveling, like being late or missing their flights, or wasting their time waiting for their baggage to be weighted. Airport check-in [1] is the process in which travelers are accepted by an airline at the airport prior to travel. The airlines typically use service counters found at airports. The check-in is normally handled by an airline itself or by a handling agent working on behalf of an airline. Travelers usually hand over any baggage that they do not wish or are not allowed to carry in to the aircraft's cabin and receive a boarding pass before they can proceed to board their aircraft.

Check-in is usually the first procedure for a traveler when arriving at an airport, as airline regulations [2] require travelers to check in by certain times prior to the departure of a flight. This duration spans from fifteen minutes to four hours depending on the destination and airline.

The main problem is standing for a long time in the baggage's queue check-in counter at the airport [2], and then waiting until boarding time which consumes a lot of time and stress. Time-consumption and stress are two of the many consequences of waiting. Moreover, traveling is harder for people who are older, pregnant, or accompanied by their children. From the airport operators' perspective, there are limited terminal space and capacity, along with optimizing income problems [3].

The aim of this paper is to introduce the implementation and design of on-demand service application called 'Baggage Check-in Handling System' that uses technology in order to make the check-in process swift and less time consuming for the traveler. 'Baggage Check-in Handling System' application is stimulated from the developments in information and communication technology [4], which have enabled new advantages that offer a wide range of real-time and demand-receptive trips.

Companies such as Lyft and Uber [5], have developed smart-phone applications to connect riders with drivers. The transport request indicates at least the pickup or drop-off location through Geographical Position System (GPS) [6]. A confirmation panel is created to present information equivalent to either the pickup or drop-off location. Through these applications, confirmation and cost of transportation, and a rating system are provided. Users' credit card information can be saved within the system to facilitate future trips. Complete description of ride sourcing is challenging though, as the services are quickly growing [7].

The contribution of our mobile phone application, 'Baggage Check-in Handling System', is that it provides transport services for the travelers' baggage after authenticating its information. The proposed system enables travelers to make an appointment before the flight's take-off, by requesting a driver to pick up the traveler's baggage from the selected location to the airport. Meanwhile, travelers can

track their baggage through Geographical Position System (GPS) [8]. Eventually after the check-in process, the driver will send a unique barcodes for the baggage to the traveler through the application. As a result, the traveler will have the choice to directly proceed to the flight's gate. Many technologies are used in our application such as the flight's booking information technology which is used to retrieve the showing booking number, or by an e-ticket. Moreover, the application is created for Android platform operating system [9], and developed in Java programming language using the Android software development kit (SDK) [10]. Finally, data between database and server have been exchanged using phpMyAdmin [11].

In fact, identifying and authenticating users can not only avoid illegal accesses, it can also reduce unnecessary redundancies of access control. 'Baggage Check-in Handling System' uses an authentication technique introduced in [12] called Secure Hash Algorithm (SHA). This algorithm was published in 2003 as the secure hash standard. SHA is a cryptographic hash function and employed in several widely used applications and protocols. This algorithm is designed to improve the scalability of authentication and reduce the overhead of access control. It is used for handling a compressed representation of a message. Given an input message, SHA produces an output called the message digest. It is claimed to be secure because it is infeasible to compute the message corresponding to a given message digest. Also, it is particularly improbable to find two messages hashing to the same value.

The remainder of the paper is organized as follows: Section II describes the proposed application scenario. The technical details to implement the suggested system and the data collection in the process of analyzing the project to gather user information is introduced in Section III. The implementation of the system, and the tools used to implement the features are presented in Sections IV and V. The usability study is introduced in Section VI. Finally, conclusions and future works are drawn in Section VII.

II. Application Scenario

The central goal for the proposed application 'Baggage Check-in Handling System' is to create a marketable mobile application based on people's needs in handling the baggage check-in process. The application consists of two main users, the driver, and the traveler. The application's scenarios for both users are introduced in Fig. 1.

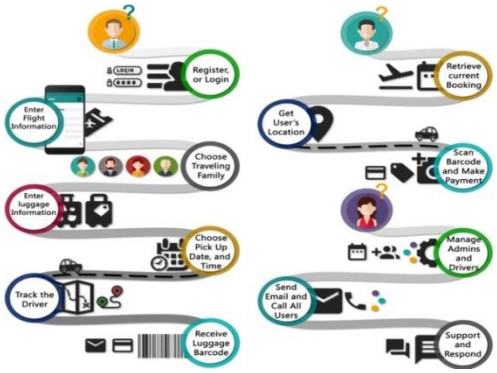


Fig. 1. Application scenarion for traveler and driver.

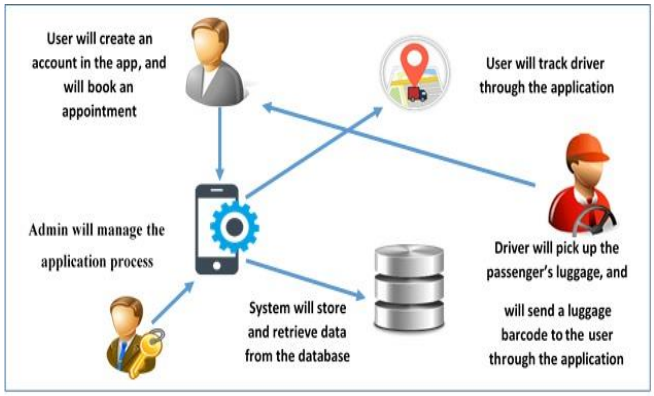


Fig. 2. High-level system architecture.

A. High-Level System Architecture

Fig. 2 introduces the high-level system architecture [13] for the proposed application, in which the traveler creates an account with all his/her and families' information. To book the service, the flight will authenticate then it will have at least three hours prior the flight. During booking, the driver will pick up the baggage while the passenger is able to track him until the check-in point. After a successful check-in process, the service's cost will be withdrawn from the passenger's pre-defined bank account, and then receive the baggage's barcodes that will be saved in his/her account in case of any problem. Finally, the passenger will proceed directly to the flight gate. All relevant information related to the traveler, and flight information will be verified through the application.

B. The Behavior and Functionality of the Application

A use case [14] shows the behavior or functionality of a system (see Fig. 3). It consists of a set of sequences of interactions between a system and a user in an environment. The first use case is the traveler login. Registration - as a new traveler - is an extended case from it. The second use case is the traveler's reservation of the pick-up baggage service. Including that, he/she will enter the booking information, select a time from the available time slot, and approve the policy. In the driver's interface, a schedule with the traveler's information, time, and location is displayed. The third use case is the traveler's ability to track the driver. Following that, the fourth use case, the driver scans the barcode then saves it and sends it to the traveler. The final use case allows the admin to log in into his/her interface, and manage the system with adding, editing, removing drivers, users, and admins; and updating the booking and scheduling to the driver.

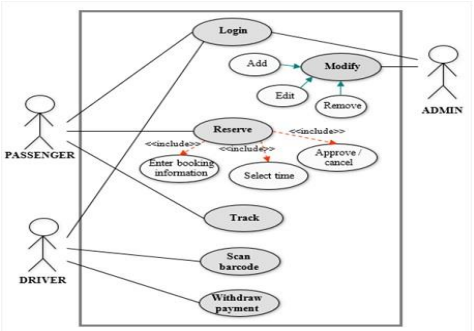


Fig. 3. Application use case.

III. Analysis and Design

This section introduces more technical details to implement the suggested system and the data collection in the process of analyzing the project to gathering user information.

A. Data Collection Techniques

The user profile questionnaire [15] includes the user's general information: age, gender, and their evaluation in using similar systems to the application. It also includes their experiences in traveling, the time they spend to check-in their baggage before their take off. Fig. 4 depicts the analysis of the questionnaire created using "Typeform" website, with a thousand and twenty-three responses. Most people worry about the security side of this application and how much trust they can put into it. Some people request for the application to have the government's permission, and that their confirmation should be displayed in the users' signing up process. Other people suggest the having a concise explanation of the procedure of the baggage's insurance in case it gets lost or harmed would make the application guaranteed, trustworthy, and reliable. The remaining statements recommend more services and techniques such as providing a video camera inside the car to watch the driver's movements, taking care of breakable/fragile baggage, adding a baggage packing service, and displaying the baggage's weight to estimate if extra payment is required.

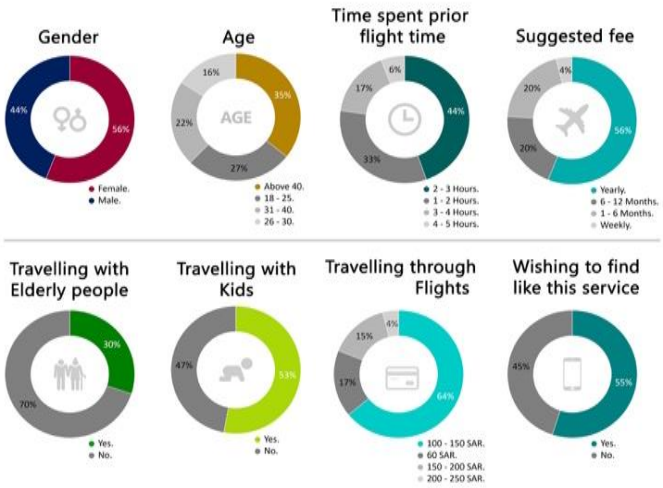


Fig. 4. Questionnaire analysis.

B. Functional Requirements

Functional requirement [16] defines the function of a system and its component. Table I represents the function as a set of inputs, outputs, and behavior for each actor.

C. Non-Functional Requirements

In requirement engineering, a non-functional requirement [16] is a necessity that specifies criteria that can be used to judge the operation of a system. The non-functional requirements of the application are introduced in Table II.

TABLE I. Functional Requirements

	No.	Requirement
Create main admin account	**1**	When using the system for the first time, an admin account should be created.
Admin	**1**	The admin shall have the ability to login to his/her account.
	2	The main admin shall have the ability to add, delete, and update drivers, and sub-admins.
	3	The admin shall have the ability to add, delete, and update users (travelers).
	4	The admin shall be able to adjust the booking schedule, and assign appointments to drivers.
	5	The admin should be able to send confirmations to the travelers via email, or messages.
	6	The admin should have the ability to view the user's feedback.
User (Traveler)	**1**	The user shall have the ability to create a new account and login to his/her account.
	2	The user shall have the ability to reserve a driver after filling his information.
	3	The user shall upload his/her boarding pass.
	4	The user should book within the time range.
	5	The user should be able to add extra services to his baggage such as wrapping, and/or breakable.
	6	The user shall view and approve the price decided by the system, and approve the policies.
	7	The user shall be able to cancel his/her reservation if the arrival time had not come yet.
	8	The user should leave his/her review of the application at the end of the process.
System	**1**	The system shall be able to generate the users' flight information.
	2	The system shall be able to assign appointments to different drivers.
	3	The system shall calculate the price.
	4	The system shall send a confirmation message to the travelers after their successful booking.
	5	The system shall send the driver's information.
	6	The system shall upload and save the boarding pass of the traveler to the driver.
	7	The system shall provide the user with the tracking of the driver.
	8	The system should notify the user about the driver's status via sending the user notifications.
	9	The system shall send and save the barcode tags to the users' profile.
	10	In case of user cancelation, the system should cancel and reassign the reservation to the driver.
	11	The system shall withdraw the predefined cost at the end of the check-in.
	12	The system should display the user's feedback.
Driver	**1**	The driver shall have the ability to login to his account.
	2	The driver should confirm that he picked up the baggage.
	3	The driver should specify his status, either if he is on the way, picked-up the traveler's baggage, arrived at the airport, or done with the check-in.
	4	The driver shall scan the barcode tags after checking in the baggage, and then send them to the traveler.
	5	The driver should shift to the appointment.

TABLE II. NON-FUNCTIONAL REQUIREMENTS

No.	Requirement	Description
1	**Usability**	- The application must be easy and simple for all types of people that vary from ages and backgrounds to use.
2	**Response Time**	- Notifications should be sent in an appropriate time. - Booking and scheduling must be implemented in the right timing, with no delays.
3	**Reliability**	- Many users can use the application at the same time, and deliver services to all users, as it was intended to.
4	**Security**	- The application contains the users' confidential data that cannot be seen or accessed into by anyone, except the users themselves, and the admin. - The application should be supported by a trusted organization. - The drivers' car will be tracked for security and safety issues.
5	**Maintainability**	- New features could be added to the application in the future.
6	**Warranty**	- The service should compensate the users' money in case any of his/her baggage got damaged or lost.

D. Software and Hardware Requirements

'Baggage Check-in Handling System' will be implemented using Android OS application. The requirements (shown in Fig. 5) are an Android Operating System, a Software developed by a Java programming language, Global Positioning System, and SQL database to store data whereas the hardware requirements are Android smart-phones with the Android operating system and Server.

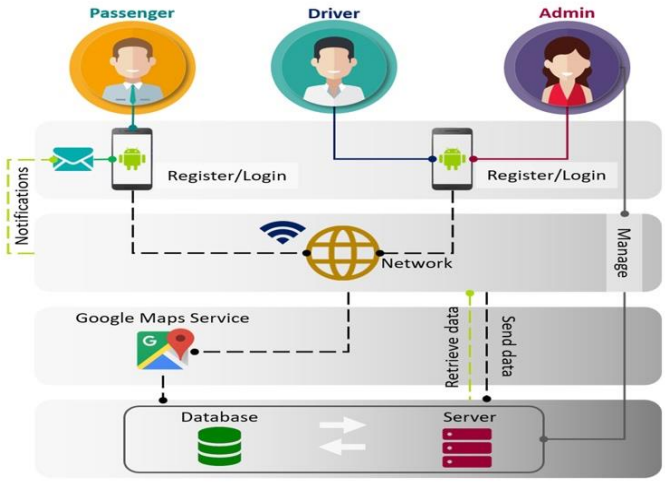


Fig. 5. Software requirements.

IV. APPLICATION METHDOLOGY

The application uses the Agile Development Methodology [17], since Agile is an Incremental Software Development Method. Furthermore, it advocates adaptive planning, evolutionary development, early delivery, and continuous improvement. It also encourages rapid and flexible response to change [17].

A. High Level Diagram

High-Level Design (HLD) explains the architecture that would be used for developing a software product [18]. Fig. 6 provides an overview of an entire system identifying the main components that would be developed for the product and their interfaces. We overviewed possible scenarios that will be followed between the main components of our system. The first scenario is for creating a new account for a traveler and validating inserted information with application database. The second scenario includes booking appointments for picking up baggage and saves the time and location into the application's database, then receives the saved barcode from it. The third scenario is retrieving the travelers' information such as their family name and phone number then saving it in the database. Finally, after the driver receives a booked appointment's information from the database, he uploads the scanned barcodes to the application's database.

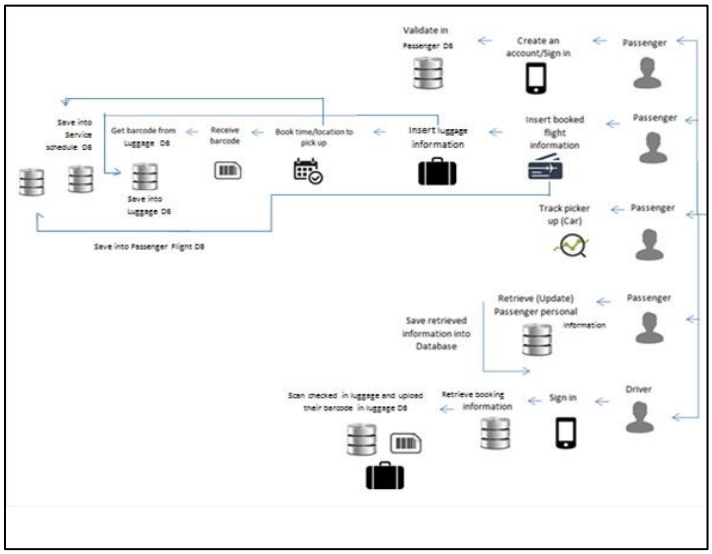


Fig. 6. High level design.

B. Sequence Diagram

The sequence diagram [19] in Fig. 7 illustrates 'Baggage Check-in Handling System' and shows the process followed between the system object. The traveler shall have an account by signing up into the application. Once the account has been created, the traveler will log into the application. After that, he\she will enter the booking information, and choose the appropriate time from the available time slot for booking. The driver will come at the specified date and time and pick up the baggage then the traveler will track his/her baggage through the map in the application until the driver checks them in. Drivers will have an account given by the admin. After that, the driver can sign in and a find booking schedule in his account and the traveler's information for each booking. The admin is given the ability to modify the traveler's, driver's and new admin's information, and edit all the appointments and assign them to the drivers.

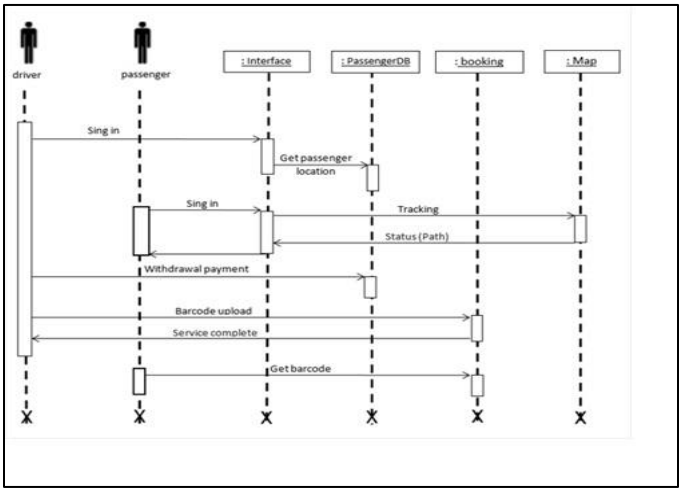


Fig. 7. Application sequence diagram.

C. Class Diagram

Mainly, Fig. 8 shows six classes of the system which are: Admin, Traveler, Driver, Appointment, Baggage, and Tracking.

login (): A function that verifies the username and password of travelers, and drivers.

addFamily (): Allows the registered traveler to add his/her family members.

setBooking (): Allows the travelers to reserve the service by setting time, date, and location.

setTrackRoute(tRoute): Allows the system to set the actual route into the system.

setScanBarcode (): Allows the driver to scan the baggage number and barcode, then send it to the traveler.

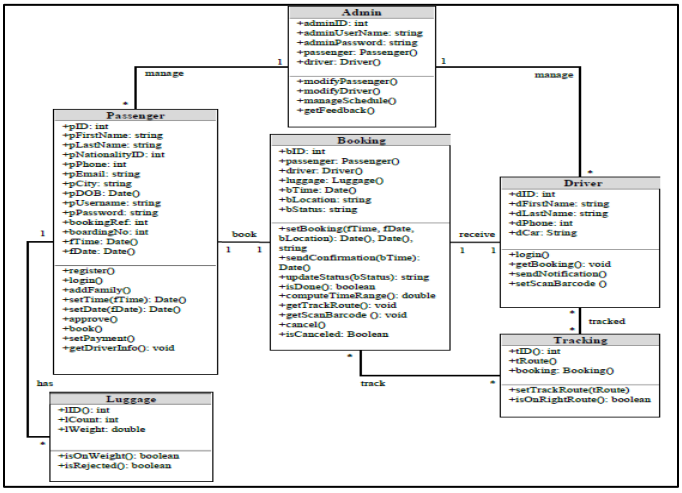


Fig. 8. Class diagram.

V. Application Implementation

This section is about the implementation of the system, and the tools used to implement tasks, such as; GPS (Global Positioning System), the database, and other functions. In some parts, android libraries are used to support and improve the functionality.

A. Objectives

The success criteria [20] of the proposed application is to minimize the possibility of missing a flight because of the time taken to check-in, also to relieve travelers from the check-in stress and decrease their effort in trying to fit in time to check-in.

B. Implementation Tools

The implementation tools that are used in the proposed application are Android Studio tool, the official integrated development environment (IDE) for Android platform development, Geographical Position System (GPS), phpMyAdmin to handle the administration of MySQL using a web browser (It can perform various tasks such as; creating, modifying or deleting databases, tables, fields or rows; executing SQL statements; and managing users and permissions) , and Graphical User Interface Tools were used which are Java programming language made for defining and activating the XML layout (The XML language used to define the main structure of the interface). The authentication technique called Secure Hash Algorithm (SHA) is used to authenticate users and to reduce the overhead of access control, it is a cryptographic hash function used for handling a compressed representation of a message. Also, it is particularly improbable to find two messages hashing to the same value.

C. Implementation Process

- Creating database tables with all relations and constrains.
- Setting the phpMyAdmin files to manipulate the database from the android studio code.
- Designing the layout of the application interfaces.
- Prepare the code for each layout.
- Filling the database with information.
- Authentication of users' information.
- Testing the application.

D. Results and Discussions

As shown in Fig. 9, the screens of the 'Baggage Check-in Handling System' are vertically positioned, with large icons and clear text. The screens reveal information about the application scenario in which during the booking's date and time the passenger will be notified by the status in the form of a notification, beginning with the driver's arrival until the successfully checked in baggage. When a passenger books an appointment, the specified driver will be notified then decides whether to approve or decline the appointment. Once the driver picks up the baggage, the passenger will track him until he reaches the airport. Passengers can contact the driver by clicking on "Contact driver". When the baggage is checked in, the driver will scan the barcodes through the application, then it will be saved to the passenger's profile.

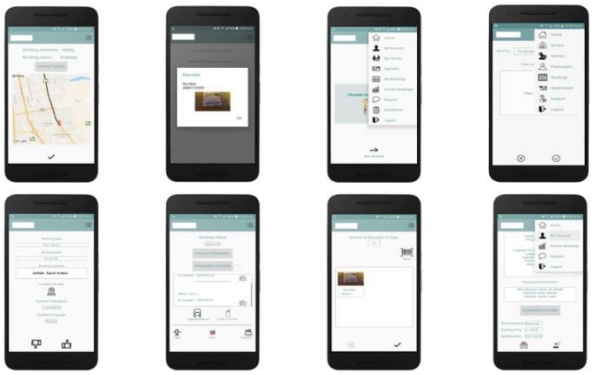


Fig. 9. Some screens of the 'Baggage Check-in Handling System'. (a) Tracking the Driver. (b) Baggage Barcode Number. (c) Traveler Menu. (d) Driver Menu. (e) Grid View of all Bookings. (f) Booking Status. (g) Scanning Barcode. (h) Traveler Information.

VI. Usability Study

Usability study is a way of seeing how easy it is to use the application by testing it with real users [21]. Users are asked to complete tasks while being observed to see where they encounter problems and experience confusion, and to check if it meets the user's expectations. If more users encounter similar problems, recommendations will be made to overcome these usability issues. In order to evaluate the usability of the proposed system, a usability study is conducted in which five participants were asked to perform a set of tasks. The focus was on the main features of the proposed system, which are registering as a new user, logging in, booking an appointment, adding a new family member, viewing booking information, tracking the driver, and adding a complaint. As results, it was observed that the executions of the different tasks revealed the ease of using the application's main functionalities.

VII. Conclusions and Future Work

In this paper, we are seeking to provide an airport check-in service for baggage through a mobile application. This service allows a traveler to make an appointment before the flight's take off by about three to five hours, by requesting a trusted driver to pick up the passenger's baggage from the selected location to the airport, so that the passenger could have the choice of directly proceeding to the flight's gate. The application is developed for all passengers but especially for people that get tired easily like elderly people, pregnant women, and families with heavy baggage. The application is created for Android platform operating system, and developed in Java programming language using the Android software development kit (SDK). Furthermore, the authentication of users is done using Secure Hash Algorithm (SHA). The results reveal the ease of using the application's main functionalities. We know that security is the main concern when using the application. The future works are supporting multiple operating systems rather than only Android, providing multiple languages, and being sponsored and adopted by a wide company to solve the lack of security.

Acknowledgments

I would like to thank the students Shahad Sait, Sarah Al-Ghamdi, Amna Moussa, Shorouq Basnawi, and Mona Abdulbaqi of King Abdulaziz University, KSA, for their contribution in the development of the 'Baggage Check-in Handling System'.

References

[1] S. Jain, R. R. Creasey, J. Himmelspach and K. P. White, "Check-in Processing: Simulation of Passengers with Advanced," in Winter Simulation Conference, AUSTRALIA, 2011.

[2] Gillen, David and W. G. Morrison, "Aviation security: Costing, pricing, finance and performance.," Journal of Air Transport Management , vol. 48, pp. 1-12, 2015.

[3] Tuan. and T. D. Cao, "Improving travel information access with semantic search application on mobile environment," in In Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia, New York, NY, USA, 2011.

[4] A. Smirnov, A. Kashevnik, N. Shilov, N. Teslya and A. Shabaev, "Mobile application for guiding tourist activities: tourist assistant-tais," in Open Innovations Association (FRUCT16), 16th Conference , 2014.

[5] MacMillan and Douglas, "Uber touts its employment opportunities," The Wall Street Journal, 2015.

[6] CIRSTEA and ALICE, "The Implications of Mobile Commerce Applications," International Journal of Scientific Knowledge, vol. 6, no. 2, May 2015.

[7] Adewumi and Adewole, "Developing a mobile application for taxi service company in Nigeria," in 2015 International Conference on Computing, Communication and Security (ICCCS), 2015.

[8] P. Verma and J. Bhatia, "Desing and Development of GPS-GSM Based Tracking System with Google Map Based Monitoring," nternational Journal of Computer Science, Engineering and Applications (IJCSEA) , vol. 3, no. 3, pp. 33-40, 2013.

[9] V. S and Kirthika.B, "Android Operating System: A Review," International Journal of Trend in Research and Development, vol. 2, no. 5, pp. 260-264, 2015.

[10] G. Pandey and D. Dani, "Android Mobile Application Build on Eclipse," International Journal of Scientific and Research Publications, vol. 4, no. 2, 2014.

[11] P. D. Team, "PHP: Security and Safe Mode - Manual," http://php.net/manual/en/features.safe-mode.php.

[12] Wade Trappe, Lawrence C. Washington.. *Introduction to Cryptography with Coding Theory*. New Jersey: Pearson Prentice Hall, 2006.

[13] A. Finkelstein, J. Kramer, B. Nuseibeh and L. Finkels, "Viewpoints: a framework for integrating multiple perspectives in system development.," International Journal of Software Engineering and Knowledge Engineering, pp. 31-57, 1992.

[14] Davis and A. M., Software Requirements: Objects, Functions, & States, Prentice-Hall, 1993.

[15] Hausman and Angela, "A multi - method investigation of consumer motivations in impulse buying behavior," Journal of Consumer Marketing, vol. 17, no. 5, pp. 403-426, 2000.

[16] Cleland-Huang and Jane, "Toward improved traceability of non-functional requirements," in TEFSE '05 Proceedings of the 3rd international workshop on Traceability in emerging forms of software engineering, California, 2005.

[17] Qusef, A. De Lucia and Abdallah, "Requirements Engineering in Agile Software Development," Journal of Emerging Technologies in Web Inteliigence, vol. 2, no. 3, August 2010.

[18] Pressman and R. S., Software Engineering: A Practitioner's Approach, 2005.

[19] JingLiu, JoshDehlinger and RobynLutz, "Safety analysis of software product lines using state-based modeling," Journal of Systems and Software, vol. 80, no. 11, pp. 1879-1892, 2007.

[20] NitinAgarwal and UrvashiRathod, "Defining 'success' for software projects: An exploratory revelation," International Journal of Project Management, vol. 24, no. 4, pp. 358-370, 2006.

[21] M.MorenoaMaria and IsabelSanchez, " Analysing the impact of usability on software design," Journal of Systems and Software,vol. 80, no. 9, pp. 1506-1516, 2007.

Prioritizing Road Maintenance Activities using GIS Platform and Vb.net

Fardeen Nodrat
Graduate School of Engineering and Science
University of the Ryukyus
Okinawa, Japan

Dongshik Kang
Department of Information Engineering
University of the Ryukyus
Okinawa, Japan

***Abstract*—One of the most important factors for the sustainable development of any country is the quality and efficiency of its transportation system. The principled and accurate maintenance of roads, in addition to having a major impact on budget savings, improves the quality and service levels of the transportation system. For this reason, road management and maintenance are the main pillars of the transportation system in any country. Nowadays, due to the increased cost of maintaining roads and the lack of funding in this area, traditional ways of managing and maintaining roads, which are more based on the experience of the experts themselves, are no longer affordable. Hence, more recent, and more systematic methods have become more popular among relevant authorities. Afghanistan is a country facing problems such as budget deficits, lack of professional experts and advanced technology in road maintenance sector. This paper presents an example of using the GIS platform and vb.net to prioritize the road maintenance and rehabilitation activities based on identified criteria. A case study conducted in an academic environment and road maintenance and rehabilitation activities prioritized. The results show that the positive criterion has the greatest impact on the ranking of road maintenance activities. The characteristic of this process is to help the decision makers to plan road maintenance requirements to effectively and efficiently allocate funds for future planning.**

Keywords*—Road maintenance; prioritization; GIS; Vb.net; Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

I. INTRODUCTION

Afghanistan is lacking proper road maintenance which resulted in a huge amount of investment loss in the past 10 years. The emphasis of the government and donors was on the development of new roads regardless of understanding and focusing the maintenance ability and budget. It is therefore over 4 billion USD investment in road assets is in the threat of loss. Recently the government and donors have recognized the issue and have focused to strengthen the ability in the maintenance sector. The Ministry of Public Works handles the management and development of national and regional highways. The expert staff previously existed in mentioned ministry has drastically decreased due to three decades of war. The increasing length of national and regional highways day by day is another challenge for the Government to maintain [1].

Road Maintenance is to preserve as closely as could reasonably be expected, the original designed condition of paved and unpaved roadways, and of traffic signs, signals and markings, in a manner most likely to minimize the total cost to society of vehicle operation and accident cost, in addition to the cost of giving the maintenance itself, under the requirements of serious asset confinements, in regard of skilled manpower, equipment and money, both local and foreign [2].

Pavement Management System (PMS) helps staff in assessing, tracking and grading pavement conditions in view of field examinations. The recurrence of roadway examination ranges from yearly to once like clockwork relying upon the kind of roadway. Comprehensive field assessments sort and evaluate pavement insufficiencies, for example, cracks, patches, and utility trench cuts. These inadequacies are inserted into the PMS Program that figures a Pavement Condition Index (PCI) for every roadway. PCI values extend from zero (very poor) to 100 (excellent) [3].

Priority ranking, as used as a part of PMS, is a procedure used to rank the pavement segments in a request of earnestness for maintenance and rehabilitation. The prioritization procedure is the fundamental phase of PMS before the decision makers take an official conclusion on the execution of maintenance program. The nature of Priority-setting is straightforwardly affecting the adequacy of accessible assets which are, much of the time, the essential justice of the decision maker. The priority ranking procedure relies upon different components like pavement condition, traffic volume, environmental effects, predicted execution standards, and budgetary requirements. Since maintenance activities influence the planning of work and assignment of assets, proper choice of such activities (priority) is critical to the most productive use of constrained assets [4].

In this paper, efforts have been made to prioritize roads maintenance and rehabilitation activities using the MRAP tool. With the growth of the alternatives, the measurement of problems increases respectively. This requires different mathematical operations to get the ideal answer. The advantages of MRAP tool are to simplify the prioritization process and also generate a different database for use in ArcGIS to generate thematic maps.

A case study in an academic environment was conducted in the study area. All prioritization processes were performed using the GIS platform and vb.net. The characteristic of this process is to help the decision makers to plan road maintenance requirements to effectively and efficiently allocate funds for future planning.

The research includes the following sections: introduction, research background, introduction of prioritization approaches, introduction of TOPSIS and Shannon entropy methods, introduction of MRAP tool, data collection, calculation of criteria values, prioritization of maintenance activities using the MRAP tool, the preparation of thematic maps using GIS, results of the experiment and conclusions.

II. Literature Review

There are numerous pieces of literature studies about the prioritization strategies utilized in road maintenance sector. Each prioritization technique has diverse methods reflected by some successive steps in ranking a set of alternatives. Frequently, there are four basic steps used as a part of prioritization process as clarified below:

1) Determining the evaluation criterion.

2) Establishing performance criteria for calculating project compliance with these criteria.

3) Somehow combining the scores of each performance measurement.

4) Project rankings in order of importance.

Each strategy can't be considered to each case and place because there will be distinctive contemplations and circumstance confronted. The following are the four methods which proposed by Hudson et al.:

1) The simple subjective ranking, this technique just depends on the judgments and encounters of decision makers and can be led by utilizing matrix and decision tree. This method is a conventional strategy in which the selections of roads depend on the encounters and subjective judgments of road engineers. In this manner, it can be conducted rapidly. The subjective ranking includes a subjective evaluation of how each task is identified with objective accomplishment by deciding cost-adequacy measures of "high, medium, and low". For this situation, there is no analytical tool utilized as a part of selecting the roads to be maintained. In this manner, the needs came about tends to be predisposition and irregularity, a long way from ideal.

2) Ranking based on parameters with scoring and/or weighting, this technique is likewise straightforward, simple to utilize, and snappy yet the outcomes might be a long way from ideal. In the field of road maintenance, there are some priority evaluation scheme in creating priority rating scores as per certain numerical composite indexes, for example, defects rating index, pavement condition index, maintenance need index, rate, priority, and fuzzy condition index. In any case, a large portion of that scheme concentrates just on the pavement condition. It causes the prioritization comes about are a long way from ideal. Hence, alternate strategies utilizing multi-criteria wind up well known. Ranking according to multi-criteria can limit the subjective components that are overwhelming in the decision-making process for planned maintenance and can build the straightforwardness of the prioritization procedure which in the end will enhance open responsibility. Along these lines, prioritization in light of parameters is superior to anything prioritization in light of a parameter.

3) Ranking based on parameters with economic analysis, this technique is the most surely understood strategy in prioritization process since this strategy is sensibly straightforward. The decision-making devices that can be utilized as a part of this strategy are benefit/cost ratio, life cycle cost analysis, or cost-effectiveness. By and by, it changes all maintenance elements to equal money related esteems, and after that uses an economic index to assess the alternative projects with the goal that it ought to be nearer to ideal. In any case, it is hard to gauge every single pertinent effect of a project in cash terms. Along these lines, this technique needs an exhaustive investigation.

4) Optimization, this strategy is very perplexing and regularly be the most tedious technique. Then again, it has the advantages of delivering the ideal decision in which it maximizes the benefit and limits the costs. Other than that, the optimization procedure considers both time (present and future) and space (whole system).

III. Prioritization Approach

A. TOPSIS Model

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multicriteria decision analysis method, which was originally developed by Hwang and Yoon in 1981 [5] with further developments by Yoon in 1987 [6], and Hwang, Lai and Liu in 1993 [7].

TOPSIS is based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS) [8].

It is a method of compensatory aggregation that compares a set of alternatives by identifying weights for each criterion, normalizing scores for each criterion and calculating the geometric distance between each alternative and the ideal alternative, which is the best score in each criterion. An assumption of TOPSIS is that the criteria are monotonically increasing or decreasing. Normalization is usually required as the parameters or criteria are often of incongruous dimensions in multicriteria problems [9], [10].

Compensatory methods such as TOPSIS allow tradeoffs between criteria, where a poor result in one criterion can be negated by a good result in another criterion. This gives a more realistic form of modeling than non-compensatory methods, which include or exclude alternative solutions based on hard cutoffs [11].

The TOPSIS method evaluates the following decision matrix which has *m* alternatives associated with *n* attributes (or criteria):

$$\boldsymbol{D} = \begin{array}{c} \\ \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \\ \vdots \\ \boldsymbol{A}_i \\ \vdots \\ \boldsymbol{A}_m \end{array} \begin{array}{c} \begin{array}{cccccc} \boldsymbol{B}_1 & \boldsymbol{B}_2 & & \boldsymbol{B}_j & & \boldsymbol{B}_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mj} & \dots & x_{mn} \end{bmatrix} \end{array} \quad (1)$$

Where,

$\boldsymbol{A}_i$ = the *i*th alternative considered,

$\boldsymbol{B}_j$ = the *j*th criterion considered,

x_{ij} = the numerical outcome of the *i*th alternative with respect to the *j*th criterion.

TOPSIS assumes that each attribute in the decision matrix takes either monotonically increasing or monotonically decreasing utility. Since all criteria cannot assume to be of equal importance, the method receives a set of weights from the decision maker. For the sake of simplicity, the proposed method will be calculated as a series of steps.

B. Shannon Entropy Method

The majority of the TOPSIS applications to real-world decision-making issues use just subjective weights defined by the decision makers. Which means, a set of weights $\boldsymbol{W} = (w_1, w_2, \dots, w_j, \dots, w_n)$, $\sum_{j=1}^{n} w_j = 1$, will be determined by the decision makers. Be that as it may, when it isn't conceivable to acquire dependable subjective weights, objective weights wind up plainly helpful. One of the techniques for getting objective weights is the use of the well-known method of Shannon entropy [12].

The entropy is a term of information theory, which is otherwise called the average (expected) measure of data contained in every criterion (each column of the decision matrix (1)). The higher the value of entropy is in a specific criterion, the lower is the differences in the ratings of alternatives regarding its criterion. This, thusly, implies this criterion gives fewer data and has a little weight. So, this criterion turns out to be less important in the decision-making process. The calculation process of the Shannon entropy as below:

1) Construct the normalized decision matrix $\boldsymbol{R} = r_{ij}$,

$$r_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \quad (2)$$

Where, $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$.

Note that x_{ij} is the performance rating of the *i*th alternative A_i, with respect to the *j*th criteria $\boldsymbol{B}_j$ and w_j represent the weight of the *j*th criteria $\boldsymbol{B}_j$.

2) Construct the vector of the Shannon entropy $\boldsymbol{e} = (e_1, e_2, \dots, e_j, \dots, e_n)$,

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^{m} r_{ij} \ln r_{ij} \quad (3)$$

Where, $r_{ij} \ln r_{ij}$ is defined as 0 if $r_{ij} = 0$.

3) Calculate the vector of diversification degrees $\boldsymbol{d} = (d_1, d_2, \dots, d_j, \dots, d_n)$,

$$d_j = 1 - e_j \quad (4)$$

The higher the degree d_j, the more important the corresponding criterion $\boldsymbol{B}_j$.

4) Calculate the vector of criteria weights $\boldsymbol{W} = (w_1, w_2, \dots, w_j, \dots, w_n)$,

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \quad (5)$$

C. MRAP Tool

The Maintenance and Rehabilitation Activity Prioritization - MRAP Tool created using Visual Studio 2015. The main purpose of the developed tool is to prioritize maintenance and rehabilitation activities using the TOPSIS model. Initially, after data entry, the tool performs the entire computing process, which is TOPSIS method and provides the user with an optimized prioritization table. While this tool prioritizes the alternatives, hence, there must be at least two alternatives with at least two criteria to use this tool. Fig. 1 shows the main screen of the tool.

Fig. 1. MRAP tool main window.

IV. Study Area

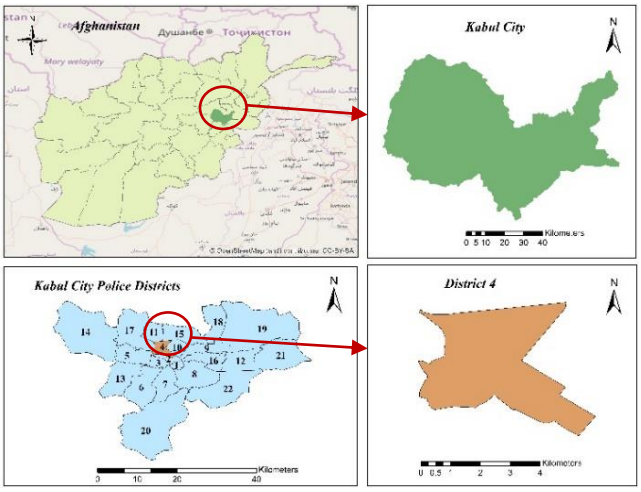


Fig. 2. Study area.

The study area covers District 4 of Kabul city. In the capital, the main residential area districts are, districts 4, 6, 11, 14, 16 respectively, and some suburban area outside of Kabul, the commercial districts are districts 1, 2, 3, 7, 9 and 10. Thus, many of the business and residential districts are located on the opposite side of the city. Because the accessibility from the west to the east is limited by only four streets, they experience significant congestion [13].

The District 4 constitutes the main extension area of the city center. The suburbanization has been continuing northwestward along Salang Watt Street from District 2 and further into District 17. This district is bordered on the planned residential areas of District 11 and District 15 to the north, and District 10 to the east [14].

The District 4 has a land area of 11.63 km2, of which 83.1% is an urban area. Moreover, 54.2% of the urban area is of high density. District 4 is, in fact, the most densely developed and populated of all the districts in the city. The district has small agricultural land covering 11.5 ha or 1.0% of the district land for rain-fed agriculture and orchards. Vacant land occupies only 6.0% due to the hills along the southwestern border. The Macro-rayon housing complex occupies 6.6 ha. Lively formal and informal markets are found in the suburbanized area along the main street. Fig. 2 shows the location of our study area.

V. Data Collection

A. Data Requirements

The next step in the preparation phase is the identification of the data requirements. All data related to research needs have been collected and extracted in connection with the achievement of research aims. Data types are:

- Road general data
- Road engineering data
- Traffic volume
- Pavement condition index
- Road maintenance data
- Road distress information data

B. Required Tools

In this study, we collected and extracted data through government agencies, via the Internet, and field surveys. Due to the lack of resources in government agencies, we must extract some data from the Internet. The following tools are used to conduct field surveys and collect physical road data as well as collect and extract data in order to achieve the research goals:

- Survey datasheet
- Tape meter
- Digital camera
- Laptop
- Internet
- Paver 5.2 (demo version)
- MRAP tool
- ArcGIS 10.4

C. Road Engineering Data

We collected and extracted the location map and the base map of the study area from the Kabul Municipality and the Internet (basically Google Earth and OpenStreetMap). Fig. 3 shows the location and base map in our study area.

Road engineering data is usually included (road name, section name, width, length, type of surface, classification, traffic volume, ..., etc.). We collected key road information from the Kabul Municipality, as well as extracted some physical road information from the Internet. The basic information about our study area is shown in Table I. Fig. 4 shows the illustration of our study area.

TABLE I. Roads Engineering Information

ID	Road's name	Sec.	Width	Length	Surface
1	Shahid Rd	1	10	207	Asphalt
2	Shahid Rd	2	10	214	″
3	Sulh Rd	1	10	132	″
4	Sulh Rd	2	10	116	″
5	Sulh Rd	3	10	126	″
6	Kulola Pushta Rd	1	10	135	″
7	Kulola Pushta Rd	2	10	158	″
8	Kulola Pushta Rd	3	10	131	″
9	Shahr e Naw St	1	7.5	131	″
10	Shahr e Naw St	2	7.5	83	″
11	Shahr e Naw St	3	7.5	78	″
12	Shahr e Naw St	4	7.5	127	″
13	Ansari 1 St	1	8.5	209	″
14	Ansari 1 St	2	8.5	204	″
15	Ansari 2 St	1	8.5	211	″
16	Ansari 2 St	2	8.5	205	″
17	Ansari 3 St	1	8.5	211	″
18	Ansari 3 St	2	8.5	205	″
19	Ansari 4 St	1	10	206	″
20	Ansari 4 St	2	10	208	″

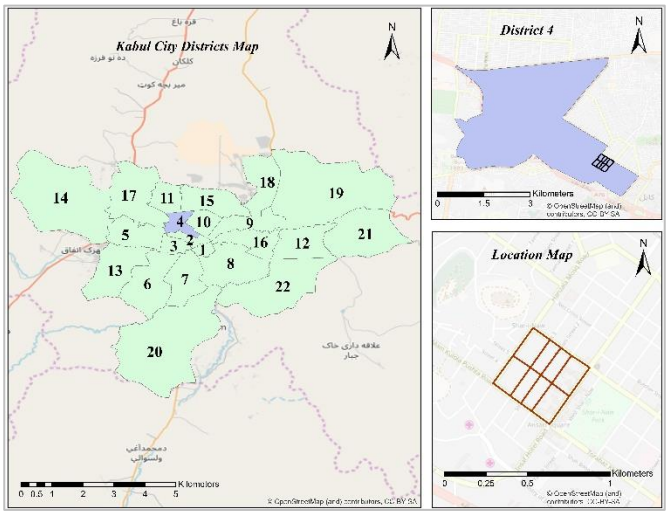


Fig. 3. Location and base map.

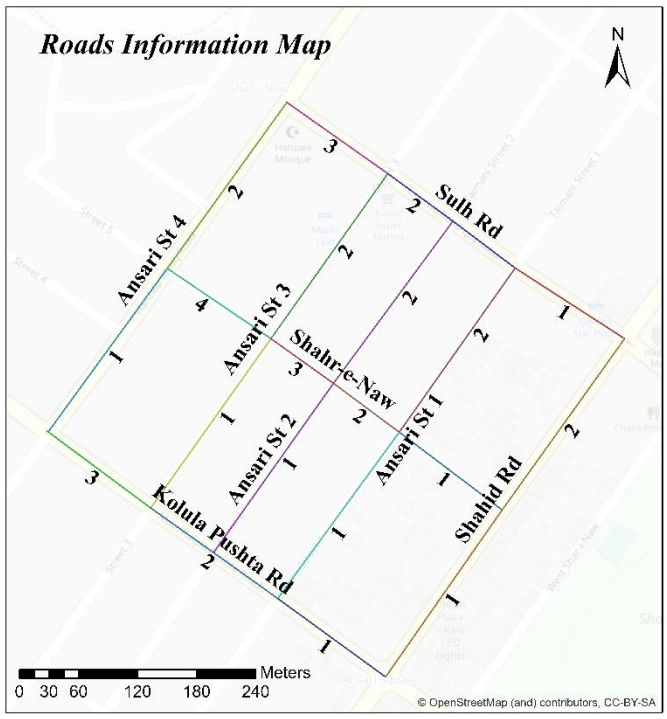


Fig. 4. Roads information map.

D. Fieldworks

Basically, the fieldwork involves a visual inspection of road distresses to calculate the PCI. To carry out the inspection activities, we prepared a survey datasheet based on the TM 5-623 form of Corps of Engineering. In the first step, after identifying the distress, we collected all the measurements (Length, area, severity, … etc.), and then the results were included in pre-prepared forms.

VI. Material And Method

A. Criteria Values Calculations

After research and studies and discussing with road maintenance experts, we have identified the following criteria for prioritizing the road maintenance and rehabilitation activities in the study area:

1) Pavement condition index
2) Traffic volume
3) Roads width
4) Political, social and cultural importance factor (IF)
5) Maintenance and rehabilitation cost

It should be noted that from the above criteria the 1-4 criteria are positive (+) and the fifth criteria are the negative (-) criteria. That means, in the positive criteria, the highest quantities have the highest priority, such as PCI. But in the negative criteria, the situation is quite the opposite. That is, in the negative criteria, the higher the number, the lower the priority.

The next step is to find the values for all the above-mentioned criteria to prepare the decision matrix.

1) Pavement condition index (PCI):

We used the Paver 5.2 (Demo version) for calculation of the PCI. After we collected the road distresses physical information, we enter those data into the Paver software, and we can easily calculate the PCI, which is a numeric value from 0-100 and the highest values show the road with good condition. As all the road constructed in 2014 so there are no major problems and based on the PCI values it shows all the roads are in good conditions with minor treatments.

2) Traffic volume:

According to a 2005 Kabul vehicle census, a total of 341,047 vehicles is registered consisting of mostly small cars (66.2%) followed by trucks. The previous 2004 percentages were approximately 49% passenger cars, 20% trucks, 15.5% motorcycles, 10% buses, 3% foreign vehicles and 1.5% rickshaws. The vehicle population increases by approximately 11% annually. Illegal importation of used vehicles is a major problem in the country. It is estimated that about 300,000 of these vehicles exist and most of them are in Kabul. The United Nations Integrated Regional Information Networks (IRIN) reports that every month 8,000 new vehicles are registered with the Kabul Traffic Department, adding to Kabul's one million vehicles. This report estimated 1,224,000 vehicles as of 2010. The narrow roadways of Kabul, built more than three decades ago are now a victim of massive traffic. The road system of Kabul was originally designed for only 25,000 to 35,000 vehicles a day and is not sufficient for the rapidly increasing number of vehicles. There is also no ring road to support the dense traffic in the city center [14]. We collected the traffic volume (TF) data from Capital Region Independent Development Authority – CRIDA.

3) Roads width:

We collected all road width from the Kabul Municipality.

4) Political, social and cultural importance factor:

In this research, the political, social & cultural importance factor in the study area, was measured through the expert views. We asked 10 experts in this area to give numbers 1 to 10 for all our study area based on their political, social and cultural aspects. Then we calculated the important factor (IF)

for all the road sections using the MS Excel, by averaging all the factor values.

5) Maintenance and rehabilitation cost:

For calculating the maintenance and rehabilitation cost for each section of the roads, depending on the type and amount the of the distresses, as well as the maintenance and rehabilitation required activity, by using the Cost-Effective Pavement Preservation Solutions for the Real-World book [15], we can easily calculate the maintenance and rehabilitation cost for each section of the roads. All the costs are in an Afghani currency with an exchange rate of 1$ = 70 Afghani (08/Jan/2018).

Table II shows the criteria values or the decision matrix for all road sections, which we achieved from the above steps while we obtained the decision matrix by calculating weights.

The next step is to calculate the weight of the decision matrix for the prioritization process. For this purpose, we perform the weighing calculation process using well-known Shannon entropy method [12] and MS Excel software. Table III shows the weight criteria for the prioritization.

TABLE II. DECISION MATRIX

ID	Road's name	Sec.	PCI	TF	W	IF	Cost
1	Shahid Rd	1	100	11074	10	7.15	1,809
2	Shahid Rd	2	99	11074	10	6.87	1,777
3	Sulh Rd	1	99	10902	10	7.28	1,809
4	Sulh Rd	2	98	10902	10	7.20	2,233
5	Sulh Rd	3	99	10902	10	7.53	1,712
6	Kulola Pushta Rd	1	95	7315	10	6.98	2,298
7	Kulola Pushta Rd	2	92	7315	10	6.98	2,162
8	Kulola Pushta Rd	3	95	7315	10	7.65	2,135
9	Shahr e Naw St	1	100	1825	7.5	7.62	897
10	Shahr e Naw St	2	100	1825	7.5	7.67	456
11	Shahr e Naw St	3	100	1825	7.5	7.24	293
12	Shahr e Naw St	4	100	1825	7.5	7.65	1,223
13	Ansari 1 St	1	100	2030	8.5	7.55	978
14	Ansari 1 St	2	94	2030	8.5	7.18	998
15	Ansari 2 St	1	97	2120	8.5	7.13	2,517
16	Ansari 2 St	2	97	2120	8.5	6.98	2,130
17	Ansari 3 St	1	100	2090	8.5	7.25	1,206
18	Ansari 3 St	2	100	2090	8.5	7.05	1,141
19	Ansari 4 St	1	94	3323	10	6.45	1,356
20	Ansari 4 St	2	98	3323	10	6.68	1,997

TABLE III. WEIGHT CRITERIA FOR PRIORITIZATION

Criteria	Weights	Mark
PCI	0.0009	+
Traffic volume	0.7279	+
Width	0.0175	+
IF	0.0029	+
Cost	0.2508	-

B. Road Maintenance Activity Prioritization

After obtaining the decision matrix values and the weights table, it is time to use the MRAP tool to prioritize the alternatives. Therefore, by entering the values of the criteria, the decision matrix will be formed, we enter the values of the weights table and identify the negative weights. Subsequently, the software begins to calculate the TOPSIS model to prioritize alternatives based on the closeness coefficient of the ideal solution. Finally, an optimized prioritization table is being prepared. Now we can export the table to different formats for use in GIS platform to prepare the desired thematic maps. Table IV shows the roads maintenance and rehabilitation activity prioritization.

TABLE IV. PRIORITIZING ROADS USING MRAP TOOL

Road's name	Sec.	Closeness coefficient	Ranking
Sulh Rd	3	0.8483	1
Shahid Rd	2	0.8463	2
Shahid Rd	1	0.8428	3
Sulh Rd	1	0.8394	4
Sulh Rd	2	0.8040	5
Kulola Pushta Rd	3	0.5607	6
Kulola Pushta Rd	2	0.5599	7
Kulola Pushta Rd	1	0.5552	8
Shahr e Naw St	2	0.2311	9
Shahr e Naw St	3	0.2154	10
Ansari 4 St	1	0.2029	11
Shahr e Naw St	1	0.1669	12
Ansari 1 St	1	0.1634	13
Ansari 1 St	2	0.1610	14
Ansari 3 St	2	0.1493	15
Ansari 3 St	1	0.1448	16
Shahr e Naw St	4	0.1368	17
Ansari 4 St	2	0.0693	18
Ansari 2 St	2	0.0539	19
Ansari 2 St	1	0.0295	20

C. Thematic Maps Preparation

Finally, after finalizing all the data collection & analysis, preparing decision matrix as well as weight table and prioritizing the maintenance and rehabilitation activity prioritization, it is time use all this information and produces the thematic maps in order to have a visual illustration of our study area. For doing this, we used the ArcGIS 10.4. Fig. 5 to 8 shows the thematic maps for various information.

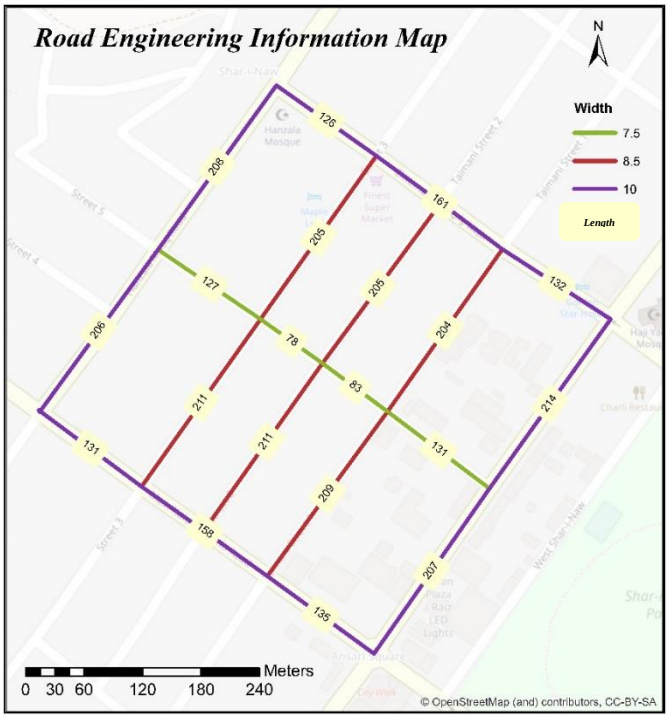


Fig. 5. Road engineering information map.

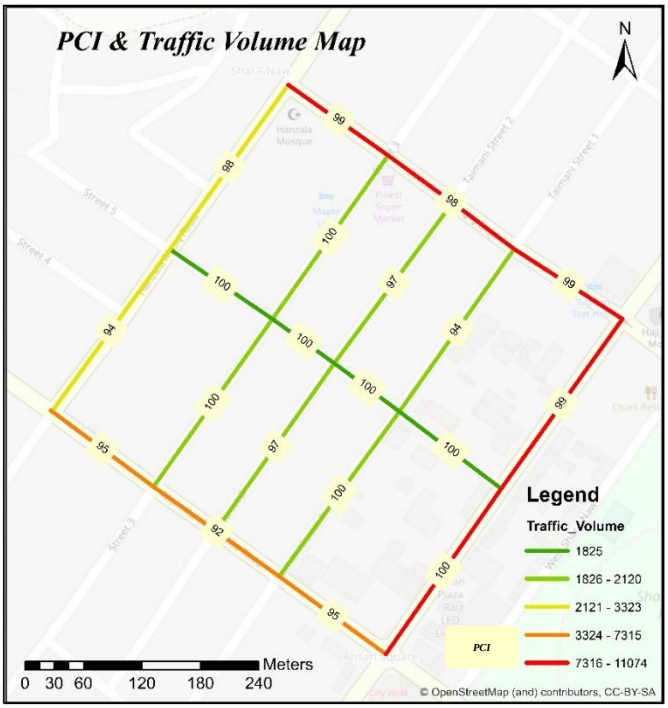


Fig. 6. Roads PCI and traffic volume map.

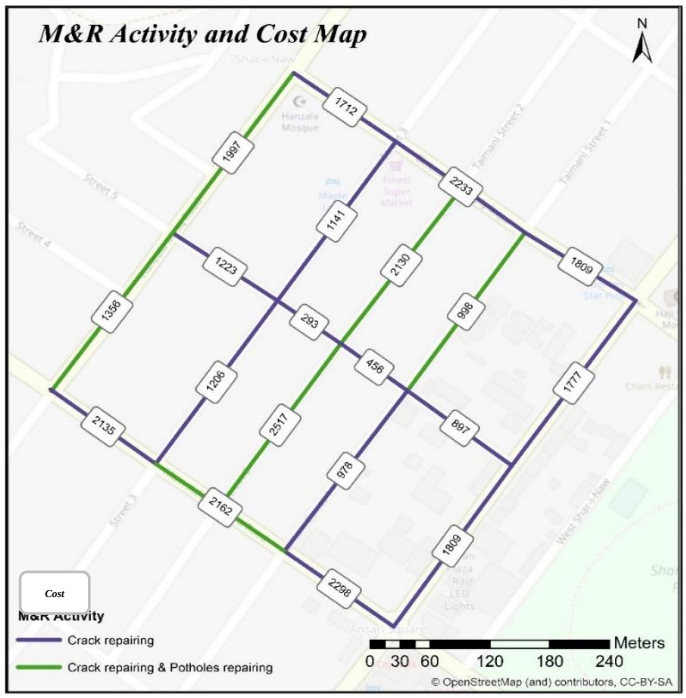


Fig. 7. Roads maintenance and rehabilitation and cost map.

Fig. 8. Roads maintenance and rehabilitation activity ranking map.

D. Results

The results show that all roads are in a good condition, with minor distresses, which, with a low budget and small maintenance activities, could return to the ideal situation. In addition, according to the results, the traffic volume on secondary roads is way more than the residential roads.

The experiment has shown that GIS software has the capability to visually display the results, which enables the decision makers to better understand the area and make better

decisions. The results also show that the MRAP tool, with a very simple process and the least resources, to prioritize maintenance activities can be a reliable tool with valid and reliable results.

According to the results, from all the criterion (PCI, width, traffic volume, IF and M&E cost) the M&E cost has a negative impact on our selection, means that, lower the cost is better for selection. While the other criterion is having a positive impact, means that the higher the values are better for our selection.

According to the results obtained, among all the alternatives the Sulh Rd, section 3, among the other top priority options, with a small margin, ranked first to be carried out due to the low cost of M&R activity, 1,712 Afghani accordingly.

From the experiments carried out, we conclude that the decision-making stage and setting of criteria for prioritizing maintenance activities are very important. That is, to the extent that the criteria are precise and decisive, as well as the ideal and valid results are obtained.

VII. Conclusions

In this case study, the main goals were to prioritize the maintenance and rehabilitation activity as well as demonstrating all the outcomes through thematic maps by using the ArcGIS software. We conclude the results and outcomes of our case study as below:

- There is a total of 3,297m roads in our study area, from secondary and residential classifications, 1,633m and 1,664m respectively.
- There are 20 roads (alternatives) with five criterions (pavement condition index, roads width, road traffic volume, political, social & cultural importance factor and maintenance and rehabilitation activities costs) for prioritizing the maintenance and rehabilitation activities.
- Survey inspections carried out in order to collect the distress information from the roads for calculation of the PCI. The Paver 5.2 (demo version) used for calculating the PCI.
- The political, social and cultural importance factor questionnaire has been developed and ten experts gave their idea for scoring the roads importance factor. Later, by using Ms. Excel software and by averaging up expert opinions, the final score for important factor was calculated.
- For obtaining the costs for maintenance activities, Ms. Excel software and the Cost-Effective Pavement Preservation Solutions for the Real-World book were used, which the total cost of maintenance in the study area is a total of AFN 31,127.
- To obtain the weights for the decision matrix, the famous Shannon Entropy method and Excel software were used.
- For prioritizing the maintenance and rehabilitation activities, the MRAP tool was used, which had reliable outcomes. Also, to use this tool, at least 2+ alternatives are required with more than two criteria.
- The integration of the TOPSIS model, as a multi-criteria decision-making method, can be used to optimize/rank the maintenance activities.
- Various thematic maps produced using ArcGIS 10.4, by using all the outcomes of the case study.

Acknowledgment

The Authors would like to thank from the University of the Ryukyus, Japan International Cooperation Agency (JICA) Project for the Promotion and Enhancement of the Afghan Capacity for Effective Development (PEACE) for providing found and giving a chance for the improvement of knowledge scales and the government of Afghanistan for its cooperation in this study.

References

[1] "Ministry of Public Work," MoPW, [Online]. Available: https://www.mopw.gov.af/fa. [Accessed 23 March 2017].

[2] "The World Road Association," PIARC, [Online]. Available: https://www.piarc.org/en/. [Accessed 23 March 2017].

[3] "Sacramento County Department of Transportation," SACDOT, [Online]. Available: https://www.sacdot.com/Pages/default.aspx. [Accessed 23 March 2017].

[4] E. A. Sharaf and F. M. Mandeel, "An Analysis of The Impact of Different Priority Setting Techniques on Network Pavement Condition," 4th International Conference on Managing Pavements, vol. 1, pp. 158-168, 1998.

[5] C. Hwang and K. Yoon, Multiple Attribute Decision Making: Methods and Applications a State-of-the-Art Survey, New York: Springer, 1981.

[6] K. Yoon, "A Reconciliation Among Discrete Compromise Solutions," Journal of the Operational Research Society, vol. 38, no. 3, pp. 277-288, 1987.

[7] Y. L. a. T. L. C.L. Hwang, "A new approach for multiple objective decision making," Computers & Operations Research, vol. 20, no. 8, pp. 889-899, 1993.

[8] T. M. a. E. A. A. Assari, "Role of public participation in the sustainability of historical city: usage of TOPSIS method," Indian Journal of Science and Technology, vol. 5, no. 3, pp. 2289-2294, 2012.

[9] K. Y. a. C. Hwang, Multiple Attribute Decision Making: An Introduction, California: SAGE Publications, 1995.

[10] A. Z. a. J. A. E.K. Zavadskas, "Evaluation of Ranking Accuracy in Multi-Criteria Decisions," Informatica, vol. 17, no. 4, pp. 601-618, 2006.

[11] R. D. J. L. a. B. E. R. Greene, "GIS-Based Multiple-Criteria Decision Analysis," Geography Compass, vol. 5, no. 6, p. 412–432, 2011.

[12] D. Shi-fei and S. Zhong-zhi, "Studies on incidence pattern recognition based on information entropy," Journal of Information Science, vol. 31, no. 6, pp. 497 - 502, 2005.

[13] J. I. C. A. (JICA), "Draft Kabul City Master Plan; Sector Report 05: Transport Infrastructure Development Plan," RECS International Inc., Tokyo, 2011.

[14] J. I. C. A. (JICA), "The Study for the Development of the Master Plan for the Kabul Metropolitan Area in the Islamic Republic of Afghanistan; Final Report, Sector Report 8: Transportation," RECS International Inc., Tokyo, 2009.

[15] W. J. Wilde, L. Thompson, and T. J. Wood, Cost-Effective Pavement Preservation Solutions for the Real World, Minnesota: Department of Transportation, Research Services & Library, 2014.

A Comparison of Usability Aspects between an Existing Hospital Website of Pakistan with a Template based on Usability Standards

[1]Muhammad Usman, [2]Mahmood Ashraf
Department of Computer Science
Federal Urdu University of Arts, Science & Technology,
Islamabad, Pakistan

Muhammad Tahir
Faculty of Computing and Information Technology
University of Jeddah,
Jeddah, Saudi Arabia

***Abstract*—More people search internet for medical and health information. Due to increase in demand for online health services, hospitals need to equip their websites with usability standards. Hospital websites should be user centered in order to increase the usability. In the instant research study, an existing public sector hospital website is compared with a designed template for healthcare website. Template was designed keeping in view the user demands for hospital websites. Usability evaluation of both websites has been performed. Twenty-one users were involved in the research study. Three representative tasks were performed by each user on each website and a questionnaire was presented afterwards to collect user opinion about the websites under evaluation. Average score was calculated against both websites for each usability component. 75% users responded positively to designed website template comparing with existing hospital website which got 33% positive responses only. Hence, it was evident that the designed template had better response for usability. The findings of this study justify the literature that user centered design can significantly improve usability of websites. This study is a step towards research which intends to understand usability problems and propose design rules for designing hospital websites of Pakistan in line with usability standards.**

***Keywords*—Usability evaluation; healthcare website; hospital website evaluation**

I. Introduction

The users are not satisfied with the health services and the number is increasing day by day [1]. More people search the internet for health information [2], [3]. A healthcare website can play a vital role in improving user satisfaction [4]. Users seems more satisfied with the availability of health information on the website of organization [5] but the quality of available information is not up to mark [6]. Usability is an important component for the successfulness of a website [7], [8]. User's involvement in the design can lead to developing more usable designs [9]. Hospital websites should be designed with the intent to address the users of all ages [10].

Despite the availability of number of health institutions, hospitals lag behind in providing online information [5]. The increased usage of internet for seeking medical information demands quality information [11]. Users turn towards hospitals when they feel satisfactory about their services [12] by searching for health related issues online [2], [13]. Therefore, hospitals need to provide satisfactory online services [2].

This research study intended to evaluate the usability of hospital websites. The study summarizes two prior studies and uses their finding to further understanding usability of hospital websites of Pakistan. Two user studies were performed to evaluate the private and public sector websites operational in Pakistan respectively prior to this study. Some design rules were formed in the previous user studies and users provided their specific requirements about the websites. User's expectations were recorded from earlier studies using questionnaire and direct observation methods and these provided bases for formulation of design rules. A website template on the basis of these design rules was designed including already existing guidelines for website development and a comparison was performed with an existing hospital website to interpret the usability.

The remaining part of the paper is described as: Section II shows the related work. Section III highlights the research methodology; experimental design, conduct of user study and procedure. Data analysis is discussed in Section IV. Section V reveals the results of research study. Discussion is done in Section VI and conclusion at the end.

II. Related Work

Website usability evaluation is getting focus of researchers lately [14]. Usability is a quality attribute [15]. Website designed in accordance with usability standards can significantly improve the level of user satisfaction [16].

Internet is searched increasingly to find health-related information by older adults and their caretakers but the information is not accurate every time [2]. A web portal was designed and evaluated for usability in order to address the requirements of older adults. 37 users were involved in the usability study. Different methods of participation were used like in person testing, telephonic conversation, video conferencing. In 70% cases, participants were unable to complete a task. The researchers emphasized that involvement of end user in the design is important. It was difficult to search information on the website and the difficult language of contents was among the key findings.

A research study was conducted by Raji, et al. (2013) to highlight the end user preferences on hospital websites inside Nigeria [11]. The study was focused on understanding the design features and contents in end user perspective. 100 participants in Group-A and another 100 in Group-B were involved in the study. Several themes were planned for testing the websites. The user responses were evaluated on the basis of usability heuristics. The researchers highlighted the importance of effective healthcare delivery system. The end users were unable to understand the instructions provided on the websites, however, they tend to prefer direct interaction.

People search the internet for medical problems, treatment and procedures [14]. The research investigated that the user expectations of patient oriented e-health tools on the hospital websites. 21 patient oriented e-health tools on the US hospitals websites were evaluated and 242 qualified participants were involved. The findings of research were that the websites lag behind the users' needs for interacting with hospital online. These findings can be applied to a hospital's planning to adopt e-health tools on their website. The paper concluded that the importance of understanding user's needs and preferences on the hospital websites cannot be neglected.

Gallant L et al. performed usability tests on teaching hospital in US using think-aloud collection protocol on 30 users. Usability test comprised of 34 tasks which were grouped in eight sections [5]. A user-centered design of hospital website was designed based on the user feedback. Collected data was analyzed using grounded theory approach. Users try to judge the trust, ease of use and usefulness as attributes of a hospital website were the key findings of the research. The researchers concluded that user-centered design in developing hospital website can become widely adopted.

A research study was developed to evaluate the website of healthcare institutions keeping in view the high demand of online health information [12]. Evaluation procedure was divided into three stages. The existing websites were evaluated against WCAG 2.0. The findings showed that proper procedures were not adopted while designing the websites. The accessibility of healthcare websites is poorly addressed which is misleading patients. Further findings depicted that people with disabilities were neglected in online designs which discourage their chances to be productive.

III. METHOD

A. Design

An existing hospital website in public sector of Pakistan was selected for evaluation along with website template designed on the basis of user responses collected in earlier studies and keeping in view user's expectations. This research study was centered to evaluate the websites as a comparison between existing (Fig. 1) and designed template (Fig. 2) in order to observer the user's understanding about these websites.

The study aimed on analysis of user data which was collected after performing representative tasks by each user and filling of post study questionnaire.

Fig. 1. Civil hospital, Karachi websites view.

Questionnaire and direct observation methods were adopted to record the usability of the websites.

A pre-study questionnaire was designed covering the consent of participant and their demographic details. The frequency of internet usage and purpose of usage was primarily focused in it. A query regarding already usage of any hospital website or not, was also included. Following representative tasks were chosen carefully keeping in view the user's common requirement for accessing the hospital website:

1) *Find the contact details of hospital*
2) *Get the online appointment*
3) *Locate the department of cardiology*

The tasks were of moderate nature to avoid the bar on user memory. A post-study questionnaire was designed based on five usability components (learnability, efficiency, memorability, error and satisfaction) developed by Nielsen [15]. An additional question for overall impression of the website design was also included in the questionnaire [17].

A five point Likert-scale was used [18], [19] for user compliance having '1' as 'strongly disagree' and '5' as 'strongly agree'.

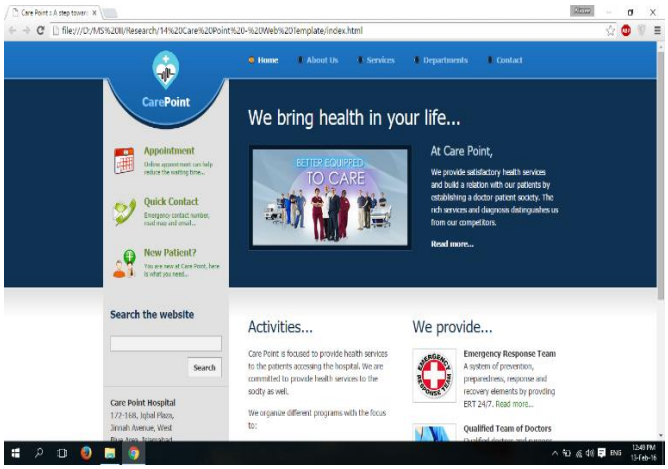


Fig. 2. Care point hospital websites template view.

TABLE I. WEB USAGE FREQUENCY

S #	Description	Frequency
1	Never	0
2	Less than 3 years	1
4	More than 2 years	20

TABLE II. AGE GROUP DIVISION OF PARTICIPANTS

S #	Description	Frequency
1	Below 20	2
2	21-30	12
3	31-40	4
4	41-50	1
5	Above 50	2

B. Participants

Twenty-one participants were involved in this empirical study. All the participants were volunteers. Different age group participants including male (48%) and female (52%) had parted in the study. Only one participant had web usage experience less than three years, rest had more than three years' experience (Table I). Fourteen participants (67%) had already used the hospital website. However, seven participants (33%) were using any hospital website for the first time.

Most of the participants (48%) were among the age group of 21-30 (Table II). Two participants were each from below 20 age group and above 50 years. Four (19%) users were from 31-40 age group and only one user had age group 40-50.

C. Experimental Design

A consent form containing demographic details was presented to the participant for collecting basic information. Representative tasks were chosen to be performed by participants. All the participants were volunteers of different ages and professional levels. The users were observed from behind when they were performing the tasks.

Questionnaire and observation methods were adopted. 20-minute time was planned for each participant to perform the tasks. Participants performed the tasks well within time. Websites were already open in the web browser. While performing the tasks, participants were free to ask about the tasks and websites. After completion of tasks, a post-study questionnaire was presented to the user.

D. Procedure

Participants were approached at their convenient locations in most of the cases. However, some of them were invited in FUUAST labs for conducting the study. Participants were observed from behind while performing tasks. Some participants were photographed from behind as well after their consent (Fig. 3).

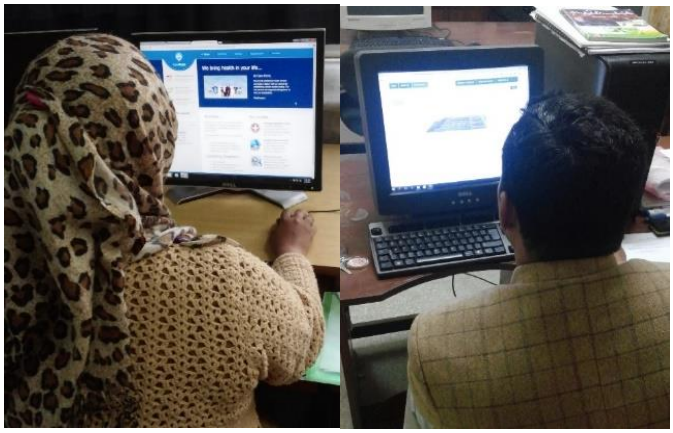


Fig. 3. Users performing tasks during user study.

When the participant declared about completion of tasks, they were presented the post study questionnaire. Participants were briefed about the questions on their demand.

IV. DATA ANALYSIS

Likert-Scale values for each website have been recorded against each usability component (Tables III and IV).

The average results showed that the participants selected "Strongly Agree" (75%) from Likert-Scale in most of the cases for Care Point Hospital Website Template (Fig. 4). However, 37% of the responses were "Disagree" and another 22% "Strongly Disagree" for usability of Karachi Civil Hospital Website, on the other hand. User tasks were performed on the Civil Hospital Karachi website before performing them on the Care Point Hospital Website template.

TABLE III. AVERAGE SCORE OF CIVIL HOSPITAL KARACHI

S #	Questions	User Responses				
		1	*2*	*3*	*4*	*5*
1	Is it easy to use the website?	9	9	2	1	0
2	I learned to use it quickly?	8	9	2	1	1
3	It is easy to remember how to use the website?	6	5	4	5	1
4	Have you found any mistake in the system while performing the tasks?	2	6	5	4	3
4.1	If so, how easy it is for you to recover from them?	0	8	3	5	5
5	It is pleasant to use?	4	10	3	4	0
6	The design of the website is beautiful?	4	7	5	5	0

TABLE IV. AVERAGE SCORE OF CARE POINT HOSPITAL TEMPLATE

S #	Questions	User Responses				
		1	*2*	*3*	*4*	*5*
1	Is it easy to use the website?	0	0	0	5	16
2	I learned to use it quickly?	0	0	0	5	16
3	It is easy to remember how to use the website?	0	1	0	7	13
4	Have you found any mistake in the system while performing the tasks?	0	0	0	1	20
4.1	If so, how easy it is for you to recover from them?	0	0	1	1	19
5	It is pleasant to use?	0	0	2	4	15
6	The design of the website is beautiful?	0	0	4	6	11

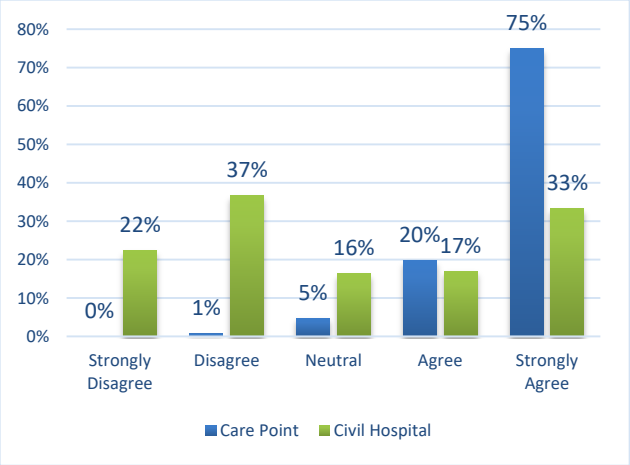


Fig. 4. Average Likert-scale selection in post study questionnaire.

In 95% cases (20 out of 21), users performed tasks without committing or facing any error on Care Point Hospital website. On the other hand, participants faced hardships in performing the tasks on Civil Hospital Karachi website as was recorded in the response in Table III.

V. RESULTS

The overall results showed (Fig. 4) that the user's responses were positively inclined towards the Care Point Hospital (CPH) Website (Template) in terms of usability. As many as 75% of the participants "strongly agreed" and another 20% "agreed" with the usability aspects of the proposed design template comparing with Civil Karachi Hospital (CHK) Website for which 22% "strongly disagreed" and 37% "disagree" responses were recorded. Only 1% responses were "disagree" and none of the participants marked "strongly disagree" for the CPH template.

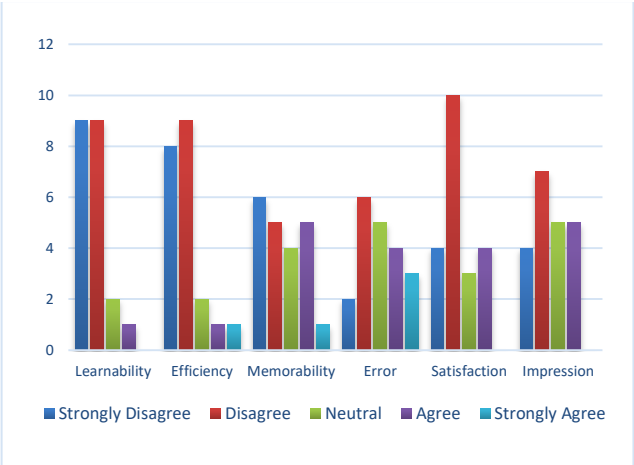


Fig. 5. Responses of participants for Civil Hospital Karachi.

For CPH template 5% responses were neutral and similarly 16% users remained neutral for CHK website. Fig. 5 portrays the user response towards CHK separately and Fig. 6 depicts the results of CPH template only.

The usability components have been evaluated as follows:

A. *Learnability*

43% users responded each in "Strongly Disagree" and "Disagree" for CHK website. 10% users remained "Neutral" about their views and only 5% "Agreed" with the learnability of the CHK website. On the contrary, 76% user "Strongly Agreed" and 24% "Agreed" with the usability aspects of CPH website template.

B. *Efficiency*

38% and 43% responses were "Strongly Disagree" and "Disagree" for CHK respectively. 10% neutral views and 5% each for "Agree" and "Strongly Agree". 76% and 24% users "Strongly Agreed" and "Agreed", respectively for CPH website template.

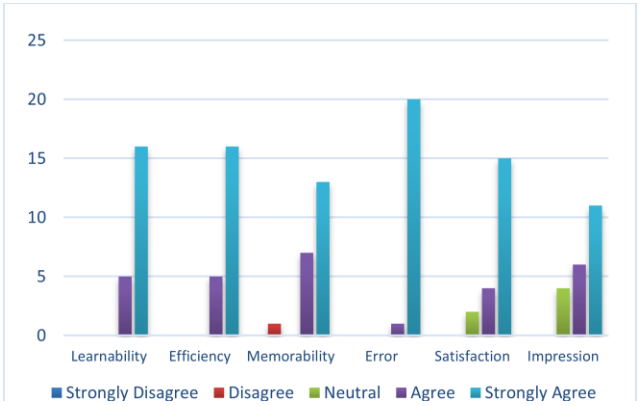


Fig. 6. Responses of participants for Care Point Hospital Website Template.

C. *Memorability*

29% users "Strongly Disagree" with the memorability of the CHK website. Another 24% "Disagree" and 19% "Neutral" responses were received. 24% "Agreed" with this component

and 5% rated "Strongly Agree" in case of CHK website. For CPH template, 62% users "Strongly Agree" with the memorability component of usability and 33% had "Agree" response. Only 5% (one instance) "Disagree" with it.

D. Error

95% users did not commit any error while performing tasks on CPH website template. On the other hand, 10% users committed many mistakes while using CHK website. 29% and 24% users marked error rate as "5 or 4" mistakes, respectively. 19% committed two mistakes and 14% committed single mistake.

E. Satisfaction

19% users remained fully unsatisfied with the CHK website as shown in Fig. 5. However, 48% responses were "Disagree". 14% "Neutral" remarks and 19% were "Agree" responses while 71% users showed their full satisfaction for CPH template. 19% responded with "Agree" to their satisfaction level and 10% had "Neutral" views.

F. Impression

52% participants appreciated the overall design of CPH website template by showing "Strongly Agree" response (Fig. 6). Another 29% showed "Agree" views. 19% had "Neutral" remarks. 19% users did not appreciate the design and responded with "Strongly Agree" for CHK website. 33% users had "Disagree" remarks and 24% were "Neutral". Another 24% users had "Agree" response.

The user responses recorded in Tables III and IV have been statistically analyzed afterwards. Paired sample t-test was applied to the data using SPSS. Care Point Hospital (CPH) website template has been assigned $\mu 1$ and Civil Hospital Karachi (CHK) website has been assigned $\mu 2$. The standard deviation has been calculated as 11.1034 and value of t was derived as $t = 17.6024$. As the value of t is very high therefore we reject null hypothesis and hence conclude that CPH website template is more usable than CHK website.

VI. Discussion

The Website Template was designed keeping in view the user's opinion and responses received from participants in User Study 1 and User Study 2. The template was named Care Point Hospital Website Template. The responses of users in earlier studies were observed and analyzed carefully hence the guidelines so proposed have been utilized in development of template. The template is based on user's requirements and approaches towards usage of hospital websites. The guidelines proposed in User Study 1 and 2 were the theme behind the design and development of template. However, more design rules and user expectations were also considered before designing the template.

The design rules such as participants suggested in earlier user studies that a website should have a search option, in case user is unable to locate specific information from the website. This will help to extract and locate such information immediately were followed in the template design.

The font face, size and color were also focused as was highlighted in the former user studies by users. The contrast of the fonts comparing with the background color was also vital in design. A hospital website poorly managed in this way has a weakened usability as was analyzed in user studies performed previously.

The important options for accessing information on the website is designed and spread in accordance with the F-Shaped Pattern for Reading Web Content [20]. The template so designed was evaluated in contrast to existing hospital website. The existing hospital website had negative usability; however, the design template had higher usability responses. Hence the results were clear. Users responded that the Care Point Hospital Websites (Template) has been more useable than the other one which has negative usability.

VII. Conclusion

Hospitals are a source of healthy society not only by providing health services but keeping their patients aware about health in particular and spreading health tips to common people in general. This aspect is more significant in Pakistan as Pakistan is a developing country where health institutions and hospitals can improve the system by playing their part. A vision of healthier society can be achieved through a system having affordable, efficient, technology appropriate and consumer friendly. A website having such features can significantly straighten the road towards goal.

A website designed by keeping in view the usability aspects can result into more satisfied users. The template under evaluation was designed in accordance with the expectation of users hence the satisfaction level of the users raised which clearly indicated in highly useable website.

This study is part of the research that intends to propose design rules for hospital websites of Pakistan. The outcome of this research will become part of planned design rules.

References

[1] Dos Anjos, T.P., et al., Usability Evaluations of Health Institutions Inspection Software. IEEE Latin America Transactions, 2016. 14(3): p. 1538-1547.

[2] Barbara, A.M., et al., The McMaster Optimal Aging Portal: Usability Evaluation of a Unique Evidence-Based Health Information Website. JMIR human factors, 2016. 3(1).

[3] Johnson, M.A. and K. Norris Martin, When Navigation Trumps Visual Dynamism: Hospital Website Usability and Credibility. Journal of Promotion Management, 2014. 20(5): p. 666-687.

[4] Raji, S., et al., Usability Evaluation of Hospital Websites in Nigeria: What Affects End Users' Preferences?, in HCI International 2014 - Posters' Extended Abstracts, C. Stephanidis, Editor. 2014, Springer International Publishing. p. 430-434.

[5] Gallant, L., C. Irizarry, and G.L. Kreps, User-centric hospital web sites: a case for trust and personalization. E-service Journal, 2007. 5(2): p. 5-26.

[6] Allam, A., P.J. Schulz, and K. Nakamoto, The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output. Journal of medical Internet research, 2014. 16(4): p. e100.

[7] Abdullah, R. and K.T. Wei, Usability measurement of Malaysia online news websites. International Journal of Computer Science and Network Security, 2008. 8(5): p. 159-165.

[8] Kaur, S., K. Kaur, and P. Kaur. Analysis of website usability evaluation methods. in Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. 2016. IEEE.

[9] Abras, C., D. Maloney-Krichmar, and J. Preece, User-centered design. Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 2004. 37(4): p. 445-456.

[10] Fink, A. and J.C. Beck, Developing and Evaluating a Website to Guide Older Adults in Their Health Information Searches A Mixed-Methods Approach. Journal of Applied Gerontology, 2015. 34(5): p. 633-651.

[11] Raji, S.O., M. Mahmud, and A. Abubakr, Evaluation of University Teaching Hospital Websites in Nigeria. Procedia Technology, 2013. 9: p. 1058-1064.

[12] Martins, J., et al., How Ill Is Online Health Care? An Overview on the Iberia Peninsula Health Care Institutions Websites Accessibility Levels, in New Advances in Information Systems and Technologies. 2016, Springer. p. 391-400.

[13] Huerta, T.R., D.M. Walker, and E.W. Ford, An Evaluation and Ranking of Children's Hospital Websites in the United States. Journal of Medical Internet Research, 2016. 18(8).

[14] Huang, E., C.-c. Angela Chang, and P. Khurana, Users' preferred interactive e-health tools on hospital web sites. International Journal of Pharmaceutical and Healthcare Marketing, 2012. 6(3): p. 215-229.

[15] Nielsen, J., Usability engineering. 1994: Elsevier.

[16] Yadrich, D.M., et al., Creating patient and family education web sites: assuring accessibility and usability standards. Comput Inform Nurs, 2012. 30(1): p. 46-54.

[17] Wynn, L. and J. Trussell, The morning after on the internet: usage of and questions to the emergency contraception website. Contraception, 2005. 72(1): p. 5-13.

[18] Norman, G., Likert scales, levels of measurement and the "laws" of statistics. Advances in health sciences education, 2010. 15(5): p. 625-632.

[19] Scales, L.-T., Analyzing and Interpreting Data From. Journal of graduate medical education, 2013: p. 541.

[20] Nielsen, J. (2006). " F-Shaped Pattern For Reading Web Content," Jakob Nielsen's Alertbox. http://www.useit.com/alertbox/reading_pattern. html.

Detection of Climate Crashes using Fuzzy Neural Networks

Rahib H.Abiyev, Mohammed Azad Omar
Department of Computer Engineering,
Near East University, Lefkosa, North Cyprus,
Mersin-10, Turkey

Boran Şekeroğlu
Department of Information Systems Engineering,
Near East University, Lefkosa, North Cyprus,
Mersin-10, Turkey

***Abstract*—In this paper the detection of the climate crashes or failure that are associated with the use of climate models based on parameters induced from the climate simulation is considered. Detection and analysis of the crashes allows one to understand and improve the climate models. Fuzzy neural networks (FNN) based on Takagi-Sugeno-Kang (TSK) type fuzzy rule is presented to determine chances of failure of the climate models. For this purpose, the parameters characterising the climate crashes in the simulation are used. For comparative analysis, Support Vector Machine (SVM) is applied for simulation of the same problem. As a result of the comparison, the accuracy rates of 94.4% and 97.96% were obtained for SVM and FNN model correspondingly. The FNN model was discovered to be having better performance in modelling climate crashes.**

Keywords*—Climate crashes; fuzzy neural networks; parallel ocean program; SVM

I. INTRODUCTION

Climate models play important role in the prediction of future climate changes. Tough climate models are offering huge benefits to the pupils. They are suffering from failure which is known as crashes or bifurcations. The failure in climate models is a result of their complex nature [1]-[4]. The scientific representation of this problem is too complex and huge, and the corresponding models involved are considered to be so complex [5]. Another important problem that has been characterised by the use of climate models is related to the software challenges. The software that is used for modelling climate takes time in changing climate conditions [6].

In the paper, the effects of ocean parameter uncertainties on climate simulation are considered. Modern tool such as uncertainty quantification (UQ) is used to solve simulation problems and improve existing climate models. Primary UQ is made up of parameters or coefficients whose values are always changing. However, Sternsrud [7] show that the changing of the parameters leads to difficulties and it became difficult to simulate the climate changes within required conditions. This can be solved by conducting the parameterization process separately [6]. The best way for climate simulation is the use of non-linear climate models. This can lead to huge changes in simulation output. But the models' properties are restricted and model sometimes fails when the adjustable parameters are amplified using small perturbations [8]. Taking into account above-mentioned it is necessary to specify the reasons and causes for climate model simulation failure. It is also needed to define the conditions that can affect the effectiveness of climate models in simulating the climate changes.

There are set of parameters that have impact on climate. The weather conditions are chaotic and affect climate simulation [9]. When the wind blows, the weather is always being in a state of disequilibrium and the climate conditions are being affected. Greenhouse gas forcing is major chaotic effect while volcanoes, sun and weather changes, etc. are smaller chaotic effects. These factors can strongly influence a simulated model. Watanabe et al. in [10] shows that climate modelling is not an easy problem as weather changes cause chaotic behaviour and are characterized by Lorenz non-linearity. This non-linearity is due to unpredictable air oscillation behaviour. Randall et al. in his paper [11] evaluates the use of climate models and their ability to predict future climate changes. The study showed that climate variables such as precipitation have lower predictability than temperature changes. The paper [12] analyzed the use of climate data to forecast future climate changes using a General Circulation Model. The study uses stochastic and generalized downscaling methods to generate the weekly data. Using various simulation models, it is possible to predict potential climate changes and their implications. The paper [13] showed that the integrated climate models could simulate climate changes. The study evaluates environmental policies targeted at reducing emissions and combines uncertainty quantification methods to simulate carbon components. The study recommends that improvements in climate models be extended to cover carbon cycle feedbacks, inertia and climate sensitivity. The paper [2] used distribution models and showed that careful selection of climate models is an important process which must not be done arbitrarily.

The climate models differ in complexity and success perspectives. These climate models consist of various subroutines, functions, algorithms (geologic, climate and biological), huge number lines of codes [11]. All these are used to describe conservative laws and equations related to momentum, energy and flow of matter within the earth's reservoirs, between the land, oceans and atmosphere. All these ideas are based on views that climate models are not always reliable and effective, and are bound to fail [3], [14]. There are no concrete reasons and concurrences about failure in climate models. For instance, [15] mentioned that the use of numerous algorithms of anthropogenic, geologic, chemical and biological nature that are used in the simulation of climate-related issues and greenhouse gases, ozone, aerosols, Sulphur, nitrogen, and

cycles of carbon is the main reason of climate model failure. Such algorithms are used in a set of circumstances and time and have solid, liquid and gaseous elements [16], [17] showed that crashes occur at a high rate. Lucas et al. have considered predictions of climate models [18]. A research is required to add and refurbish existing information about crashes in climate models. Bifurcations or crashes are common in any situation irrespective of its complexity and went to establish that intermediate climate models are also prone to crashes. This study aims to examine and predict the failure of parameter-induced simulation crashes in climate models. The accurate prediction of climate crashes is very important. For this purpose, in this paper, FNN is used to predict the failure probability and improve prediction results. The paper is organised as follows. Section 2 presents fuzzy neural networks used for detection of climate crashes. Section 3 presents simulation study. Section 4 gives conclusions.

II. Fuzzy Neural Networks for Detection Climate Crashes

The Fuzzy neural networks (FNN) model conducts a fuzzy reasoning process using the neural network structure [19]-[22]. Here, problem is to determine the accurate values of the parameters of the FNN model. This is obtained through evaluation of the error response of the designed classification system. TSK-type fuzzy rules are basically used for designing the fuzzy systems. TSK fuzzy rules include fuzzy antecedent and crisp consequent parts. These fuzzy systems approximate nonlinear systems with linear ones and have the following form:

If x_1 is A_{11} and x_2 is A_{21} and … and x_m is A_{m1} Then

$$y_1 = b_1 + \sum_{i=1}^{m} a_{i1} x_i$$

if x_1 is A_{12} and x_2 is A_{22} and … and x_m is A_{m2} Then

$$y_2 = b_2 + \sum_{i=1}^{m} a_{i2} x_i \qquad (1)$$

If x_1 is A_{1n} and x_2 is A_{2n} and … and x_m is A_{mn} Then

$$y_n = b_n + \sum_{i=1}^{m} a_{in} x_i$$

where x_i and y_j are input and output signals of the system respectively, $i=1,...,m$ is the number of input signals, $j=1...r$ is a number of rules. A_{ij} are input fuzzy sets, b_j and a_{ij} are coefficients. Fuzzy sets are applied for the description of A_{ij} parameters of the antecedent parts of the fuzzy rules.

The structure of FNN used for prediction of the climate crushes is given in Fig. 1. The input layer (block) is used for distributing of the coming x_i signals. In next block the membership degrees of input signal for ach linguistic value are calculated. Linguistic values are represented by Gaussian membership functions that are characterized by the width and center parameters.

$$\mu 1_j(x_i) = e^{-\frac{(x_i - c_{ij})^2}{\sigma_{ij}^2}}, \quad i=1..m,\ j=1..r \qquad (2)$$

where, c_{ij} *and* σ_{ij} are centre and width of membership functions, correspondingly. These signals are inputs for the next rule layer.

The output signals of the rule layer are computed through the use of t-norm min (AND) operation:

$$\mu_j(x) = \prod_i \mu 1_j(x_i), \quad i=1,..,m,\ j=1,...,r \qquad (3)$$

where, Π is the min operation. These $\mu_j(x)$ signals are input signals for the output layer. The consequent layer includes n linear systems. In this layer, at first the values of the rules' output are determined as

$$y1_j = b_j + \sum_{i=1}^{m} a_{ij} x_i \qquad (4)$$

The output signals of the rule layer are multiplied by the output signals of the consequent layer. The output of j-th node is calculated as $y_j = \mu_j(x) y1_j$

After calculating y_j, the output signals of FNN are determined as

$$u_k = \frac{\sum_{j=1}^{r} w_{jk} y_j}{\sum_{j=1}^{r} \mu_j(x)} \qquad (5)$$

where, u_k are the output signals of FNN, ($k=1,..,n$). After calculating the output signal, the training of the parameters of the network starts. The algorithm described in [23]-[25] is used for learning the parameters of FNN.

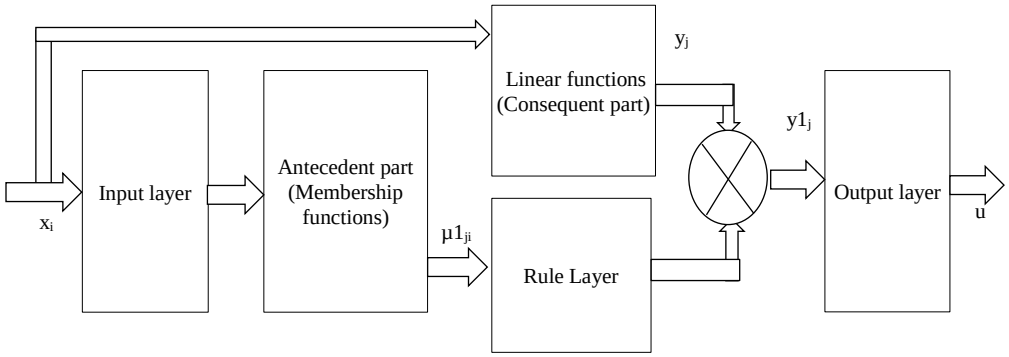


Fig. 1. The structure of FNN based prediction system.

III. SIMULATION

In the paper, POP2 model is used to select ocean model parameters. These model parameters were subjected to different parameterizations on a sub-grid scale. The main emphasis behind such parameterization was to determine the resultant outcome of vertical and horizontal oceanic turbulent after simulation [18], [26], [27]. Table I provides details of the uncertainty ranges of the model parameters used in this study.

The parameters are taken according to the [18]. These are: spatial anisotropic viscosity that was used to determine the horizontal momentum and was represented by the parameters 13 to 18, isopycnal eddy-induced transport of the horizontal mixers that were for parameters 10 to 12, the parameters 7 to 9 that can be used to simulate mixed layer eddies and submesoscale, and were used for the abyssal tidal mixing. Further prescriptions were K-profile parameterization associated with vertical mixing and convection and these corresponded to parameters 1 to 6. The examination of the ensembles was done in three different stages with simulations amounting to 180. The first and second studies were used to program machine learning algorithms so that they can track and analyze simulation crashes. The third study was conducted so as to determine their potential to forecast simulation crashes. 46 failures were observed out of the 540 simulations that were done. The recorded failures were observed at different intervals of the integration phase. 18 POP2 parameter values were examined using a Latin hypercube method. This was also important as it resulted in the establishment of an ensemble. In addition, normalized log-uniform probability functions were also employed to represent the model parameters' high and low values.

Statistical data were collected as a result of 540 simulations. During simulation, 494 successes and 46 failures occurring at the various times were observed. During simulation, Latin hypercube method is used to sample the values of the 18 POP2 parameters (Table I). The parameters of the model are represented with standard uniform and log-uniform probability distribution functions normalised in the interval [0,1],

Using statistical data, the training of FNN was performed. The problem is the accurate prediction of failures. The fragment of data set is given in Table II. In the table, the data from 1 to 18 are the values of input parameters. The data of number 19 are the values of output, that are 1 is the success, 0 is the failure.

The data sets include 18 inputs and one output. FNN is used for prediction purpose. At first, the parameters of FNN system is initialised randomly, then gradient descent algorithm is applied for training. The training is carried out using 10 fold cross-validation approach. During the design of FNN prediction system training, evaluation and test results are obtained. During training, evaluation and test stages root mean square error and recognition rate are used to measure FNN performance. RMSE is computed as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^d - y_i)^2} \qquad (6)$$

where y_i^d and y_i are the target and current output signals, the number of samples is represented by N. RMSE is applied for the training of the network. Accuracy rate is used to measure the performance of FNN using test data set.

During training, the input data sets are fed to the FNN input. Using formulas (1)-(5) the output of the network is computed. On the output, the deviation of current output from target signal is determined. This value is used to determine RMSE. Using RMSE, the training of FNN system is performed. The training is performed for 500 epochs. The simulation is performed using 8, 16, 24 fuzzy rules (hidden neurons). Root mean square errors (RMSE) indicate the difference between the actual values and the predicted values. Fig. 2 depicts the plot of RMSE values of FNN based system for 500 epochs. The simulation results for three cases are given in Table III. In all of the cases of the fuzzy neural algorithm that were conducted, success rates were observed to be averaging high above 93.15% for 8 rules. The accuracy rate for 16 rules 96.11%, for 24 rules 97.96% were obtained. The RMSE values for test data were 0.3607, 0.3146 and 0.2609 for 8, 16 and 24 fuzzy rules correspondingly. The highest accuracy rate can be observed to be associated with the activity of 24 rules with a success rate of 97.96% and is composed of 16 neurons. The obtained results are obtained by averaging of the simulations.

FNN with the 24 neurons was established to be the best model in terms of accuracy and this follows a recorded accuracy rate of 97.96% while model FNN with 8 neurons had the lowest accuracy rate of 93.15%.

The performance of FNN classifier is evaluated using Sensitivity, Specificity and Precision. These factors can be computed using true positive, true negative, false positive and false negative parameters. The FNN classifier that correctly predicts successes and failures are denoted true positives (TP) and true negatives (TN), respectively. The classifier that incorrectly predicts current output failures and successes are denoted as false negatives (FN) and false positives (FP), respectively. Using these parameters we can determine true positive rate (TPR), true negative rate (TNR) and positive predictive value (PPV). These variables are used to evaluate sensitivity, specificity and precision, correspondingly. Here,

$$TPR = \frac{TP}{(TP+FN)};\quad TNR = \frac{TN}{(TN+FP)};\quad PPV = \frac{TP}{(TP+FP)} \qquad (6)$$

When classifiers are predicting all output values perfectly then the values TPR and TNR (or sensitivity and specificity) become equal to 1. The values of TPR, TNR and PPV for FNN classifier for climate crashes prediction are given in Table IV.

For comparative analysis, the same problem is solved using support vector machine (SVM). Table V includes a fragment from the set of simulations. Six cases were used for the SVM algorithm with a 10-fold cross-validation and all the cases have attained accuracy rates that are above 91%. The highest accuracy rate of 94.1% can be noted to be in line with a quadratic SVM while the lowest rate of 91.5% is recorded for Fine Gaussian SVM and Coarse Gaussian SVM. The best one is Quadratic SVM, Accuracy with 10-Fold Cross-Validation is 94.4%.

As shown in Tables III and V, the recognition rates of the FNN system with 16 and 24 rules are better than SVM.

After training, the models' performances are estimated to show which model has the best overall score. Such a score is termed the root mean square error (RMSE) on the validation set alternatively, it can be said to be useful in estimating the performance of the trained model on new data. Response plots were used to determine which model offers the best performance in terms of the predictability power. The decision criteria is to accept that the model is a good model and can forecast or predict the actual when the resultant distance or margin between the actual and predicted values is small.

TABLE I. CCSM4 OCEAN MODEL PARAMETERS

	Description	Module	Scale[1]	(low, default, high)	Parameter[2]
1	Ration of background diffusivity and vertical viscosity	Vmix_kpp	Log	(4.0, 10.0, 20.0)	Prandt1
2	Max PSI induced diffusion	Vmix_kpp	Log	(0.1, 0.13, 0.5)	Bckgrnd_vdc_psim
3	Equatorial diffusivity	Vmix_kpp	Log	(0.01, 0.01, 0.5)	Bckgrnd_vdc_cq
4	Banda sea diffusivity	Vmix_kpp	Lin	(0.5, 1.0, 0.5)	Bckgrnd_vdc_ban
5	Base background vertical diffusivity	Vmix_kpp	Log	(0.032, 0.16, 0.8)	Bckgrnd_vdc1
6	Mixed diffusion coefficients	Vertical_mix	Log	(1.0, 10.0, 50.0) x 10^3	Convect_corr
7	Convect_visc (momentum) and convect_diff (tracer)	Tidal	Log	(2.5, 5.0, 20.0) x 10^4	Vertical_decay_scale
8	Tide induced turbulence's vertical decay scale	Tidal	Log	(25.0, 100.0, 200.0)	Tidal_mix_max
9	Tidal mixing threshold	Mix_submeso	Lin	(0.05, 0.07, 0.01)	Efficiency_factor
10	Submesoscale eddies' efficiency factor	Hmix_gm	Log	(0.05, 0.03, 0.03)	Slm_corr
11	Slm_r (redi terms) and slm_b bolus' maximum slope	Hmix_gm	Lin	(2.0, 3.0, 4.0) x 10^7	Ah_bolus
12	Bolus mixing's diffusion coefficient	Hmix_gm	Lin	(2.0, 3.0, 4.0) x 10^7	Ah_corr
13	Ah_bkg_srbl (horizontal diffusivity within the surface boundary) and Ah (redi mixing's diffusion coefficient and background)	Hmix_aniso	Lin	(30.0, 45.0, 60.0)	Vconst_7
14	Variable viscosity parameter	Hmix_aniso	Lin	(2, 3, 5)	Vconst_5
15	Variable viscosity parameter	Hmix_aniso	Log	(0.5, 2.0, 10.0) x 10^{-8}	Vconst_4
16	Variable viscosity parameter	Hmix_aniso	Lin	(0.16, 0.16, 0.02)	Vconst_3
17	Variable viscosity parameter	Hmix_aniso	Log	(0.25, 0.5, 2.0)	Vconst_2
18	Variable viscosity parameter	Hmix_aniso	Lin	(0.3, 0.6, 1.2) x 10^7	Vconst_corr

[1] Logarithmic and linear scales were applied for parameters whose ratios were between the range high/low ≥ 5 and high/low <5,
[2] Individual correlated pair of parameters were denoted by numbers 1, 7, 9 and 13

TABLE II. FRAGMENT FROM DATA SET

No	Input parameters' values													
1	0.8590	0.6060	0.9976	0.7834	0.4062	0.0414	0.1611	0.4153	0.1668	0.6556	0.5900	0.8819	0.9610	0.1725
2	0.9278	0.4577	0.3732	0.1041	0.5132	0.6290	0.5488	0.8987	0.3530	0.4139	0.2937	0.4249	0.9769	0.0136
3	0.2529	0.3594	0.5174	0.1975	0.0618	0.3034	0.1536	0.9318	0.9881	0.8053	0.4235	0.9032	0.8579	0.6234
4	0.2988	0.3070	0.5050	0.4218	0.6358	0.8134	0.6544	0.9166	0.2871	0.1635	0.3298	0.1733	0.6150	0.5191
5	0.1705	0.8433	0.6189	0.7421	0.8448	0.2228	0.1403	0.3991	0.5636	0.8619	0.4578	0.7910	0.6155	0.2545
6	0.7359	0.9349	0.6056	0.4908	0.4415	0.9712	0.7966	0.0094	0.4027	0.9476	0.8298	0.4762	0.3528	0.3669
7	0.4283	0.4446	0.7462	0.0055	0.1919	0.6098	0.4058	0.8463	0.3809	0.5466	0.4978	0.6812	0.8340	0.0557
8	0.5679	0.8280	0.1959	0.3921	0.4875	0.6478	0.6626	0.6838	0.4792	0.4261	0.1594	0.9058	0.0951	0.5185
9	0.4744	0.2966	0.8157	0.0100	0.3585	0.7379	0.0494	0.3973	0.0602	0.4171	0.9711	0.8284	0.2309	0.3713
10	0.2457	0.6169	0.6794	0.4715	0.5515	0.4409	0.5785	0.8868	0.2365	0.9456	0.4923	0.0837	0.9548	0.8538
11	0.1042	0.9758	0.8034	0.5979	0.7439	0.0360	0.2649	0.5224	0.2905	0.3254	0.0843	0.5206	0.5772	0.3470
12	0.8691	0.9143	0.6440	0.7617	0.3123	0.6159	0.9592	0.6948	0.3918	0.6665	0.9743	0.0729	0.7835	0.9473
13	0.9975	0.8452	0.7184	0.3628	0.6502	0.0175	0.6981	0.8865	0.2549	0.3743	0.9264	0.9481	0.5304	0.5973
14	0.4486	0.8642	0.9248	0.9128	0.5223	0.9323	0.4674	0.4117	0.4884	0.1003	0.2954	0.9996	0.1752	0.4288
15	0.3075	0.3467	0.3154	0.9780	0.0435	0.3293	0.6371	0.4811	0.0537	0.2133	0.8042	0.7285	0.5445	0.4014
16	0.8583	0.3566	0.2506	0.8459	0.3767	0.9541	0.0113	0.9265	0.8622	0.2229	0.8708	0.2859	0.0814	0.8204
17	0.7970	0.4384	0.2856	0.6994	0.2801	0.1354	0.1473	0.0264	0.4151	0.0073	0.5463	0.2109	0.7330	0.5996
18	0.8699	0.5123	0.3659	0.4760	0.1323	0.2948	0.2138	0.0927	0.4871	0.4200	0.8849	0.8336	0.5314	0.1357
19	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0	1.0000

TABLE III. FNN MODEL RESULTS

No.	Neurons	epoch	SSE_{Train}	$RMSE_{Train}$	$RMSE_{Evaluation}$	SSE_{Test}	$RMSE_{Test}$	Accuracy
1	8	500	635.8686	0.3617	0.3614	70.2851	0.3607	93.15%
2	16	500	477.7429	0.31353	0.321350	53.4488	0.3146	96.11%
3	24	500	333.2924	0.261875	0.261879	36.7607	0.2609	97.96%

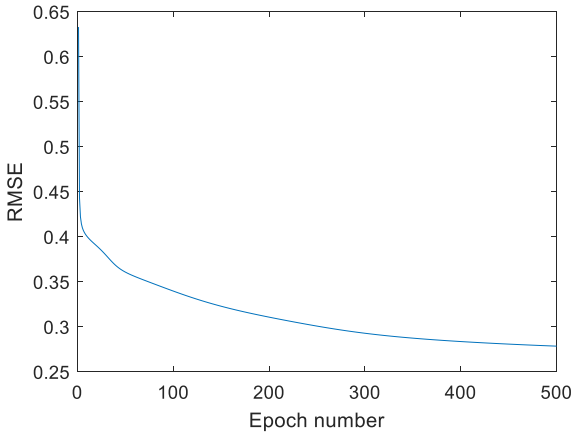


Fig. 2. Plot of RMSE.

TABLE IV. FNN MODEL RESULTS

No.	Neurons	Sensivity	Specifity	Precision
1	8	0.93032	1	1
2	16	0.96673	1	1
3	24	0.97821	1	1

TABLE V. SVM RESULTS

No.	Types SVM	Cross-Validation	Accuracy
1	Linear SVM	10-Fold	93.1%
2	Quadratic SVM	10-Fold	94.4%
3	Cubic SVM	10-Fold	94.1%
4	Fine Gaussian SVM	10-Fold	91.5%
5	Medium Gaussian SVM	10-Fold	91.7%
6	Coarse Gaussian SVM	10-Fold	91.5%

IV. CONCLUSIONS

The main emphasis of the study was to determine if there are any crashes or failure that are associated with the use of simulation models as well as conditions that can cause climate models to fail by determining the chances that POP2 simulation will fail. For this purpose, the fuzzy neural network was applied to determine chances of failure of the models. The simulation crashes were based on the idea that they may either succeed or fail (binary problem) and failure probabilities were quantified using FNN based machine learning classification. The quantification process was based on the 18 model parameters and the simulations based on cross-validation techniques. Conclusions can be made that the occurrence of the crashes is as a result of several numerical reasons which are caused by changes in the combination of parameter values used in the simulation process. Based on the obtained accuracy rate, conclusions can be made that climate models have a high predictive capacity to simulate climate changes. It can also be finally concluded that the fuzzy-neural network performs better in modelling climate crashes as compared to SVM.

REFERENCES

[1] V.Artale, S.Calmanti1, A.Carillo, A.Dell'Aquila, M.Herrmann, G.Pisacane, etc. An atmosphere–ocean regional climate model for the Mediterranean area: assessment of a present climate simulation. Climate Dynamics, vol.35, no.5, 2010, pp.721-740.

[2] L.J.Beaumont, L.Hughes, and A.J.Pitman, Why is the choice of future climate scenarios for species distribution modelling important?. Ecology letters, vol.11, no.11, 2008, pp.1135-1146.

[3] S.M.Easterbrook, Climate change: a grand software challenge, in: FoSER, edited by: G.-C.Roman, K.J.Sullivan, 2010, pp.99–104, ACM.

[4] S.Rugaber, R.Dunlap, and S.Ansari, Managing software complexity and variability in coupled climate models. IEEE software, vol.28, no.6, 2011, pp.43-48.

[5] P. E.Farrell, M.D.Piggott, G.J.Gorman, D.A.Ham, C.R.Wilson and T.M.Bond, Automated continuous verification for numerical simulation, Geosci. Model Dev., vol.4, 2011, pp.435–449.

[6] J.T.Kiehl, and C.A.Shields, Climate simulation of the latest Permian: Implications for mass extinction. Geology, vol.33, no.9, 2005, pp.757-760.

[7] Stensrud, D. J. (2009). Parameterization schemes: keys to understanding numerical weather prediction models. Cambridge University Press.

[8] P. R.Gent, Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z.-L., & Zhang, M. The Community Climate System Model Version 4, J. Climate, vol.24, 2011, pp.4973–4991.

[9] L. R.Leung, Y.Qian, and X.Bian, Hydroclimate of the western United States based on observations and regional climate simulation of 1981–2000. Part I: Seasonal statistics. Journal of Climate, vol.16, no.12, 2003, pp.1892-1911.

[10] M.Watanabe, T.Suzuki, R.O'ishi, Y.Komuro, S.Watanabe, S.Emori, and K.Takata, Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. Journal of Climate, vol.23, no.23, 2010. pp.6312-6335.

[11] D.A.Randall, R.A.Wood, S.Bony, R.Colman, T.Fichefet, J.Fyfe, and R.J.Stouffer, Climate models and their evaluation. In Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR), pp.589-662, 2007, Cambridge University Press

[12] Jones, P. G., Thornton, P. K., & Heinke, J. (2009). Generating characteristic daily weather data using downscaled climate model data from the IPCC's Fourth Assessment.

[13] D.P.Van Vuuren, J.Lowe, E.Stehfest, L.Gohar, A.F.Hof, C.Hope, and G.K.Plattner, How well do integrated assessment models simulate climate change?. Climatic climate system model (CCSM) and community earth system model (CESM). Rep. LAUR-01853, 141, pp.1-140, 2011.

[14] W.M.Washington, and C.L.Parkinson, An introduction to three dimensional climate modeling, University Science Books, 2nd Edn, 2005.

[15] S.M.Easterbrook and T.C.Johns, Engineering the software for understanding climate change, Comput. Sci. Eng., vol.11, 2009, pp. 65–74.

[16] N.R.Edwards, D.Cameron, and J.Rougier, Recalibrating an intermediate complexity climate model, Clim. Dynam., vol.37, 2011, pp.1469– 1482.

[17] T.Clune, and R.Rood, Software testing and verification in climate model development, IEEE Software, vol.28, 2011, pp.49–55.

[18] D.D.Lucas, R.Klein, J.Tannahill, D.Ivanova, S.Brandon, D.Domyancic, and Y.Zhang, Failure analysis of parameterinduced simulation crashes in climate models, Geosci. Model Dev., vol.6, 2013, pp.1157–1171.

[19] R. Abiyev, N.Akkaya, E.Aytac, I.Günsel, and A.Çağman, Brain-computer interface for control of wheelchair using fuzzy neural networks, BioMed research international, 2016.

[20] R.H.Abiyev, Fuzzy Wavelet Neural Network for Prediction of Electricity Consumption. AIEDAM: Artificial Intelligence for Engineering Design, Analysis and Manufacturing. Vol. 23, Issue: 2, 2009, pp:109-118.

[21] R.H. Abiyev, Time Series Prediction Using Fuzzy Wavelet Neural Network Model, ICANN-2006, Lecture Notes in Computer Sciences, Springer-Verlag, Berlin Heidelberg, 2006, pp.191-200

[22] R.H. Abiyev, Fuzzy Wavelet Neural Network for Control of Dynamic Plants, International Journal on Computational Intelligence, Vol. 1, No. 2, 2004, pp.139-143

[23] R.H. Abiyev, S. Abizade, Diagnosing parkinson's diseases using fuzzy neural system, Computational and Mathematical Methods in Medicine. Volume 2016 (2016), Article ID 1267919

[24] R.H.Abiyev, Controller based on Fuzzy Wavelet Neural Network for Control of Technological Processes. In proceeding of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, IEEE CIMSA 2005, pp.215-219, Giardini Naxos - Taormina, Sicily, Italy, 20-22 July 2005.

[25] R.H.Abiyev, T.Al-shanableh. Neuro-Fuzzy Network for Adaptive Channel Equalization. 5th Mexican International Conference on Artificial Intelligence. IEEE CS press. MICAI 2006. Apizaco, Mexico November 13-17, 2006

[26] J.P.Evans, and M.F.McCabe, Regional climate simulation over Australia's Murray-Darling basin: A multitemporal assessment. Journal of Geophysical Research: Atmospheres, vol.115(D14), 2010.

[27] H.Shiogama, M.Watanabe, M.Yoshimori, T.Yokohata, T.Ogura, J.D.Annan, and R.Nobui, Perturbed physics ensemble using the MIROC5 coupled atmosphere–ocean GCM without flux corrections: experimental design and results. Climate dynamics, vol.39, no.12, 2012, pp.3041-3056.

Norm's Trust Model to Evaluate Norms Benefit Awareness for Norm Adoption in an Open Agent Community

Al-Mutazbellah Khamees Itaiwi
College of Graduate Studies
Universiti Tenaga Nasional
Kajang, Selangor, Malaysia

Mohd Sharifuddin Ahmad, Alicia Y. C. Tang
College of Computer Science &
Information Technology
Universiti Tenaga Nasional
Kajang, Selangor, Malaysia

***Abstract*—In recent developments, norms have become important entities that are considered in agent-based systems' designs. Norms are not only able to organize and coordinate the actions and behaviour of agents but have a direct impact on the achievement of agents' goals. Consequently, an agent in a multi-agent system requires a mechanism that detects specific norms for adoption while rejecting others. The impact of such norms selection imposes risks on the agent's goal and its plan ensuing from the probability of positive or negative outcomes when the agent adopts or reject some norms. In an earlier work, this predicament is resolved by enabling an agent to evaluate a norm's benefits if it decides to adopt a particular norm. The evaluation mechanism entails a framework that analyzes a norm's adoption ratio, yield, morality and trust, the unified values of which indicates the norm's benefits. In this paper, the trust parameter of the mechanism is analyzed and a norm's trust model is proposed and utilized in the evaluation of a norm's benefits for subsequent adoption or rejection. Ultimately, the norm's benefits are determined as a consequence of a favorable or unfavorable trust value as a significant parameter in a norm's adoption or rejection.**

***Keywords*—Norm's benefits; norm's trust; norm detection; normative multi-agent systems; intelligent software agent**

I. INTRODUCTION

Trust is one of the most important aspects in human relations. In its absence, we face problems with those around us, because trust is the basis of relations in all its forms. There are many connotations of trust in a social context [1]. Thus, trust is defined as a relationship of dependence between two parties; the first party (trustor) has the confidence to rely on another party (trustee) to adopt its actions [2], [3]. Therefore, relationships between people can be inferred from trust. Conceptually, trust is also referred to relationships within and between social groups (families, friends, communities, organizations, companies, nations, etc.). It is a popular approach to frame the dynamics of group interactions in terms of trust [4].

In sociology and psychology, trust is the subject of continuous research to measure the degree of trust to another, which is the extent of belief in honesty from the other party. According to Romano [5] who views trust from the standpoint of multiple disciplines, *"trust is a subjective assessment of another's influence in terms of the extent of one's perceptions about the first-rate and significance of another's influence on one's consequences in a given situation, such that one's expectation of, openness to, and inclination towards such influence grant a sense manage over the achievable outcomes of the situation".*

Trust can be seen as betting on potential contracts, which may bring benefits. Once the bet has been determined (i.e., confers trust), the trustor suspends his/her disbelief and does not consider the possibility of taking any negative action at all. Because of this, trust acts as a redactor of social complexity [6]. This phenomenon [7] can be compared with studies on social actors and their decision-making process, in the expectation that the understanding of this process (and modelling) permits the emergence of trust. Therefore, trust is part of the idea of social influence and on this basis, trust can be seen as a personal trait that increases personal relationships.

In an earlier work [8], it is proposed that intelligent agents should adopt or reject norms based on their awareness of the norms' expected benefits or losses rather than by sanctions or imitating other agents. Consequently, a framework constituting agents' awareness of norms' benefits is proposed, which is a formulation of Norm's Adoption Ratio, Yield, Trust, and Morality. With these parameters, agents compute the benefits of detected norms and subsequently determine whether the norms increase or decrease their utilities for eventual adoption or rejection.

Norm's Trust (NT) is one parameter in the formulation that motivates an agent to adopt a norm when the agent is able to compute a norm's trust value. A norm's trust refers to the degree of an agent's belief in a norm that influences other agents to adopt the norm. If the trust value of a particular norm is high, it increases the possibility of adopting the norm.

The motivation in this work stems from the need for software agents to detect and recognize the norms that are prevailing in a society of agents. In open normative-MAS, agents adopt norms to increase their utilities.

Implementations for such adoption are manifested by mechanisms, which are based on sanction, imitation, or social learning. However, without analyzing these norms, agents

ultimately adopt the norms, 'unaware' of its benefits for its adoption. However, in real world situations, a number of agents persistently violate the norms for their benefits, which may offer advantages in its quest to achieve their goals. Hence, it is proposed, in this work, that intelligent agents should adopt norms based on their 'awareness' of the norms' expected benefits on their utilities and not merely by sanctions or imitating other agents.

In open-MAS, numerous types of norms are enacted in many multi-agent societies. Consequently, a visitor agent must be able to evaluate all norm variations in these societies. To avoid the adverse effect of failure to comply with a society's norm, an agent must be able to evaluate a norm's trust, which is one of the factors that is perceived as beneficial for the agent in achieving its goals [9].

In this paper, the work-in-progress of the research in norm's benefits awareness is presented. It discusses the final parameter in formulating a norm's benefit, which is the norm's trust. The paper is organized as follows: Section II reviews the literature in this area. Section III discusses the development process. Section IV introduces the concept of a norm's benefit. Section V explains the concept of norm's trust. Section VI discusses the evaluation of the norm's trust. Sections VII and VIII present the social simulation and Section IX concludes the paper.

II. Literature Review

Norms are essential for the conduct of a society to establish order and harmony. Generally, people exercise the norms when they are in a new society, and occasionally, violations of the norms may be subjected to punishment or rejection by the community [10], [11]. Conversely, rewards are conferred in some cases of norms compliance. For example, when we are in a foreign country and want to use a train, we may notice people queuing, sitting and loitering while waiting for the arrival of the train. It comes to mind whether the norm (queuing, sitting or loitering) is trusted or avoiding it will lead to the failure to embark the train and have to wait for the next train? [7]. In this case, it is possible to rely on certain sources to ascertain the trustworthy of this norm. One of these sources is to enquire the authorized people at the station about that norm and whether it is trusted or distrusted [7], [12].

Occasionally, we need to know information about some things in our society and usually, we ask competent authorities. For example, if we want to know the difference between Einstein's General Theories and Special Theories of Relativity, we will certainly ask people with a specialty in Physics. This is also the case if we want to know a trusted norm in a society and how trustable it is to apply it in that environment. It is better to ask information from the authorized people in that environment. Van Dijke shows in his study how an authority affects the behaviour of workers and increase their trust in high-level authority [13].

Another reliable source is the reputation of a norm. For example, if we are looking for a new dishwasher, we would probably pick up a copy of the Consumer Report, or we may ask our friends or neighbors if they are happy with a particular brand and that would help us to choose the right one. Similarly, we use the reputation of a norm if we do not have sufficient information as to whether or not the prevailing norm is trusted. In the context of the Semantic Web, Van Dijke et al. shows an overview about the difference between the reputation metrics and explains that the reputation metrics are of two types, which are global and local reputation metrics [13]. Kiefhaber et al. shows that an entity can ask their neighbors about the reputation of another entity, their opinion of the target entity that will get transferred to their neighbors and so on [14].

Many scholars differ in their definitions of the concept of trust. Some define trust as part of the social and cognitive aspects of an organization, and many of the literature refers to it as one of the most important components of society [5], [15], [16]. Trust is an interactive relationship and a complex organizational structure between two or more parties. It arises from the urgent need to interact with members of a community. This relationship requires reliance on the others to achieve a specific goal. To establish this trust, the relationship between the parties must be free from anxiety. It is to trust or rely on someone's ability or involvement.

III. Discussion

The literature provides useful information for the development and computation of the norm's benefits concept that incorporates trust as a computational element. Topics in norms, norms detection, trust and reputation are reviewed, which provide general and basic ideas that are important to build the trust model.

While there are many techniques of norms detection that have been proposed by researchers, the issue of open MAS has made the problem somewhat complex when dealing with similar norms in multi-agent societies. Consequently, the concept of norm's benefits is chosen to enable agents to compute specific factors that contribute to the objective determination of norms for adoption in these societies.

IV. Concept of Norm's Benefits

The parameters that constitute the norms' benefits are identified from the review and analysis of the literature. In a previous work [8], these parameters are proposed to include the Norm's Adoption Ratio, Norm's Yield, Norm's Morality, and Norm's Trust. The significance of these parameters is justified by assessing the influence of each of the parameter on the decision of agents to adopt or reject a norm:

- **Norm's Adoption Ratio (N_{AR}):** It is the ratio of agents enacting a particular norm to the population of agents in a community. If P is the agents' population, and N_a is the number of agents enacting a particular norm, then $N_{AR} = N_a:P$. A high ratio is obtained when a majority of agents enact a norm while experiencing its benefits. Such experience reinforces an agent's decision to enact the norm and gain the expected benefits or violate the norm to avoid expected losses. For example, in an elevator scenario, if a majority practices the norm of *excusing* oneself when exiting the

elevator, an agent expects that the benefit from adopting such norm increases its reputation.

- **Norm's Yield (N_Y):** A norm's yield is the expected gain received from adopting a norm arising from the norm's return on an agent's utility. When an agent discovers the yield of a particular norm, it infers the benefits of adopting the norm. If the norm possesses high yield, it motivates the agent to adopt it. For example, reading news online becomes the norm of many communities because it is inexpensive and convenient.
- **Norm's Morality (N_M):** This refers to the state of a norm (good or bad) with reference to a moral code. The morality of a norm allows an agent to check whether the norm conforms to its moral code. If it conforms, the probability of adopting the norm is high and vice versa. For example, talking loudly or shouting is generally considered as a low morality norm for many communities. But if it is computed as a strong norm in a particular community, an agent has the option to accept or reject the norm basing on the norm's expected benefits.
- **Norm's Trust (N_T):** A norm's trust refers to the degree of an agent's belief in a norm that influences other agents to adopt the norm. If the trust value of a particular norm is high, it increases the possibility of adopting the norm. Andrighetto et al. [17] exemplify a bus stop scenario of a particular community, in which when people arrive at the bus stop, they do not form a queue but sit on a bench and memorize who came earlier than them. In such situation, because people highly trust the norm, they adopt the norm.

If an agent is able to determine the values of the above parameters, it can compute the norm's benefits, which offers a more elegant method to adopt or reject the norm.

Fig. 1 shows a proposed norm's benefits model. A visitor agent observes and evaluates the parameters' values (i.e., Norm's Adoption Ratio, Norm's Yield, Norm's Trust, and Norm's Morality). Having determined the parameters' values, e.g. high; medium; or low, the agent's belief is influenced by these values, which in turn influence its decision to adopt or ignore the norm.

V. Concept of Norm's Trust

Norm Trust, as a research topic, has several meanings. For example, McKnight and Chervany [2] refer trust to one party who is willing to rely on the actions of another party. For the purpose of this research:

Definition 1: A Norm's Trust is the degree to which an agent can be expected to rely on the social norms that are believed, applied and followed without adversely affecting its objectives while reaping the norm's benefits.

A. The Norms' Trust Model

This concept is validated by proposing a norm's trust model based on an agent's belief about Authority, Reputation, and Adoption for adopting the norms in a new environment.

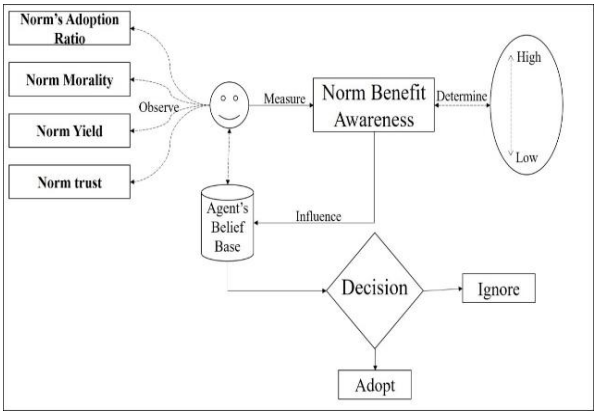


Fig. 1. Evaluating the Norm's benefit awareness.

B. Authority

A factor that determines the trust value of a particular norm is observing authorized agents, which is one of the resources for a new agent when joining a society. Authorized agents represent their societies and have the authority to reward or sanction a society's member. Therefore, authorized agents are trusted and its norm has a high trust value. The verification is justified by an agent, which endorses the norm indicating that the norm is trusted by the authorized body.

Therborn [18] states that the acceptance of a particular norm is significantly greater if the individual views the source as being credible, such as an accredited/prestigious organization, parents or people in authority. However, we exploit the agent's authority level proposed by Abdul Hamid et al. [19], who divide the trust level into three categories; low, medium and high. While Abdul Hamid et al. [19] divide the trust level into three categories, we exploit only two categories: Trust (1) and Distrust (0).

Definition 2: Authority, $\mathfrak{A}_A$, is a set of agents in the Domain D, which have the power that derives its legitimacy by respecting cultural patterns and existing rules and regulations, such as Governments.

If $\mathfrak{A}_A$ is a set of Authority agents, and $\mathbb{A}_A$ is an Authority agent, then,

$$\mathfrak{A}_A = (\mathbb{A}_{A1}, \mathbb{A}_{A2}, \dots\dots, \mathbb{A}_{An})$$

If $\mathbb{A}_A$ is an authorized agent in the domain, D, then

$$\mathbb{A}_A \in \mathfrak{A}_A \iff Authorised\ (\mathbb{A}_A, D) \wedge \mathbb{A}_A \in D \qquad (1)$$

This means that $\mathbb{A}_A$ belongs to authorized agents $\mathfrak{A}_A$, if and only if, $\mathbb{A}_A$ is authorized in D and $\mathbb{A}_A$ belongs to D. In this regard, a visitor agent asks the authorized agents about a candidate norm, $\acute{C}_{\acute{\eta}}$ whether or not it is trusted and determines the summation of authorised agents $\mathfrak{A}_A$, and if the summation is ONE then return value ONE (1), otherwise, return ZERO (0).

$$\mathfrak{A}_A(\acute{C}_{\acute{\eta}}) = \begin{cases} \acute{C}_{\acute{\eta}} = 1 & if\ \sum_{i=j}^{j} \mathbb{A}_A\,(j) = 1 \\ \acute{C}_{\acute{\eta}} = 0 & otherwise \end{cases} \qquad (2)$$

C. Reputation

Reputation is not an expectation without bounds but learning of the past. A sociologist, Barbara Misztal [20], states that reputation is a memory fixed to a particular personality. Simply, a strong reputation builds trust and thus a type of social evaluation. It is a conviction about other's assessment. Josang et al. [21] describe reputation as an opinion about an entity, therefore, interactions between people generate reputation. Experience gained from interactions between members of a society sets reputation values for others.

Shinji [22] shows that agents will be motivated due to reputation formation. Abdul Hamid et al. [19] believe that the reputation of an agent, which practices a norm in a new environment, impacts the norm's trust value. The Neighbour-Trust Algorithm is exploited to calculate the reputation score of each agent [14].

$$r_{VN} = \frac{\sum_{i \in neighbor(N)} w_{Vi} . t}{\sum_{i \in neighbor(N)} w_{Vi.}} \quad (3)$$

where r_{VN} is the reputation value, t is the direct trust values of N neighbouring agents, and w_{Vi} are the weights that represent the personal opinion of the requesting agent. These weights are normally independent of the context of the direct trust values the neighbors provide.

For example, if a visitor agent, A, wants to get information about agent C, agent A asks agent B about its opinion on agent C. In this case, w_{Vi} is the trust weight that agent A gives based on the information which agent B provides. t is the direct trust value agent B has about agent C. Later, when agent A might have a direct experience with agent C, the trust value is represented by t only. To get a more accurate value, agent A should ask many more neighbour agents.

D. Adoption Ratio

A Norm Adoption Ratio (N_{AR}) is the ratio of agents practicing a particular norm to the population of agents in a community. To calculate the N_{AR}, a formula proposed by Mahmoud et al. [10] is used. The formula is called a Norm Strength (NS). In their work, they assume that an agent observes a society's members' activities, collects episodes and add these to a record file to be analyzed for detecting the potential norms. The episode is a set of events that an agent enacts in a domain to achieve its goal. For example, in a restaurant domain, the episode might be "arrive, sit, order, eat, pay, tip, and depart" [23].

The calculation of the Norm Strength according to Mahmoud et al. [10], is as follows, where n is a norm:

$$NS(n) = \frac{Number\ of episodes\ which\ include\ n}{Total\ number of episodes} \quad (4)$$

From Fig. 2, there is an agent and a number of norms. The agent first (1) observes the norms of an environment. Then, it (2) detects the potential norm and (3) evaluates the norm based on Authority, Reputation, and Adoption to obtain the norm's trust value. The agent then (4) updates the norm's trust value of the detected norms to its belief base (5). The agent can reason and decide to comply with or even adopt the potential norm.

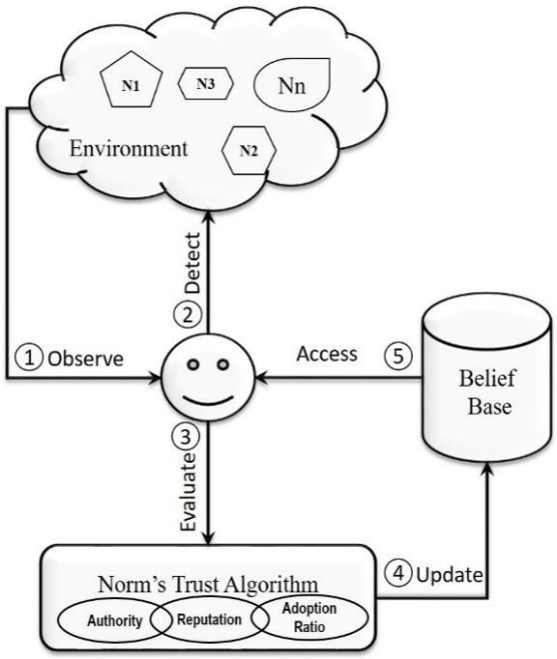


Fig. 2. The Norm's Trust Model.

The norm's trust algorithm assesses the Authority, Reputation and Adoption Ratio of the potential norm to evaluate the norms' trust value. The norms' trust value contributes to the adopt/reject decision.

VI. Norm's Trust Evaluation Model

Abdul Hamid et al. [19] propose a norm's trust concept, which is based on the transitive trust of a visitor agent who trusts a local agent's information of another local agent enacting a detected norm. This concept is exploited using the three factors associated with the process: Authority, Reputation, and Adoption Ratio.

Fig. 3 illustrates the trust inference process that applies to a particular norm. Agent A firstly observes a set of behaviours which agents B, C and D perform. Then, agent A infers the norm's trust value of the norm, n1, if agents B, C, and D perform the norm, n1. Through the three filters that influence the norm's trust, agent A evaluates the trustworthiness of the agents B, C and D and infers the norm n1's trust value.

Based on literature relating to trust and reputation models of MAS [17], [24], a number of information sources, namely anecdotal evidence, personal/direct experience, witness accounts, and social studies data that play a role in affecting a trust value are ascertained. For the purpose of evaluating the trustworthiness of agents, the trust factor as defined in the context of these models, have been used. In this research, the motives for adopting the norms, together with analysis of the mentioned sources, are both given due importance.

Based on these analyses, three main factors are categorized that influence norms existence in a society, which are Authority, Reputation, and Adoption Ratio that are mentioned earlier.

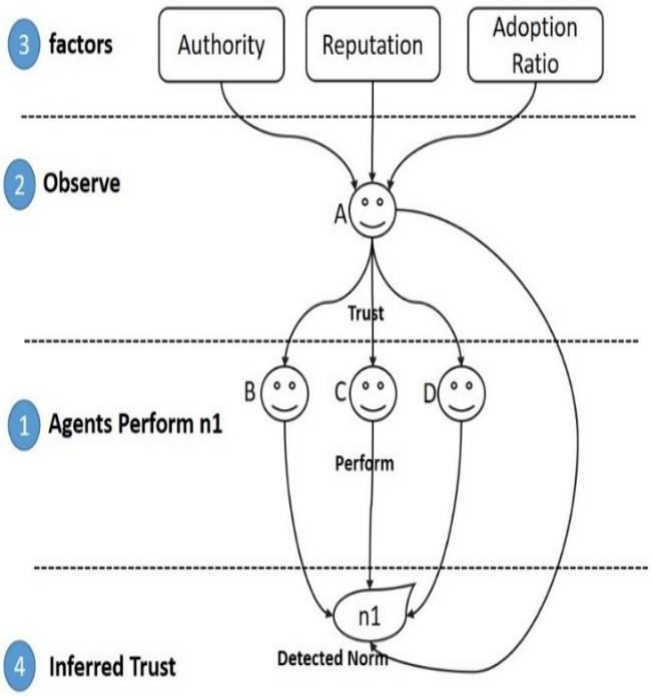


Fig. 3. Trust Inferences through Filters.

A trust value influences the decisions which can be determined from the identified factors. To determine the norms' trust (NT) value, we consider the three factors (Authority, A; Reputation, R; and Adoption Ratio, AR). We assume that the threshold value for a norm trust value, NT = 0.5. While Abdul Hamid et al. [19] describe three levels of a norm's trust, in this work only two levels are exploited:

- **Trust, NT_F:** A norm is fully trusted when all the three parameters (A, R, AR) each holds a value that jointly produces a high value of the norm's trust. There is no conflict between the values of the parameters and the agent positively verifies the norm with all factors. An agent, α, entirely trusts the norm, η, if and only if all the three parameters indicate high values of trust in the norm, η:

$$NT_F(\alpha,\eta) \leftrightarrow \text{trust}\big((A \wedge R \wedge AR),\eta\big) \qquad (5)$$

- **Distrust, NT_D:** A norm is distrusted when all the three parameters negatively produce a very low value. This means that the agent, α, distrusts the norm, η, if and only if all the three parameters indicate low values of trust in the norm, η.

$$NT_D(\alpha,\eta) \leftrightarrow \text{distrust}\big((A \wedge R \wedge AR),\eta\big) \qquad (6)$$

Therefore, the formulation of a decision to Trust or Distrust is as follows:

For an agent, α, the detected norm, η, a Trust decision is 1, and Distrust decision is 0:

$$NT(\eta a) = \begin{cases} 1, NTF \geq 0.5 \\ 0, NTD < 0.5 \end{cases} \qquad (7)$$

TABLE I. THE SUMMARY OF NORM ADOPT/REJECT DECISION

Condition	**Norm's Trust (NT) Level**	
		Decision
$NT = 1$	Trust	The agent *will* adopt a norm if its norm's trust value is equal to the highest possible value, 1. NT_F: NT = 1
$NT = 0$	Distrust	An agent *will* reject a norm if its norm's trust value is equal to the lowest possible value, 0. NT_D: NT = 0

These decisions are shown as a willingness matrix that portrays the adoption or rejection of a norm. The willingness level to adopt or reject depends on the NT threshold value (0.5). Table I shows the summary of the decision's options.

VII. SOCIAL SIMULATION

An example of a social simulation is presented, in which a visitor agent, *A*, enters a train station to take a train to another station. Agent *A* observes other local agents' behaviours in the domain and through its norm detection function, agent *A* detects three different behaviours practiced by the local agents which are; 11 agents queue and wait behind a yellow line (N_1), five agents wait while sitting on a bench (N_2), and four agents loiter around the platform (N_3). Agent *A* has to decide which behaviour it has to trust and adopt.

In this example, the first stage in a norm's trust evaluation, agent *A* evaluates its neighbours' norm trust values based on the reputation scores using (3) and the authority level [18]. Based on the Neighbour-Trust Algorithm [14] to calculate the trust level for norm n1, agent *A* evaluates the reputation score for Agent1 at this stage, by asking the neighbour agents' opinions about Agent1. It is assumed that the visitor agent *A* obtains all the reputation values, t . It then assigns the corresponding weights, w_{Vi} as shown in Table II below for each of the neighbor agents. Based on (3) the visitor agent calculates the reputation score of the potential norm.

TABLE II. REPUTATION SCORE OF NEIGHBOUR AGENTS

Agent1 Neighbor's	w_{Vi}	t	$w_{Vi}.t$
Agent2	0.99	0.92	0.9108
Agent4	0.88	0.80	0.7040
Agent8	0.89	0.88	0.7832
Agent11	0.77	0.90	0.6930
Agent15	0.66	0.88	0.5808
Agent16	0.75	0.88	0.6600
Agent19	0.95	0.88	0.8360
Sum	5.89	6.14	5.1678
	r_{VN}		0.87739

From the table, the reputation score of Agent1 is 0.87739, which is a high reputation.

In the second stage, agent *A* evaluates the authority level of Agent1 based on agent *A*'s database. Consequently, the Authority is (1). Then, in the third stage, agent *A* evaluates the Adoption Ratio. As mentioned earlier, the trust value of the potential norms (NT) is calculated based on its Adoption Ratio, AR. Using (4), the list of Reputation Scores and Authority for each neighbour and the Adoption Ratio for each potential norm is as listed in Table III. This shows the values of Reputation, Authority and the adoption Ratio of each potential norm practiced by the neighbour agents. Consequently, the visitor agent decides to adopt the norm, n1, as it is the only trusted behavior.

The trust model is validated as a simulation of the train station scenario by using Netlogo, which is a programmable agent-based modelling environment for simulating natural and social phenomena. The simulation is run five times and each run has a new environment with a different number of norms (see Fig. 4). In each run, the visitor agent observes and detects the norms in the environment, calculates and evaluates the trust value for the potential norm and decides whether to trust or distrust it.

Based on these premises, Table IV shows the simulation results. The results show that in Runs 1 and 3, the trusted norm is SIT, while in Runs 2 and 4, QUEUE is the trusted norm. Hence a visitor agent may adopt these two norms in this particular environment.

TABLE III. THE TRUST VALUE OF POTENTIAL NORMS

Practicing Agents	Norm, n_i	Neighbor, *Ni*	Reputation Score	Authority Level	Adoption Ratio, AR	Trust Level
Agent1	n1	N1	**0.87**	1	0.55	Trusted
Agent2	n1	N2	0.45	0	0.55	Distrust
Agent3	n1	N3	0.4	0	0.55	Distrust
Agent4	n1	N4	0.43	0	0.55	Distrust
Agent5	n1	N5	0.49	0	0.55	Distrust
Agent6	n1	N6	0.43	0	0.55	Distrust
Agent7	n1	N7	0.49	0	0.55	Distrust
Agent8	n1	N8	0.45	0	0.55	Distrust
Agent9	n1	N9	0.81	1	0.55	Trusted
Agent10	n1	N10	0.43	0	0.55	Distrust
Agent11	n1	N11	0.39	0	0.55	Distrust
Agent12	n2	N12	0.36	0	0.41	Distrust
Agent13	n2	N13	0.33	0	0.41	Distrust
Agent14	n2	N14	0.38	0	0.41	Distrust
Agent15	n2	N15	0.31	0	0.41	Distrust
Agent16	n2	N16	0.44	0	0.41	Distrust
Agent17	n3	N17	0.49	0	0.33	Distrust
Agent18	n3	N18	0.45	0	0.33	Distrust
Agent19	n3	N19	0.42	0	0.33	Distrust
Agent20	n3	N20	0.23	0	0.33	Distrust

TABLE IV. SIMULATION RESULTS

Simulation Runs	Potential Norm	Adoption Ratio	Authority	Reputation	Trust Value	Decision
Run 1	SIT	1	1	1	1	Trust
	QUEUE	0	1	0	0	Distrust
	LOITER	1	0	1	0	Distrust
Run 2	SIT	0	0	0	0	Distrust
	QUEUE	1	1	1	1	Trust
	LOITER	0	0	1	0	Distrust
Run 3	SIT	1	1	1	1	Trust
	QUEUE	0	1	1	0	Distrust
	LOITER	0	0	0	0	Distrust
Run 4	SIT	0	0	1	0	Distrust
	QUEUE	1	1	1	1	Trust
	LOITER	0	1	1	0	Distrust
Run 5	SIT	0	1	1	0	Distrust
	QUEUE	0	0	0	0	Distrust
	LOITER	0	0	1	0	Distrust

The findings in this research are significant in that they offer an elaborate approach to norms' analysis and computation for an eventual norm's adoption or rejection in normative multi-agent systems. The norm's adoption or rejection is based on the computation of the norms' factors which manifest the benefits that the norms would entail to achieve the agents' goals. Consequently, these findings significantly contribute to the literature in normative multi-agent systems.

VIII. SIMULATION MODEL

In this simulation model, using NetLogo designed by Uri Wilensky (1999), a virtual environment is created for calculating the norm's trust. The virtual environment is a train station which is represented by the passengers (people) and the inspector (visitor agent). The virtual environment has the functions to create a new domain, select and run a domain, and set the variables of the domain. Fig. 4 shows the user interface after opening and running a model from the Models Library. It has three parts which are:

- The top left part of the window shows the train station environment, which consists of passenger agents, senior agents, authorized agents and the visitor agent.
- The left part of the window (text box) shows the results of the norms that the agent detected. The text box shows the values of all the potential norms. It also shows the procedure of the Norm's Trust calculation for the potential norms in the domain.
- The bottom left part of the window shows the simulation buttons for controlling the simulation model and has a few boxes and buttons which are:

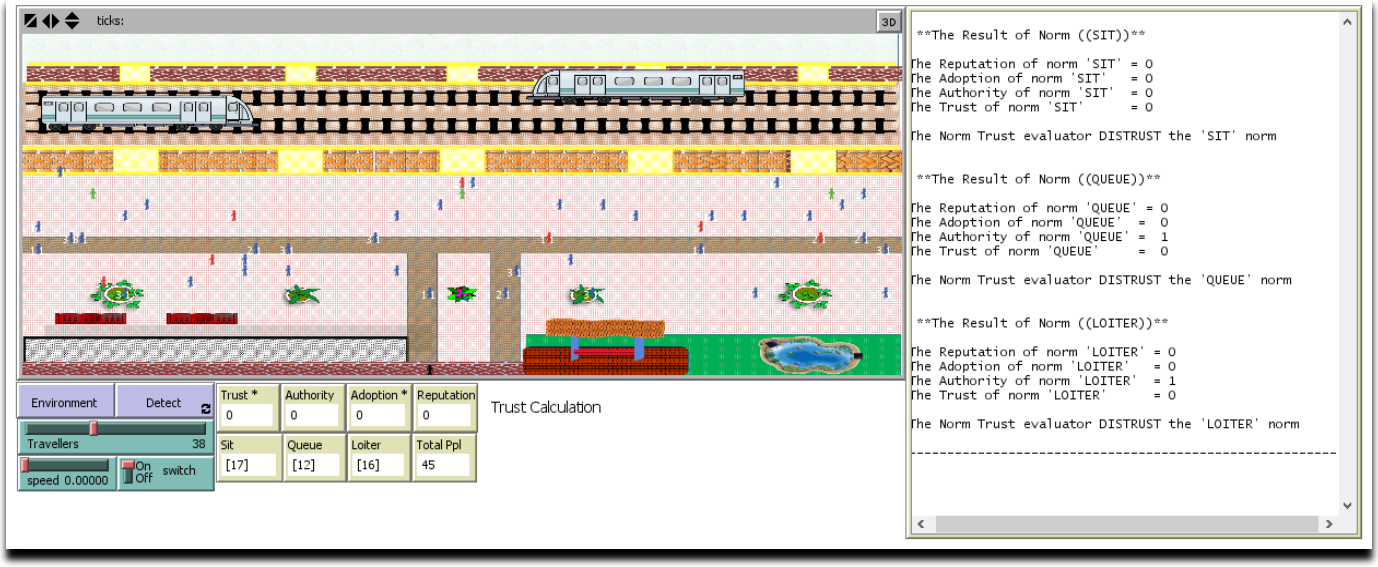


Fig. 4. Trust modelling simulation.

- **Environment:** To create a new instance of the train station. By clicking the button, a new environment is configured. The number of passengers is selected from a slider, labeled as Traveler. A number of senior agents is created randomly. Three policemen are deployed on the station platform. In addition, a visitor agent is also created. Every time the button is clicked, it creates a new instance of the domain.
- **Detect:** To enable the visitor agent to detect the potential norms and calculate the norms' trust values. When the button is clicked, the visitor agent tours the station platform and collects as much information as possible about the enacted norms by the passengers and return to its original location to analyze the collected data.
- **Travelers:** To set the number of passengers (people) at the station. By using this slider the number of passengers on the platform from 0 up to 100 passengers is created.
- **Speed:** To set the speed of the simulation using the slider.
- **Trust:** Shows the result of a trusted norm of the potential norms based on the values of Authority, Adoption Ratio, and Reputation of the potential norms. Thus, the trust is either "1" which means the potential norm is a Trusted norm or "0" which means the potential norm is a Distrusted norm.
- **Authority:** Shows the norm's trusted value of the potential norm according to authorized agents' opinion. Based on that recommendation, the result is either "1" which means the norm is trusted or "0" which means the norm is distrusted.
- **Adoption:** Shows the result of adoption ratio of the potential norm based on the number of people who trust and practice the potential norm. So, if the number of agents who trusts the potential norm is more than 50% then the adoption result shows "1" or "0" if the number of agents who trust the potential norm is less than 50%.
- **Reputation:** Shows the norm's reputation value of the potential norm according to senior agents' belief. Based on that, it shows "1" if the potential norm has a high reputation or "0" if it has a low reputation.

IX. Conclusions and Further Work

In this paper, a norm's trust model is proposed to facilitate agents' decision-making process in norm adoption or rejection. The model constitutes a technique that assists agents in determining the norm's benefits to improve agents' decisions in adopting or rejecting the norms. A norm's trust formula is exploited based on Abdul Hamid et al. [19], but a new architecture is proposed for calculating the norm's trust. The model is validated by a simulation, in which a visitor agent observes other local agents' behaviours in a train station and detects three different behaviours enacted by the local agents. The simulation results indicate that the trust model imparts a trustable value for the detected norms, which the agent can use to adopt or reject the norms.

In cases where agents encounter multiple norms, the norms' trust levels indicate how much they can be relied upon in fulfilling the normative goals (generated from the adopted norms), neither conflicting with the agents' internal structures nor interfering with their intended goals.

This paper is a part of the authors' research in agent 'awareness' of norms' benefits. A norm's trust is an important factor, whose value is needed to be determined as a parameter in the formulation of a norm's benefits. The benefits are a measure with which a decision is made whether to adopt or reject a detected norm. The other parameters in an earlier publication [8] are Norm's Adoption Ratio, Norm's Yield, and Norm's Morality.

In future works, all these parameters shall be included in the formulation of the norm's benefits and a comprehensive simulation to validate the formulation shall be developed.

REFERENCES

[1] S. Parsonsa, K. Atkinsonb, K. Haighc, K. Levittd, P. M. J. Rowed, M. P. Singhf, et al., "Argument schemes for reasoning about trust," Computational Models of Argument: Proceedings of COMMA 2012, vol. 245, p. 430, 2012.

[2] D. H. McKnight and N. L. Chervany, "The meanings of trust," 1996.

[3] J.-M. Seigneur and P. Dondio, "Trust and reputation for successful software self-organisation," in Self-organising Software, ed: Springer, 2011, pp. 163-192.

[4] R. Falcone and C. Castelfranchi, "Social trust: A cognitive approach," in Trust and deception in virtual societies, ed: Springer, 2001, pp. 55-90.

[5] D. M. Romano, "The nature of trust: conceptual and operational clarification," Louisiana State University, 2003.

[6] S. C. Currall and M. J. Epstein, "The Fragility of Organizational Trust:: Lessons From the Rise and Fall of Enron," Organizational Dynamics, vol. 32, pp. 193-206, 2003.

[7] N. H. A. Hamid, M. S. Ahmad, A. Ahmad, A. Mustapha, M. A. Mahmoud, and M. Z. M. Yusoff, "Trusting Norms: A Conceptual Norms' Trust Framework for Norms Adoption in Open Normative Multi-agent Systems," in Distributed Computing and Artificial Intelligence, 12th International Conference, S. Omatu, Q. M. Malluhi, S. R. Gonzalez, G. Bocewicz, E. Bucciarelli, G. Giulioni, et al., Eds., ed Cham: Springer International Publishing, 2015, pp. 149-157.

[8] A.-M. K. Itaiwi, M. S. Ahmad, M. A. Mahmoud, and A. Y. Tang, "Norm's Benefit Awareness in Open Normative Multi-agent Communities: A Conceptual Framework," in Distributed Computing and Artificial Intelligence, 11th International Conference, 2014, pp. 209-217.

[9] A. Artikis, M. Sergot, and J. Pitt, "Specifying norm-governed computational societies," ACM Transactions on Computational Logic (TOCL), vol. 10, p. 1, 2009.

[10] M. A. Mahmoud, A. Mustapha, M. S. Ahmad, A. Ahmad, M. Z. M. Yusoff, and N. H. A. Hamid, "Potential norms detection in social agent societies," in Distributed Computing and Artificial Intelligence, ed: Springer, 2013, pp. 419-428.

[11] B. T. R. Savarimuthu, S. Cranefield, M. A. Purvis, and M. K. Purvis, "Identifying prohibition norms in agent societies," Artificial intelligence and law, vol. 21, pp. 1-46, 2013.

[12] J. Duffy, H. Xie, and Y.-J. Lee, "Social norms, information, and trust among strangers: theory and evidence," Economic theory, pp. 1-40, 2013.

[13] M. Van Dijke, D. De Cremer, and D. M. Mayer, "The role of authority power in explaining procedural fairness effects," Journal of Applied Psychology, vol. 95, p. 488, 2010.

[14] R. Kiefhaber, S. Hammer, B. Savs, J. Schmitt, M. Roth, F. Kluge, et al., "The neighbor-trust metric to measure reputation in organic computing systems," in Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2011 Fifth IEEE Conference on, 2011, pp. 41-46.

[15] R. Falcone, C. Castelfranchi, H. L. Cardoso, A. Jones, and E. Oliveira, "Norms and Trust," in Agreement Technologies, ed: Springer, 2013, pp. 221-231.

[16] P. Faulkner, "Norms of trust," 2010.

[17] G. Andrighetto, D. Villatoro, and R. Conte, "Norm internalization in artificial societies," Ai Communications, vol. 23, pp. 325-339, 2010.

[18] G. Therborn, "Back to norms! On the scope and dynamics of norms and normative action," Current Sociology, vol. 50, pp. 863-880, 2002.

[19] N. H. A. Hamid, M. S. Ahmad, A. Ahmad, A. Mustapha, M. A. Mahmoud, and M. Z. M. Yusoff, "Trusting Norms: A Conceptual Norms' Trust Framework for Norms Adoption in Open Normative Multi-agent Systems," in Distributed Computing and Artificial Intelligence, 12th International Conference, 2015, pp. 149-157.

[20] B. Misztal, Trust in modern societies: The search for the bases of social order: John Wiley & Sons, 2013.

[21] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," Decision support systems, vol. 43, pp. 618-644, 2007.

[22] S. Teraji, "A theory of norm compliance: Punishment and reputation," The Journal of Socio-Economics, vol. 44, pp. 1-6, 2013.

[23] B. T. R. Savarimuthu, "Mechanisms for norm emergence and norm identification in multi-agent societies," University of Otago, 2011.

[24] I. Pinyol and J. Sabater-Mir, "Computational trust and reputation models for open multi-agent systems: a review," Artificial Intelligence Review, vol. 40, pp. 1-25, 2013.

AUTHORS' PROFILES

Al-Mutazbellah K. Itaiwi has obtained his B.Sc. in Computer Science from the College of Computer, University of Anbar, Iraq in 2007. he obtained his Master of Information Technology at the College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Malaysia in 2012 and enrolled in the PhD of Information and Communication Technology program since 2013 at the College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Malaysia. During his studentship at UNITEN, he conducted additional laboratory work for degree students at the College of Engineering. his current research interests include software agents and multi- agent systems

Mohd Sharifuddin Ahmad is currently the Head of Center for Agent Technology (CAT) at the College of Computer Science and Information Technology, Universiti Tenaga Nasional (UNITEN). He obtained his MSc. in Artificial Intelligence from Cranfield University, UK in 1995. He obtained his PhD. in Artificial Intelligence from Imperial College, London, UK in 2005. His research interests include Software Agents and Knowledge Management.

Alicia Y.C. Tang currently works at the Department of Systems and Networking, Universiti Tenaga Nasional (UNITEN). Alicia does research in Agents & Autonomous Systems, Data Mining and Artificial Intelligence. Their current project is 'i-VSM and i-RAM based on the concept of Agents of Things (AoT).'

Day-ahead Base, Intermediate, and Peak Load Forecasting using K-Means and Artificial Neural Networks

Lemuel Clark P. Velasco, Noel R. Estoperez, Renbert Jay R. Jayson,
Caezar Johnlery T. Sabijon, Verlyn C. Sayles
Mindanao State University-Iligan Institute of Technology
Iligan City, The Philippines

***Abstract*—Industries depend heavily on the capacity and availability of electric power. A typical load curve has three parts, namely, base, intermediate, and peak load. Predicting the three (3) system loads accurately in a power system will help power utilities ensure the availability of the supply and to avoid the risk for over- or under- utilization of generation, transmission, and distribution facilities. The goal of this research is to create a suitable model for day-ahead base, intermediate and peak load forecasting of the electric load data provided by a power utility company. This paper presents an approach in predicting the three (3) system loads using K-means clustering and artificial neural networks (ANN). The power utility's load data was clustered using K-means to determine the daily base, intermediate and peak loads that were then fed into an ANN model that utilized Quick Propagation training algorithm and Gaussian activation function. It was found out that the implemented ANN model generated 2.2%, 1.84%, and 1.4% as the lowest MAPE for base, intermediate, and peak loads, respectively, with highest MAPE below the accepted standard error rate of 5%. The results of this study clearly suggest that with the proper method of data preparation, clustering, and model implementation, ANN can be a viable solution in forecasting the day-ahead base, intermediate, and peak load demand of a power utility.**

Keywords*—*K-means clustering; artificial neural networks; base intermediate and peak load; day-ahead load forecasting

I. INTRODUCTION

Electricity needs to be consumed the moment it is generated. Thus, electric companies should plan how much energy is needed to be bought from suppliers in order to meet consumers' demands [1]-[3]. Failure to provide industries and consumers will result to tremendous loss of resources while obtaining surplus amount of energy will result to underutilization of electric utility resources. If the power utility cannot provide the proper amount of electricity as soon as the consumer demanded it, the power quality would lead into service interruption [1], [3]. Hence, it is a pivotal point to obtain an accurate forecast of electric power systems in order to meet the changing power consumption consumers. A typical load curve has three parts namely base, intermediate, and peak loads. Base load is the minimum level of electricity required for a period of twenty four hours and provides power that keeps running constantly [4]. Intermediate and peak load are the next to be brought in-line whenever the demand increases above the base electric load.

Predicting the three system load: base, intermediate, and peak load accurately in a power system will help power utility companies ensure the availability of the supply as well as avoid the risk for over or under utilization of generation, transmission, and distribution facilities [4], [5]. Having accurate estimates of base load demand in a power systems will also aid electric companies to possibly meet the demand of the market as this contributes to the continuous large amount of electricity in any system. Moreover, predicting the precise peak energy usage will further help the company in determining the need of purchasing the new and existing resources, as well as the type of resources that are necessary to meet the consumers demand. Furthermore, having knowledge of the peak energy usage will help the electric utility plan for the extension of existing facilities and installation of new power plants to reliably meet consumer demand [5]. K-means clustering technique and Artificial Neural Networks (ANN) model are suggested as suitable approach in clustering and forecasting load profile that can determine base, intermediate and peak loads [2], [3], [6], [7]. K-means is a widely used unsupervised learning algorithm that solves clustering problems or data categorizing. Moreover, it follows the simple and easy way in classifying datasets into assumed *k* clusters [6], [7]. While ANN is a mathematical model that has powerful classification and gained a significant performance in most researches mainly load forecasting due to its flexibility and its generality. Being one of the popular machine learning technique, supervised ANN has been widely used and has been proven to predict promising forecast of electric load but there are still gaps in coming up with a serial process in its usage along with unsupervised K-means [1], [2], [6].

In the Philippines, a power utility company faces a major problem in determining the base, intermediate, and peak load in their decision making since estimating these system loads is currently done by assigning assumed values resulting to inaccurate guess. This paper attempted to present a new technique for data preparation that is based on K-means clustering in order to determine the three system loads. Furthermore, an ANN model was also introduced and implemented to predict the day-ahead base, intermediate and peak loads. With proper data analysis along with appropriate

data clustering for data to be fed into a prediction model, this study aims develop a forecasting tool which power utilities can use to augment the gap of supply and demand in electric load.

II. METHODOLOGY

A. Load Data Preparation and K-means Clustering

This study used monthly electric load data for three years from 2012-2014 that has been used from an existing system of an electric power utility company. As shown in Table I, the worksheet of the data contain three sheets corresponding to the three metering points of the power utility company which contains the metering point name, the date, time, kilowatt delivered (KW_DEL), kilowatt per hour delivered (KWH_DEL) and kilovolt amps reactive hours delivered (KVARH_DEL).

TABLE I. FORMAT OF THE RAW LOAD DATA

SEIL	BDATE	TIME	KW_DEL	KHW_DEL	KVARH_DEL
XXX	XXX	XXX	XXX	XXX	XXX
XXX	XXX	XXX	XXX	XXX	XXX

Electric load data was grouped in terms of its metering points by creating three databases while eliminating columns that was not be used in the study. Before importing the data into the database, corrections were made through filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies in order to avoid erroneous data that could make the model uncertain [8]. The three metering points were then aggregated to come up with the total load duration of the entire locality represented by the three metering points.

Cluster centers by day of the electric load data was then determined using K-means clustering method. K-means clustering is a partitional clustering method that attempts to find the number of clusters (*k*), which are represented by centroids [6], [9]. Cluster parameters such as number of clusters (*k*) and maximum number of iterations (*i*) were then determined. Choosing the number of *k* clusters has no agreed upon solution, thus finding the number of *k* clusters varies on the number of clusters desired. Three basic steps in K-means clustering were then conducted [6], [7], [9]. Firstly, data items were partitioned into I initial cluster, where data consists of data points from x_1, x_2..,x_n and each point is assigned to the nearest centroid. Thereafter, it identified the group of data points and assigned each point to one group or one cluster. Secondly, assignment was done to an item to which the mean or the centroid was the nearest. Similarity of clusters was computed using mean value of the objects which were also considered as cluster centers or centroids. Lastly, an iterative process of assigning the centroids until no more reassignments of cluster centers was done. This process will stop iterating when there will be no point to navigate each cluster to another and when the centroid will remain the same [7], [10]. A graphical representation of the resulting cluster centers was generated with the load curve and load duration curve in a day in order to present the percentages of the base, intermediate, and peak load.

The clustered data was then partitioned into two data sets: training set and testing set. The training set was used for training and adjusting the weights on the neural network while testing set was used for testing the design of the network to confirm the actual predictive power. If the input data of the predictive model is not normalized, the training of the network would be slow [11]. There are various normalization methods that would produce values either from a range of 0 to 1 or -1 to 1. However, in this study only Min-max normalization method was used as suggested by authors dealing with load prediction using ANN [8], [12]. After normalizing the training set, the normalized load data, that contains the inputs and their corresponding expected outputs, will then be fed into the neural network for training. On the other hand, the normalized dataset will be inputted into the model, trained and tested, the resulting outputs of the neural network will undergo denormalization as shown in (1) to show the actual value where $x = x_1, x_2,...x_n$, y is the denormalized data and z_i is the normalized data.

$$y = z_i\,(max(x) - min(x)) + min(x) \quad (1)$$

B. ANN Model Implementation

An ANN model with multilayer perceptron as an architecture having input layer, hidden layer, and output layer was implemented in a load prediction system. Fig. 1 shows that the ANN model has eight input neurons consisting of the day of the week, holidays or non-holidays, weekends or weekdays, week number, and month. The ANN used four hidden neurons along with Quick Propagation training algorithm and Gaussian activation function. The signals of the ANN were multiplied with bias weights and mapped into three output nodes indicating the day-ahead base, intermediate, and peak load consumption. The ANN model was implemented through desktop-based software with the use of Encog library in order to achieve the training and testing results. Encog library is a Java-based library which provides interchangeable models with efficient, internal implementations and supports machine learning models with choice of training algorithms [12]. System features and use cases that will carry out the processes needed for the clustering of the actual data up to the load prediction, starting from data loading, clustering, and ending in generating of the predicted values. The functions were grouped into classes such as the data scaling, clustering, database querying and the ANN model class for the training, testing, and forecasting functions. Other features such as export of clustered datasets and charts were also included for the purpose of easier comparison of the data.

A validation set from January 2015, outside the training and testing sets was used to test the model's accuracy. Clustering of the validation set was also conducted in order to determine their actual base, intermediate and peak loads while actual load of every past day was appended to the model to predict its day-ahead load. An important criterion in evaluating the prediction accuracy of the forecasting model is to compute the measure of error. The accuracy of the prediction model can often be defined as forecasting error, which is the difference between the actual load and the predicted load [1], [3], [5], [11]. In order to evaluate the performance of the neural

network model quantitatively, error measures was calculated in this study. Denormalization of the ANN output was conducted and used Mean Absolute Percentage Error (MAPE) which measures the error in terms of percentage and calculated as the average percentage error to compare the denormalized data to the actual data [2], [11], [12]. After denormalizing the data, the forecasted values were evaluated by comparing it to the clustered data to check if the forecasted value is close enough to the clustered data. A graphical representation of the computations was then generated for the purpose of illustrating the comparison between the clustered actual and predicted base, intermediate and peak load values.

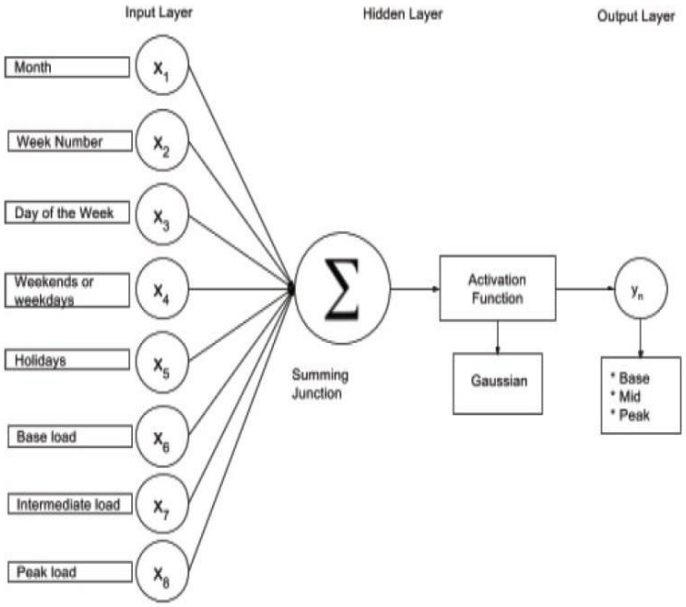


Fig. 1. Block diagram of the implemented ANN Model.

III. Results and Discussion

A. Load Data Preparation and K-means Clustering Results

Based on the raw load dataset, the data used as input for clustering were the date, time and the kilowatt delivered (KW_DEL) which is the consumed load. The same columns were also used by researches in predicting electric load which disregarded SEIL, KWH_DEL and KVARH_DEL [2], [12]. Certain parameters were set in clustering the raw electric load data such as the number of clusters (*k*) and the maximum number of iterations (*i*). The number of clusters was set into three which represents base, intermediate, and peak load. Using K-means clustering technique, the intervals associated with the base, intermediate, and peak load were calculated. In the first step, the number of clusters was assumed to be three and initial cluster centers were determined.

Table II shows the first *i*=1 iteration that calculated the three cluster centers. These cluster centers has the sample values of 17,402.2929 KW for base, 20,023.5783 KW for intermediate, and 25,302.8772 KW for peak. These cluster centers will be reassigned until the groupings of data points will stay the same as exemplified by groups of authors [6], [7], [10].

TABLE II. Sample Cluster Centers Iteration

i	Base	Intermediate	Peak
1	17, 402.2929	20,023.5783	25,302.8772
2	16,820.9705	20,538.0019	26,137.5521
3	16,820.9705	20,645.4180	26,202.7391
4	16,820.9705	20,645.4180	26,202.7391

Fig. 2 depicts the *k* cluster grouping points from the 96 observations of a day's 15-minute load intervals. Blue represents the base clusters, orange represents the intermediate clusters, and red represents the peak cluster points while the star sign represents the initial cluster center (*k*) of the data points. In the second step, K-means converges at 4th iteration, *i* = 4 depicting the iteration of the resulting cluster centers. K-means converges faster whenever the initial cluster centroids are selected [7], [13]. It shows that *k* clusters converge at iteration 4 and after 4th iteration there is no reassignment of cluster centers. The final cluster centers were determined after there was no more changes occurring in clustering the datasets and when final cluster centers have found the natural grouping of points in each cluster centers with the same values. Final cluster centers has the value of 16,820.9705 KW for base, 20,645.4180 KW for intermediate, and 26,202.7391 KW for peak, this means that these values will no longer be reassigned.

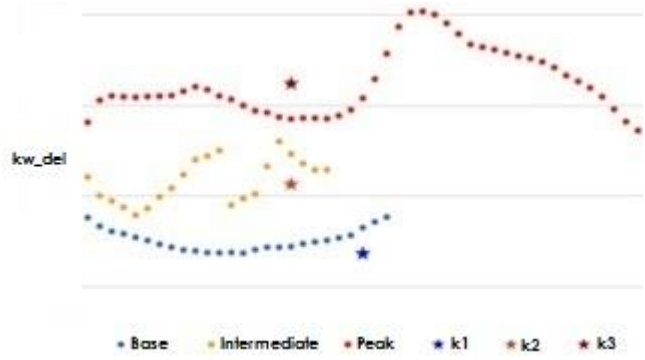


Fig. 2. Sample *k* clusters.

Load duration curve is the arrangements of all load levels into descending order of magnitude. It can be beneficial for electric power system in economic dispatching, system planning, and reliability evaluation. Moreover, it helps to determine if there is a need to come up with replacement decisions due to an over or under loading condition [14]. With the three clustered data produced, it was then mapped out together with the load duration curve to effectively show how much base, intermediate, and peak load demand is required in the electric power system. Fig. 3 describes the sample load duration curve. It is show that for a particular day, the electric power system requires 59.01% for base electric load. This means that 59.01% will be used for the continuous supply of large amount of electricity. Coal-fired plants and nuclear units

are appropriate for the base load station [4], [14]. On the other hand, intermediate electric load requires 27.83% of the time period. Combine cycle units are used in the intermediate power plants. Moreover, the system requires 13.16% peak electric load. Diesel, hydro and gas turbines belongs in the peak load station categories [15].

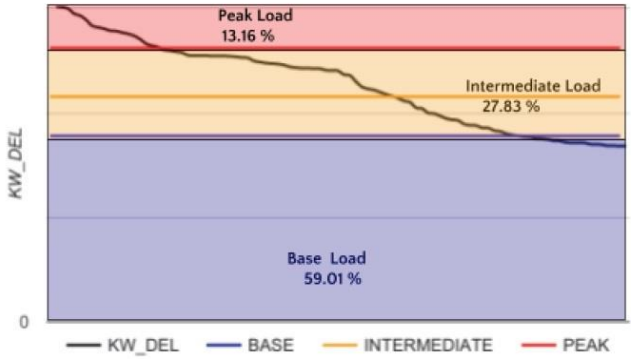


Fig. 3. Sample load duration curve.

Out of the 101,760 observations, 3,180 emerged after clustering. The original dataset having the three metering points has 96 observations each day representing the 15-minute interval load data each day. After the clustering process, the aggregated consumption of the locality was grouped in 1,060 days of three data points representing the base, intermediate, and peak clusters. The clustered load data was then divided into training set and testing where 70% of the data given was partitioned into training data comprising the clustered load data from 2012 to 2013 and the remaining 30% of the clustered 2014 load data was set for testing data . As shown in Table III, the clustered data was normalized using min-max normalization that produces an output by the range of 0 to 1 since no negative values was observed in the clustered data. Min-Max normalization performs a linear transformation on the original data and preserves the relationship among data values [11]. This normalization technique was used in order for the data to be fed into the neural network for training dataset and to test the accuracy of the prediction.

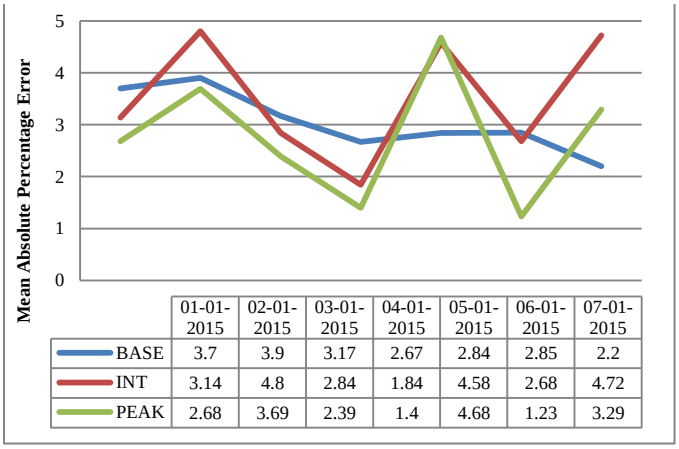
TABLE III. SAMPLE NORMALIZED DATA RESULTS

Date	Base	Intermediate	Peak
XXX	0.240666598	0.23465889	0.238462481
XXX	0.239323519	0.232076194	0.245447171

B. ANN Model Implementation Results

Using clustered data of a validation set, MAPE was calculated for a week's daily prediction in order to evaluate the performance of the model. As shown in Fig. 4, highest MAPE were 3.90%, 4.80%, and 4.68%, respectively for base, intermediate and peak loads. Lowest MAPE was 2.2%, 1.84%, and 1.4% respectively for base, intermediate and peak loads with peak loads having majority of the lowest MAPE should the three be compared. The MAPE of the clustered and the predicted peak load has always been the lowest among the three except on January 5 where it also has its highest MAPE. The clustered intermediate and peaks loads seem to show a trend different from that of base load. It can also be observed that except on January 6 and 7, all of the three loads were all together increasing or decreasing in MAPE values until base load has broken the uniformity on the last two days. Base load is generally steady as validated in the figure showing intermediate load adjusting with the performance of the base and the peak load [5], [14]. According to studies, acceptable error range for load prediction is between 3%-3.5% while the corporate tolerance error of the power utility being tested is below 5% [2], [11], [12], [15]. We can then infer that the performance of the clustering and the prediction is acceptable.

The forecasted values were denormalized and compared to the clustered values of the validation set. In order to assess the forecasted result, visualization was made in each category to compare the forecasted results and the clustered data of the base, intermediate, and peak load. As shown in Fig. 5, the forecasted values were compared to clustered data of the validation set's base load which are generally bought by power utility companies on bilateral and long-term contracts [4], [14]. It can be observed that on January 2, 2015, the highest MAPE for the base load of 3.90% shows the only instance that predicted base load is below the actual clustered base load.



	01-01-2015	02-01-2015	03-01-2015	04-01-2015	05-01-2015	06-01-2015	07-01-2015
BASE	3.7	3.9	3.17	2.67	2.84	2.85	2.2
INT	3.14	4.8	2.84	1.84	4.58	2.68	4.72
PEAK	2.68	3.69	2.39	1.4	4.68	1.23	3.29

Fig. 4. Performance of the implemented forecasting model.

Fig. 5. Comparison between clustered base load vs. predicted base load.

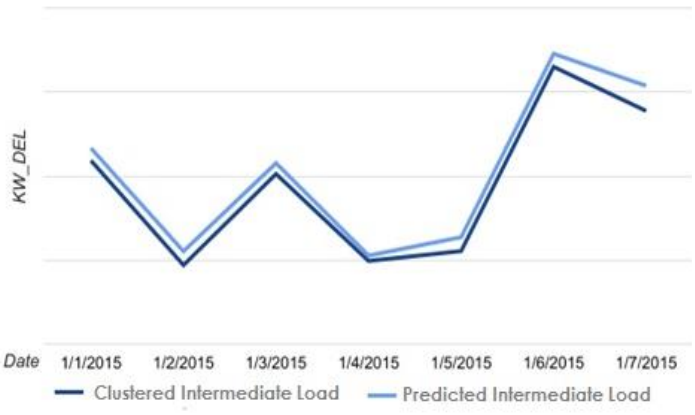


Fig. 6. Comparison between clustered intermediate load vs. predicted intermediate load.

Fig. 6 shows the predicted intermediate load being compared to clustered intermediate load which are usually bought by power utility companies on spot markets because of affordability. As shown, the forecasted intermediate values for the intermediate load were close enough to the actual intermediate values. Since MAPE values can either be above or below the predicted and the actual, it shows here that predicted load is always above the clustered intermediate load.

Fig. 7 shows the comparison between the forecasted and the clustered peak load. It can be observed that on January 1, 2, and 3 predicted peak load is higher than the clustered peak load while in the January 4, 5, and 6 shows the clustered peak load higher than the predicted peak load. January 5, 2015 has the highest MAPE of 4.68% followed by the lowest MAPE of 1.23%. Although peak load has the same trend with that of the intermediate load, this does not always mean uniformity in occurrence on that the predicted is also higher than the actual since MAPE are absolute zero values. What is notable is that regardless whether whichever is higher or lower than the other, the trend of the intermediate load adjusts with that of the peak load being evident on the MAPE [5], [14], [16].

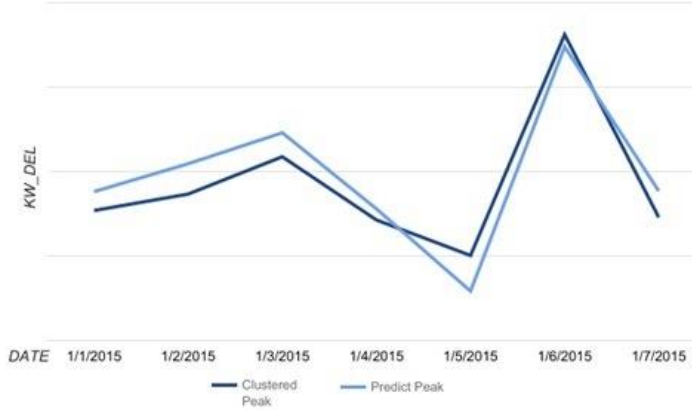


Fig. 7. Comparison between clustered peak load vs. predicted peak load.

IV. Conclusion and Recommendations

This study attempted to develop a forecasting tool that will predict the day-ahead base, intermediate and peak loads by clustering the historical delivered daily electric loads using K-means and employing ANN to forecast the day-ahead clustered loads. The goal was achieved by data preparation, K-Means clustering, ANN implementation, and comparison of the forecasted results with the derived clustered base, intermediate, and peak loads. After successfully determining the daily base, intermediate and peak loads using K-Means, a multilayer perceptron neural network with eight input neurons, four hidden neurons along with Quick Propagation training algorithm and Gaussian activation function was implemented in Encog. Highest MAPE of the forecasting tool were 3.90%, 4.80%, and 4.68% while lowest MAPE were 2.2%, 1.84%, and 1.4% respectively for base, intermediate and peak loads. Techniques for the clustering the daily loads other than K-means clustering technique could be investigated to determine the daily base, intermediate and peak loads.

For future work, it is recommended that performance analysis on different training algorithms and activation functions be conducted to develop a more optimized model. Aside from month, day of the week, weekend/weekdays, holiday/non-holidays indicators that this study used as input neurons, other additional factors that can affect the training process of the neural network such as temperature and weather variables can be considered by future researches as essential factors of the prediction model to generate better results of the forecasting model. This study aims to help electric system decision makers by discussing concepts of developing a close to accurate forecasting technique in predicting the base, intermediate, and peak loads in order for power utilities to come up with better short, medium and long term decisions.

References

[1] K.Y. Lee, Y.T. Cha and J.H. Park, “Short-term load forecasting using an artificial neural network”, IEEE Transactions on Power Systems, Vol. 7, Issue 1, February 1992.

[2] L.C. Velasco, C. Villezas, P.N. Palahang, and J.A. Dagaang, “Next day electric load forecasting using Artificial Neural Networks”, International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management, IEEE DOI: 10.1109/HNICEM.2015.7393166, December 2015.

[3] H. Quan, D. Srinivasan, and A. Khosravi, “Short-Term Load and Wind Power Forecasting Using Neural Network-Based Prediction Intervals”, IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, Issue 2, pp: 303 – 315, February 2014.

[4] M. Ndlovu and T. Lekalakala, “Renewable energy options and base load stations for South African community”, 8th Conference on the Industrial and Commercial Use of Energy, IEEE, October 2011.

[5] L. Jin, L. Ziyang, S. Jingbo, S. Xinying, “An efficient method for peak load forecasting”, The 7th International Power Engineering Conference, IEEE, Print ISBN: 981-05-5702-7 , DOI: 10.1109/IPEC.2005.206877, December 2005.

[6] S. Nuchprayoon, “Electricity load classification using K-means clustering algorithm”, 5th Brunei International Conference on Engineering and Technology, IEEE, Print ISBN: 978-1-84919-991-9, DOI: 10.1049/cp.2014.1061, June 2015.

[7] A, Alkilany, A. Ahmed, H. Said, and A. Abu Bakar, “Application of the k-means clustering algorithm to predict load shedding of the Southern Electrical Grid of Libya”, DOI: 10.1109/INTECH.2014.6927763, October 2014.

[8] L.C. Velasco, A.R. Granados,J. M. Ortega and K.V. Pagtalunan, “Medium-Term Water Consumption Forecasting using Artificial Neural Networks” 17th Conference of the Science Council of Asia, National Research Council of the Philippines, June 2017.

[9] S.A. Azad, A. B. M. Shawkat Ali, and P. Wolfs, "Identification of typical load profiles using K-means clustering algorithm", Asia-Pacific World Congress on Computer Science and Engineering, IEEE, DOI: 10.1109/APWCCSE.2014.7053855, November 2014.

[10] M. Ghofrani, M. de Rezende, R. Azimi, and M. Ghayekhloo, "K-means clustering with a new initialization approach for wind power forecasting", Transmission and Distribution Conference and Exposition, IEEE, 10.1109/TDC.2016.7519931, July 2016.

[11] L.C. Velasco, P.N. Palahang, C. Villezas, and J. A. A. Dagaang, "Performance Analysis of Different Combination of Training Algorithms and Activation Functions in Predicting the Next Day Electric Load", 16th Philippine Computing Science Congress, Computing Society of the Philippines, March 2016.

[12] L.C. Velasco, P. Bokingkito, and J. T. Vistal, "Week-Ahead Load Forecasting using Multilayer Perceptron Neural Network for a Power Utility", 17th Conference of the Science Council of Asia, June 2017.

[13] W. L. Zhao, C.H. Deng, and C.W. Ngo, "Boost K-Means", Computer Vision and Pattern Recognition, Cornell University Library, December 2016.

[14] A. Salimi-beni, D. Farrokhzad, M. Fotuhi-Firuzabad, S.J. Alemohammad, "A New Approach to Determine Base, Intermediate and Peak-Demand in an Electric Power System", International Conference on Power System Technology, IEEE, DOI: 10.1109/ICPST.2006.321928, October 2006.

[15] A. Singh and V.K. Tripathi, "Load Forecasting Using Multilayer Perceptron Neural Network", International Journal of Engineering Science and Computing, DOI 10.4010/2016.1135, ISSN 2321 3361, pp. 4548 - 4551, May 2016.

[16] F. Ueckerdt and R.Kempener, "From Base load to peak: renewables provide a reliable solution", International Renewable Energy Agency, 2015.

Proposed an Adaptive Bitrate Algorithm based on Measuring Bandwidth and Video Buffer Occupancy for Providing Smoothly Video Streaming

Saba Qasim Jabbar
School of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China

Dheyaa Jasim Kadhim
Electrical Engineering Department, College of Engineering, University of Baghdad, Baghdad, Iraq

Yu Li
School of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China

***Abstract*—Dynamic adaptive streaming via HTTP (DASH) has been popular disseminated over the Internet especially under the circumstances of the time varying network, which it is currently the most challenging for providing smoothly video streaming via high quality. In DASH system, after completing the download of one segment, the player estimates the available network bandwidth by calculating the downloading throughput and then adapting the video bitrate level based on its estimations. However, the estimated bandwidth in the application layer is not accurate due to off-intervals appearance during the downloading process. To avoid the unfairness of bandwidth estimation by the clients, this work proposes a logarithmic approach for received network bandwidth, which includes increasing or decreasing this bandwidth logarithmically to converge the fair share bandwidth (estimated bandwidth). After obtaining the measured bandwidth, an adaptive bitrate algorithm is proposed by considering this measured bandwidth in addition to video buffer occupancy. The video buffer model is associated with three thresholds (i.e. one for initial startup and two for operating thresholds). When the video buffer's level stays between the two operating thresholds, the video bitrate will keep unchanged. Otherwise, when the buffer occupancy is too high or too low, an appropriate video bitrate is chosen to avoid buffer overflow/underflow. Simulation results show that the proposed scheme is able to converge the measured bandwidth to the fair share bandwidth very quickly. Also the proposed scheme is compared with conventional scheme, we found that our proposed scheme outperforms in achieving the best performance in terms of efficiency, stability and fairness.**

***Keywords*—DASH; video streaming; video buffer; video adaptive bitrate algorithm, QoE**

I. INTRODUCTION

In last few years, HTTP adaptive video streaming has been widely used in Internet video technologies such as Microsoft Smooth Streaming, Netflix, Apples HLS, Adobes HDS and Akamai HD [1], [2], since the use of HTTP over TCP is easy to configure and simplify the traversal of firewalls. Moreover, the standard of HTTP based adaptive video streaming is the Dynamic Adaptive Streaming over HTTP (DASH) [3]. In DASH systems, each video is encoded into multiple representations of different bitrates and each representation is divided into multiple segments (2-10 seconds of video time). At the client side, a client requests and receives video segments from DASH servers that own the segments continuously. One of the most important features in DASH system is adapting the video bitrate to a varying network bandwidth dynamically. In such way, DASH clients' can enjoy with video at maximum quality possible since they receive video segments from different versions each of which being encoded with specific bitrates in a way throttling the visual quality to match the available network bandwidth.

It is important challenging to satisfy the user experience during a whole video session under time-vary network conditions. Without a powerful adaptive bitrate algorithm, DASH's client may face frequent interruptions and which it is degraded his video quality. For example, if the selected video bitrate is higher than the available bandwidth, it will cause network congestion. On the other hand, if the video bitrate is lower than the available bandwidth, the visual video quality would not reach the maximum allowed by the available bandwidth. Besides, smooth video bitrates and stable bitrates switching are preferred through video playback [4]. Due to the On-Off phase which it is a natural phenomenon in DASH system (ON means downloading a segment and OFF means staying idle), the competing clients may face difficultly in estimating their bandwidths which usually leads to several performance problems such as inefficiency, instability, and unfairness [5].

There are many research problems about video streaming for multiple DASH clients over TCP network still open and challenging. As an example, the tradeoff between the stability and efficiency which it is very important issue to have smooth video bitrates. Under time-varying bandwidth condition, requesting a low video bitrate will produce more opportunities for rate selection, and therefore well ensure high stability (smoothness) with continuous video playback but it also causes a low video quality with low bandwidth usage (inefficiency). Another tradeoff challenging is between sensitivity and stability, since the channel bandwidth is inherently different in time, the high sensitivity of the bitrate control technique usually makes the video bitrate identical with the bandwidth and hence leading to high instability. Such these challenges become more troublesome when multiple clients compete on the bottleneck link because each client will try to optimize the quality of the video without looking at others. Besides, the fairness problem arises for multiple clients in DASH

technology since they are deployed via HTTP/TCP network. The existing bitrate adaptive algorithms for DASH system are aiming to either achieve the efficiency of high bandwidth usage for video adaptation bitrate to match available bandwidth, or maintain continuous video playback by homogeneity the video bitrate to avoid buffer overflow/underflow such as bandwidth-based approaches discussed in [6]-[8] and buffer-based approaches in [9][10].Recently, the problems with video streaming for multiple DASH clients competing over a common bottleneck have been studied in many researches such as [11]-[13].

In [11] and [12], the authors studied problems of bitrate adaptation and determined the causes of many unwanted interactions that arise as a result of modifying the video bitrate over HTTP. Furthermore, a series of techniques have been developed that have tried to systematically guide the tradeoffs between stability, fairness and efficiency. However, these techniques did not look well in showing the factors that affect the quality of experience, such as the video bitrate oscillation and video playback interruptions. While in [13], a Markov Decision Process (MDP) is used to deal with the stochastic decision problem, which reduces both the number of playback interruptions and the number of quality level switches while increases the quality of experience. Besides rate adaptation, there are some other research works addressing the fairness problem from several aspects such as the work in [14] aims at achieving relative fairness at the packet level through the implementation of the weighted fair list. At [15], a new protocol is proposed in the context of complex multi-server adaptation, aimed at improving the user experience by providing the best fairness, efficiency and stability. The problem of fairness has been addressed by the application of a server on traffic shaping as in [16]. In our work, a bitrate adaptive approach is designed on the basis of stability and fairness of DASH system under the scenario that many clients compete for network resources.

In this work, to avoid the unfair bandwidth that was estimated by the client due to off intervals during the downloading process, an increment method is designed based on estimated bandwidth for converging the measured bandwidth to the estimated bandwidth through increasing scheme. Exclusively, when the measured bandwidth is smaller than the estimated bandwidth (i.e. estimated bandwidth is equal to the size of a segment divided by the time it takes to download this segment), a logarithmic law based increment scheme is designed for converging the measured bandwidth toward share bandwidth in efficient way among competing clients. In contrast, when measured bandwidth is higher than estimated bandwidth, a conservative reducing scheme is designed for reducing the measured bandwidth to avoid congestion. On other side for stable and smooth video bitrate, we propose an adaptive bitrate algorithm that takes the measured bandwidth based on increment scheme and the buffer occupancy level into account for keeping a continuous video playback and selecting the best video bitrate to download the next segment. In our approach video buffer is associated with three predefined thresholds for keeping the buffer in stable state. Using this model, clients will download video segments continuously without the need for OFF interval. Especially, when the share bandwidth is higher than the selected video bitrate (with the condition the video bitrate not exceed the measured bandwidth), the buffer occupancy will increase and converge toward the maximum level. The adaptive algorithm will switch the bitrate in a way preventing buffer overflow, not idle and vice versa. Through simulation results, our approaches outperform the existing algorithms in measuring the fair share bandwidth, achieving fairness, buffer stability and reducing the number of video bitrates switching.

II. System Design

Generally HTTP clients operate the media content that transferred from streaming server via available network bandwidth. The Server-Client model for streaming system is shown in Fig. 1 below, where the network bandwidth represents the data producing rate and video content's bitrate represents the data consuming rate to the video buffer of the client.

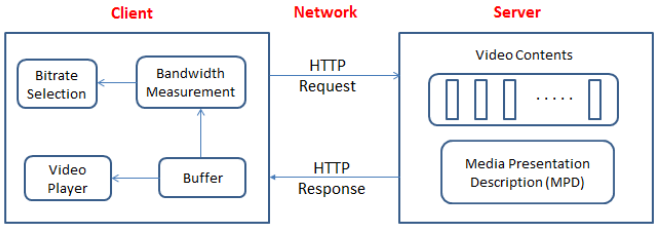


Fig. 1. Server-client model for streaming system.

Assuming at the video playing time, the current duration being playbacked is fully loaded, the next duration's playback time is from t_s to te as shown in Fig. 2, and T_i denotes the network bandwidth. While V_{Ri+1} defines the video bitrate for the next duration i+1 to guarantee there is no playback interruption during the future interval $[t_s, t_e]$.

At the beginning of each download stage, an adaptation algorithm will select the suitable bitrate of the next segment to download $V_{Ri+1} \in V$, $V = \{ V_R^1, ..., V_R^n, ..., V_R^L \}, 1 \le n \le L$ and $V_R^1 < V_R^2 < ..... < V_R^L$, which it will specify how much time is giving for the current segment download until the next download request, i.e. $t^{est}(i+1)$, which it is the estimated required time for downloading the next segment). Hence the client initiates an HTTP GET request to the server for the segment of sequence number i+1 with V_{Ri+1} then the downloading is starting. Let $t^m=[t_e- t_s]$ be the measured time that is required to complete the download. Assuming that no pipelining of downloading is involved, the next download step starts after the following time that can be expressed as follows:

$$t^{act}(i+1) = \max(t^{est}(i+1), t^m(i+1)) \quad (1)$$

where $t^{act}(i+1)$ is the actual time to download $(i+1)^{th}$ segment. If the download duration t(i+1) is shorter than the $t^{est}(i+1)$ then a client must wait time $[t^{est}(i+1)- t^m(i+1)]$ which is the off-interval before starting the next downloading step (case A), otherwise the client starts the next download step immediately after the current download is completed (case B).

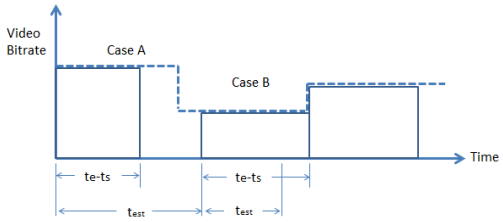


Fig. 2. HTTP segment downloading process.

III. Shared Bandwidth Approach

To explain the effect of bandwidth consumption on adapting the suitable video bitrates, three different scenarios are considered: ideal participated, more participated and less participated which they represent respectively the amount of bandwidth asked by client to be equal to, more than and less than the shared bandwidth. We assume a temporal overlap of the ON-OFF intervals among two competing clients. Suppose the available bandwidth is T and two active connections share this bandwidth fairly, i.e. each has T/2.So T_1 and T_2 are the bandwidth received by client1 and client2 respectively. In ideal participated, the total amount of traffic is requested by the two clients perfectly filling the link, i.e. each client exploits the available bandwidth during its ON period $T_1 = T_2 = T/2$. If the shared bandwidth is overestimated, new segments with high video bitrates maybe asked by the clients who lead to congestion. In this case, the current estimated bandwidth is less than the previous estimation, and hence clients will turn to ask segments with lower video bitrates. This fluctuation in video quality level leads to video buffer instability as shown in Fig. 3(a). Fig. 3(b) shows the case of more participated where the ON duration of one client falls in the ON duration of the other client. This happened when the clients are still asking video segments for different video bit-rates. In this situation, the monitored throughput by the clients is $T_1 > T/2$ and $T_2 = T/2$, i.e., first client overestimates the fair share of bandwidth.

When only one client overestimates the fair share of bandwidth, the two clients will converge to a stable but unfair equilibrium that the client, who overestimates the fair share of bandwidth, will request a higher video bitrate, causing unfairness. While Fig. 3(c) discusses the problem of imperfect usage of the shared bandwidth when the ON duration for the two clients are aligned, the available bandwidth at the server side is discrete and limited in this case clients estimate the fair share of bandwidth correctly, it still may cause imperfect utilization problem. For example if the available shared bandwidth is not T/2, clients will ask segments with video bitrates smaller than T/2 to avoid buffer underflow. The OFF periods should adopt to solve the buffer overflow case and imperfect utilization problem happens.

Accordingly, the bandwidth oscillation will impact on video bitrates stability when the bandwidth is in more participated case (i.e., the congestion occurs), then the bandwidths are estimated by the clients approximately equal, and fairness can be obtained. When congestion occurs, the video buffer at the client side will be drained, causing playback stop since clients may ask segments with bitrates more than the shared bandwidth, which may degrade the video quality of experience.

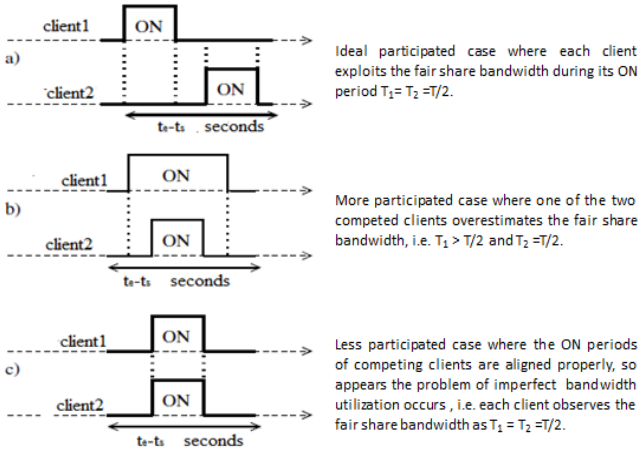


Fig. 3. Two clients are competing shared bandwidth during download a segment.

When the bandwidth is in less participated case, the requested video bitrate is smaller than the available bandwidth, and OFF periods is needed to stop the transmission so as to avoid buffer overflow. In this case, when the client buffer reaches the maximum level Bmax, this will lead to bandwidth overestimation and the video bitrates will oscillate accordingly. But the available shared bandwidth cannot continue with the same video bitrate, then video buffer falls below Bmax, so the client returns to the buffer growing mode and the off-intervals disappear, so the link again is going to be in more participated case and the measured throughput starts to converge to the fair-share bandwidth. Finally, due to the quantization impact, the requested video bitrate falls below the fair-share bandwidth, and the client buffer starts growing again, completing one oscillation cycle.

IV. Video Bitrate Adaptive Approach

In this section, we will describe a four stages model for DASH bitrate adaptation algorithm, which includes both the conventional algorithms (e.g., [17], [18]) and our proposed bitrate adaptation algorithm in this work, which will serve as a benchmark. Essentially, a bitrate adaptation algorithm proceeds in the following four stages:

***a)* Estimating Stage:** The algorithm begins the first step in adaptation process by estimating the network throughput which be used in selecting the video bitrate. The current shared bandwidth equates to the previous TCP throughput that monitors during the ON interval, i.e.:

$$\hat{T}_i = \tilde{T}_{i-1} \tag{2}$$

***b)* Filtering Stage:** Then a filtering process is done by using a smoothed function fs() with the aim of removing outliers. The algorithm takes the measured history as an input i.e.:

$$\tilde{T}_i = f_s(\{\hat{T}_j : j \le i\}) \tag{3}$$

***c)* Adapting Stage:** The continuous measured throughput is adapted to get the suitable video bitrate. In this step a discrete video bitrate is selected based on side history

information such as video buffer occupancy, previous selected video bitrate V_{Rj}, …, etc. Then the current video bitrate is given by:

$$V_{Ri+1} = f_a(\tilde{T}_{i+1};\ V_{Rj}, B(j) : B(j) < B_{max}, j \le i) \quad (4)$$

***d)* Next Segment Downloading Stage:** A scheduling time for download the next segment is determined by the adaptive algorithm as a mechanical function based on video buffer size. If the current buffer occupancy B(i) is less than the maximum level B_{max}, then the estimated time for next download is set to 0 and the download time for next segment begins after the current download is finished. Then the estimated time for next down is given by: But the time is τ (. i.e. :

$$t^{est} = \begin{cases} 0, & B(i) < B_{max} \\ \tau, & B(i) = B_{max} \end{cases} \quad (5)$$

where τ is the video segment period to stop the video buffer from growing.

In general, all of today's commercial DASH clients implement the first stage and last stage of the rate adaptation algorithm in a similar way, though they may differ in their implementation of the second stage and third stage of the algorithm. In this work we make some modifications to these stages since the off intervals may become a source of ambiguity for a client to correctly perceive its fair share of the network bandwidth, thus preventing the client from making accurate rate adaptation decisions. To ensure fairness, we focus on preventing the happenstance of OFF phenomenon, that is, the bandwidth is either in the ideal participated or more participated. In the case of more participated, fairness is guaranteed, while in the ideal of participated, there is an unlimited number of bandwidth-sharing modes. It is worth to note that although it is easy to avoid OFF phenomenon by keeping on downloading without break with them, it is impossible even to distinguish if the bandwidth is in more participated or in the ideal participated.

Moreover, because the available video bitrate is limited and discrete, instead of using the estimated bandwidth, we suggest an effective bandwidth increment scheme to guide the adaptive video bitrate. The stability is taken into account through proposing double threshold of buffer level, and then the frequency of filling the buffer occupancy and the impact of bandwidth variations on the adaptive video bitrate can be reduced. As per during the rate adaptation process, the smoothness is also taken into account and it will be discussed deeply in next section. In DASH system, the video bitrate can be changed only when a segment is completely downloaded. Then, the video bitrate for the next segment to be downloaded is calculated accordingly, we first estimate the network bandwidth and filter it. Then, with this estimated bandwidth, we update the incremented bandwidth which will be used to guide the rate adaption. According to the buffer occupancy, the video bitrate is adapted when the buffer occupancy is between the two predefined thresholds to avoid buffer overflow and underflow.

V. Smoothing Network Bandwidth

In this section, we will describe the principles of our approach to smooth the network bandwidth estimation and how filtering it, also we will describe our Trial increment approach based on bandwidth estimation and then we will explain in details our proposed stable and smooth adaptive bitrate algorithm.

A. Network Bandwidth Estimation and Filtering

Assume the encoding of a video segment is done into different versions of L, with playback of different video bitrates $V = \{V_R^1, ..., V_R^n, ..., V_R^L\}$ in V set. All video versions are divided into segments of equal length, each of which consumes the same playing time τ. For each client, the streaming process is divided into sequential segment downloading steps i=1, 2, 3… then, the required bandwidth to download $(i+1)^{th}$ segment, is estimated as [6]:

$$\hat{T}_{i+1} = \frac{\tau\, V_{Ri+1}}{t_m^{i+1}} \quad (6)$$

where τ is segment duration. The downloaded segments are stored in the video buffer. Furthermore, to remove the impact of noise and interference during the receiving bandwidth estimation, a noise-filter step is adopted [19]. Then, we have:

$$Y_{i+1} = \partial \hat{T}_i + (1-\partial) Y_i \quad (7)$$

where Y_{i+1} is the available modified bandwidth after filtering step. ∂ is a smooth coefficient factor that reflects the changing network conditions; it is measured as:

$$\partial = \left| \frac{x_i}{z_i} \right| \quad (8)$$

where x_i and z_i represent the absolute and smoothed values of the throughput deviation (e_i) respectively and they are given by:

$$\begin{aligned} x_i &= \rho e_i + (1-\rho)\, x_{i-1} \\ z_i &= \rho |e_i| + (1-\rho)\, z_{i-1} \end{aligned} \quad (9)$$

where ρ is a constant with $0 \le \rho \le 1$. At any instant e_i is determined as:

$$e_i = \hat{T}_i - Y_i \quad (10)$$

where e_i is difference between the measured value and adjusted value. If ∂ is a small value, the previous adjusted bandwidth may play a more important role in predicting the bandwidth.

As ∂ gets larger, the adjusted bandwidth is closer to the current measured value. Thus, if the variation of bandwidth is large, we should decrease ∂, and if the variation of bandwidth is small, we can increase ∂ to more accurately reflecting the change of the network conditions.

B. Trial Increment Approach

Due to the on-off phenomenon in the DASH system, the bandwidth estimated by the client is discretionary and cannot be used directly for the adjustment rate. Additionally, only when bandwidth is under more participated case, the congestion happens, so bandwidth estimated by client equals to a fair-share bandwidth since all clients see almost the same bandwidth as available. On the other hand, when congestion occurs, the required video bitrate cannot be supported by bandwidth, and video freezing may occur. Thus, how to get a free bandwidth quota without congestion is a critical issue to improve the performance of the rate adaptation for DASH system.

In this work, we propose a logarithmic increase scheme based on received network bandwidth, which includes increasing the bandwidth logarithmic and decrease it if the increment in bandwidth more than the fair share bandwidth for making the measured bandwidth quickly converge to the fair share bandwidth. The aim behind this scheme is that the estimated bandwidth by a client is always the upper bound of the fair share bandwidth due to the off intervals. Thus, during the increment phase, the measured bandwidth will continuously increase until it exceeds the estimated bandwidth expressed in (7). Then, when it is higher than the estimated bandwidth, the client will switch to the decrease phase for avoiding congestion.

Let $\tilde{Y}_{i+1}$ be the measured bandwidth which would be used for adapting the suitable video bitrate to $(i+1)^{th}$ segment and it is initialized to zero that $\tilde{Y}_{init} = 0$. Whenever the downloaded process of segment is completed, the estimated bandwidth is update according to expression (7). Then, the measured bandwidth is updated as follows:

1) Increment Phase:

$$\tilde{Y}_{i+1} = \tilde{Y}_i + \max\left(\frac{Y_i - \tilde{Y}_i}{2}, \wp\right), \quad \text{if } \tilde{Y}_i < Y_i \tag{11}$$

where $\wp$ is a constant to avoid slow convergence, the measured bandwidth will quickly approach to the estimated bandwidth in the logarithmic way. However, when the gap between $\tilde{Y}_i$ and Y_i is small, i.e. $\left(\frac{Y_i - \tilde{Y}_i}{2} < \wp\right)$, the measured bandwidth will be additively increased by $\wp$ rather than by logarithmic approach. Otherwise when the gap between the measured bandwidth and the fair share bandwidth is large, an aggressive way should be employed to increase the measured bandwidth quickly. Practically, the client fair share bandwidth in practice cannot get, so instead its upper bound bandwidth (i.e., the estimated bandwidth) is used.

2) Decrement Phase:

$$\tilde{Y}_{i+1} = \tilde{Y}_i + \beta\left(Y_i - \tilde{Y}_i\right), \quad \text{if } \tilde{Y}_i \geq Y_i \tag{12}$$

where β is a positive constant satisfying $\beta > 1$. When the measured bandwidth exceeds the estimated bandwidth, a conservative way should be employed to control $\tilde{Y}_i$ to be not higher than Y_i i.e. when the measured bandwidth is higher than the fair share bandwidth, congestions may happen, so leading to playback freezing. In this case, a conservative way is designed to decrease the measured bandwidth guaranteeing that no congestion happens.

C. Proposed Adaptive Bitrate Algorithm

At the client aspect, the video player uses an adaptive algorithm to determine the suitable bitrate to be selected for download the next segment. Every segment that is downloaded placed in a buffer of maximum size B_{max}. The buffer is associated with three thresholds (B_0, B_{low}, B_{high}) where $B_0 < B_{low} < B_{high}$ and V represents the set of available bitrates as cleared in Fig. 4. After getting the measured bandwidth in the previous section, to obtain a smooth and stable video bitrate with high bandwidth utilization a bitrate adaptation scheme is designed based on the buffer occupancy with the measured bandwidth. In this work, the buffer occupancy is denoted by the buffered video duration considering that the client buffer may contain segments from different versions, i.e., different video bitrates and there is no longer a direct mapping between the buffered video size and the buffered video duration.

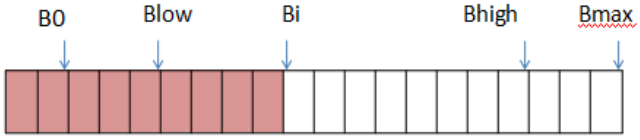


Fig. 4. Client video buffer.

The smooth bitrate adaptation scheme is designed with the aim to provide a smooth video quality and avoid buffer overflow/underflow. The impact of short term bandwidth oscillations on video bitrates can be eliminated by using thresholds for buffer occupancy. At the beginning of the video session (or after an interruption due to buffer empty state), the number of segments in the buffer is below B_0 and the video playback begins during which the lowest bitrate is selected. This step is necessary because reducing the startup time is important to prevent the user from giving away the video session.

When the share bandwidth link is in more participated case, i.e., the congestion occurs, the bandwidths estimated by the clients are approximately equal, and then fairness can be obtained. When congestion occurs, the video buffer at the client side will be underflow, causing playback stop since clients may ask segments with bitrates more than the share bandwidth, which degrades video quality of experience. So the rate adaptation scheme decreases the video bitrate to minimum level. Once the video buffer exceeds B_0 and lower than B_{low}, to avoid buffer draining and provide a continuous playback, the video bitrate should be selected no higher than the measured bandwidth. Similarly, when the buffer occupancy is higher than B_{high}, to ensure no buffer overflow occurs while achieving high bandwidth utilization, a high video bitrate can be selected. When the video buffer occupancy falls between the two thresholds B_{low} and B_{high}, the risk of buffer overflow/underflow is low, and the video bitrate should be maintained unchanged so that smooth quality is provided. Consider a video content is coded into L representations with different video bitrates. The

set of video bitrates is denoted by $V = \{ V_R^1, ..., V_R^n, ..., V_R^L \}, 1 \le n \le L$. B_i is the current buffer occupancy after download i^{th} segment and measured bandwidth $\tilde{Y}_{i+1}$ for next segment. The video bitrate for $(i+1)^{th}$ segment is selected as:

$$V_{Ri+1} = \begin{cases} \min\limits_{n=1} \{V_R^n | V_R^n \in V,\ V_R^n < \tilde{Y}_{i+1}\} & \text{if } B_i \le B_0 \\ \max\limits_{1 \le n \le L} \{V_R^n | V_R^n \in V, \frac{\tau V_R^n}{\hat{T}_{i+1}} \le B_i - B_0,\ V_R^n \le \tilde{Y}_{i+1}\} & \text{if } B_i < B_{low} \\ \min\limits_{1 \le n \le L} \left\{ V_R^n | V_R^n \in V, \frac{\tau V_R^n}{\hat{T}_{i+1}} \le B_i - B_{low},\ V_R^n \ge \tilde{Y}_{i+1} \right\} & \text{if } B_i > B_{high} \\ V_{Ri} & \text{if } B_{low} < B_i < B_{high} \end{cases} \quad (13)$$

At the starting of the video session, the number of segments in the buffer is below B_0 and the rate adaptation scheme decreases the video bitrate to minimum level. Once the buffer level goes beyond B_0, the amount of time required to load the next segment is less than the B_i - B_0, the maximal video bitrate which is no higher than the measured bandwidth is selected so as to guarantee a continuous video playback. When the buffer occupancy is higher than the B_{high}, the minimal video bitrate which is no lower than the measured bandwidth is selected so as to improve video quality; otherwise when the buffer occupancy between B_{low} and B_{high} the current bit rate is preserved.

VI. Simulation Results

The proposed scheme is evaluated using ns-3 network simulator and the topology implemented in this work is shown in Fig. 5 below. The topology consists of an HTTP server, two or more HTTP clients and a pair of network elements (i.e. x>=2). To achieve adaptive streaming, the HTTP server offers the client seven levels of representations to adapt the video rates these are V={ 356, 500, 800, 1200, 1500, 2400 and 3500Kbit/s}. The length of video segment and video buffer is 2s and 35 sec, respectively. The values of other parameters B_{low}, B_{high}, B_0, β, ρ, $\wp$ are 15s, 30s, 5s, 1.25, 0.5, 32kbps respectively.

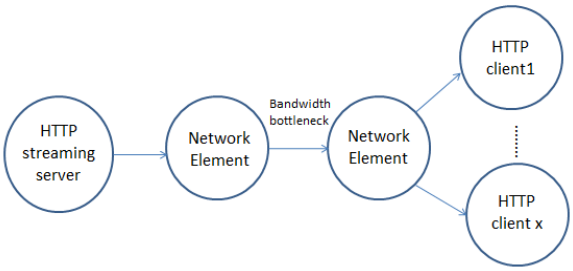
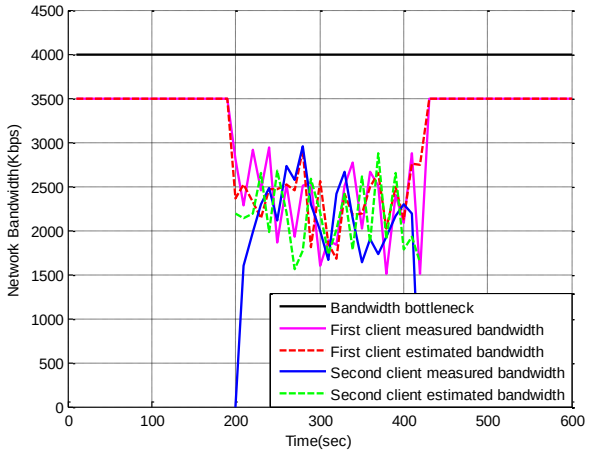


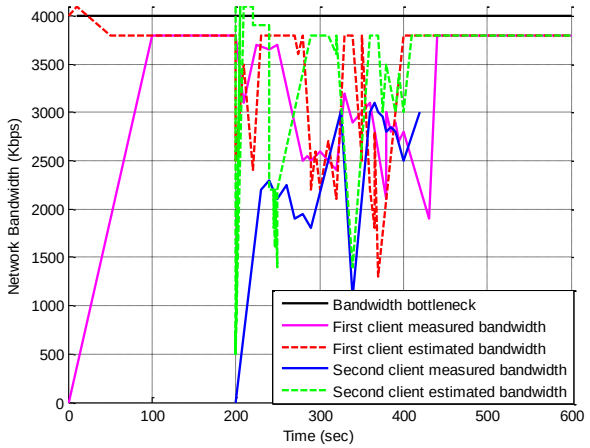
Fig. 5. Network topology configuration in the simulation.

To evaluate the performance of the proposed scheme, we compare it with a conventional scheme [6] as in Fig. 6 and 7 in term of measured bandwidth, video bitrate and buffer occupancy, respectively. The available network bandwidth is equal to 4Mbps with two clients compete on it. The first client enters at the time 0s while the second client enters the network at 200s and leaves at 400s. Fig. 6(a) and 6(b) show the ability of increment scheme in converging the measured bandwidth to the estimated bandwidth (share bandwidth) quickly through the accurate logarithmic law based increment scheme.

We can note from Fig. 6(a) and 6(b), when the second client enters the system, it needs only about 10s to follow the fair share bandwidth well, the first client detects the change in the available network bandwidth such that their measured bandwidth converge to the fair share bandwidth quickly. While the other scheme spends a long time to converge the measured bandwidth to the estimated bandwidth in linear increasing scheme based on estimated bandwidth. The figures also demonstrate that when two clients compete with each other, the compared scheme couldn't track the fair-share bandwidth well because of the slow start and frequent fluctuations, when the bandwidth changes.



a: Proposed scheme

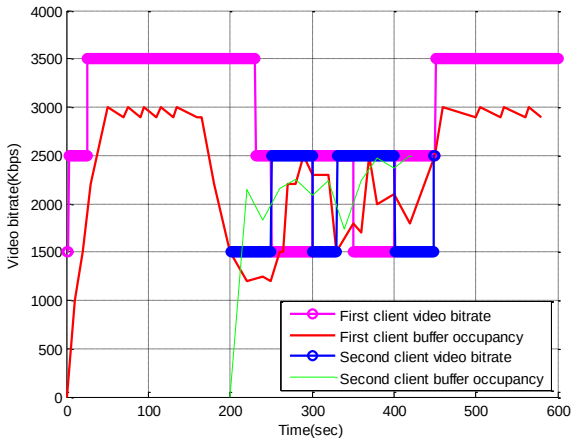


b: Conventional scheme

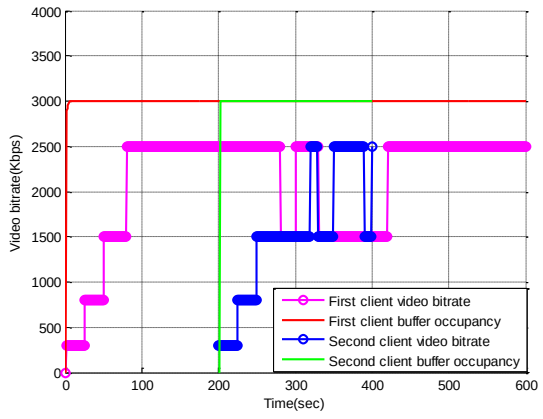
Fig. 6. Network bandwidth computed by two clients.

In Fig. 7(a) and 7(b) the video bitrate under the proposed scheme is smooth and stable than the second scheme. The conventional scheme uses the linear increasing method in measuring the share bandwidth which is used for adapting the video bitrate hence this leads to video bitrate oscillation for matching the measured bandwidth. In addition the video buffer

occupancy is always high to ensure continuous video playback. Since the video bitrate is separate, the selection of suitable bitrate is usually lower than the measured bandwidth which leads to low bandwidth utilization efficiency. While the video bitrate of our scheme is smoother due to many reasons; at the beginning the adaptation of bitrate isn't depending on the measured bandwidth only, it associates with the buffer occupancy. The video buffer is related with three operating thresholds for avoiding buffer overflow/underflow and compacting the short term bandwidth oscillation that effects on video bitrates. Beside the bitrate adaptive algorithm takes many factors into account that impact on the QoE for video streaming, involving maximize the video bitrates and minimize the video start time, the number of buffering events, and the number of bitrate switching events. So from figure 8a the bitrate curve is smoother than the other scheme. Finally, no buffer overflow/underflow occurs, i.e., video playback without break is guaranteed.



a: Proposed scheme



b: Conventional scheme

Fig. 7. Buffer occupancy and video bitrate of two clients under bandwidth bottleneck.

In Fig. 8, we compare our scheme with other scheme in term of efficiency, stability, and fairness based on the following metrics [11]:

***a)* Instability metric:** Clients are likely to be sensitive to frequent and important video bitrate switches as indicated by some studies. The instability metric is defined as:

$$M_{instability} = \frac{\sum_{a=0}^{z-1} \left| V_{Ri,i-a} - V_{Ri,i-a-1} \right| \omega(a)}{\sum_{a=1}^{z} V_{Ri,i-a}\, \omega(a)} \quad (15)$$

The instability metric equals to the weighted sum of all bitrate switch steps monitored within the last 10 segments divided by the weighted sum of bitrates in the last z=10 segments. The weight function $\omega(a) = i-a$.

***b)* Unfairness metric:** At time t, the unfairness metric is defined as:

$$M_{unfairness} = \sqrt{1 - jainFair_t} \quad (16)$$

where JainFairt is the index of Jain fairness at time t and is calculated based on the bitrates V_{Ri} over all clients.

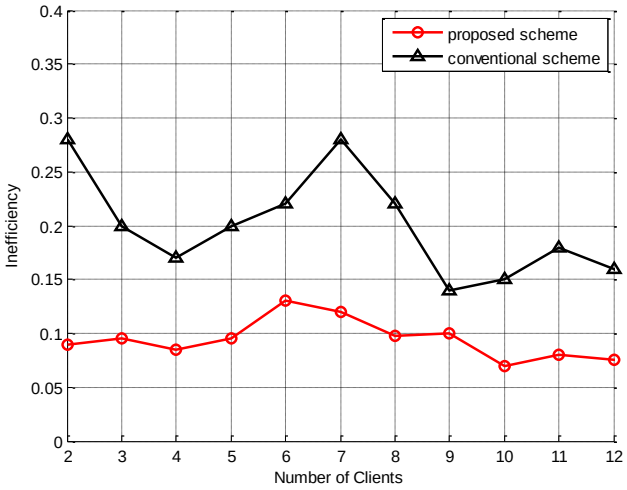
***c)* Inefficiency metric:** This metric is calculated as:

$$M_{inefficiency} = \left| \frac{\sum_{j} V_{Ri,j}}{T} \right| \quad (17)$$

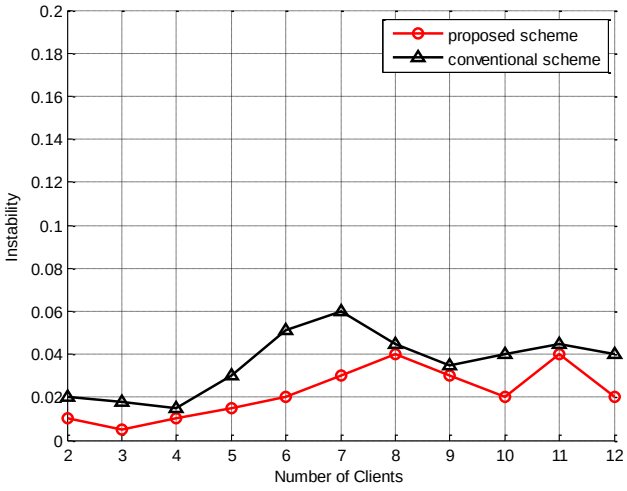
where $V_{Ri,j}$ is the video bitrate of i^{th} segment for j^{th} client and T is the available bandwidth. A value close to zero means that the clients in average are using as high an average video bitrate as possible to improve video quality.

In Fig. 8, the number of clients is changed from two to twelve with the available network bandwidth fixed at 8Mbps in order to evaluate the performance of our scheme with other scheme. From the figure the proposed scheme is outperforming due to the increment scheme based on bandwidth measurement also video buffer model is associated with three operating thresholds for preventing buffer underflow /overflow and the robust of the adaptive algorithm based on estimating the download time of next segment based on segment size.

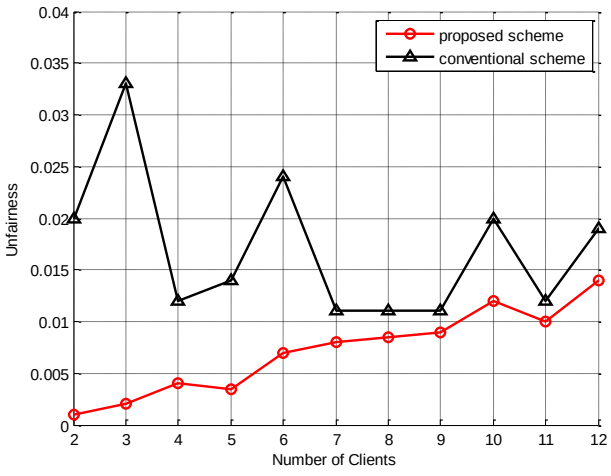
In Fig. 9, our scheme always achieves the lowest inefficiency, instability, and unfairness, i.e., compared with conventional scheme. The proposed scheme can provide higher and smoother video bitrate, and besides, it can also better guarantee the fairness between the clients. As the available bandwidth increases, the performance does not linearly increase or decrease. This is because the video bitrate is discrete so that for a given available bandwidth, if the fair-share bandwidth is approximately equal to the video bitrate, generally better performance can be achieved, and vice versa.



a: Inefficiency

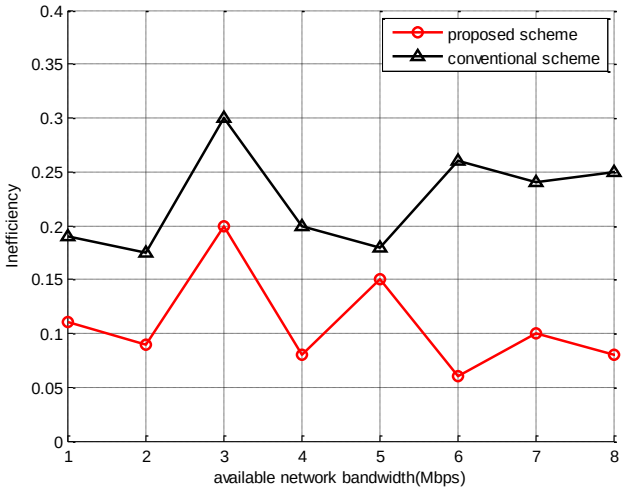


b: Instability

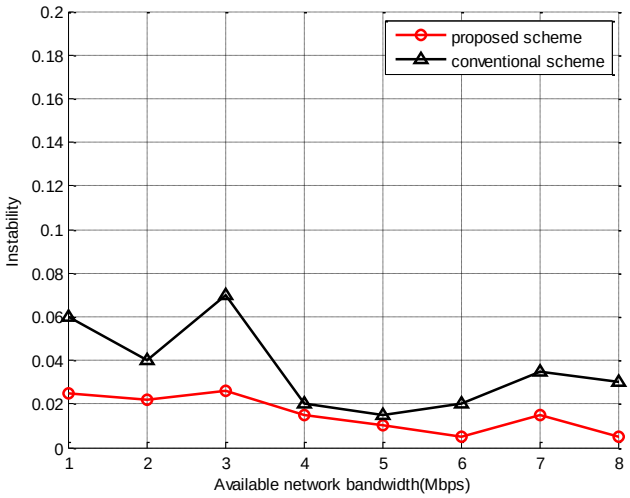


c: Unfairness

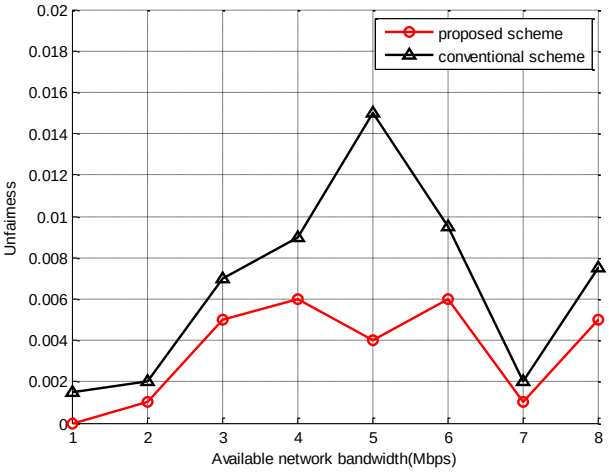
Fig. 8. Performance evaluation when the available network bandwidth is fixed at 8 Mbps, and the number of competing clients varies from 2 to 12.



a: Inefficiency



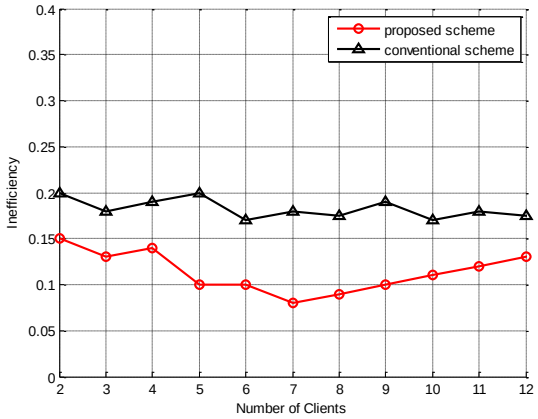
b: Instability



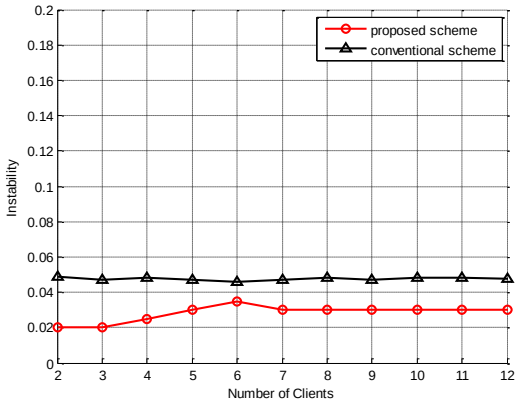
c: Unfairness

Fig. 9. Performance evaluation for the two schemes when the available network bandwidth is changing from 1Mbps to 8Mbps with two competing clients.

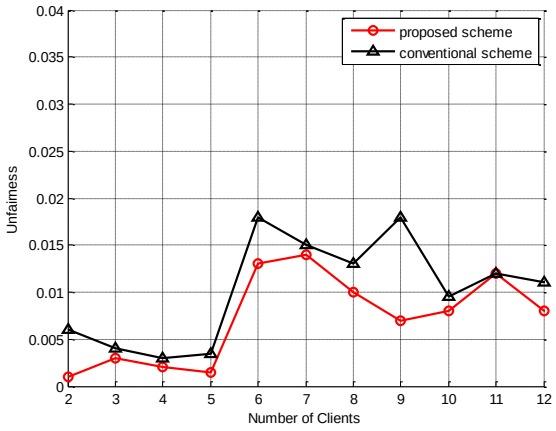
From the figure our scheme has better performing in all situations as in (a) and (b) , the other scheme has inefficient bandwidth utilization, instability video bitrate because the decision of video bitrate is based on the measured fair share bandwidth, which is always the same (equal to 1 Mbps) in this experiment . In term of fairness, the other scheme fluctuates with increasing the no. of clients while the proposed scheme performs better since fairness is easy to be impacted with increasing the number of competing clients (Fig. 10).



a: Inefficiency



b: Instability



c: Unfairness

Fig. 10. The performance assessment of two schemes under changing both the number of competing clients and available network bandwidth.

VII. Conclusion

In this work, a bitrate adaptive approach is designed on the basis of stability and fairness in DASH system under the scenario that many clients compete for network resources. The proposed scheme is based on estimated bandwidth to obtain the fair-share bandwidth (i.e. the measured bandwidth) by trial increment mechanism. Then collecting the measured bandwidth with the video buffer occupancy, we propose an adaptive video bitrates algorithm based on time needed for downloading the next segment to achieve smooth, stable video quality and avoid buffer underflow/overflow. From simulation results, our scheme is able to quick converge the measured bandwidth to the fair-share bandwidth even with increasing the number of competing clients, hence achieving fairness bandwidth utility. Besides, the video bitrate is allowed to be higher than the measured bandwidth, thereby achieving a higher bandwidth utilization efficiency and higher average video bit-rate compared with a conventional scheme.

References

[1] Cisco, V.N.I., 2016. Forecast and Methodology, 2015-2020. White Paper, Cisco.

[2] Pantos, R. and May, W., 2010. HTTP Live Streaming draft-pantos-http-live-streaming-05. Apple Inc., IETF draft, 23.

[3] Sodagar, I., 2011. The mpeg-dash standard for multimedia streaming over the internet. IEEE Multimedia, 18(4), pp.62-67.

[4] Kim, J., Caire, G. and Molisch, A.F., 2016. Quality-aware streaming and scheduling for device-to-device video delivery. IEEE/ACM Transactions on Networking, 24(4), pp.2319-2331.

[5] Akhshabi, S., Anantakrishnan, L., Begen, A.C. and Dovrolis, C., 2012, June. What happens when HTTP adaptive streaming players compete for bandwidth?. In Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video (pp. 9-14). ACM.

[6] Liu, C., Bouazizi, I. and Gabbouj, M., 2011, February. Rate adaptation for adaptive HTTP streaming. In Proceedings of the second annual ACM conference on Multimedia systems(pp. 169-174). ACM.

[7] Zhou, B., Wang, J., Zou, Z. and Wen, J., 2012, January. Bandwidth estimation and rate adaptation in HTTP streaming. In Computing, Networking and Communications (ICNC), 2012 International Conference on (pp. 734-738). IEEE.

[8] Lin, Q., Liu, Y., Shen, Y., Shen, H., Sang, L. and Yang, D., 2014, December. Bandwidth estimation of rate adaption algorithm in DASH. In Globecom Workshops (GC Wkshps), 2014 (pp. 243-247). IEEE.

[9] Huang, T.Y., Johari, R., McKeown, N., Trunnell, M. and Watson, M., 2015. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. ACM SIGCOMM Computer Communication Review, 44(4), pp.187-198.

[10] Kim, S., Yun, D. and Chung, K., 2016, January. Video quality adaptation scheme for improving QoE in HTTP adaptive streaming. In Information Networking (ICOIN), 2016 International Conference on (pp. 201-205). IEEE.

[11] Jiang, J., Sekar, V. and Zhang, H., 2012, December. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. In Proceedings of the 8th international conference on Emerging networking experiments and technologies (pp. 97-108). ACM.

[12] Yin, X., Bartulović, M., Sekar, V. and Sinopoli, B., 2017, May. On the efficiency and fairness of multiplayer HTTP-based adaptive video streaming. In American Control Conference (ACC), 2017 (pp. 4236-4241). IEEE.

[13] Zhou, C., Lin, C.W. and Guo, Z., 2016. mDASH: A markov decision-based rate adaptation approach for dynamic HTTP streaming. IEEE Transactions on Multimedia, 18(4), pp.738-751.

[14] Ma, H., Hao, J. and Zimmermann, R., 2014, July. Access point centric scheduling for dash streaming in multirate 802.11 wireless network. In

Multimedia and Expo (ICME), 2014 IEEE International Conference on (pp. 1-6). IEEE.

[15] Zhang, S., Li, B. and Li, B., 2015, June. Presto: Towards fair and efficient HTTP adaptive streaming from multiple servers. In Communications (ICC), 2015 IEEE International Conference on (pp. 6849-6854). IEEE.

[16] Dubin, R., Dvir, A., Hadar, O., Shalala, R. and Ahark, O., 2015, January. Video complexity hybrid traffic shaping for HTTP Adaptive Streaming. In Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE (pp. 683-688). IEEE.

[17] Tian, G. and Liu, Y., 2012, December. Towards agile and smooth video adaptation in dynamic HTTP streaming. In Proceedings of the 8th international conference on Emerging networking experiments and technologies (pp. 109-120). ACM.

[18] Miller, K., Quacchio, E., Gennari, G. and Wolisz, A., 2012, May. Adaptation algorithm for adaptive streaming over HTTP. In Packet Video Workshop (PV), 2012 19th International (pp. 173-178). IEEE.

[19] Dolinsky, K. and Celikovsky, S., 2012, June. Kalman filter under nonlinear system transformations. In American Control Conference (ACC), 2012 (pp. 4789-4794). IEEE.

Software Bug Prediction using Machine Learning Approach

Awni Hammouri, Mustafa Hammad, Mohammad Alnabhan, Fatima Alsarayrah
Information Technology Department
Mutah University, Al Karak, Jordan

***Abstract*—Software Bug Prediction (SBP) is an important issue in software development and maintenance processes, which concerns with the overall of software successes. This is because predicting the software faults in earlier phase improves the software quality, reliability, efficiency and reduces the software cost. However, developing robust bug prediction model is a challenging task and many techniques have been proposed in the literature. This paper presents a software bug prediction model based on machine learning (ML) algorithms. Three supervised ML algorithms have been used to predict future software faults based on historical data. These classifiers are Naïve Bayes (NB), Decision Tree (DT) and Artificial Neural Networks (ANNs). The evaluation process showed that ML algorithms can be used effectively with high accuracy rate. Furthermore, a comparison measure is applied to compare the proposed prediction model with other approaches. The collected results showed that the ML approach has a better performance.**

Keywords*—Software bug prediction; faults prediction; prediction model; machine learning; Naïve Bayes (NB); Decision Tree (DT); Artificial Neural Networks (ANNs)

I. Introduction

The existence of software bugs affects dramatically on software reliability, quality and maintenance cost. Achieving bug-free software also is hard work, even the software applied carefully because most time there is hidden bugs. In addition to, developing software bug prediction model which could predict the faulty modules in the early phase is a real challenge in software engineering.

Software bug prediction is an essential activity in software development. This is because predicting the buggy modules prior to software deployment achieves the user satisfaction, improves the overall software performance. Moreover, predicting the software bug early improves software adaptation to different environments and increases the resource utilization.

Various techniques have been proposed to tackle Software Bug Prediction (SBP) problem. The most known techniques are Machine Learning (ML) techniques. The ML techniques are used extensively in SBP to predict the buggy modules based on historical fault data, essential metrics and different software computing techniques.

In this paper, three supervised ML learning classifiers are used to evaluate the ML capabilities in SBP. The study discussed Naïve Bayes (NB) classifier, Decision Tree (DT) classifier and Artificial Neural Networks (ANNs) classifier. The discussed ML classifiers are applied to three different datasets obtained from [1] and [2] works.

In addition to, the paper compares between NB classifier, DT classifier and ANNs classifier. The comparison based on different evaluation measures such as accuracy, precision, recall, F-measures and the ROC curves of the classifiers.

The rest of this paper is organized as follow. Section 2 presents a discussion of the related work in SBP. An overview of the selected ML algorithms is presented in Section 3. Section 4 describes the datasets and the evaluation methodology. Experimental results are shown in Section 5 followed by conclusions and future works.

II. Related Work

There are many studies about software bug prediction using machine learning techniques. For example, the study in [2] proposed a linear Auto-Regression (AR) approach to predict the faulty modules. The study predicts the software future faults depending on the historical data of the software accumulated faults. The study also evaluated and compared the AR model and with the Known power model (POWM) used Root Mean Square Error (RMSE) measure. In addition to, the study used three datasets for evaluation and the results were promising.

The studies in [3], [4] analyzed the applicability of various ML methods for fault prediction. Sharma and Chandra [3] added to their study the most important previous researches about each ML techniques and the current trends in software bug prediction using machine learning. This study can be used as ground or step to prepare for future work in software bug prediction.

R. Malhotra in [5] presented a good systematic review for software bug prediction techniques, which using Machine Learning (ML). The paper included a review of all the studies between the period of 1991 and 2013, analyzed the ML techniques for software bug prediction models, and assessed their performance, compared between ML and statistic techniques, compared between different ML techniques and summarized the strength and the weakness of the ML techniques.

In [6], the paper provided a benchmark to allow for common and useful comparison between different bug prediction approaches. The study presented a comprehensive comparison between a well-known bug prediction approaches, also introduced new approach and evaluated its performance by building a good comparison with other approaches using the presented benchmark.

D. L. Gupta and K. Saxena [7] developed a model for object-oriented Software Bug Prediction System (SBPS). The study combined similar types of defect datasets which are available at Promise Software Engineering Repository. The study evaluated the proposed model by using the performance measure (accuracy). Finally, the study results showed that the average proposed model accuracy is 76.27%.

Rosli et al. [8] presented an application using the genetic algorithm for fault proneness prediction. The application obtains its values, such as the object-oriented metrics and count metrics values from an open source software project. The genetic algorithm uses the application's values as inputs to generate rules which employed to categorize the software modules to defective and non-defective modules. Finally, visualize the outputs using genetic algorithm applet.

The study in [9] assessed various object-oriented metrics by used machine learning techniques (decision tree and neural networks) and statistical techniques (logical and linear regression). The results of the study showed that the Coupling Between Object (CBO) metric is the best metric to predict the bugs in the class and the Line Of Code (LOC) is fairly well, but the Depth of Inheritance Tree (DIT) and Number Of Children (NOC) are untrusted metrics.

Singh and Chug [10] discussed five popular ML algorithms used for software defect prediction i.e. Artificial Neural Networks (ANNs), Particle Swarm Optimization (PSO), Decision Tree (DT), Naïve Bayes (NB) and Linear Classifiers (LC). The study presented important results including that the ANN has lowest error rate followed by DT, but the linear classifier is better than other algorithms in term of defect prediction accuracy, the most popular methods used in software defect prediction are: DT, BL, ANN, SVM, RBL and EA, and the common metrics used in software defect prediction studies are: Line Of Code (LOC) metrics, object oriented metrics such as cohesion, coupling and inheritance, also other metrics called hybrid metrics which used both object oriented and procedural metrics, furthermore the results showed that most software defect prediction studied used NASA dataset and PROMISE dataset.

Moreover, the studies in [11], [12] discussed various ML techniques and provided the ML capabilities in software defect prediction. The studies assisted the developer to use useful software metrics and suitable data mining technique in order to enhance the software quality. The study in [12] determined the most effective metrics which are useful in defect prediction such as Response for class (ROC), Line of code (LOC) and Lack Of Coding Quality (LOCQ).

Bavisi et al. [13] presented the most popular data mining technique (k-Nearest Neighbors, Naïve Bayes, C-4.5 and Decision trees). The study analyzed and compared four algorithms and discussed the advantages and disadvantages of each algorithm. The results of the study showed that there were different factors affecting the accuracy of each technique; such as the nature of the problem, the used dataset and its performance matrix.

The researches in [14], [15] presented the relationship between object-oriented metrics and fault-proneness of a class. Singh et al. [14] showed that CBO, WMC, LOC, and RFC are effective in predicting defects, while Malhotra and Singh [15] showed that the AUC is effective metric and can be used to predict the faulty modules in early phases of software development and to improve the accuracy of ML techniques.

This paper discusses three well-known machine learning techniques DT, NB and ANNs. The paper also evaluates the ML classifiers using various performance measurements (i.e. accuracy, precision, recall, F-measure and ROC curve). Three public datasets are used to evaluate the three ML classifiers.

On the other hand, most of the mentioned related works discussed more ML techniques and different datasets. Some of the previous studies mainly focused on the metrics that make the SBP as efficient as possible, while other previous studies proposed different methods to predict software bugs instead of ML techniques.

III. USED MACHINE LEARNING ALGORITHMS

The study aims to analyze and assess three supervised Machine Learning algorithms, which are Naïve Bayes (NB), Artificial Neural Network (ANN) and Decision Tree (DT). The study shows the performance accuracy and capability of the ML algorithms in software bug prediction and provides a comparative analysis of the selected ML algorithms.

The supervised machine learning algorithms try to develop an inferring function by concluding relationships and dependencies between the known inputs and outputs of the labeled training data, such that we can predict the output values for new input data based on the derived inferring function. Following are summarized description of the selected supervised ML algorithms:

- *Naïve Bayes (NB)*: NB is an efficient and simple probabilistic classifier based on Bayes theorem with independence assumption between the features. NB is not single algorithms, but a family of algorithms based on common principle, which assumes that the presence or absence of a particular feature of the class is not related to the presence and absence of any other features [16], [17].
- *Artificial Neural Networks (ANNs)*: ANNs are networks inspired by biological neural networks. Neural networks are non-linear classifier which can model complex relationships between the inputs and the outputs. A neural network consists of a collection of processing units called neurons that are work together in parallel to produce output [16]. Each connection between neurons can transmit a signal to other neurons and each neuron calculates its output using the nonlinear function of the sum of all neuron’s inputs.
- *Decision Tree (DT)*: DT is a common learning method used in data mining. DT refers to a hierarchal and predictive model which uses the item’s observation as branches to reach the item’s target value in the leaf. DT is a tree with decision nodes, which have more than one branch and leaf nodes, which represent the decision.

TABLE II. DS1 - THE FIRST SOFTWARE FAULTS DATASET

D_i	F_i	T_i		D_i	F_i	T_i
1	2	75		24	2	8
2	0	31		25	1	15
3	30	63		26	7	31
4	13	128		27	0	1
5	13	122		28	22	57
6	3	27		29	2	27
7	17	136		30	5	35
8	2	49		31	12	26
9	2	26		32	14	36
10	20	102		33	5	28
11	13	53		34	2	22
12	3	26		35	0	4
13	3	78		36	7	8
14	4	48		37	3	5
15	4	75		38	0	27
1`6	0	14		39	0	6
17	0	4		40	0	6
18	0	14		41	0	4
19	0	22		42	5	0
20	0	5		43	2	6
21	0	9		44	3	5
22	30	33		45	0	8
23	15	118		46	0	2

IV. DATASETS AND EVALUATION METHODOLOGY

The used datasets in this study are three different datasets, namely DS1, DS2 and DS3. All datasets are consisting of two measures; the number of faults (F_i) and the number of test workers (T_i) for each day (D_i) in a part of software projects lifetime. The DS1 dataset has 46 measurements that involved in the testing process presented in [1]. DS2, also taken from [1], which measured a system faults during 109 successive days of testing the software system that consists of 200 modules with each having one kilo line of code of Fortran. DS2 has 111 measurements. DS3 is developed in [2], which contains real measured data for a test/debug program of a real-time control application presented in [18]. Tables I to III present DS1, DS2 and DS3, respectively.

The datasets were preprocessed by a proposed clustering technique. The proposed clustering technique marks the data with class labels. These labels are set to classify the number of faults into five different classes; A, B, C, D, and E. Table IV shows the value of each class and number of instances that belong to it in each dataset.

In order to evaluate the performance of using ML algorithms in software bug prediction, we used a set of well-known measures [19] based on the generated confusion matrixes. The following subsections describe the confusion matrix and the used evaluation measures.

TABLE III. DS2 - THE SECOND SOFTWARE FAULTS DATASET

D_i	F_i	T_i		D_i	F_i	T_i		D_i	F_i	T_i
1	5	4		38	15	8		75	0	4
2	5	4		39	7	8		76	0	4
3	5	4		40	15	8		77	1	4
4	5	4		41	21	8		78	2	2
5	6	4		42	8	8		79	0	2
6	8	5		43	6	8		80	1	2
7	2	5		44	20	8		81	0	2
8	7	5		45	10	8		82	0	2
9	4	5		46	3	8		83	0	2
10	2	5		47	3	8		84	0	2
11	31	5		48	8	4		85	0	2
12	4	5		49	5	4		86	0	2
13	24	5		50	1	4		87	2	2
14	49	5		51	2	4		88	0	2
15	14	5		52	2	4		89	0	2
16	12	5		53	2	4		90	0	2
17	8	5		54	7	4		91	0	2
18	9	5		55	2	4		92	0	2
19	4	5		56	0	4		93	0	2
20	7	5		57	2	4		94	0	2
21	6	5		58	3	4		95	0	2
22	9	5		59	2	4		96	1	2
23	4	5		60	7	4		97	0	2
24	4	5		61	3	4		98	0	2
25	2	5		62	0	4		99	0	2
26	4	5		63	1	4		100	1	2
27	3	5		64	0	4		101	0	1
28	9	6		65	1	4		102	0	1
29	2	6		66	0	4		103	1	1
30	5	6		67	0	4		104	2	1
31	4	6		68	1	3		105	0	1
32	1	6		69	1	3		106	1	2
33	4	6		70	0	3		107	0	2
34	3	6		71	0	3		108	0	1
35	6	6		72	1	3		109	1	1
36	13	6		73	1	4		110	0	1
37	19	8		74	0	4		111	1	1

A. Confusion Matrix

The confusion matrix is a specific table that is used to measure the performance of ML algorithms. Table V shows an example of a generic confusion matrix. Each row of the matrix represents the instances in an actual class, while each column represents the instance in a predicted class or vice versa. Confusion matrix summarizes the results of the testing algorithm and provides a report of the number of True Positive (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

TABLE IV. DS3 - THE THIRD SOFTWARE FAULTS DATASET

D_i	F_i	T_i	D_i	F_i	T_i	D_i	F_i	T_i
1	4	1	38	9	2	75	1	2
2	0	1	39	7	2	76	11	2
3	7	1	40	12	2	77	1	2
4	10	1	41	12	2	78	0	2
5	13	1	42	15	2	79	2	2
6	8	1	43	14	2	80	2	2
7	13	1	44	7	2	81	4	2
8	4	1	45	9	2	82	1	2
9	7	1	46	11	2	83	0	2
10	8	1	47	5	2	84	4	2
11	1	1	48	7	2	85	1	1
12	6	1	49	7	2	86	1	1
13	13	1	50	14	2	87	0	1
14	7	1	51	13	2	88	2	3
15	9	1	52	14	2	89	0	1
16	8	2	53	11	2	90	0	2
17	5	2	54	2	1	91	1	1
18	10	2	55	4	1	92	1	1
19	7	2	56	4	2	93	0	1
20	11	2	57	3	2	94	0	2
21	5	2	58	6	2	95	0	1
22	8	2	59	6	2	96	0	1
23	13	2	60	2	2	97	1	2
24	9	2	61	0	1	98	0	1
25	7	2	62	0	1	99	1	1
26	7	2	63	3	1	100	0	1
27	5	2	64	0	1	101	0	1
28	7	2	65	4	1	102	0	2
29	6	1	66	0	1	103	0	1
30	6	1	67	1	1	104	2	1
31	4	1	68	2	1	105	0	1
32	12	2	69	0	2	106	1	2
33	6	2	70	1	2	107	0	2
34	7	2	71	2	2	108	2	2
35	8	2	72	5	2	109	0	2
36	11	2	73	3	2			
37	6	2	74	2	2			

TABLE V. NUMBER OF FAULTS CLASSIFICATION

Faults Class	Number of Faults	Number of Instances		
		DS1	*DS2*	*DS3*
A	0-4	30	76	57
B	5-9	5	23	33
C	10-14	5	4	18
D	15-19	2	3	1
E	More than 20	4	5	0

TABLE VI. THE CONFUSION MATRIX

Predicted	Actual	
	Class X	*Class Y*
Class X	TP	FP
Class Y	FN	TN

B. Accuracy

Accuracy (ACC) is the proportion of true results (both TP and TN) among the total number of examined instances. The best accuracy is 1, whereas the worst accuracy is 0. ACC can be computed by using the following formula:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

C. Precision (Positive Predictive Value)

Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions. The best precision is 1, whereas the worst is 0 and it can be calculated as:

$$Precision = TP / (TP + FP) \quad (2)$$

D. Recall (True Positive Rate or Sensitivity)

Recall is calculated as the number of positive predictions divided by the total number of positives. The best recall is 1, whereas the worst is 0. Generally, Recall is calculated by the following formula:

$$Recall = TP / (TP + FN) \quad (3)$$

E. F-measure

F-measure is defined as the weighted harmonic mean of precision and recall. Usually, it is used to combine the Recall and Precision measures in one measure in order to compare different ML algorithms with each other. F-measure formula is given by:

$$F\text{- measure}= (2* Recall * Precision)/(Recall + Precision) \quad (4)$$

F. Root-Mean-Square Error (RMSE)

RMSE is a measure for evaluating the performance of a prediction model. The idea herein is to measure the difference between the predicted and the actual values. If the actual value is X and the predicted value is XP then RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (Xi - XPi)^2} \quad (5)$$

V. EXPERIMENTAL RESULTS

This study used WEKA 3.6.9, a machine learning tool, to evaluate three ML algorithms (NB, DT and ANNs) in software bug prediction problem. A cross validation (10 fold) is used for each dataset.

The accuracy of NB, DT and ANNs classifiers for the three datasets are shown in Table VI. As shown in Table VI, the three ML algorithms achieved a high accuracy rate. The average value for the accuracy rate in all datasets for the three classifiers is over 93% on average. However, the lowest value appears for NB algorithm in the DS1 dataset. We believe this is because the dataset is small and NB algorithm needs a bigger dataset in order to achieve a higher accuracy value. Therefore, NB got a higher accuracy rate in DS2 and DS3 datasets, which they are relatively bigger than the DS1 dataset.

TABLE VII. ACCURACY MEASURE FOR THE THREE ML ALGORITHMS OVER DATASETS

Datasets	NB	DT	ANNs
DS1	0.898	0.951	0.938
DS2	0.950	0.972	0.954
DS3	0.954	0.990	0.963
Average	**0.934**	**0.971**	**0.951**

TABLE VIII. PRECISION MEASURE FOR THE THREE ML ALGORITHMS OVER DATASETS

Datasets	NB	DT	ANNs
DS1	0.956	1	1
DS2	0.989	0.990	0.981
DS3	0.990	1	0.990
Average	**0.978**	**0.996**	**0.990**

TABLE IX. RECALL MEASURE FOR THE THREE ML ALGORITHMS OVER DATASETS

Datasets	NB	DT	ANNs
DS1	1	1	1
DS2	0.905	1	0.990
DS3	0.972	1	0.981
Average	**0.959**	**1**	**0.990**

The precision measures for applying NB, DT and ANNs classifiers on DS1, DS2 and DS3 datasets are shown in Table VII. Results show that three ML algorithms can be used for bug prediction effectively with a good precision rate. The average precision values for all classifiers in the three datasets are more than 97%.

The third evaluation measure is the recall measure. Table VIII shows the recall values for the three classifiers on the three datasets. Also, herein the ML algorithms achieved a good recall value. The best recall value was achieved by DT classifier, which is 100% in all datasets. On the other hand, the average recall values for ANNs and NB algorithms are 99% and 96%, respectively.

In order to compare the three classifiers with respect to recall and precision measures, we used the F-measure value. Fig. 1 shows the F-measure values for the used ML algorithms in the three datasets. As shown the figure, DT has the highest F-measure value in all datasets followed by ANNs, then NB classifiers.

Finally, to evaluate the ML algorithms with other approaches, we calculated the RMSE value. The work in [2] proposed a linear Auto Regression (AR) model to predict the accumulative number of software faults using historical measured faults. They evaluated their approach with the POWM model [20] based on the RMSE measure. The evaluation process was done on the same datasets we are using in this study.

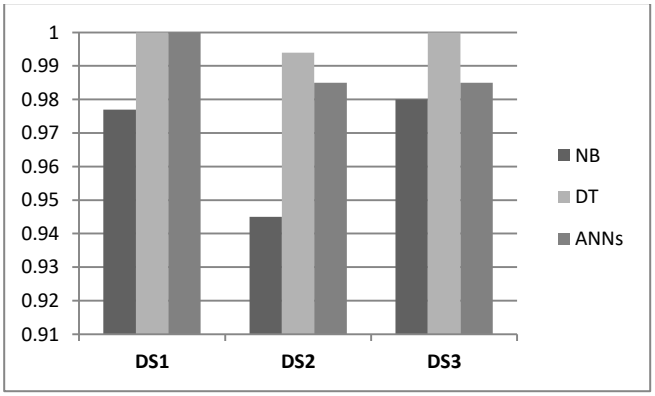


Fig. 1. F-measure values for the used ML algorithms in the three datasets.

TABLE X. RMSE VALUES FOR THE THREE ML ALGORITHMS, AR MODEL, AND POWM MODEL

	Machine Learning Algorithms			Approaches Presented in [2]	
Datasets	***NB***	***DT***	***ANNs***	***AR Model***	***POWM Model***
DS1	0.163	0.082	0.151	4.096	14.060
DS2	0.199	0.104	0.130	0.687	150.075
DS3	0.120	0.062	0.162	3.567	152.969

Table IX presents the RMSE measure for the used ML algorithms, as well as, AR and POWM models over the three datasets. The results show that NB, DT, and ANNs classifiers have better values than AR and POWM models. The average RMSE value for all ML classifiers in the three datasets is 0.130, while the average RMSE values for AR and POWM models are 2.783 and 105.701, respectively.

VI. CONCLUSIONS AND FUTURE WORK

Software bug prediction is a technique in which a prediction model is created in order to predict the future software faults based on historical data. Various approaches have been proposed using different datasets, different metrics and different performance measures. This paper evaluated the using of machine learning algorithms in software bug prediction problem. Three machine learning techniques have been used, which are NB, DT and ANNs.

The evaluation process is implemented using three real testing/debugging datasets. Experimental results are collected based on accuracy, precision, recall, F-measure, and RMSE measures. Results reveal that the ML techniques are efficient approaches to predict the future software bugs. The comparison results showed that the DT classifier has the best results over the others. Moreover, experimental results showed that using ML approach provides a better performance for the prediction model than other approaches, such as linear AR and POWM model.

As a future work, we may involve other ML techniques and provide an extensive comparison among them. Furthermore, adding more software metrics in the learning process is one possible approach to increase the accuracy of the prediction model.

REFERENCES

[1] Y. Tohman, K. Tokunaga, S. Nagase, and M. Y., “Structural approach to the estimation of the number of residual software faults based on the hyper-geometric districution model,” IEEE Trans. on Software Engineering, pp. 345–355, 1989.

[2] A. Sheta and D. Rine, “Modeling Incremental Faults of Software Testing Process Using AR Models ”, the Proceeding of 4th International Multi-Conferences on Computer Science and Information Technology (CSIT 2006), Amman, Jordan. Vol. 3. 2006.

[3] D. Sharma and P. Chandra, "Software Fault Prediction Using Machine-Learning Techniques," Smart Computing and Informatics. Springer, Singapore, 2018. 541-549.

[4] R. Malhotra, "Comparative analysis of statistical and machine learning methods for predicting faulty modules," Applied Soft Computing 21, (2014): 286-297

[5] Malhotra, Ruchika. "A systematic review of machine learning techniques for software fault prediction." Applied Soft Computing 27 (2015): 504-518.

[6] D'Ambros, Marco, Michele Lanza, and Romain Robbes. "An extensive comparison of bug prediction approaches." Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on. IEEE, 2010.

[7] Gupta, Dharmendra Lal, and Kavita Saxena. "Software bug prediction using object-oriented metrics." *Sādhanā* (2017): 1-15..

[8] M. M. Rosli, N. H. I. Teo, N. S. M. Yusop and N. S. Moham, "The Design of a Software Fault Prone Application Using Evolutionary Algorithm," IEEE Conference on Open Systems, 2011.

[9] T. Gyimothy, R. Ferenc and I. Siket, "Empirical Validation of Object-Oriented Metrics on Open Source Software for Fault Prediction," IEEE Transactions On Software Engineering, 2005.

[10] Singh, Praman Deep, and Anuradha Chug. "Software defect prediction analysis using machine learning algorithms." 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, 2017.

[11] M. C. Prasad, L. Florence and A. Arya, "A Study on Software Metrics based Software Defect Prediction using Data Mining and Machine Learning Techniques," International Journal of Database Theory and Application, pp. 179-190, 2015.

[12] Okutan, Ahmet, and Olcay Taner Yıldız. "Software defect prediction using Bayesian networks." Empirical Software Engineering 19.1 (2014): 154-181.

[13] Bavisi, Shrey, Jash Mehta, and Lynette Lopes. "A Comparative Study of Different Data Mining Algorithms." International Journal of Current Engineering and Technology 4.5 (2014).

[14] Y. Singh, A. Kaur and R. Malhotra, "Empirical validation of object-oriented metrics for predicting fault proneness models," Software Qual J, p. 3–35, 2010.

[15] Malhotra, Ruchika, and Yogesh Singh. "On the applicability of machine learning techniques for object oriented software fault prediction." Software Engineering: An International Journal 1.1 (2011): 24-37.

[16] A.TosunMisirli, A. se Ba¸ S.Bener,"A Mapping Study on Bayesian Networks for Software Quality Prediction", Proceedings of the 3rd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, (2014).

[17] T. Angel Thankachan1, K. Raimond2, "A Survey on Classification and Rule Extraction Techniques for Data mining",IOSR Journal of Computer Engineering ,vol. 8, no. 5,(2013), pp. 75-78.

[18] T. Minohara and Y. Tohma, "Parameter estimation of hyper-geometric distribution software reliability growth model by genetic algorithms", in Proceedings of the 6th International Symposium on Software Reliability Engineering, pp. 324–329, 1995.

[19] Olsen, David L. and Delen, " Advanced Data Mining Techniques ", Springer, 1st edition, page 138, ISBN 3-540-76016-1, Feb 2008.

[20] L. H. Crow, "Reliability for complex repairable systems," *Reliability and Biometry, SIAM*, pp. 379–410, 1974.

Long-Term Weather Elements Prediction in Jordan using Adaptive Neuro-Fuzzy Inference System (ANFIS) with GIS Techniques

Omar Suleiman Arabeyyat
Computer Engineering Department
Al-Balqa Applied University, BAU
Al-Salt, Jordan

***Abstract*—Weather elements are the most important parameters in metrological and hydrological studies especially in semi-arid regions, like Jordan. The Adaptive Neuro-Fuzzy Inference System (ANFIS) is used here to predict the minimum and maximum temperature of rainfall for the next 10 years using 30 years' time series data for the period from 1985 to 2015. Several models were used based on different membership functions, different methods of optimization, and different dataset ratios for training and testing. By combining a neural network with a fuzzy system, the hybrid intelligent system results in a hybrid Neuro-Fuzzy system which is an approach that is good enough to simulate and predict rainfall events from long-term metrological data. In this study, the correlation coefficient and the mean square error were used to test the performance of the used model. ANFIS has successfully been used here to predict the minimum and maximum temperature of rainfall for the coming next 10 years and the results show a good consistence pattern compared to previous studies. The results showed a decrease in the annual average rainfall amounts in the next 10 years. The minimum average annual temperature showed the disappearance of a certain predicted zone by ANFIS when compared to actual data for the period 1985-2015, and the same results behavior has been noticed for the average annual maximum.**

***Keywords*—Rainfall prediction; hybrid intelligent system; Adaptive Neuro-Fuzzy Inference System (ANFIS); GIS; time series prediction; long-term weather forecasting; climate change**

I. Introduction

Rainfall plays an enormous role in climate classification. As a result, the climate in any area is strongly influenced by rainfall [5]. It is found that the rate of warming varies from one region to another on the earth surface, and the precipitation shows either increasing or decreasing rates in various regions on the earth's surface. Several decades of warming and a variety of hydrologic and landscape responses have already occurred, and these changes are expected to accelerate during the 21st Century unless the greenhouse-gas emissions are brought under control and are even reversed [12].

Climate change studies in Jordan show that the minimum temperature has increased twice the rate of the maximum temperature increase [10]. One of the essential steps in climate change studies is the analysis of trends in the available records of climatological data for the selected stations. Weather forecasting is the application of science and technology to predict the state of the atmosphere for a future time and a given location [13]. One of the main fields of weather forecasting is rainfall prediction, which is important for the food production plan, water resource management and all activity plans in nature. Rainfall forecasting is a non-linear forecasting process that varies according to area. It is strongly influenced by climate change. Another parameter which is considered an element of climate is the temperature, in its both average annual maximum temperatures and the average annual minimum temperature; both temperatures can also give a good indication about the behavior of the weather. In this study, both temperatures are also introduced in the simulation process for the predication as well as for the rainfall.

In Jordan, and according to the rainfall pattern, different zones or regions have been created and grouped into three major categories and as follows, with the highest rainfall (400-600 mm/yr) occurring in the northwest of the country in upper land areas. The lower mean annual rainfall of (250-350) mm occurs in central Jordan. Mainly, the region is far north of the Jordan Valley. The lowest rainfall (less than 170 mm/yr) occurs in the lower land regions of the east and south of the country [10].

In this study, the main aim is to study the effectiveness of using the Adaptive Neuro-Fuzzy Inference System (ANFIS) in rainfall and temperature predictions for the coming two decades in a region like Jordan which is classified as a semiarid and dry region. Weather elements data are multi-dimensional, non-linear, and dynamic. Therefore, to search for an appropriate model, a comparative evaluation is conducted to evaluate the proposed system against others found in the literature; the obtained results show the competiveness and the power of ANFIS technique for this study.

The remaining part of this study is structured as follow: the second section provides a definition for the ANFIS technique. The third section presents the research methodology followed in this study. The fourth section discusses the results. The final section concludes the study and suggests future work.

II. Adaptive Neuro Fuzzy Inference System (ANFIS)

The Adaptive Neuro-Fuzzy Inference System (ANFIS), which is the main interest of this study, was first introduced by

Jang in 1993 [7]. ANFIS algorithm is the fuzzy-logic based paradigm that grasps the learning abilities of ANN to enhance the intelligent system's performance by using the knowledge gained after learning. Using a given input-output data set, ANFIS constructs a fuzzy inference system whose membership function parameters are tuned or adjusted using a hybrid type of neural algorithms [8]. Several techniques can be used for prediction such as neural network or fuzzy logic but ANFIS has largely extended the capabilities of both technologies in hybrid intelligent systems. The advantages of neural networks in learning and adaptation and those of fuzzy logic systems in dealing with the issues of human-like reasoning on the linguistic level, transparency and interpretability of the generated model as well as in handling uncertain or imprecise data, enable building higher level intelligent systems [14].

III. Research Methodology

This study uses various combinations of data sets of Long-term rainfall readings from 26 stations in Jordan obtained from the Jordan Meteorological Department (JMD). These data represent annual historical rainfall readings for more than 30 years. Before starting the work, it is necessary to pre-process the data and remove outliers in order to compute average annual rainfall readings for the whole stations. GIS is a very useful tool which is used for computing and representing these data as shown in Fig. 1.

The fuzzy inference technique used in this study is the Mamdani method, which was proposed by Mamdani and Assilian. In Mamdani's model, the fuzzy implication is modeled by Mamdani's minimum operator. The conjunction operator is min, the t-norm from compositional rule is min, and for the aggregation of the rules, the max operator is used. The ANFIS Architecture used in this study is illustrated in Fig. 2. There are five layers in ANFIS which have one input layer with a neural training layer. The Fuzzification layer, which transfers the original crisp values into fuzzy ones, chooses the value with the maximum membership degree to participate in the process of the neural training layer according to the maximum membership degree principal. One output layer is similar to that in ANN but with a fuzzy number output.

Developing an index based on the fuzzy logic necessitates the comprehension of three important parts of the fuzzy inference system, including membership functions, fuzzy set operations and inference rules. Each selected input or input set has a domain called the universe of discourse that is divided into subsets which are expressed by linguistic terms. The relationships between the subsets of inputs and outputs, as well as those among the subsets of inputs, are defined by, if and then by rules and fuzzy set operators [6].

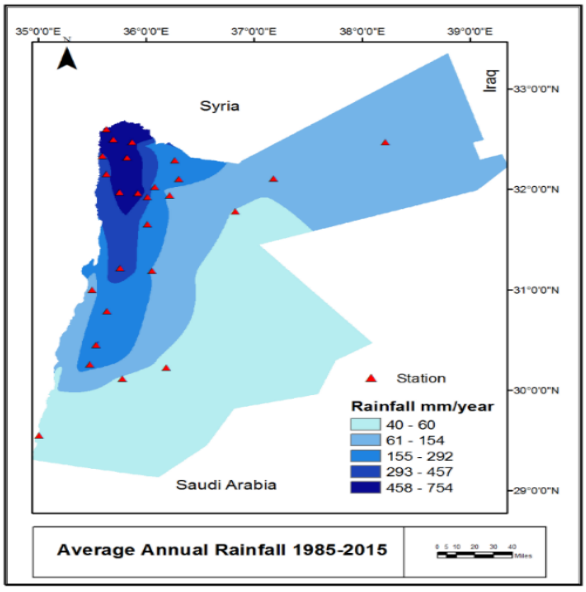


Fig. 1. Interpolated average annual rainfall for the period 1985-2015 of the study area.

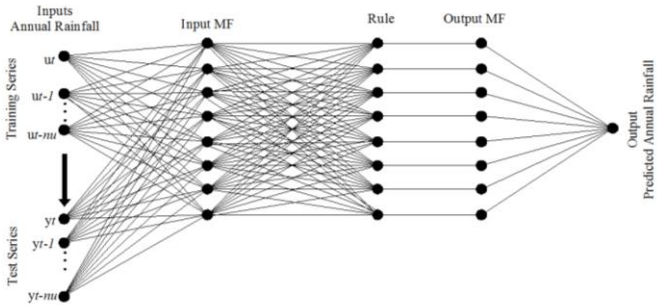


Fig. 2. General architecture of the used ANFIS.

In the Fuzzification of entries step, all variables are assigned to a membership function in order to be transformed from numerical to linguistic subsets such as small, poor, excellent or high ones. A Membership Function (MF) is a function that defines how each point in the input space is mapped to a membership value between 0 and 1 [15]. It is necessary to know that not only can a single map have more than one fuzzy membership function, but also several different maps can have membership values for the same proposition or hypothesis [2].

Defining the rules is an important step in the Fuzzy logic approach by which we link the hypothesis with the conclusion through a certainty factor. These rules are based on the form "if …then and". The knowledge in a problem-solving area can be represented by a number of rules. The task of rules' definition is usually accomplished by experts with general knowledge on the specific field. There is no need for assigning weights in the criteria used.

The processing of the rules or inference step in Mamdani's approach consists of three stages [4]:

a) Aggregation, which returns the fulfillment of hypothesis for every rule individually (max, sum).

b) Implication, which combines the aggregation's results to the rule's certainty factor (min, prod).

c) Accumulation, which brings together the individual results of the variables.

The Fuzzy gamma operation is used in this study, which is defined in terms of the fuzzy algebraic product and the Fuzzy algebraic sum by (1):

$$\mu(x) = (\text{Fuzzy Sum})^{\gamma} * (\text{Fuzzy Product})^{1-\gamma}$$
$$, 0 \leq \gamma \leq 1 \qquad (1)$$

where (γ) is a parameter chosen in the range (0 , 1). The wise choice of the (γ) produces output values that ensure a flexible compromise between the "increasing" tendencies of the Fuzzy algebraic sum and the "decreasing" effects of the Fuzzy algebraic product where (γ) is a parameter chosen in the range (0, 1). When (γ) is 1, the combination is the same as the fuzzy algebraic sum; and when (γ) is 0, the combination equals the fuzzy algebraic product [9].

The defuzzification of the output fuzzied values is the transformation of the fuzzy set results into a linguistic expression or a crisp value [3]. This transformation can be done by several methods such as centroid, bisector, Middle, Smallest, and Largest of Maximum; the used method in this study is the Centroid defuzzification.

Centroid defuzzification returns the center of area under the curve; this method selects the output crispy value corresponding to the center of gravity of the output membership function. The only disadvantage of this method is that it is computationally difficult for complex membership functions [16].

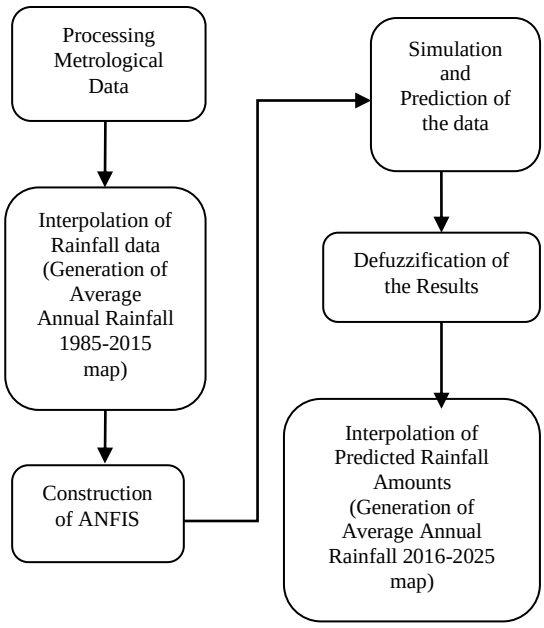


Fig. 3. Flowchart of the methodology followed in conducting the study.

In this study, several membership functions were tested. The Gaussian shape membership function is used for representing the situation of rainfall in the study area. For the used Fuzzy Gamma operator, Bonham-Carter [2] discussed the effect of variations in (γ) for the case of combining two values μA =0.75 and μB =0.5. In this study, several trials were performed to determine the value of gamma (γ) which yields the best reliable rainfall map. γ =0.85 is the most satisfactory value for that.

The methodology followed in conducting this study is presented in Fig. 3. It show the processes in order, start with data processing using the GIS tool, after that the data will be processed using the simulator tool using the ANFIS.

IV. Results

In order to have the best results from the ANN models we change the number of the neurons, hidden layers and the learning and training algorithms which we did until reached the best results of the data. To study the effectiveness of ANFIS, three weather elements have been studied here, the rainfall, the maximum annual average temperature, and the minimum one. The following results were obtained for three elements simulated by ANFIS.

A. ANFIS and Average Annual Rainfall Prediction

The actual data for the annual average rainfall during the period 1985-2015 is shown in Fig. 1. As the figure illustrates, the output results for the study area have come up with five zones according to the average precipitation, which are: *Very Poor, Poor, Moderate, High, and Very High*. In this study, the results will be discussed and analyzed according to these classifications. Using the GIS tools, the area for each zone has been calculated from Fig. 1.

The total area for the whole country is calculated by the same way using GIS tools, and then the percentage of each zone to the total country's area can be easily determined. Table I gives the calculated percentage of the actual data for the period from 1985-2015. ANFIS is introduced here to predict the average annual rainfall during the period 2016-2025 according to the method introduced before. The predicted results using ANFIS simulation are shown in Fig. 4.

Comparing between the predicted results in Table I and Fig. 1 and 4, it is clear that the area that has a *very high rainfall* zone intensity will witness a shrinkage to about 2.81% from 3.29% in comparison to the study area during the period 1985-2015. This means that the rainfall will be less in comparison to the actual data during the period 1985-2015.

The highest percentage portion of the study area, which is about 87.2%, belongs to *very poor rainfall* intensity as shown in Table I. This means that the Country will witness a drawback in rainfall and more dryness as a lack of rainfall will be expected. These obtained results are in great consistency with previous studies which are related to rainfall predictions ([1], [11], [16], [17]).

TABLE II. RAINFALL CLASSIFICATION RESULTS FOR BOTH ACTUAL AND SIMULATED DATA USING ANFIS

Rainfall Class	Very High	High	Moderate	Poor	Very Poor
Area% (1985-2015) Actual data	3.29	4.37	9.64	39.88	42.82
Area% (2016-2025) predicted using ANFIS	2.81	2.03	3.41	4.55	87.20

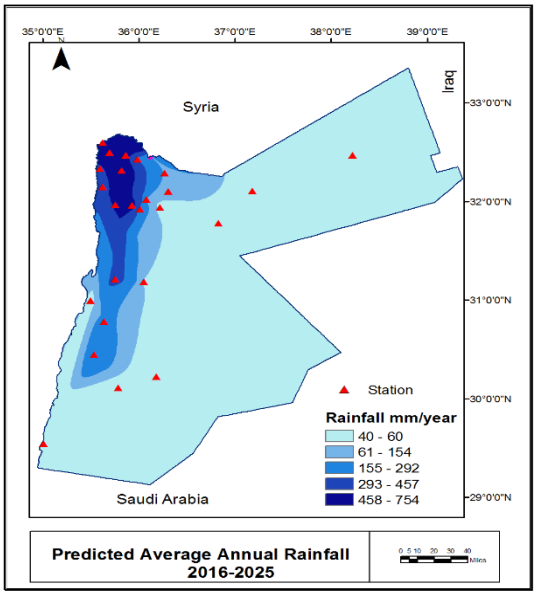


Fig. 4. Interpolated predicted average annual rainfall for the period 2016-2025 of the study area.

B. ANFIS and Minimum Average Temperatures Prediction

The total average annual minimum temperature for actual data during the period 1985-2015 and the predicted simulated data according to ANFIS during the period 2016-2025 are both shown in Fig. 6 after converting it to a map using GIS. Table II also shows the percentage of the actual data for the minimum temperature distributed all over the study area. The average minimum temperatures according to the actual data have also been classified into five zones, same as the rainfall pattern introduced before, and it is better to keep the same zones classification for rainfall pattern or comparison purposes. According to Fig. 5, the pattern zones for the average annual minimum temperature can be classified in ^{o}C as follows:

Very High: 14.9-19.8,
High: 13.1-14.8,
Moderate: 11.8-13.0,
Low: 10.2-11.7, and
Very Low: 9.0-10.1.

From Table II, it is clear that the ANFIS prediction shows that there is a change pattern in a minimum temperature at the country level, in comparison to the two periods 1985-2015 and 2016-2025. There is a clear decrease in the very high zone temperature for the period 2016-2025, which means that the potential for getting a colder weather among the area in the Rift valley –Middle West of the country- and north region in the country is possible. Since the calculated percentage of this area is very high it has decreased from 4.37% to 1.18% as indicated in Fig. 5 by red color zones. It is very interesting to see that the zone of ***low*** has tangibly shrink in a way that almost disappeared in the middle region zone in the country, and most of this zone has moved to the *moderate* zone which has been increased from 54.39% to 85.31% as illustrated in Table II. This will give an indication from this prediction that this middle region of the country will witness a higher pattern in its average annual minimum temperature compared to the period 1985-2015. This zone is indicated by the yellow color in Fig. 5.

TABLE III. AVERAGE ANNUAL MINIMUM TEMPERATURE RESULTS FOR BOTH ACTUAL AND SIMULATED DATA USING ANFIS

Annual Minimum Temperature (^{o}C)	Very High 14.9-19.8	High 13.1-14.8	Moderate 11.8-13.0	Low 10.2-11.7	Very Low 9.0-10.1
Area% (1985-2015) Actual data	4.37	7.67	54.39	30.4	3.17
Area% (2016-2025) predicted using ANFIS	1.18	8.21	85.31	4.73	0.57

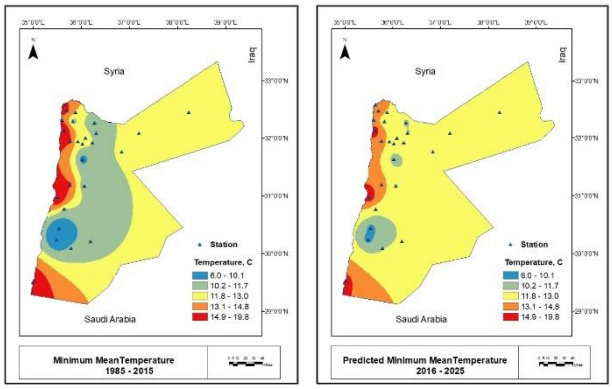


Fig. 5. Interpolated predicted average annual minimum temperature actual and predicted by using ANFIS.

C. ANFIS and Maximum Average Temperatures Prediction

The total average annual maximum temperature for the actual data during the period 1985-2015 and predicted simulated data according to ANFIS during the period 2016-2025 are both shown in Fig. 6 after using GIS tools. The classification of temperature zones based on the actual data plotted by GIS tools has come up with five zones same as rainfall and minimum temperature classification and will keep the same classification for the easiness of comparison as follows in (^{o}C):

Very High: 29.0-31.4,
High: 26.0-28.9,
Moderate: 25.0-26.9,
Low: 23.0-24.9, and
Very Low: 19.0-22.9

Table III gives a clear indication that the percentage changes in the average annual maximum temperature has not significantly changed in its patterns compared to the average annual minimum temperature shown in Table II. For example, the change in the *moderate* zones from period 1985-2015 to period 2016-2025 is only 5% in difference when it is compared to the average annual minimum temperature in Table II. Moreover, there will be an increase in the southern region of the country for the *Very High* zone, which means that these areas, like the Gulf of Aqaba, will witness a temperature that is higher than the one witnessed in the previous period 1985-2015. This region is indicated in Fig. 6 with the red color. It is also clear that the *low* zone has clearly disappeared during the period 2016-2025, which means that this region will witness a higher temperature during the predicted period, and this is indicated in Fig. 6 with the dark green color. It is clear that the *moderate* zone has not changed significantly after the simulation and prediction. In other words, this region will not be affected by the weather changes in its pattern of the average annual maximum temperature. This region is indicated in Fig. 6 with the orange color and is almost known in Jordan as the desert area.

TABLE IV. AVERAGE ANNUAL MAXIMUM TEMPERATURE RESULTS FOR BOTH ACTUAL AND SIMULATED DATA USING ANFIS

Annual Maximum Temperature (°C)	**Very High 29.0-31.4**	**High 26.0-28.9**	**Moderate 25.0-26.9**	**Low 23.0-24.9**	**Very Low 19.0-22.9**
Area% (1985-2015) Actual data	4.16	61.01	13.9	15.6	5.33
Area% (2016-2025) predicted using ANFIS	7.03	69.89	18.07	4.94	0.07

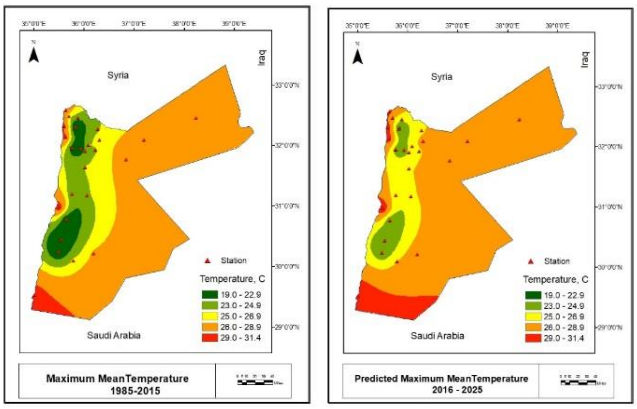


Fig. 6. Interpolated predicted average annual maximum temperature actual and predicted by using ANFIS.

V. CONCLUSIONS AND FUTURE RESEARCH

In this study it is clear that the model ANFIS has presented good results for simulating and predicting rainfall precipitation in the arid regions. The pattern according to the ANAFIS model has classified the study area into reasonable zones, which are consistent with the actual case study according to rainfall and temperature predictions. The predicted results have come up with a conclusion that there will be a decrease pattern in rainfall at the country level. For the temperature prediction results, the country will also witness a lower in minimum average annual temperature, and some regions have moved from lower region temperature zones to higher ones. The same behavior has been noticed on the average annual maximum temperature as some regions have a higher percentage area compared to the actual data during the period 1985-2015 which will mean a pattern changes in weather elements, while some regions has moved from the ***high zone*** to the ***moderate zone***, which means a lower maximum temperature.

The initial investigation of applying the ANFIS technique to weather elements prediction shows a good performance, in the future, other artificial intelligence technique like NARX-ANN model can be used, also other factors of weather elements can be included like humidity, speed of the wind and height to show more valuable results.

REFERENCES

[1] A. Dahamsheh and H. Aksoy, Structural Characteristics of Annual Precipitation Data in Jordan, Theoretical and Applied Climatology, Vol. 88, No. 3-4,2007, pp. 201-212.

[2] B. Carter, F. Graeme, Geographic Information Systems for Geoscientists, Modelling with GIS, Oxford; Pergamon Press, 1994.

[3] C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

[4] E. Mandelas, T. Hatzichristos and P. Prastacos, A Fuzzy Cellular Automata Based Shell for Modeling Urban Growth – A Pilot Application in Mesogia Area .10th AGILE International Conference on Geographic Information Science , Aalborg University, Denmark,2007.

[5] G. Alfarisy and W. Mahmudy, Rainfall forecasting in Banyuwangi using adaptive neuro fuzzy inference system. Journal of Information Technology and Computer Science Vol.1, No. 1, 2016, pp. 65 – 71.

[6] H. Garibi, A. Mavi, R. Nabizadeh, H. Arabalibeik, M. Yunesian and M. Sowlat, A novel approach in water quality assessment based on fuzzy logic .Journal of Environmental Management, Vol. 112, 2012, pp.87-95.

[7] J. Jang, ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Trans. Syst. Man Cybern. Vol. 23, 1993, pp. 665–685.

[8] J. Patel and F. Parekh, Forecasting Rainfall Using Adaptive Neuro-Fuzzy Inference System (ANFIS). International Journal of Application or Innovation in Engineering & Management (IJAIEM). Vol. 3, No. 6, 2014, pp 262- 269.

[9] J. Rather and Z. Raouf, Fuzzy Logic Based GIS Modeling for Identification of Groundwater Potential Zones in the Jhagrabaria Watershed of Allahabad District, Uttar Pradesh, India. International Journal of Advances in Remote Sensing and GIS, Vol.1, No.2, 2012.

[10] M. Freiwan and M. Kadioglu, Spatial and temporal analysis of climatological data in Jordan. International Journal of Climatology, Vol 28, No. 4, 2008, pp.521–535.

[11] M. Matouq, T. El-Hasan, H. Al-ilbisi, M. Abdelhadi, M. Hindiyeh, S. Eslamian and S. Duheisat, The climate change implication on Jordan: A case study using GIS and Artificial Neural Networks for weather forecasting, Journal of Taibah University for Science, Vol. 7, No. 2, 2013, pp. 44-55

[12] M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, IPCC, Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2007.

[13] N. Bushara and A. Abraham, Computational Intelligence in Weather Forecasting: A Review. Journal of Network and Innovative Computing, Vol. 1, No. 1,2013, pp. 320-331.

[14] N. Bushara and A. Abraham, Using Adaptive Neuro-Fuzzy Inference System (ANFIS) to Improve the Long-term Rainfall Forecasting. Journal of Network and Innovative Computing,Vol. 3, No.1,2015, pp. 146-158.

[15] P. Riad, M. Billib, A. Hassan and M. Omar, Overlay Weighted Model and Fuzzy Logic to Determine the Best Locations for Artificial Recharge of Groundwater in a Semi-Arid Area in Egypt , Nile Basin .Water Science & Engineering Journal, Vol. 4, No.1, 2011.

[16] S. Naaz, A. Alam, and R. Biswas, Effect of different defuzzification methods in a fuzzy based load balancing application. International Journal of Computer Science Issues, Vol.8, No. 5, 2011.

[17] T. Alwada'n, B. Zahran, A. Mesleh, M. Matouq and O. Al-Heyasat, Rainfall Prediction in Semi-Arid Regions in Jordan Using Back Propagation Neural Networks International Journal on Engineering Applications, Vol. 3, No. 6, 2015 pp.158-163.

Detection Capability and CFAR Loss Under Fluctuating Targets of Different Swerling Model for Various Gamma Parameters in RADAR

Md. Maynul Islam, Mohammed Hossam-E-Haider
Department of Electrical, Electronic and Communication Engineering
Military Institute of Science and Technology
Dhaka, Bangladesh

***Abstract*—Target detection of RADAR deals with different and manifold problems over few decades. The detection capability is one of the most significant factors in RADAR system. The main aim of detection is to increase probability of detection while decreasing rate of false alarm. The threshold of detection is modified as a function of the receiver noise level to keep a fixed rate of false alarm. Constant False Alarm Rate (CFAR) processors are used to maintain the amount of false alarm under supervision in a diverse background of interference. In Signal to Noise Ratio (SNR) level, a loss can be occurred due to CFAR processor. Gamma function is used to determine the probability of false alarm. It is assumed in adaptive CFAR that the interference distribution is familiar here. This type of CFAR also approximates the unknown parameters connected with various interference distributions. CFAR loss depends on gamma function. Incomplete gamma function plays an important role in maintaining threshold voltage as well as probability of detection. Changing the value of gamma function can improve the probability of detection for various Swerling Models which are proposed here. This paper has also proposed a technique to compare various losses due to CFAR in terms of different gamma function in presence of different number of pulses for four Swerling Models.**

***Keywords*—Swerling model; Constant False Alarm Rate (CFAR) loss; false alarm; gamma function; probability of detection**

I. Introduction

In presence of non-stationary background noise (or noise plus clutter) the detection of radar return signals becomes complicated. Function of time can represent a radar target depending on the huge number of real targets whose return changes in magnitude from low to high. Probability of detection in radar depends on many parameters, incomplete gamma function is one of them. Varying the value of gamma parameter, the detection capability can be improved. In Constant False Alarm Rate, the measurement of the noise power levels from the leading and the trailing reference windows are dependent on the Cell Averaging (CA) technique [1]-[4]. The efficiency of CA detector is evaluated in the situations when the operating environment is perfect and when it includes some of fallacious targets along with the target of interest. The primary and the secondary targets are considered to be fluctuating in terms of four Swerling models. The theoretical results show that for various False Alarm rates the probability of detection will be different for various gamma parameter. Four types of Swerling Model have various CFAR rate for changing number of pulses in presence of different gamma parameters.

II. System Model

When only noise present in the radar, the probability of false alarm P_{fa} is defined when a sample exceed the threshold voltage V_T. The detection probability P_D is the probability that a sample can surpass the threshold voltage having noise plus signal. It can be written as,

$$P_D = \int_{V_T}^{\infty} \frac{r}{\Psi^2} I_o\left(\frac{rA}{\Psi^2}\right) \exp[-(r^2 + A^2)/2\Psi^2] dr \qquad (1)$$

Where r is the envelope of the threshold voltage, A is the amplitude of the return signal with variance of noise Ψ^2. For a radar signal of sine waveform having amplitude A, the power of the signal will be $A^2/2$.

The Chi-square distribution is applied to a wide range of targets, its *pdf* can be written as,

$$f(\sigma) = \frac{m}{\Gamma(m)\sigma_{avg}} \left(\frac{m\sigma}{\sigma_{avg}}\right)^{m-1} e^{-\frac{m\sigma}{\sigma_{avg}}} \qquad (2)$$

Where, $\Gamma(m)$ is the gamma function of argument m and σ_{avg} is the average value. As the degree gets larger the distribution corresponds to constrained Radar Cross Section (RCS) values. The limit m tends to ∞ corresponds to a constrained RCS target.

Detection of signals threshold is constantly balanced as a function of the receiver noise level in different cases to maintain a constant false alarm rate [5]. In Signal to Noise Ratio (SNR) level a loss of 1 dB can be occurred due to CFAR processor.

In order to maintain a fixed predetermined probability of false alarm, the threshold of detection is calculated. A relationship between the threshold value V_T and the probability of false alarm P_{fa} can be shown as:

$$V_T = \sqrt{2\Psi^2 \ln\left(\frac{1}{P_{fa}}\right)} \qquad (3)$$

If the noise power Ψ^2 is assumed to be constant, then a fixed threshold can satisfy the above equation. However, due to many reasons this condition is rarely true. In order to maintain a constant probability of false alarm the threshold value must

be continuously updated based on the estimates of the noise variance. The method of continuously changing the threshold value to maintain a fixed probability of false alarm is known as Constant False Alarm Rate (CFAR) [6].

The Swerling models were introduced to model a variety of target reflections that occur over the radar integration interval. In Swerling model I & II where the signal amplitudes are fully correlated over the incoherent integration interval but are independent from one integration interval to the next. In Swerling model II & IV the signal amplitudes are uncorrelated from pulse to pulse throughout the integration interval [7].

The probability of false alarm corresponding to a fixed threshold was derived earlier. When CA-CFAR is implemented, then the probability of false alarm can be derived from the conditional false alarm probability, which is averaged over all possible values of the threshold in order to achieve an unconditional false alarm probability. The conditional probability of false alarm when $y = V_T$ can be written as [6]

$$P_{fa}(V_T) = e^{-\left(\frac{y}{2\Psi^2}\right)} \tag{4}$$

As a result, unconditional probability of false alarm is [8]

$$P_{fa} = \int_0^{\infty} P_{fa}(y) f(y) dy \tag{5}$$

Where, $f(y)$ is the *pdf* of the threshold value.

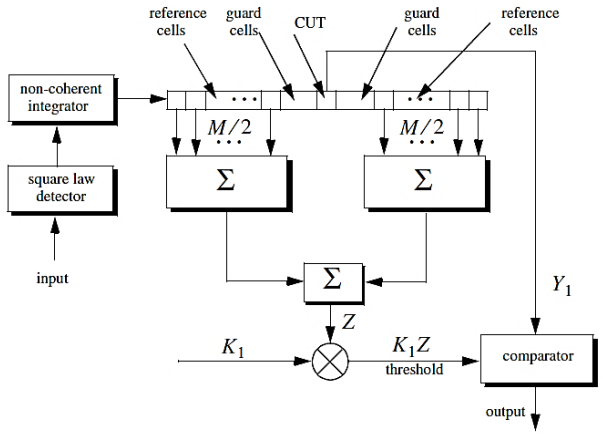


Fig. 1. Conventional CA-CFAR with non-coherent integration.

Practically, CFAR averaging is often implemented after non-coherent integration and the output of each reference cell is the sum of squared envelopes (Fig. 1). It follows that the total number of summed reference samples is Mn_p. The output Y_1 is also the sum of n_p squared envelopes.

When noise alone is present in the Cell Under Test (CUT), Y_1 is random variable whose *pdf* is a gamma distribution with $2n_p$ degrees of freedom. Additionally, the summed output of the reference cells is the sum of Mn_p squared envelopes. Thus, Z is also a random variable which has a gamma *pdf* with $2Mn_p$ degrees of freedom [9].

The probability of false alarm is then equal to the probability that the ratio Y_1/Z exceeds the threshold. More precisely,

$$P_{fa} = Prob\left\{\frac{Y_1}{Z} > K\right\} \tag{6}$$

In target detection, threshold V_T can be determined from probability of false alarm, P_{fa}. For any number of pulses and non-coherent integration DiFranco & Rubin give a standard form relating threshold & probability of false alarm [9].

$$P_{fa} = 1 - \Gamma_I\left(\frac{V_T}{\sqrt{n_p}}, n_p - 1\right) \tag{7}$$

Where, Γ_I is used to denote the incomplete gamma function and it can be expressed as [9].

$$\Gamma_I\left(\frac{V_T}{\sqrt{n_p}}, n_p - 1\right) = \int_0^{V_T\sqrt{n_p}} \frac{e^{-\gamma}\gamma^{n_p-1-1}}{(n_p-1-1)!} d\gamma \tag{8}$$

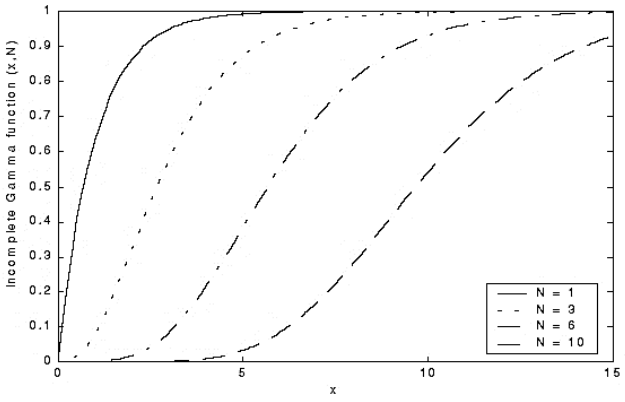


Fig. 2. The incomplete gamma function for different values of N.

Fig. 2 shows the behaviour of incomplete gamma function with respect to independent variable which can be varied due to *x* and *N* [10].

III. Simulation

For different Swerling Model Simulation has been carried out in terms of fluctuating target. Conventional Cell Averaging CFAR has been used here. Probability of detection is different for different false alarm rate if the Signal to Noise Ratio (SNR dB) varies firmly. Due to loss of constant false alarm rate (CFAR) [11]-[15], probability of detection can be changed in radar detection for fluctuating target. For every model of Swerling, CFAR loss has been simulated and compared. Detection probability of target detection can be found in terms of Signal to Noise Ratio (SNR). From probability of false alarm, loss occurred due to Constant false alarm rate was calculated and compared with respect to gamma function for different swerling model on fluctuating target.

IV. Result and Discussion

In this paper, comparison has been shown for CFAR loss vs gamma function for four types of Swerling model. From Fig. 3 it is seen that the value of CFAR loss is decreasing with increasing of gamma function for Swerling model I. In this case CFAR loss can be reduced if the number of pulse is comparatively less.

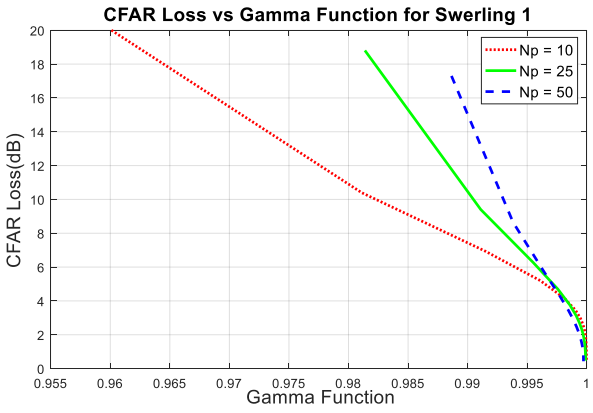


Fig. 3. CFAR Loss vs Gamma Parameter for Swerling Model I.

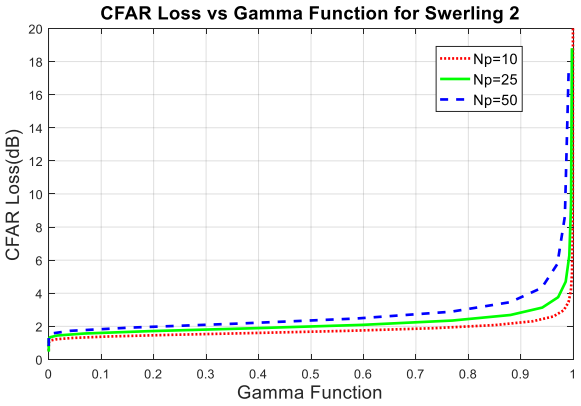


Fig. 4. CFAR Loss vs Gamma Parameter for Swerling Model II.

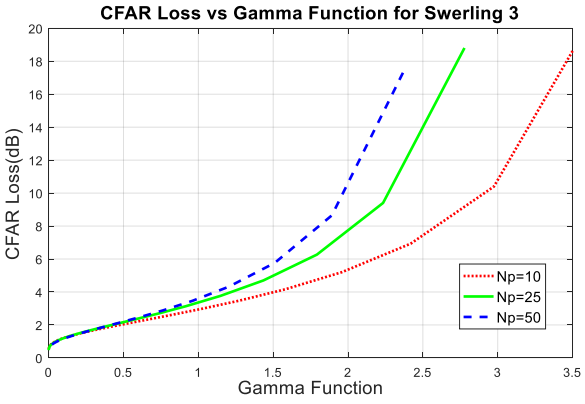


Fig. 5. CFAR Loss vs Gamma Parameter for Swerling Model III

From Fig. 4 it is clear that the value of CFAR loss is sharply increasing after a certain period of gamma function for Swerling model II. The curve of CFAR loss slightly varies from others for different number of pulses.

It is shown in Fig. 5 that CFAR loss is increased if the value of gamma function increases for Swerling model III. CFAR loss is comparatively low for less number of pulses.

For Swerling model IV from Fig. 6 it is seen that the curve of CFAR loss shows rapid response in terms of gamma function. For a small value of gamma CFAR loss can be increased sharply upto 20 dB.

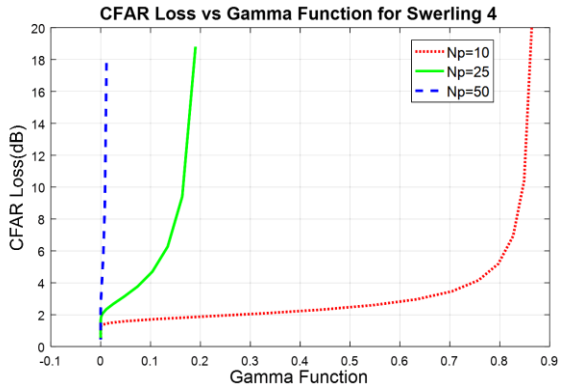


Fig. 6. CFAR Loss vs Gamma Parameter for Swerling Model IV.

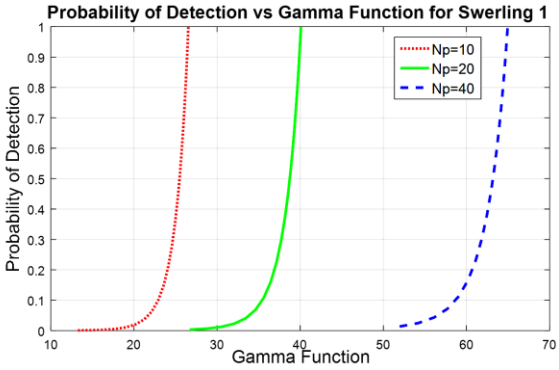


Fig. 7. Pd vs Gamma Parameter for Swerling Model I.

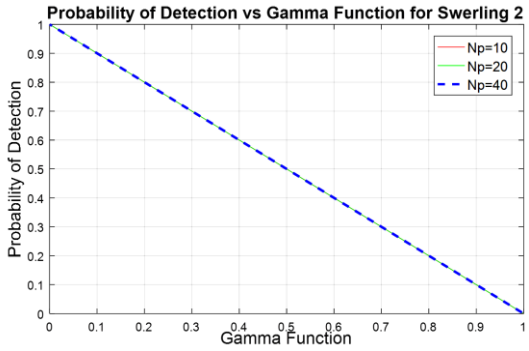


Fig. 8. Pd vs Gamma Parameter for Swerling Model II

Probability of detection also depends on gamma parameter. For Swerling model 1 detection capability of radar increases rapidly for a little change of gamma function (Fig. 7). For different number of pulses the range of gamma parameter is different.

The fluctuation of targets is independent from pulse to pulse rather than from scan to scan for Swerling model 2. From Fig. 8 it is clear that probability of detection decreases linearly with increasing gamma function. It happens for any number of pulse.

Probability of detection also decreases with increasing gamma function in Swerling model 3. But the behavior of this curve is almost linear in nature. If the number of pulse is higher, the detection capability improves for any values of gamma parameter which is shown in Fig. 9.

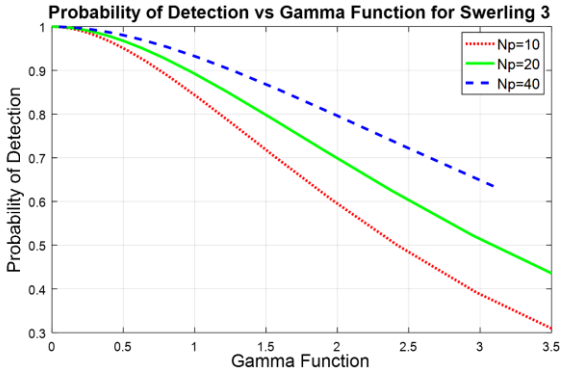


Fig. 9. Pd vs Gamma Parameter for Swerling Model III.

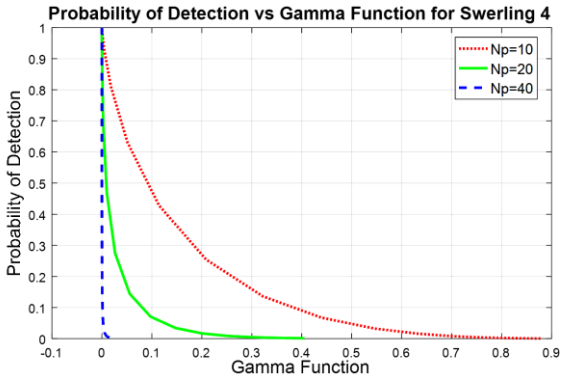


Fig. 10. Pd vs Gamma Parameter for Swerling Model IV.

For Swerling model 4, the detection capability rapidly decreases from its highest value with incrasing gamma parameter. For different number of pulses the curve shows various behaviours which are mentioned in Fig. 10.

It is clear from the above figures that there happens more CFAR loss for Swerling model II, III & IV if the value of gamma function is increased. For Swerling model I, CFAR loss is less for increasing value of gamma function. Again, gamma function is related to probability of detection in RADAR. At low value of gamma, the CFAR loss is minimum for Swerling model III. It is clear that Swerling model III has the second highest detection capability where the loss due to Constant False Alarm Rate (CFAR) is lower. For Chi square distribution used in probability of detection, gamma is inversely proportional to the pdf. For Swerling model II, III & IV the probability of detection decreases with increasing gamma parameter. For better detection capability, the value of gamma function should be kept as low.

V. Conclusion

This paper presents an analytical method for comparison of CFAR loss for various value of gamma function & method of improving detection capability in RADAR technology. Differences among four Swerling model have been simulated in case of fluctuating target. It is observed that lower CFAR loss can give better accuracy in target detection. For probability of detection lower gamma function is also desirable. It has been revealed that decreasing the value of gamma as well as increasing Cell Array can be the better solution where targets are fluctuated. In future analysis, comparison between CFAR Loss and Cell Array can be a new dimension of this research. Due to inherent nature of coherent and non-coherent integration, the proposed system is flexible and easy to implement.

References

[1] G. Gigli and G. A. Lampropoulos, “A new maximum likelihood generalized gamma CFAR detector,” in *Geoscience and Remote Sensing Symposium, 2002*, Toronto, Ontario, Canada, 2002.

[2] C. J. Willis, “Target modeling for SAR image simulation,” *The International Society for Optical Engineering*, vol. 9243, Oct. 2014.

[3] B. Thomas and S. L. Donnie, “Improved RCS model for censored Swerling III and IV target models,” in *IEEE Aerospace Conference Proceedings*, 2013.

[4] E. M. Mohamed, “Analytical performance evaluation of adaptive detection of fluctuating radar targets,” *Radioelectronics and Communications Systems*, vol. 56(7), pp. 321–334, Jul. 2013.

[5] K. Lingjiang, C. Guolong, Y. Xiaobo, and W. Bing, “Constant false alarm rate performance prediction for non-independent and non-identically distributed gamma fluctuating targets,” *IET Radar Sonar Navigation*, vol. 10(5), Oct. 2015.

[6] Mahafza, Bassem, and Atef Elsherbeni. *"Radar Detection"*, Electrical Engineering Handbook, 2009.

[7] D.A. Shnidman, "Expanded swerling target models", *IEEE Transactions on Aerospace and Electronic Systems, 2003.*

[8] Mahafza, *"Radar Detection"*, Radar Systems Analysis and Design Using MATLAB, 2000.

[9] B. R. Mahafza, “*Radar Systems Analysis and Design Using MATLAB*.” CHAPMAN & HALL/CRC, 2000.

[10] Mahafza, "Target Detection and Pulse Integration", Radar Signal Analysis and Processing Using MATLAB, 2008.

[11] B. El Mashade, Mohamed. "Partially-Correlated χ2 Targets Detection Analysis of GTM-Adaptive Processor in the Presence of Outliers", *International Journal of Image Graphics and Signal Processing, 2014.*

[12] Md. Maynul Islam, Md. Roman Sarker, Md. Tanjilul Alam, and Mohammed Hossam-E-Haider, "Comparison of analog and digital pulse compression technique and reduction of side lobes using transversal filter", *International Conference on Electrical Engineering and Information & Communication Technology, 2014.*

[13] W. Bing, C. Guolong, K. Lingjiang, and Y. Xiaobo, “Performance prediction of the CFAR detector for generalized Swerling-Chi fluctuating targets,” in *IEEE National Radar Conference Proceedings*, 2013, vol. 52(1), pp. 1–5.

[14] E´. Magraner, N. Bertaux, and P. Re´fre´gier, “A new CFAR detector in gamma-distributed nonhomogeneous backgrounds,” presented at the 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, 2008.

[15] L. H., Z. S., and Y. J., “An integrated target detection and tracking algorithm with constant track false alarm rate,” *Journal of European Industrial Training*, May 2016.

Intelligent Transportation System (ITS) for Smart-Cities using Mamdani Fuzzy Inference System

[1]Kashif Iqbal
Government College University
Lahore, Pakistan

[2] Muhammad Adnan Khan,
Sagheer Abbas, Zahid Hasan
School of Computer Science
National College of Business
Administration & Economics,
Lahore, Pakistan

[3] Areej Fatima
Department of Computer Science
Lahore Garrison University,
Lahore, Pakistan

***Abstract*—It is estimated that more than half of the world population lives in cities according to (UN forecasts, 2014), so cities are vital. Cities, as we all know facing with complex challenges – for smart cities the outdated traditional planning of transportation, environmental contamination, finance management and security observations are not adequate. The developing framework for smart-city requires sound infrastructure, latest current technology adoption. Modern cities are facing pressures associated with urbanization and globalization to improve quality-of-life of their citizens. A framework model that enables the integration of cloud-data, social network (SN) services and smart sensors in the context of smart cities is proposed. A service-oriented radical framework enables the retrieval and analysis of big data sets stemming from Social Networking (SN) sites and integrated smart sensors collecting data streams for smart cities. Smart cities' understanding is a broad concept transportation sector focused in this article. Fuzzification is shown to be a capable mathematical approach for modelling traffic and transportation processes. To solve various traffic and transportation problems a detailed analysis of fuzzy logic systems is developed. This paper presents an analysis of the results achieved using Mamdani Fuzzy Inference System to model complex traffic processes. These results are verified using MATLAB simulation.**

Keywords*—Information Communication Technology (ICT); Internet of Things (IoT); Intelligent Transportation System (ITS); Fuzzy Inference System (MFIS); Traffic Congestion Conditions (TCC); SNA; MF; Mamdani Fuzzy Inference System (MFIS)

I. INTRODUCTION

It is the time of Social Networking, Cloud Computing and explosion of smart sensors deployed everywhere [1]. According to UN survey in 2014, more than half of world's population now living in urban areas [2] and increasing surely alerting city planners. Connected cities emerge when Internet of Things (IoT) technologies and socially-aware network systems aggregate administrations over a whole connected metropolitan territory. When thinking of connected urban areas, one may think of high tech cities that have the prominent cutting-edge technologies for their citizens like Copenhagen, London, New York, Chicago, Stockholm or Amsterdam. However, small residential communities have also been benefiting from interfacing individuals, administrations, city infrastructure and services. This article investigates city transportation problem and a portion of the difficulties that are involved with developing widespread IoT techniques. The coalition of world-class IoT improvement anticipates working with each of these smart urban communities that enable citizens to make technology utilization more sensible, adaptable and sustainable. Many urban cities and towns around the globe are turning to socially connected smart devices to solve urban problems [3], for example, traffic congestion, environmental contamination, healthcare, security surveillance to enhance the living standards for their general public everyday comforts. Smart sensors that are installed throughout the city, in vehicles, in buildings, in roadways, in control monitoring systems, security surveillance and applications and devices that are utilised by individuals who are living or working in the city [4]. Delivering information to the public that is utilizing through these high tech smart cities opportunities. The big-data analytics utilized to decide on how public spaces are planned, how to make the best utilization of their assets and how to convey administrative notifications more proficiently, viable and appropriately [5].

Therefore, most urban cities have embraced huge investments during recent decades in Information Communication Technology (ICT) infrastructure including computers, broadband availability and some sensing frameworks [6]. These infrastructures have engaged various inventive administrations in territories, for example, demographic sensing, urban coordination and real information that makes living ones close. Such administrations have been widely sent in a few urban cities, accordingly exhibiting the potential advantages of ICT frameworks for organisations and the natives themselves [7]. During most recent years it has

additionally seen a blast of sensor distribution, along with the development of adaptive systems, internet-of-things [8] current advancements of sensor-based systems have emerged. Currently, the advantages of social communication and internet-of-things distributions for smart urban areas have likewise been exhibited [9].

Current Smart City data analysis implies complex stream analytics for a comprehensive set of activities aiming to turn into real actionable outcomes [10]. The analysis comprises of following contributions:

1) Analysis of thousands of traffic blockage cases, road capacity measures, traffic signalling and dynamic, consistent information to give a better message to the citizens.

2) Events, episodic road examination, utilising real-data gathered by citizens, devices and sensors.

3) Turning web-based into social media information, important city events analysis, assumptions, examinations, and numerous other things. Consolidating information from physical (sensors/devices) and social sources (social organisations) can give full, essential information and adds to better assessment and bits of knowledge.

Over-all speaking, smart cities realization is a broad concept so, the transportation sector is focused in this article. Fuzzy logic is one of the strongest candidate solution for mathematical based modelling. In this article fuzzy logic-based solution is proposed for transportation problem. The input parameters are: Vehicular Speed (VS), Road Capacity (RC), Traffic Signals (TS), Trip Riding Distance (RD) and Distance Traffic Signals (DTS). A detailed Transportation fuzzy logic system is developed based on rule-based inferencing to solve the traffic congestion issues. Analysis of the results obtained using Mamdani Fuzzy Inference System is verified using MATLAB Simulation

The objective of this paper is to analysis key issues and the solutions about traffic congestion in a smart city in the light of critical inducing aspects. The rest of the paper is structured as follows: Section II gives an overview of related and similar works that can be found in the international literature. Section III presents the fundamental architecture and approach. Sections IV and V presents technical details, sentiment analysis of problems and a conceptual model for smart-cities. Section VI provides a proposed Mamdani Fuzzy Inference System (MFIS) based results analyses, the work is planned in the context of simulation and Section VII contains conclusion and future work to be planned in the context of smart drive mobile apps.

II. Literature Review

A smart IoT system which automatically notifies necessary information of passengers after triggering of shock detector sensors to lowering loss rates in accidents and alert nearby local public safety organization about the physical location of accident suggested by Nasr *et al.* [11]. Rizwan et al. industrialize a smart traffic management system roadside unit. It carries alternate routing to avoid traffic blocking and increase traffic flow through IoT and lower traffic density, offers predictive analytic technique (Big-data techniques) [12]. Scalable Enhanced Road Side Unit, SERSU, proposed by Al-Dweik *et al.* used wireless communication network and radio frequency adaptive traffic control system, pollution detection system and weather information system. SERSU components were placed on the roadsides with various breaks, capturing generated sensor signals by vehicle sensors module [13]. Modern techniques in cars, internet and their current and future relationship, detail history of usage of electronic devices in automobiles, and social implication of these technologies briefly studied by Goggin [14]. Joshi *et al.* made infrared-based sensor system, which to monitor traffic flow and provides alternate road traffic routing path to drivers for the avoidance of traffic crowding capture infrared radiations emitted by vehicles on road surface [15]. Handte *et al.* designed IoT enabled the navigational system for real transport facility, provided complete guidance of routes to bus riding passengers for urban bus riders in Madrid, which were assisting in micro-navigation, expects massive aware routes. A system to communicate with onboard sensors to sense the presence of onboard passengers, this system was based on mobile devices. Their system collected real-world bus user's response for better accessibility of travel information [16]. Zanella *et al.* advised web-based service approach for IoT service architecture to resolve integration issues for different end node devices connected to IoT system Zanella *et al. also* evaluated key ideas, facilities and solution are currently available for implementation of IoT based smart cities [4]. Technological challenges and socio-economic opportunities in developing and designing of future smart cites discussed key by Theodoridis *et al.*, they also suggested 3-tier IoT nodes and 3-plane architecture model. Further, they develop a city scale test bed for future internet and IoT experimentation [17]. A hierarchy which combines smart homes and smart cities described by Skouby *et al.*, they also proposed a four-layered model to join end nodes IoT devices, communication technologies like distributed artificial intelligence and cloud of things [6]. Gubbi *et al.* presented Radio Frequency Identification (RFID's) a user-centric cloud-based vision of implementation of IoT, by the interaction of public and private clouds, major research trends, IoT application domain, current and future enabling technologies etc. that will drive IoT shortly [14]. Base Station arrangement, based architecture sensor system for intelligent traffic light system (TLS) suggested by Chong *et al*. They designed intelligent software, implemented on TLS which continuously communicates with the base station and calculates green light time, and provide monitoring of traffic by officers [18].

Internet of Vehicle (IoV), a unique solution for smart traffic management is discussed by Dandala *et al*. They argued that IoV can be an effective solution conventional IoT based traffic management technique to overcome traditional traffic issues. Further, they described to be a reality which is a vehicle to vehicle's owner that IoV needs four types of communication, a vehicle to vehicle, a vehicle to centralize server and vehicle to the third party like police patrol, ambulance, etc. [19]. Cognition was used for user authentication in vehicles [26]. Sagheer *et al.* proposed a fuzzy inference system to avoid traffic congestion using bio-inspired method [27].

Density-based signalling to overwhelmed issues raised by fixed time signalling for example in fixed time signalling method the traffic lights have predefined periodic time system suggested by Thakur *et al.* provides intelligent signalling by assigning the greener signal to dense traffic region to avoid congestion by continuously evaluating traffic density [20].

Ramchandra *et al.* proposed a comparable system which device traffic lights by using average speed of vehicles dynamically according to the density of traffic. In this proposed system every vehicle is equipped with On-Board

Device (OBD) distribute data to centralise server using Zigbee protocol which acquires vehicle speed data process [21].

Chowdhury *et al.* proposed intelligent traffic light system for messaging between emergency vehicles infrastructure and to reduce traffic congestion and increase reliability to traffic signals. The proposed system considers the priority of vehicle depends on the type of incident and to secure signals from hacking [22]. Some shortcomings in the traditional intelligent transportation system and argued to prefer Radio Frequency Identification (RFID) pointed out by Ou *et al.*, sensor system and networking technologies to overwhelmed traditional intelligent transportation systems [23].

Information-Centric Networking to project and device Future Internet Architecture proposed by Amadeo *et al.* In Information-Centric Networking which uses IoT submissions to access data of every end node device having unique location name [24].

III. Smart-City Key Features Realization

The availability of smart solutions for cities has risen quickly over the most recent years. Therefore, technical solutions exist for each city to become smarter. The challenge today is mostly to execute proper solutions proficiently, as opposed to just concentrating on innovations. Smart city areas cannot be developed through a patchwork approach, yet by the well-ordered adoption of incremental changes. The most proper way of smart-city realisation is introducing a smart system working group of volunteers characterize its manageability vision and afterwards lays out an electronic well-ordered guide and execution design. The capacity to distinguish the acutest bottlenecks to send coordinated and flexible solutions and afterwards to use these outcomes into other smart community's activities requires involvement and strong specialised expertise.

Smart City Key Resources: Transportation, Climate Change, Energy, Utilities, Security Surveillance, Healthcare, Business Management, etc.

Connected cities enhance the experience of workers by analyzing data and smart city coordinators by breaking down information from reporting frameworks including sensors, roadside cameras, brilliant monitoring systems and speed check signs. Applying IoT innovations to solve urban community's issues includes gathering the information that is collecting from sensors, recordings by cameras, interpersonal organizations and brilliant devices that are examining real-data. This data is delivered noteworthy bits of knowledge that are utilized straightforwardly to trigger actuators that are associated with smart devices. For example, versatile smart city assets, connected by implications, to illuminate choices on policy and to streamline jobs. In smart urban communities, these arrangements include monitoring geographic information from Global Positing System (GPS) trackers and RFID labels on vehicles [15], structures, buildings and power stations, breaking down the proceed of vehicles to recognize occurrences or blockages. Smart buildings security, interpersonal organizations, city administrations are straightforwardly modifying frameworks continuously to control the activity stream in city events, security observation investigation and reduce traffic delays. Authentic analysis of city traffic, security investigation and movement blockage and roadside sensors information can likewise be utilized to alter time delays, misinterpreting security observation, speed cutoff points and city toll tax, control security monitoring and activity stream in the more flows for long-term outcomes. To route movement around incidents, sensors additionally write about the state of streets conditions, weather updates, buildings structures, road lights and extensions with the goal that support to schedule maintenance when required.

Smart cities will make emerging activities in transportation, utilities, smart buildings and smart security. Smart city design plan leaders shaped a working group of ecosystem system accomplices to evaluate robust city community's abilities and guide a long-term vision that coordinates with the city's future planning. Smart city planners have endorsed digitalising citywide assets like fast travel framework, smart buildings, smart security, electric transport and is additionally pushing ahead for far-reaching IoT hub that will pioneer digital city infrastructure.

Designing a roadmap for smart-cities is based on four core pillars: Connectivity is the foundational layer of a smart-city. In real-time data is collected about peoples, places and things by smart sensors and this data are stored on cloud application servers to analyze and utilised to take better real-time decisions and planning as shown in Fig. 1.

Mobility means moving peoples, goods and information efficiently and efficiently. The economic-mobility means regardless of circumstances online job seekers in smart-cities find maximum jobs available that are not handy via public transportation.

Next is security improving public and private places security, data protection and cyber-security while using latest ICT's technologies on-line and off-line.

Sustainability, of course, is focusing on sustainable practices in critical sectors of cities such as transportation, energy consumption, climate change, utilities, security observations, and financial services.

Implementation of smart-cities solutions may have three things every day for their citizen, i.e., creates values, generates revenues and cut costs depending on value exchange smart systems and smart projects.

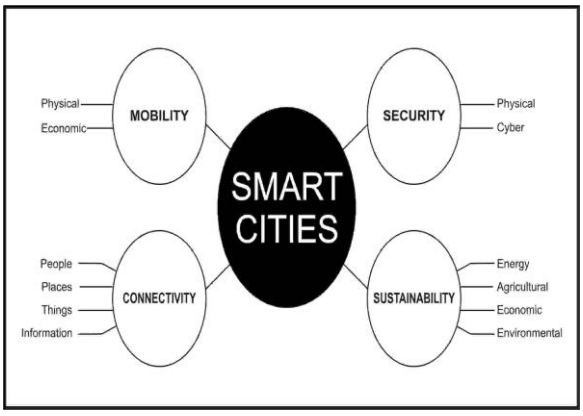


Fig. 1. Road-map four pillars of a smart-city.

A. Smart Cities Framework Model Overview

The adaptive data analysis stage is plotted out in the background. It is made out of different layers, bring down a level (devices, communication planes), middle layers (data, information examination) at higher layers (application, dashboard planes). At each layer, distinctive programming code chunks perform specific operations, related to collecting data, messaging, data accessing, semantic annotation, examination or perception where applications can join segments from different layers in light of their specific pre-requisites. Along these advances toward getting to be plug and play and can be mainly used in smart-city sectors applications. The present extraordinary cutting-edge advances of portability smart-phones, interpersonal organization services and objects are coordinated together for a new time machine to machine and person to person communication correspondence [17].

1) Main Components of the Model

The layered framework model of a smart city as shown in Fig. 2 is having four main layers described below as:

a) Sensing Layer

Sensor Layer comprises tens to thousands of sensor hubs connected using smart remote technologies. They gather data from the environment and convey it to other connected devices that pass the data to the cloud server over the Internet.

b) Communication Layer

Wireless heart innovative technology gives excellent remote protocol access to the full range of processors capacity, control, and resource management to applications. DigiMesh is an exclusive shared systems networking topology for use in remote end-point network connectivity through the physical Internet.

c) Data Layer

The capacity and processing of data should be possible on the edge of the networks itself or in a cloud server. If any preprocessing of data is a need, then it is typically done at either the sensor or some other proximate device.

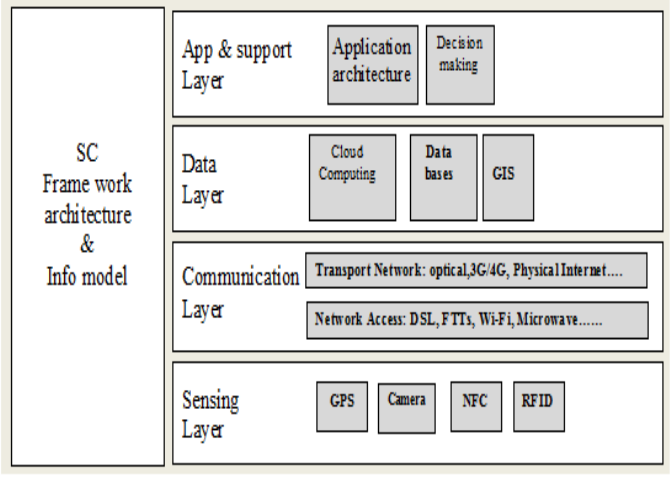


Fig. 2. Layered Model of a smart-city.

The processed data is then regularly sent to a remote server. The capacity and processing abilities of an IoT object are additionally controlled by the assets accessible, which are regularly exceptionally compelled because of constraints of capacity, vitality, control, and computational ability.

d) Application Layer

The application layer is responsible for data organization and presentation. The application layer on the Internet is regularly in light of Hyper Text Transfer Protocol (HTTP)/ File Transfer Protocol (FTP) standards. The proposed events in this examination are sharing of dynamic data to customers using mobile phones as a particular device. It may be HTTP is not reasonable in resource enabling situations since it is relatively verbose and this manner brings about a significant parsing overhead. Many other innovative conventions have been produced for IoT resources, for example, Message Queue Telemetry Transport (MQTT) and Constrained Application Protocol (CoAP).

Along with these four layers following components move toward becoming plug and play real-time integration and stream-analytics that can be utilized explicitly by specific smart applications framework by adaption of these technical modules given below:

a) Data Wrapper

It is a program that extracts the content from a particular information source and translates it into an organization format. Using sensory meta-data, it extends a generic way to describe features of sensors, about the data stream that containing general information. A semantic annotation module annotates the sensory parsed data.

b) Data Aggregation

For data aggregation, the source information originates from public records online databases. The information is packaged into aggregate reports this information is useful for business, marketing, local and government organizations. It reduces the large volume of data, i.e. the size of raw sensory observations delivered by the data wrappers by using data compression techniques and time series analysis.

c) Data Federation

Answers to user queries, according to the requirements it first finds the relevant stream. It then translates the user request into Resource Description Framework-Stream Processing (RDF-RSP) queries and obtains results accordingly. As fast changing real-world data from sensors and online services evolves IoT-based smart environment monitoring, real-time processing and analytics based on RSP semantics. RDF query language manages continuous data streams SPARQL, and CQELS languages support RDF reasoning.

d) Event Detection

The event detection is the identification of items, events and observations, i.e. constraints on what defines an event is relaxed or usually modelled as a set of thresholds or probabilities. In city sectors, it provides tools or web software's applications that monitor urban areas events such as the need for clearing transport deadlock, emergency

facilities, irrigation facilities, pest identification in crops and growing view of the traffic congestion form real-time annotated and aggregated data streams.

IV. Conceptual Framework for Smart-Cities

The proposed technique in this study is sharing of dynamic information to users using smartphones as a communication device. Current day advanced modern technologies of knowledge mobilization, cloud-servers and the smart-city app are integrated into a new era of quick communication. Advanced smart-phones may have limited sensing capabilities but enhanced computational strength, lesser cost, excessive usage, availability of Global System for Mobile communication (GSM) and mobile internet signals, availability of different sensors in smartphones like a gyroscope, digital compass, proximity sensor, etc. Services available like Google map, Google weather, IBM live streaming analytics etc. is prime motivation to use a smartphone as sensor I/O device in the proposed system. Moreover, specialised and more accurate sensors like accelerometer, Global Positioning System (GPS), and shock sensor etc. services are realised. Also, specific and more accurate sensors like accelerometer, global positioning system, and amaze sensor so on so forth are outlined and created on various stages and new technologies integrating with existing technologies in a single integrated system that is beyond the scope of the current proposes a study as shown in Fig. 3.

The proposed system aims to provide efficient and effective smart cities traffic infrastructure. In this study, we show the concepts of cloud computing, bid-data analysis, internet of things, human-computer interactions, software engineering paradigm etc. can be the realization of smart-cities traffic framework to improve the living standards of their citizens.

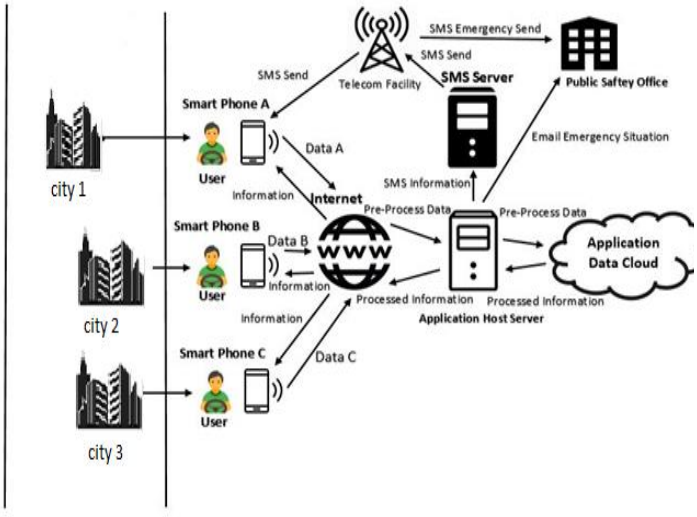


Fig. 3. Proposed Model of Smart City App.

Studies suggest that smart cities need specific information for experiencing globalization by making efficient smart city decisions like smart transportation, smart energy distribution, demo-graphic information, smart utilities, healthcare services, etc.

Travelers from one city to another city have very little information about nearby pinpoint spatial locations, safety organisations, emergency services and government building and necessary information for traveller and visitors with them. In case of any emergency situation, even local public safety organizations have no personal, medical records or emergency contact numbers for any situation. On any highways, peoples hesitate to do or accept any help from others travellers or unknown peoples because of no information. To overcome these issues with smart sensors, surveillance cameras, Wireless Fidelity (Wi-Fi) devices may help the citizen of smart-cities by the following critical technical innovations:

- Users will get real-time and dynamic information about the city routes with other cities in a particular range by Google Map as shown in Fig. 4. Smart mobile-app collected device and personal mobile information for notification, newsfeeds in text or audio format.
- Smart city will alert their citizens about road congestion in the form of text, audio/video format.
- Provide necessary information of other cities around and also provides a platform to communicate with these connected cities by text messages.
- Track record of smart cities user is or travellers from start to destination and generates alerts of essential places nearby like fuel stations, restaurants, hospitals, emergency stations, etc.
- Smart cities are real-data streaming analytics which provides complete details of their citizen with utility services to transportations facilities.
- Users can send and receive with one push any emergency messages to other cities users (using GSM/GPS, other Internet services) as well as inform nearby emergency organizations in the form of an email as shown in Fig. 5.
- Users can comfortably offer or accept emergency pickup, health-care services, nearby building info and share visits from other travellers especially on highways because smart cities will keep track of these connected cities.
- Reporting of any crime, security surveillance, weather forecast, misconduct to authorities nearby (if witnessed) with proper privacy.
- The user receives text messages as well as audio/videos format to prevent mental divergence.

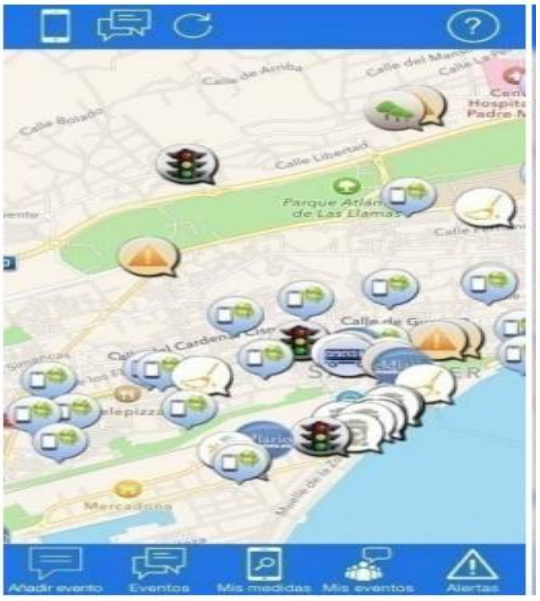


Fig. 4. Sample prototype of Smart-Map.

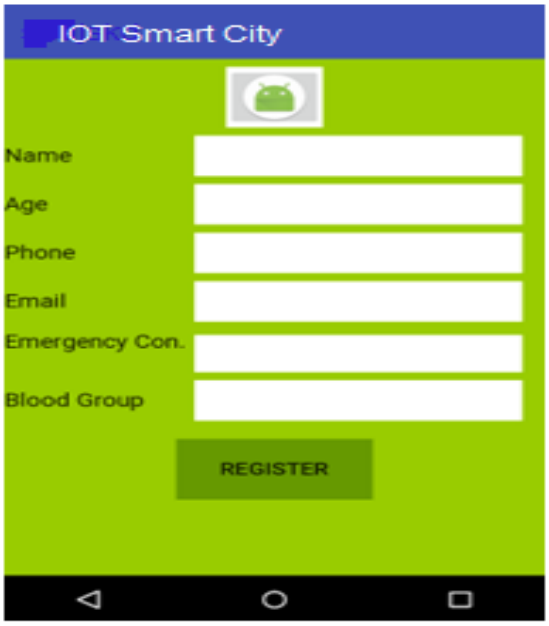


Fig. 5. Prototype of Smart-City App.

Smart cities users are using user's smartphone as sensing as well as a communication device. In the proposed system smartphones will act as wireless sensor network's node. Internet and Mobile telecommunication GSM signals will act a medium of communication between all wireless networked sensor nodes. The application server will host smart-cities application and is connected to Short Message Service (SMS) server which will generate text messages, and it is also an interface between end nodes and application cloud-server. The application server will also send e-mails to public safety organizations in case of an emergency. The application server will provide all the necessary computations. Microsoft Azure IoT cloud server will be used because of its enhanced features for smart cities utility services, transportation conditions, environmental conditions and security features realization.

V. Social Analyses

In this section social analysis of smart cities as a sample has been performed in the form of graphical representation. A graph is a data structure which consisting of a finite number of edges and nodes. There are many ways to represents a node, edge graph, for example, adjacency matrix, graph ML format, CSV files.

The adjacency matrix is a two-dimensional square matrix whose size is equal to the number of nodes in the graph. However, if input graph contains a large number of nodes and less number of edges then the adjacency matrix became sparse and space consuming. Fig. 6 represents a sample connected cities graph and Table I represents an adjacency matrix 6*6 of connected cities.

In the graph illustrated in Fig. 6 nodes represent cities at different ranges and edges for instance roads between cities, paths and connectivity or relationship between cities are in different ranges.

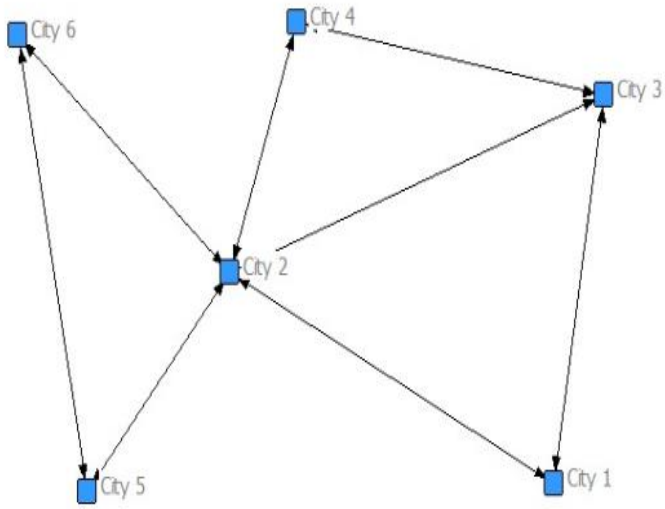


Fig. 6. Connected smart cities graph.

Edges define the relationship between different users or cities resources, a directed edge from city 1 to city 2 represents that city one can communicate with city 2 and city 3. City 2 can communicate with city 5, city 4 and city 3 and so on for every city connection in a graphical format having close centrality measures. The adjacency matrix is represented in Table I in this "0" represents no relationship, and "1" represents the positive relationship. If city one users want to communicate with city four user's, they can communicate with the help of city two based on shortest path algorithm between two nodes traversed.

TABLE I. ADJACENCY MATRIX BETWEEN CITIES 6 X 6

	City 1	City 2	City 3	City 4	City 5	City 6
City 1	0	1	1	0	0	0
City 2	1	0	1	1	1	1
City 3	1	0	0	1	0	0
City 4	0	1	1	0	0	0
City 5	0	1	0	0	0	1
City 6	0	1	0	0	1	0

This phenomenon is used on the higher level as well in computational intelligence. In which every node represents a smart city a cluster of cities and edges represents any one of cites which can reside or act as the interaction between two groups or clusters. The central city would be helpful for communication between cities in different geographical location city areas might be other cities. This technique will enhance the range of communication between two distanced cities. The model of communication of distant (out of the range) cites clusters shown in Fig. 7.

The adjacency matrix of connected cities is shown in Table II.

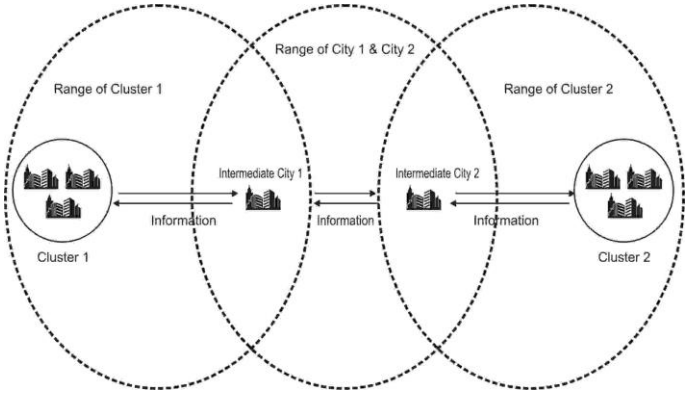


Fig. 7. Communication of cities clusters.

TABLE II. ADJACENCY MATRICES OF DISTANT CITIES

	Cluster 1	Inter City1	InterCity 2	Cluster 2
Cluster 1	0	1	0	0
Intermediate City1	1	0	1	0
Intermediate City2	0	1	0	1
Cluster 2	0	0	1	0

VI. PROPOSED MFIS BASED SOLUTION

This section explains in detail Mamdani Fuzzy Inference System (MFIS) based on smart-city Traffic Congestion Conditions (TCC) controls. The facts given below explain the measuring of TCC for the Smart-city for smart drive facility which is based on Mamdani Fuzzy logic principles.

In this article, the planned MFIS which is capable of measuring the TCC for the city algorithm is given in Table III. The five inputs and one output MFIS is proposed to calculate TCC.

In this method five inputs that are: Vehicle Speed (S), Load Capacity (C), Traffic Signals (T), Distance between Signal (D), Riding Distance (R) are taken. These inputs are used to build up a lookup table given in Table IV to decide TCC for a respective algorithm for input-output relation given by MFIS. Its mathematical representation is shown in (1).

$$\mu_{CG} = \text{MFIS}\ [\mu_{VS}, \mu_{RD}, \mu_{TS}, \mu_{DTS}, \mu_{RD}] \quad (1)$$

In this article, the Intelligent Transportation System (ITS) is measured using Mamdani Fuzzy Inference System (MFIS). Table I shows the proposed MFIS Based ITS algorithm. The I/O surface for MFIS is given in Fig. 1.

TABLE III. PROPOSED MFIS BASED TCC ALGORITHM

1. Inputs: In this system 5, Fuzzy Input variables are used which are the following (Vehicle Speed, Road Capacity, Traffic Signal, Distance Traffic Signal and Riding Distance)
2. Each Fuzzy Input variable has different types of membership functions.
3. Every Fuzzy membership function is used to build fuzzy inference rules.

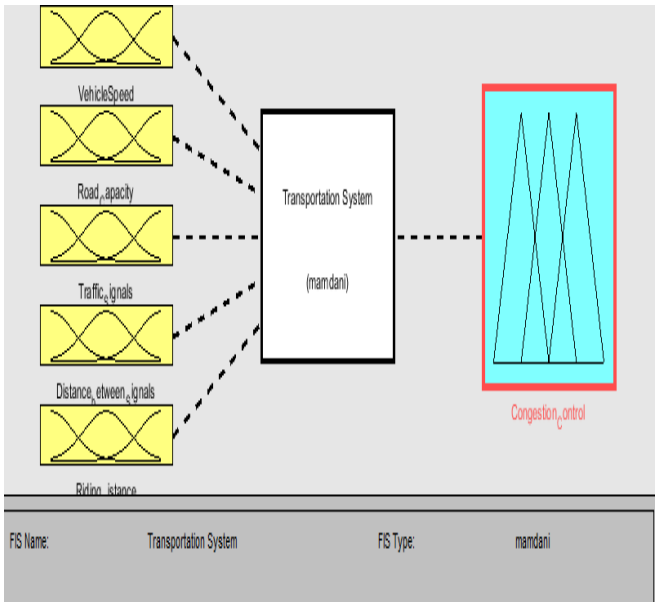


Fig. 8. Input and output surface for MFIS.

TABLE IV. MATHEMATICAL AND GRAPHICAL MF OF MFIS INPUT VARIABLES

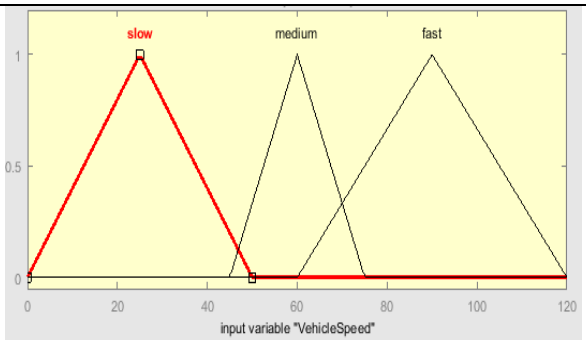
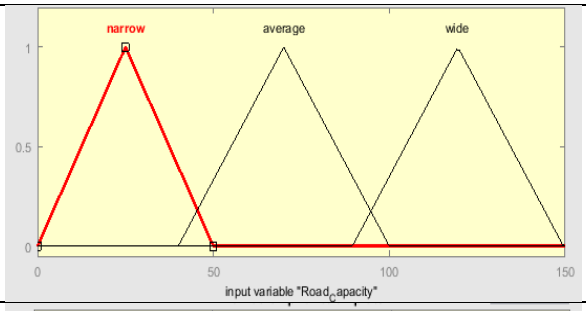
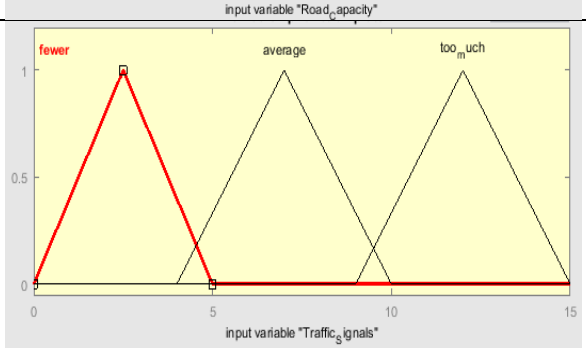
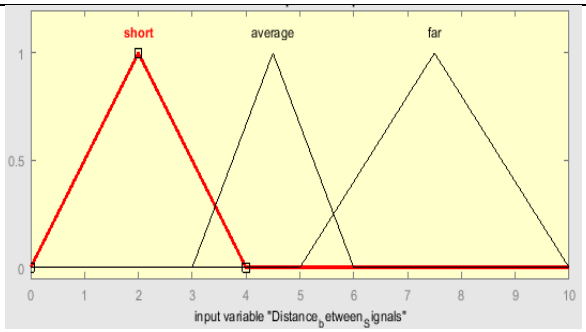
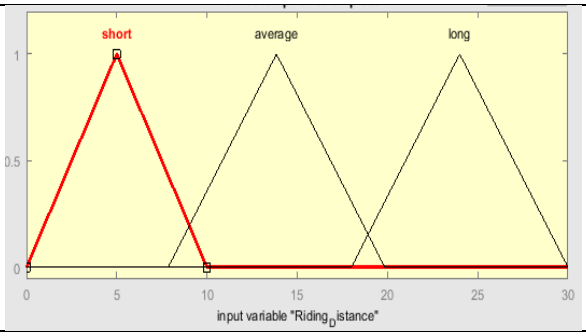
Sr. No.	Input	Membership Function(MF)	Graphical Representation of MF
1	VS (μ_{VS}(S))	$\mu vs, slow \quad (S) = \begin{cases} \frac{s}{25} & if\ S \in [0, 25] \\ \frac{50-s}{25} & if\ S \in [25, 50] \end{cases}$ $\mu vs, medium \quad (S) = \begin{cases} \frac{s-45}{15} & if\ S \in [45,60] \\ \frac{75-s}{15} & if\ S \in [60,75] \end{cases}$ $\mu vs, fast \quad (S) = \begin{cases} \frac{s-60}{30} & if\ S \in [60,90] \\ \frac{20-s}{5} & if\ s \in [90,120] \end{cases}$	
2	RC (μ_{RC}(C))	$\mu rc,,narrow \quad (C) = \begin{cases} \frac{c}{25} & if\ C \in [0, 25] \\ \frac{50-c}{25} & if\ C \in [25, 50] \end{cases}$ $\mu rc, average \quad (C) = \begin{cases} \frac{c-40}{30} & if\ C \in [40,70] \\ \frac{100-c}{30} & if\ C \in [70,100] \end{cases}$ $\mu rc,wide \quad (C) = \begin{cases} \frac{c-90}{30} & if\ C \in [90,120] \\ \frac{150-c}{30} & if\ C \in [120,150] \end{cases}$	
3	TS (μ_{TS}(T))	$\mu TS, fewer \quad (T) = \begin{cases} \frac{T}{2.5} & if\ T \in [0,2.5] \\ \frac{5-T}{2.5} & if\ T \in [2.5,5] \end{cases}$ $\mu TS, average \quad (T) = \begin{cases} \frac{T-4}{3} & if\ T \in [4,7] \\ \frac{10-T}{3} & if\ T \in [7,10] \end{cases}$ $\mu TS, much \quad (T) = \begin{cases} \frac{T-9}{3} & if\ T \in [9,12] \\ \frac{15-T}{3} & if\ T \in [12,15] \end{cases}$	
4	DTS (μ_{DTS}(D))	$\mu DTS, short \quad (D) = \begin{cases} \frac{D}{2} & if\ D \in [0, 2] \\ \frac{4-D}{2} & if\ D \in [2, 4] \end{cases}$ $\mu DTS, average \quad (D) = \begin{cases} \frac{D-3}{1.5} & if\ D \in [3,4.5] \\ \frac{6-D}{1.5} & if\ D \in [4.5,6] \end{cases}$ $\mu DTS, far \quad (D) = \begin{cases} \frac{D-5}{2.5} & if\ D \in [5,7.5] \\ \frac{15-D}{2.5} & if\ D \in [7.5,10] \end{cases}$	
5	RD (μ_{RD}(R))	$\mu RD, short \quad (R) = \begin{cases} \frac{R}{5} & if\ R \in [0,5] \\ \frac{10-R}{5} & if\ R \in [5,10] \end{cases}$ $\mu RD, average \quad (R) = \begin{cases} \frac{R-8}{6} & if\ R \in [8,14] \\ \frac{20-R}{6} & if\ R \in [14,20] \end{cases}$ $\mu RD, long \quad (R) = \begin{cases} \frac{R-18}{6} & if\ R \in [18,24] \\ \frac{30-R}{6} & if\ R \in [24,30] \end{cases}$	

TABLE V. INPUT VARIABLE RANGES

Sr #	Input Parameters	Ranges	Semantic sign
1	VS	0-50 45-75 60-120	Slow Medium Fast
2	Cap	0-50 40-100 90-150	Narrow Average Wide
3	TS	0-5 4-10 9-15	Fewer Normal Too much
4	DTS	2-4 3-6 5-10	Nearer Average Far
5	RD	0-10 8-20 18-30	Nearer Center Far

A. Input Fuzzy Sets

Fuzzy input variable is statistical values that are used to calculate the Traffic Congestion Condition in smart cities.

In this article, five different types of fuzzy variables are used for the analysis of congestion in smart cities. The detail of these input variables is given in Table V.

B. Fuzzy Output Variable

Fuzzy output variable Traffic Congestion Control (TCC) is used to calculate the result by the values of input variables in the world of discourse. The details of output are shown in Table VI.

TABLE VI. OUTPUT VARIABLE RANGES

Sr #	Output of MFIS	Ranges	Semantic sign for Congestion
1	Congestion Control	0 - 0.5 0.2 - 0.7 0.5 - 1	No delay (Less) Average delay (Medium) Much delay (High)

C. Membership Functions

Membership function gives curve value between 0 and 1, and it provides a mathematical function which provides statistical values of input and output variable. Membership functions are also available in MATLAB tool. The propose solution uses the membership function which is as follows:

- Trim

Trim is a triangular curve built-in MATLAB function. To calculation of this function, three scalar parameters are used in the proposed solution which is Low, Medium, and High. The mathematical equations and graphical representation of membership function are given in Table IV.

D. Rule-Based

In this system most, suitable rules for system understanding are applied. This rule-base system contains about 81 input-output rules, the system complexity increased if the number of rules increased. The Mamdani Fuzzy Inference rules are shown in Fig. 9.

E. Inference Engine

The Mamdani Inference Engine is used to map five inputs to one output (TCC) as shown in Fig. 8.

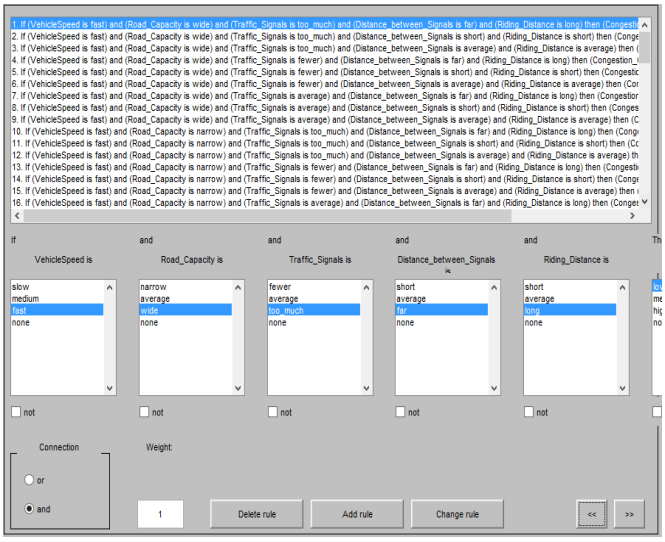


Fig. 9. I/O Rules for ITS.

F. De- Fuzzifier

In this article centroid, De- Fuzzifier is used. Fig. 10 to 12 represents rule surface of Proposed ITS using MFIS.

Fig. 10 shows that if Vehicle Speed is between 1-80 km/s and Traffic Signals are lies in the range of 10 to 15, then Traffic Congestion is approximately 80%, which is high.it also shown that, if Vehicular Speed between 80 – 120 km/s and Traffic Signal is 10-15, then Congestion is low approximately 10%.

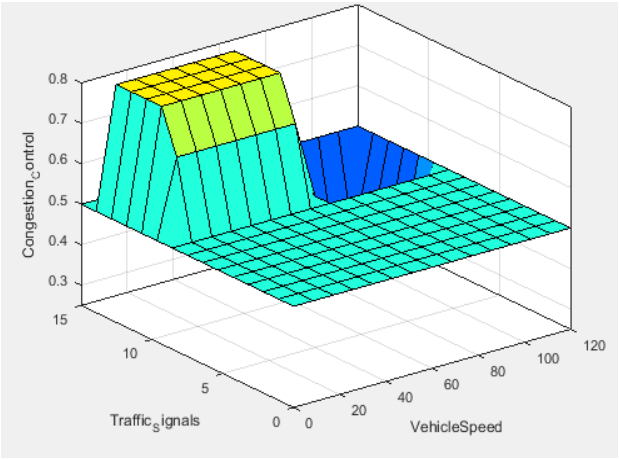


Fig. 10. Rule surface for traffic signals and vehicle speed.

Fig. 11 shows Congestion Control using input variables Traffic signals and Riding Distance. It is observed that congestion is approximately 80% when Traffic signals are 10-15 and Riding Distance between Source to Destination lies in

the range of 8 to 20 km. Congestion is approximately 60% if Traffic signals are 9-10 and Riding Distance grater then 8 km.

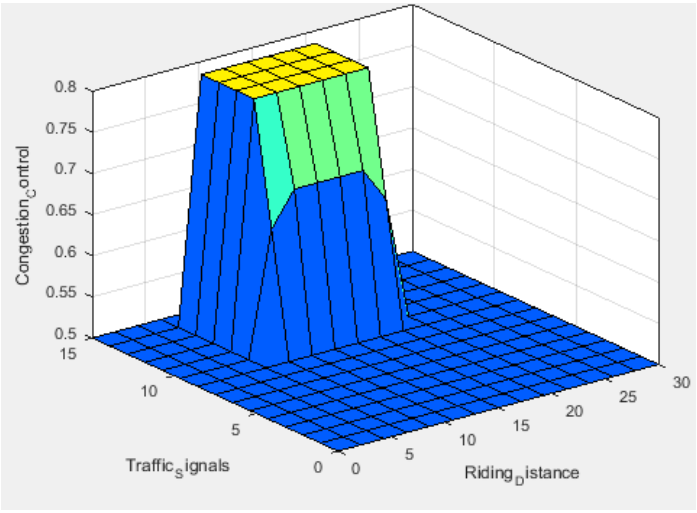


Fig. 11. Rule surface for traffic signals and riding distance.

Fig. 12 shows Congestion depends upon input variables Road Capacity and Riding Distance between source to destination. Approximately, there is no Congestion when road capacity is extensive (110 to 120 vehicles on the road) and Riding Distance is 9 to 20 km. If Road Capacity between 100 to 110 approx and Riding Distance between 10 to 20 km then congestion is increased upto 20% increase. If road capacity is less than 90 (narrow road), then the congestion is up to 50%.

So, it concludes that Congestion inversely proportional to Road Capacity.

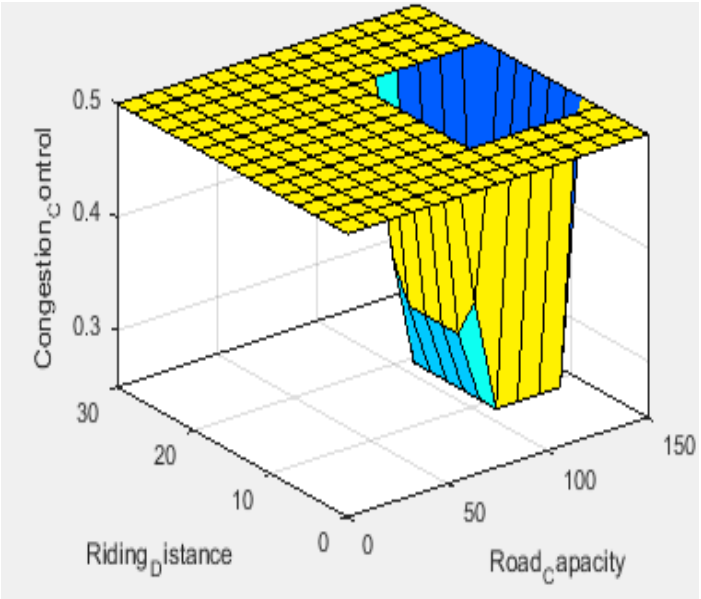


Fig. 12. Rule surface for riding distance and road capacity.

G. Simulation Results

For simulation results, MATLAB R2017a tool is used. MATLAB is also used for modelling, simulation, algorithm development, prototyping and many other fields. MATLAB is an efficient tool for programming, data analysis, visualisation and computing. For simulation results, five inputs and one output Congestion Control variable is used.

TABLE VII. CONGESTION CONTROL BASED ON RULES DEFINED

Input Variables					Output
VS	**RC**	**TS**	**DTS**	**RD**	**Congestion**
H	W	L	H	H	Less Delay
H	N	H	L	L	Medium Delay
L	N	H	L	L	High Delay

Table VII explains the rules of Proposed Congestion Control system. Fig. 13 to 15 shows the proposed system evaluation.

Fig. 13 shows the congestion is less if vehicular speed is high and road capacity is wide. It further depicts that if a traffic signal is few and distance between signals is high and riding distance is far congestion is low.

Fig. 14 shows that congestion is Medium if the vehicle speed is high and Road Capacity is Narrow, and traffic signals are too much, and the distance between signals is low, and riding distance is far then congestion is medium.

Fig. 15 explains congestion is high if the vehicle speed is Low and Road Capacity is Narrow, and traffic signals are too much, and the distance between signals is small, and riding distance is also low than congestion is high.

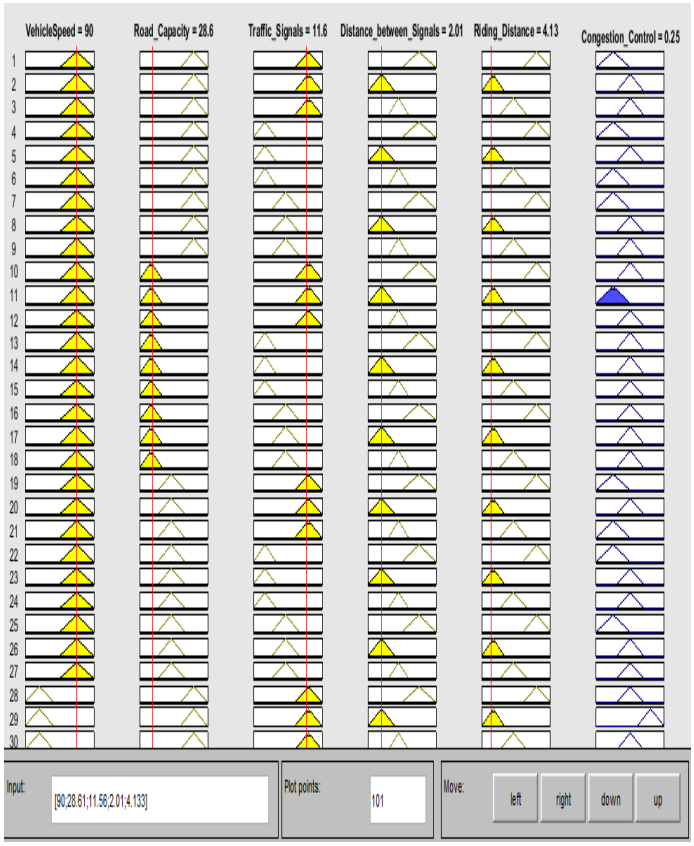


Fig. 13. Lookup diagram for low TCC.

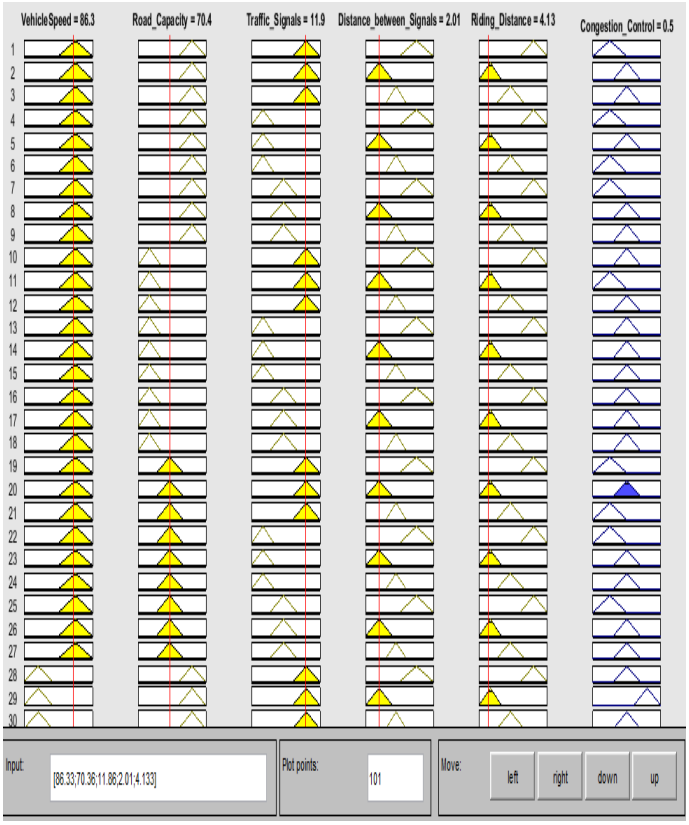


Fig. 14. Lookup diagram for medium TCC.

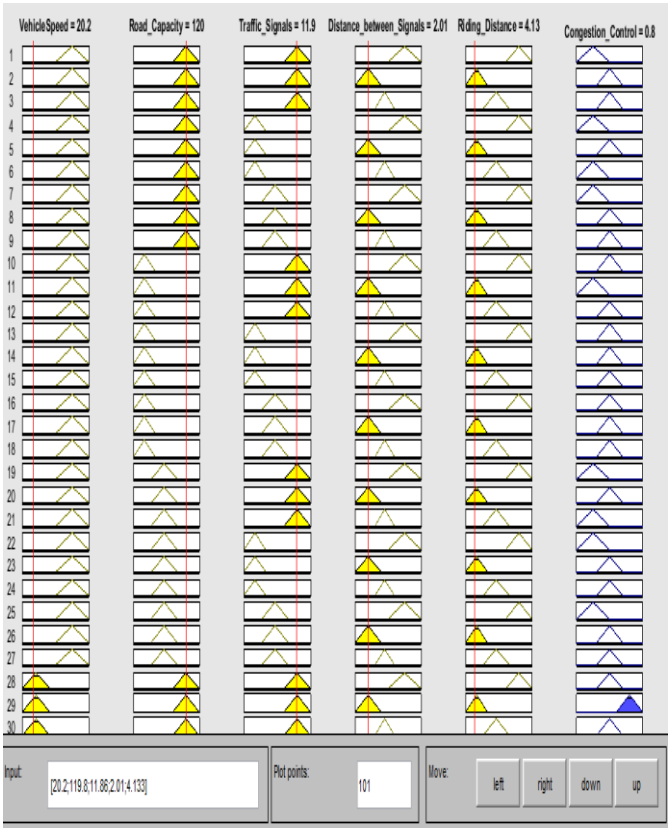


Fig. 15. Lookup diagram for high TCC.

VII. Conclusion and Future Work

Implementation of smart-cities infrastructure and services is a long-game process [25]. The advantages of smart-city to communities will not likely be quick and will probably be incremental in the first step. Nevertheless, to accomplish smart-cities infrastructure through the utilization of SNA, Information Communication Technologies and IoT's to scale their city framework and extend services reasonably while offering substantial financial advantages. This TCC fuzzy expert system is designed with the help of 5 input and one output variable. Mamdani Fuzzy Inference System (MFIS) is used to evaluate the Traffic Congestion Conditions (TCC) in smart-city. The proposed system design of TCC has been beneficial to determine the city traffic congestion. Through this system, anybody can check any traffic congestion. In future, MFIS would be used to evaluate the performance of the other resources of smart-city like Environmental Conditions, Energy Consumption, Healthcare and Security Surveillance, etc.

References

1. Han, Q., Liang, S., & Zhang, H. (2015). Mobile cloud sensing, big data, and 5g networks make an intelligent and smart world. IEEE Network, 29(2), 40-45.
2. https://news.un.org/en/story/2014/07/472752-more-half-worlds-population-now-living-urban-areas-un-survey-finds.
3. Carmona, Matthew. Public places, urban spaces: the dimensions of urban design. Routledge, 2010.
4. Kitchin, Rob. "Big Data, new epistemologies and paradigm shifts." Big Data & Society 1.1 (2014): 2053951714528481.
5. Misuraca, Gianluca, Francesco Mureddu, and David Osimo. "Policy-making 2.0: Unleashing the power of big data for public governance." Open government. Springer, New York, NY, 2014. 171-188.
6. Albino, Vito, Umberto Berardi, and Rosa Maria Dangelico. "Smart cities: Definitions, dimensions, performance, and initiatives." Journal of Urban Technology 22.1 (2015): 3-21.
7. Gabri Malek, Chunlin LI, Z. Yang, NajiHasan.A.H and X.Zhang,' Improved the Energy of Ad hoc On-Demand Distance Vector Routing Protocol', International Conference on Future Computer Supported Education, Published by Elsevier, IERI, pp. 355-361,2012.
8. Lakshmi, I. "A literature survey on Big Data Analytics in Service Industry." International Journal of Engineering and Computer Science 5.4 (2016). Breslin, J., Decker, S., 2007.
9. Dimitrakopoulos, George, and George Bravos. Current Technologies in Vehicular Communication. Springer, 2016.
10. Psomakelis12, Evangelos, et al. "BIG IOT AND SOCIAL NETWORKING DATA FOR SMART CITIES."
11. Nasr, Elie, Elie Kfoury, and David Khoury. "An IoT approach to vehicle accident detection, reporting, and navigation." In Multidisciplinary Conference on Engineering Technology (IMCET), IEEE International, pp. 231-236. IEEE, 2016
12. Rizwan, Patan, K. Suresh, and M. Rajasekhara Babu. "Real-time smart traffic management system for smart cities by using the Internet of Things and big data." In Emerging Technological Trends (ICETT), International Conference on, pp. 1-7. IEEE, 2016.
13. Al-Dweik, Arafat, Radu Muresan, Matthew Mayhew, and Mark Lieberman. "IoT-based multifunctional Scalable real-time Enhanced Road Side Unit for Intelligent Transportation Systems." In Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference on, pp. 1-6. IEEE, 2017.
14. Goggin, Gerard. "Driving the internet: mobile internets, cars, and the social." Future Internet 4, no. 1 (2012): 306-321.
15. Joshi, Yashashree, Ashwini Joshi, Neha Tayade, Priyanka Shinde, and S. M. Rokade. "IoT Based Smart Traffic Density Alarming Indicator."

(2016).Handte, Marcus, Stefan Foell, Stephan Wagner, Gerd Kortuem, and Pedro José Marrón. "An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders." IEEE Internet of things journal 3, no. 5 (2016): 735-744

16. Theodoridis, Evangelos, Georgios Mylonas, and Ioannis Chatzigiannakis. "Developing an IoT smart city framework." In Information, intelligence, systems and applications (Lisa), 2013 fourth international conference on, pp. 1-6. IEEE, 2013.
17. Chong, Hon Fong, and Danny Wee Kiat Ng. "Development of IoT device for traffic management system." In Research and Development (SCOReD), 2016 IEEE Student Conference on, pp. 1-6. IEEE, 2016
18. Chong, Hon Fong, and Danny Wee Kiat Ng. "Development of IoT device for traffic management system." In Research and Development (SCOReD), 2016 IEEE Student Conference on, pp. 1-6. IEEE, 2016
19. Thakur, Tanvi Tushar, Ameya Naik, Sheetal Vasari, and Manjiri Gogate. "Real-time traffic management using the Internet of Things." In Communication and Signal Processing (ICCSP), 2016 International Conference on, pp. 1950-1953. IEEE, 2016
20. Ramachandra, Sujit H., K. Nitesh Reddy, Vivek R. Vellore, Sumanth Karanth, and Tareesh Kamath. "A novel dynamic traffic management system using on-board diagnostics and Zigbee protocol."
21. In Communication and Electronics Systems (ICCES), International Conference on, pp. 1-6. IEEE, 2016
22. Chowdhury, Abdullahi. "Priority-based and secured traffic management system for an emergency vehicle using IoT." In Engineering & MIS (ICE), International Conference on, pp. 1-6. IEEE, 2016
23. Ou, Haoyuan, Jianming Zhang, and Yi Wang. "Development of intelligent traffic control system based on internet of things and fpga technology in proteus." traffic 20 (2016): 2.
24. Amadeo, Marica, Claudia Campolo, Jose Quevedo, Daniel Corujo, Antonella Molinaro, Antonio Iera, Rui L. Aguiar, and Athanasios V. Vasilakos. "Information-centric networking for the internet of things: challenges and opportunities." IEEE Network 30, no. 2 (2016): 92-100.
25. Coutard, Olivier, et al. "Urban Megatrends: Towards a European research agenda." (2014): 1-17.
26. Saeed, Yousaf, et al. "Impact of Cognition on User Authentication Scheme in Vehicle using Fuzzy Logic and Artificial Neural Network." International Journal of Computer Science and Information Security 14.10 (2016): 285.
27. Abbas, Sagheer, et al. "Bio-inspired neuro-fuzzy based dynamic route selection to avoid traffic congestion." International Journal of Scientific and Engineering Research2.6 (2011): 284-289.

Customer Satisfaction Measurement using Sentiment Analysis

Shaha Al-Otaibi, Allulo Alnassar, Asma Alshahrani, Amany Al-Mubarak,
Sara Albugami, Nada Almutiri, Aisha Albugami
Information Systems Department
College of Computer and Information Sciences
Princess Nourah Bint Abdulrahman University
Riyadh, Saudi Arabia

***Abstract*—Besides the traditional methods of targeting customers, social media presents its own set of opportunities. While companies look for a simple way with a large number of responses, social media platforms like Twitter can allow them to do just that. For example, by creating a hashtag and prompting followers to tweet their answers to some question they can quickly get a large number of answers about a question while simultaneously engaging their customers. Additionally, consumers share their opinions about services and products in public and with their social circles. This valuable data can be used to support business decisions. However, it is huge amounts of unstructured data that is difficult to extract meaningful information out of them. Social Media Analytics is the field which makes insights out of social media data and analyzes its sentiment rather than just reading and counting text. In this article, we used Twitter data to get insight from public opinion hidden in data. The support vector machine algorithm is used to classify sentiment of tweets whether it is positive or negative and the unigram applied as a feature extraction method. The experiments were conducted using large set of training dataset and the algorithm achieved high accuracy around 87%.**

Keywords*—Social media analytics; sentiment; classification; support vector machine; unigram

I. INTRODUCTION

In the few recent years, the social media platforms have been growing while people build a global communication network on the Internet via many social media applications. Daily a huge volume of media is created on the social networks. For example, in Twitter - one of the most popular social media application - there are over 500 million tweets or posts per day[1]. It is a revolution of how media is created and distributed by sharing, and realizing messages without any control. Social media has an important impact on the field of business, advertisement, and e-commerce as it explains consumer behavior and feedback about particular business proposals, services and products. Opinions and purchase decisions of the people and organizations are now affected and sometimes taken as a response to the content of social media before going to the market and actually test the product. In social media, all data from posts, comments and replies needs measuring results and concluding insights out of them rather than just reading the opinions of others, this is known as social media analytics. Social media analytics are the practice of gathering data from social media platforms and analyzing that data to make business decisions. The most common use of social media analytics is to mine customer sentiment in order to support marketing and customer service activities. The importance of social media analytics is intuitive and flexibly used by companies, organizations and individuals to know the insight of the market. It helps companies to know customers' viewpoints and their comments on the quality of the products and services to make successful business decisions. The typical objectives include increasing revenues, reducing customer service costs, getting feedback on products and services, as well as improving public opinion of a particular product or business division [1], [2].

To clarify the concept of social media analytics, we should present the problem from two viewpoints: the business problem and the technical issues. As business problem, the pre-sale means knowing the activity of the competitors in the market. Hence, companies need to know the right time to release their products or services in the market. Additionally, they need to check the state of the market if there is a product similar to its product or service that will be launched in the market and compare with each other, as well as determine what is the positive and negative about those products or services and try to improve it. Then, they will able to add a competitive advantage in their product or service. After-sale, the companies need to check the social media feedback and customers' opinions about the product or service, they want to know how many of the followers, interacting, re-tweet, fans and replies about company's account and products. Finally, it helps companies to understand the experiences of others with the product. Second, the technical issues related to the difficulty in extracting information and data about a particular product and deciding whether it is negative or positive of the marketing products. Moreover, the social media analytics require accessing the Internet and needs a large space to store the collected data for processing. They also need to filter and clean massive data, wide, variety, noise and unrelated data sources in social media content as well as, the extraction of keywords and show off all Hashtag that works mention to the product account and others [1].

In this article, we aim to support the organizations and individuals in decision-making through providing analysis of products information, customer's opinions, and the reviews of

[1] https://blog.twitter.com/

products in the social media. Indeed, the issues arises the need of providing such analytics include, knowing about the competitors from other companies as well as the need to solve the lack of means and tools to evaluate the products on the market. The proposed system will help companies and organizations to get benefit information about their products and services. It will save beneficiary's time and serve them by learning more about their products and stimulate the work of producing a lot of quantity of products through knowing the viewpoints of their customers as well as, the information about competitors' products. As initial step of this work, we will cover the textual data about products and services in Twitter and apply the characteristics of the intended users of the system, such as age, gender, education, number of followers, etc.

The article is organized as follows: Section II presents the background information and related work. Section III demonstrates the design issues and the implementation details related to analyzing social media contents. Then, the experimental evaluation standard is presented in Section IV. Finally, we conclude this work in Section V.

II. Background Information and Related Work

This section starts by presenting the background information. Then, we review the literature that related to the social media analytics.

A. Background Information

Data is the currency of social media marketing and the understanding of social media analytics is essential for making data useful. Hence, the analytics allow marketers to identify sentiments and trends in order to better meet their customer's needs [2]. Facebook, Twitter, Pinterest and other social networks continue to spread a torrent of data, and organizations need to measure the business value. Now if customer wants to buy a product, he/she is no longer limited to asking his/her friends and families because there are many product reviews on the Internet which give opinions of existing users of the product. For a company, it may no longer be necessary to conduct surveys, organize focus groups or employ external consultants in order to find consumer opinions about its products and those of its competitors because the user-generated content on the Web can already give them such information [3]. Businesses often struggle to measure consumer interest and to determine what social data is actually useful for them to collect. By utilizing sentiment analytics complemented with human intelligence, companies can filter out noise and—with the help of machine-learning technology—identify the critical data that advances their business.

This section presents the social media analytics framework, techniques, types of audience, social media network choices, and features of media analytics tools.

1) Social media analytics Framework: The typical framework involves three-stage process: capture, understand, and present [4]. During the work of Chong et al., they develop CUP framework that add the identify stage to allow the identification of posts/tweets prior to the capture stage [5]. This identification is done using keywords which are determined by users. These keywords are then used in the automated scripts query requests to social network's API, e.g. Twitter API, collects posts/tweets containing those keywords. Therefore, the steps include: the identify stage is the data accessing stage that involves identifying relevant keywords to use in collecting social media data. Then, the capture stage is the data cleaning step that involves obtaining relevant social media data by listening to various social media sources, archiving relevant data and extracting pertinent information, hence not all data captured will be useful. Next, the understand which is the data analysis stage that selects relevant data for modeling, removing noisy, low quality data, and employing various advanced data analytic methods to analyze the data retained and gain insights from it. Finally, the present is the data visualization stage that deals with displaying findings from the understand stage in a meaningful way [6].

2) Social Media Analytics Techniques: Many techniques can be used for social media analytics. First, the Supervised Classification, where the classification is the separation or ordering of objects into classes. Text classification is automatically assign the texts into the predefined categories. In this machine learning technique, the classifier learns how to classify the categories of documents based on the features extracted from the set of training data. The supervised classification includes: Support Vector Machine (SVM), Naïve Bayes, Neural Network, K-nearest Neighbor, and Decision tree [7], [8]. A detailed review of the above classifiers along with their advantages and disadvantages are explained in [8] and [9]. Typical text classification process has the following steps: collect data, normalize data, analyze the input data, train the algorithm, test the algorithm, and apply on the target data [9]. Second, Unsupervised Text Mining/Clustering: Text clustering is unsupervised learning, where no label or target value is given for the data. It is a method of gathering items or (documents) based on some similar characteristics among them. It performs categorization of data items exclusively based on similarity among them. Most clustering algorithms need to know the number of categories in advance. Some researchers use clustering instead of classification in topic detection because it hard to find data set for new topics [10].

3) Types of Audience: Twitter subscribers are older in age and count way more than Facebook's [10] so they are likely generated more trustworthy opinions. Also, people share their opinions publicly on Twitter unlike Facebook where social interactions are often private [11]. For these reasons, we selected Twitter as data source. However, follow/friend action in Twitter is not mutual like in Facebook so social circle of a user is not clear.

4) Features of Media Analytics Tools: Most of media analytics tools accomplish goals like:

- *Competitive benchmarking*: The ability to view profile and content information for other accounts like competitors.
- *Centralized analytics*: A single place to see and compare statistics and metrics for all (or most) of your social media accounts.
- *Influencer identification*: A list of the accounts or people that engage (share, comment, etc.) with your content most frequently.
- *Tracking of common social activities*: Tracking of customer service related interactions, or other common social network activities.
- *Dashboards*: Pre-made or custom dashboards so that you can easily keep tabs on the accounts, competitors, and metrics that matter the most to you.
- *Reporting*: Exportable reports and data often coupled with scheduling and email delivery [12].

5) Review of Data Analytics Systems: The social media analytics systems can be divided into two types: Platform tools and Cross-Platform tools. Platform tools are provided by the official social media networks such as Twitter, Facebook, YouTube, ... etc. while Cross-Platform are commercial tools allow the user to analyze different social networks types [10].

First, we will list the data analytics platform tools which are provided by the official social media networks:

- Twitter Analytics has a built-in analytics platform. It is available to individuals as well as businesses. Number of tweets, tweet impressions, visits to profile, mentions and followers are all tracked. There are monthly statistics on most popular tweets, mentions, and followers for that month. User can click on any Tweet to see the impressions, likes, retweet, and engagements, but no sentiment analysis is provided.
- Facebook Insights are available to any of the admins of company page once user has over 30 fans. It displays detailed metrics about posts and the engagement they earn. Audience analysis can help to understand who is engaging more, and includes demographic and locations breakdown. Engagement metrics can be seen for each of posts, helping user to understand what type of content works best. There are also metrics on video views, actions taken on user page and the reach of your posts. Also here, no sentiment analysis is provided.
- YouTube Analytics provides an in-house analytics tool so anyone who has uploaded videos can understand their performance. The tool displays performance metrics, engagement metrics, and demographics. It helps user understands how people found videos, how much they watched, if they clicked through to user's website, and who they were.
- Google Alerts/Analytics While not strictly a social media analytics tool. It allows user to monitor the web for new content, mentions of brand, competitors, or industry thought leaders. Creating an Alert means you will receive email notifications when Google finds new results on the topic across blogs, forums and news sites. Google analytics is primarily a web analytics tool, but it provides a small but important role in social media analysis: a breakdown of which social sites are driving traffic to user's website.

Second, the Cross-platforms which are commercial tools allow the user to analyze different social networks types.

- SimplyMeasured is a paid tool, but has various free individual reports for Instagram, Twitter, Facebook, Vine and Google+. A follow of the SimplyMeasured Twitter account is enough to secure user's report. They provide insights such as Facebook content, competitor and fan page analysis, key Twitter analytics, Instagram engagement, content and trends analysis and many others.
- Quintly covers Facebook, Twitter, Google+, LinkedIn, Instagram and YouTube, and it has a free tool for Facebook analytics. Quintly comes with a standard dashboard that can be customized with widgets to suit user's needs and track the metrics that matter to user.
- Brandwatch crawls millions of sites and allows user to build flexible and accurate searches through advanced Boolean queries. Brandwatch categories, rules, and tags allow users to slice and dice the data any way they want.

B. Related Work

There are many surveys for data analytics and related topics, some of them will be presented in this paragraph. Bo and Lillian presented a survey that covered the techniques and approaches for opinion mining and sentiment analysis to promise enabling opinion-oriented information seeking systems. It provides a discussion of available resources, benchmark datasets, and evaluation campaigns were provided [13]. Isaac presented a survey of different social network analysis techniques employed in many applications interpreting social media data, e.g. Twitter. It focuses on two main approaches to sentiment analysis: supervised learning and unsupervised learning techniques used for natural language processing, classification and prediction. Major statistical packages such as SAS and SPSS include dedicated sentiment analysis modules used in [10]. Additionally, a review of text classification on social media data is to discuss the different types of classifiers and their advantages and disadvantages [8]. Moreover, a comparison of the most popular packages, e.g. R, Matlab, SciPy, Excel, SAS, SPSS, and Stata that are typically used for data analysis was presented in [14]. A book was published of mining data from the social web such as Facebook, Twitter, LinkedIn, Google+, GitHub and More. This book provided an explanation on how to acquire, analyze, and summarize data from social media networks, email, websites, and blogs by employing the Natural Language Toolkit, NetworkX, and other scientific computing tools to mine popular social web [15]. Moreover, Twitter data analytics book presented an understanding of the basics of collecting, storing, and analyzing Twitter data. The

first half of this book discusses collection and storage of data. The second half is focused on analysis. It provided the common measures and algorithms that are used to analyze social media data [16]. Finally, the text mining and analysis book covered the practical methods, examples, and case studies using SAS. It delivered a comprehensive theoretical reference for text mining as well as many practical examples and case studies using the Statistical Analysis System (SAS) [17].

There are many datasets used for data analytics provided in the literature, for example: The datasets of customer reviews, pros and cons as well as comparative opinions [18]. The MPQA opinion corpus provided opinion datasets, e.g. Subjectivity Lexicon [19]. Additionally, Twitter sentiment analysis training data contains corpus of already classified tweets in terms of sentiment analysis training and testing where it contains more than 1,500,000 classified tweets, each row is marked as "1" for positive sentiment and "0" for negative sentiment [20]. Moreover, Sanders-Twitter sentiment corpus designed for training and testing Twitter sentiment analysis algorithms. It consists of 5513 hand-classified tweets. These tweets were classified with respect to one of four different topics [21].

In this paragraph, we will present some of researches for data analytics tools. First, the sentiment analysis and text mining for social media microblogs using open source tools. It presents an empirical study that used R package to perform text mining and sentiment analysis for Twitter online reviews about two retail stores in UK [6]. Second is the experiment on binary classification for Twitter sentiment analysis. This experiment demonstrates how to use Microsoft Azure Machine Learning Studio to train a text sentiment classification engine using the Two-Class SVM [22]. Third, the emotion classification of social media posts for estimating people's reactions to communicate alert messages during crises. This article describes a methodology for analyzing tweets about Sandy hurricane and annotating them with four emotional labels. Two classification algorithms were experimented: Naïve Bayes and SVM classifiers. The results show that the algorithm achieves the best results with about 60% accuracy [23]. Fourth, the localized Twitter opinion mining using sentiment analysis analyzes tweets about iPhone 6 using SentiWordNet, part of SNLP which is an open source natural language processing tool developed by Stanford University [24]. Finally, the data mining and analysis on Twitter++ study start with a few discussions of how geo-tagged tweets in Twitter can be used to identify useful user features and behaviors as well as identify places of interests. Then, it presents a clustering analysis and proposes different similarity measures to detect communities [25].

Many tutorials describe how to analysis Twitter data, for example, step-by-step practical tutorials build Twitter analytics tool with R package included in [26]-[28]. Additionally, the tutorials designed to build Twitter mining tool with Python are included in [29] and [12]. Finally, the practical tutorials build Twitter mining tool with MATLAB are included in [11] and [30]. Table I illustrates a comparative analysis of some presented works.

TABLE I. A COMPARATIVE ANALYSIS OF SIMILAR WORKS

Ref.	Sentiment Analysis	Network Analysis	Customer Service	Application Data	Social Media Network	Analysis Method	Programming Package
[6]	Yes	No	No	Customer Review	Twitter (Facebook applicable)	Data Mining (association rules) (Lexicon-based)	R Package (Twitter)
[22]	Yes	No	No	Sentiment140 dataset	Twitter	Classification (SVM) 2 classes	Microsoft Azure Machine Learning Studio + R Package
[23]	Yes	No	No	Sandy hurricane	Twitter	Classification (SVM) 4 classes	Python Package (tweetstream)
[24]	Yes	No	No	Reviews on iPhone6	Twitter	Natural Language Processing	Stanford NLP (SentiWordNet)
[25]	Yes	Yes	No	geo-tagged	Twitter	Data Mining (Clustering)	MATLAB (Twitter 4j)
[33]	Yes	Yes	No	Slahdot Lexicon (MPQA)	Slashdot Website (comments on news)	Data Mining Predictive Analytic	KNIME

III. PROPOSED SOLUTION OF DATA ANALYTICS

Our proposed solution started by data collection which is an important aspect of any type of research study. Hence, the choice of data collection method is influenced by the data collection strategy, the type of variable, the accuracy required, the collection point and the skill of the source. The main data collection methods we used: first, the literature review and tools analysis. It supported us for collecting set of requirements regarding the analysis algorithm, analysis metrics as well as user interface design. Second, set of interviews were conducted with the respondent and notes are subsequently interpreted for further analysis. We conducted set of interviews with clients selling their home-made products, such as accessories and crafts, using different social media networks. Answers from respondents mainly raised the issue that searching within social media is very difficult for

them to target particular categories with people such as customers/competitors existing in particular country, are of particular age, females as well as customers which are influencers and having high number of followers. Targeting the right customers and monitoring the right competitors will bring them higher profits. Hence, we included one requirement about filtering input data against the criteria they mentioned by them. Third, we used questionnaires which are completed and returned by respondents. We have used GoogleForms to design a questionnaire that contains 16 questions of many types (yes/no, multiple-choice and open answer) and directed to different categories of people, i.e. students, tutors, business owners and consumers. The result of questioner let us focus on analyzing Twitter data since it will be more useful to target large number of people. We plan to satisfy the following SW/HW requirements:

- Libraries to communicate with Twitter API to authenticate added Twitter accounts and retrieve of Twitter data.
- Statistical and Machine learning development packages such as LibSVM, WEKA or R.
- Benchmark tweets database for customer reviews on an arbitrary product.
- Lexicon dictionary of sentiment words classified as positive and negative.
- Laptop machine with at least 8 GB of RAM and no less than Terabyte disk.
- Public server with high quality feature to upload the system and accommodate huge amounts of data.

Moreover, the nonfunctional requirements that should be satisfied are:

- Security/privacy by providing access permissions for system data, i.e. login/logout, valid emails and authorized Twitter accounts.
- Availability: The system available for service when requested by users.
- Usability: Simple UI to provide easy-to-learn end system.
- Reliability: The ability of a system to perform its required functions with accuracy no less than 80%.
- Visualization: The system should display metrics visually as well as numerically. Visual presentation includes keyword cloud, bar charts, pie charts, trend graphs and comparative graphs while numerical includes totals and percentages in addition to specific scores.
- Sentiment analyzed tweets are marked in different colors for negative and positive.

A. *System Design of Proposed Solution*

This section illustrates the design of the proposed solution and its architecture including the structure, the description about each structure component and the used system design tools. The system will be implemented in five-tier server-client architecture model consisting of presentation layer, business logic layer and data access layer for internal components. However, the additional integration layer and data Source layer are used to describe external components. Fig. 1 illustrates the main system's architecture and components.

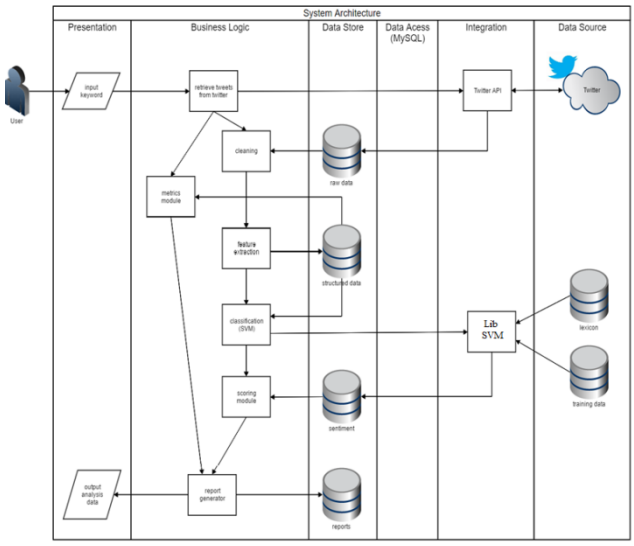


Fig. 1. System architecture.

Presentation Tier, this layer contains the user oriented functionality responsible for managing user interaction with the system, and generally consists of components that provide a common bridge into the core business logic encapsulated in the business layer [31]. In the proposed system, the presentation layer does two tasks: accepts user's input data such as keyword list and the type of analysis report and the other task is to later visualize the analysis results.

Business Logic Tier, it implements business functionality of the system. For example, it moves and processes data between the two surrounding layers. In our proposed system, the business logic layer consists of the following tasks:

1) **Tweets retrieval,** Twitter is the most exaggerated part of social networking site, it consists of various blogs which are related to various topics worldwide. Instead of taking whole tweets, we will rather search on particular keywords and store all tweets in the form of text files by using mining tool i.e. WEKA/R/LibSVM which provides sentiment classifiers.

2) **Cleaning and Pre-processing of extracted data,** after retrieval of tweets, sentiment analysis tool is applied on raw of tweets but in most of cases, it gives poor performance. Therefore, preprocessing techniques are necessary for obtaining better results. We extract tweets, i.e. short messages from Twitter which are used as raw data. This raw data needs to be preprocessed. So, preprocessing involves following steps:

- Exclude tweets with non-English languages.
- Remove emoticons and substitute with their textual meanings.
- Remove URL's, hashtag mark "#", mentions "@" and retweet prefix, i.e. "RT".
- Remove punctuation marks and articles such as "a", "an" and "the".
- Remove stop words such as "to", "of", "is", "are", "this" and "for".
- Normalize elongated words, e.g., happyyyyyy, by only one or two occurrences only.

3) **Feature extraction,** in feature extraction method, we extract the aspects from the processed dataset. Later this aspect is used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram and bigram. Additionally, the machine learning techniques require representing the key features of text or documents for processing. These key features are handled as feature vectors which are used for the classification task. The features extraction method that will be considered in this system is unigram.

4) **Sentiment Classification,** the selected **s**entiment classifier is the SVM as it scores higher than other approaches according to [6]. Training of classifier data is the main motive of this step. A reference model is derived based on the analysis of a set of training data. Training data consists of data objects whose class labels are known. The derived model can be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification process is done in a two-step process as illustrates in Fig. 2. First step is Training in which we will build a model from the training set. Second step is Prediction in which we will check the accuracy of the model and use it for classifying new data

5) **Sentiment Scoring Module,** we use the lexicon/dictionary that applied in [6] in which English language words assigns a score to every word, between 1 (Negative) to 3 (Positive). So, this scoring module is going to determine score of sentiments in the sentiment analysis of data. Based on the dictionary assignment of score, the system interprets whether the tweet is positive, negative or neutral.

6) **Computing metrics,** this component is irrelevant to sentiments, however, it computes meaningful measurements about tweets and Twitter users. The raw data that comes from Twitter API contains the following parameters of each tweet which later can be used to calculate metrics:

- **Likes:** list of people liked this tweet. It's usually positive in sentiment.
- **Followers**: list of people that are currently subscribed to this tweet.
- **Mentions**: list of @username included in this tweet.
- **Replies**: list of responses to this tweet that begins with tweet writer's @username.
- **Retweet (RT):** list of users who shared this tweet.

Data Access Tier, the data Access Tier communicates with the database. In the proposed system, we are going to use MySQL DBMS to manage data storage, querying and retrieval.

Integration Tier, this tier is responsible for communicating with external resources and systems such as data stores, API's and legacy applications. The business tier is coupled with the integration tier whenever the business objects require data or services that reside in the resource tier. The components in this tier can use some proprietary middleware to work with the resource tier [4]. In the proposed system, this layer contains components interacting with Twitter API in order to access Twitter data in addition to open source library such as WEKA or R in order to use their functions and classes implementing machine learning algorithms, e.g. SVM.

Data Source Tier, this tier contains the business data and external resources such as Twitter network, training data source and lexicon benchmark.

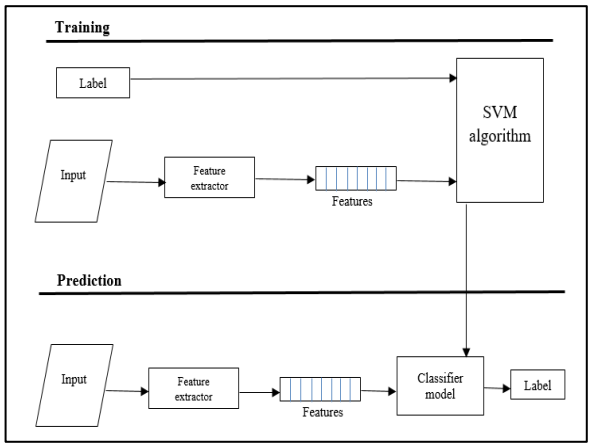


Fig. 2. Training SVM vs. Prediction.

B. System Implementation

In this phase, we take the determined system specifications and code them. The implementation requirements include the following software and hardware specification.

1) Hardware Requirements

- Laptop with processor Intel core i5, minimum speed of 1.7 GHz and 8.00 GB of RAM for faster running and better performance.
- Internet connection to access Twitter network and bring Twitter data online.

2) Software Requirements

- Window 10, 64-bit operating system or similar alternatives.
- As for user interface design, we used a bootstrap HTML5 free template called SIMINTA as a ground for our design.

- The website of proposed system was implemented using HTML5/CSS3 for page design, PHP for server-side scripts. Therefore, an Apache server distribution, such as XAMPP, was needed to execute PHP scripts.
- XAMPP also includes MySQL server which we used to store users' accounts and analysis reports' data. The database was managed using PHPMyAdmin module in XAMPP.
- Registered our application on Twitter Application Management to get Twitter access tokens and authorization. For the implementation of Twitter API interface, the Twitter -API-PHP library applied while it was recommended by Twitter developers' page.
- In addition, we used an executable software to run SVM classification algorithm from LIBSVM, which is a library developed for Support Vector Machines.
- For preprocessing, algorithm training and testing, we utilize different training datasets and some lingual dictionaries including stop word list, acronyms dictionary, positive and negative tweets provided by [16].
- To draw charts, we use classes from PHPLOT free php library.

3) System Major Services

The proposed system provides the following major services to its users:

- Account Analysis: user can search for specific Twitter account and analyze its author's activity rate in addition to followers' engagement with this account for the last ten days. For example, user can monitor his own product's account or a competitor's public account.
- Keyword Analysis: user can search Twitter social media network for any keyword, hashtag or mention of interest to check the public opinion and other valuable indicators about it. Keyword Analysis is on three types:

a) Sentiment Analysis: The percentage of latest positive vs. negative tweets talked about this search term. Search term can be a company name or a product brand for example.

b) Compute Metrics: The strength and reach of this search term in the public. Top hashtags and top keywords accompanied with this search term as well as and top Twitter users who are most interested about this search term.

c) Comparative Analysis: Providing sentiment analysis and metrics for two opposed search terms.

- Reports: Results of keyword analysis including individual analysis and comparative analysis can be stored in the database and retrieved back as needed. It can also be printed out or saved as PDF.

4) Implementation Details

In this section, we will give a brief description of how the proposed system was actually implemented using the specified software and hardware requirements. Firstly, we mention that our proposed website named TweetAdvisor.

- **System Website Registration in Twitter Apps:** Twitter, as many other social networks, have its own web services API (Application Programming Interface) that applications, such as our website, can work with. However, in order to use Twitter web services API, the first step is to register our website on Twitter's Application Management. After that, it will be provided the necessary access and authentication tokens to access Twitter data and services.
- **Twitter REST API:** After we registered the system's website to Twitter apps world, we need to access and call the appropriate Twitter 's web services to handle the website functions. The REST APIs provide programmatic access to read and write Twitter data. Read user profile, timeline or search Twitter data, and more. The REST API identifies Twitter applications and users using OAuth; responses are in JSON format. Basically, we used the following three GET web services from Twitter:

*a).*GET search/tweets: returns a collection of Tweets matching a query.

*b).*GET users/show: returns profile information about user specified by the user_id or screen_name parameter in the query.

*c).*GET statuses/user_timeline: returns a collection of the most recent Tweets posted by the user indicated by the screen_name or user_id parameters in the query.

- **Sign Up/Sign In:** The importance of creating an account is that private analysis reports conducted by the user can be saved and retrieved. User can create an account that providing basic personal information such as username, email and password.
- **Keyword *Analysis: Tweets Fetch*:** User can search Twitter social media network for any keyword, hashtag or mention of interest to check the public opinion and other valuable indicators about it as illustrates in Fig. 3. Keyword Analysis is on three types Sentiment Analysis, Compute Metrics and Comparative Analysis. Additionally, the proposed system allows the user to determine the following parameters:

a) **Exclude**: it returns all the tweets that don't contain specified words/phrases.

b) **From**: it retunes all the tweets coming from the specified user's screen name.

c) **min_followers_count**: it returns all the tweets only written by users who have a minimum number of followers, i.e. target influencers or famous users.

After specifying the required parameters, a query will be sent to Twitter API in order to retrieve tweets result. The request to Twitter web service is accepted only if the authentication via access tokens passed. Access tokens are given after successful app registration as explained in previous section. The search query will return the result data in JSON

tree format which is converted into an array object and then saved in PHP session for the next step; the preprocessing. Before preprocessing, the raw result array is filtered to only include tweets which are more than 20 characters in length and exclude retweets and redundant tweets as appears in Fig. 4. For connection with Twitter API we use a PHP Twitter -API-PHP library recommended by Twitter developers' page.

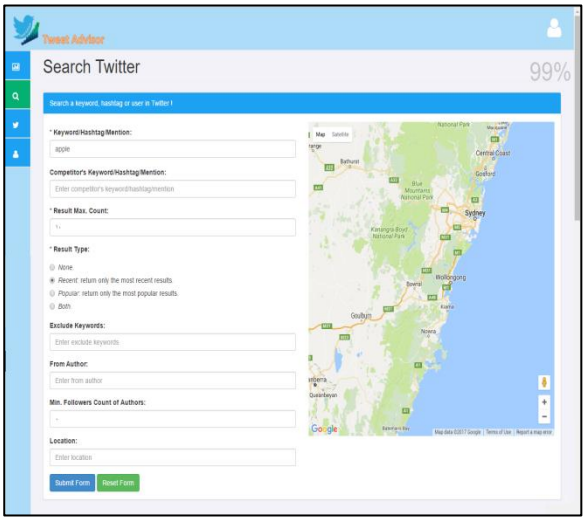


Fig. 3. Search criteria in Twitter.

- **Keyword Analysis: Preprocessing:** Preprocessing is the step needed to clean the data from noise, standardize and convert it to a structured format before extracting distinct features from it. In this work, we used four external resources in order to preprocess the data and provide prior score for some of the commonly used words:

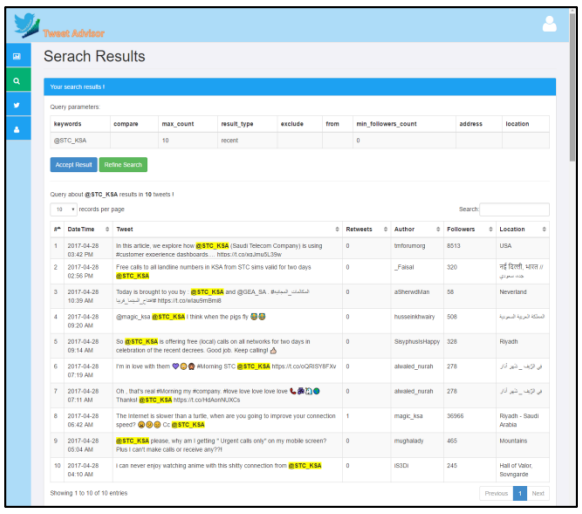


Fig. 4. Search results screen.

a) **Emoticon Dictionary**: Each emoticon is annotated into one of four classes: a) Extremely-Positive; b) Positive; c) Extremely-Negative; d) Negative as given in [32].

b) **Acronym Dictionary**: We used the acronym expansion list as given in [32].

c) **SentiWord List**: is a list of English words classified by its POS (short for position in sentence) and rated for valence with an integer between minus eight (negative) and plus eight (positive). POS types are noun (N), verb (V), adjective (A) and adverb (R). We used the SentiWord list as given in [32].

d) **Stop Words**: is a list such as a, is, the, with, and, or, I, you, etc. which occur in high frequency in a sentence but don't carry any sentiment information and thus are of no use to us. We used the stop words list as given in [32].

After building these lingual and sentiment dictionaries, preprocessing of tweets starts. We following preprocessing steps: remove extra whitespaces, replace each acronym with its expansion, tokenize each tweet, i.e. split into an array of separate words. then for each word in a tweet do the following, remove 'RT' prefix, lowercase , replace url, with ||U||, replace hashtag sign, i.e. #, with a ||H||, replace exclamation mark with ||EXC||, user mention replaced it with ||T||, remove punctuations, remove all digits, if emoticon, replace it with its equivalent sentiment into one of ||N|| for 'Negative', ||XN|| for 'Extremely-Negative', ||P|| for 'Positive'; or ||XP|| for 'Extremely-Positive', replace each "n't", "no", "not", "never", "cannot" with ||NOT|, if a stop word, remove it, replace tag word with its position of sentence + ||POS|| + word. An example of a preprocessed tweet is given in Table II.

After preprocessing, tweets are stored in array session so feature extraction phase starts.

Keyword Analysis: Unigram Feature Extraction: Feature vector is the most important concept in implementing a classifier. A good feature vector directly determines how successful the classifier will be. The feature vector is used to build a model which the classifier learns from the training data and further can be used to classify previously unseen data. In tweets training data, consisting of positive and negative tweets, we can split each tweet into words and add each word to the feature bag. Adding individual (single) words to the feature bag is referred to as 'unigrams' approach, see Table III.

So in unigram features, each feature is a single word found in a tweet. If the feature is present, the value is 1, but if the feature is absent, then the value of this feature is just not included. The entire feature vector of each tweet will be a combination of each of these feature words and based on this pattern, a tweet is labeled as positive or negative. See, Tables IV and V.

TABLE II. AN EXAMPLE OF A PREPROCESSED TWEET

Tweet	#Wearables | Apple Music reportedly strikes deal with popular social media app Musical.ly :D :D https://t.co/mobfOW1YIS… https://t.co/9bJGxs986r
Pre-processed	||H||wearables N||POS||apple N||POS||music R||POS||reportedly V||POS||strikes V||POS||deal A||POS||popular A||POS||social N||POS||media N||POS||application ||P|| ||P|| A||POS||musical N||POS||love ||U|| ||U||

TABLE III. UNIGRAM APPROACH – BAG OF WORDS

Tweet (Positive)	#Wearables | Apple Music reportedly strikes deal with popular social media app Musical.ly: D: D https://t.co/mobfOW1YIS… https://t.co/9bJGxs986r
Pre-processed	||H||wearables N||POS||apple N||POS||music R||POS||reportedly V||POS||strikes V||POS||deal A||POS||popular A||POS||social N||POS||media N||POS||application ||P|| ||P|| A||POS||musical N||POS||love ||U|| ||U||

Unigram Feature Bag	{
	wearables, apple, music, reportedly, strikes, popular, social, media, application, love
	}

Tweet (Negative)	My brother lost his phone in his room and my mom calling me trynna get me to do the find my phone shit. □
Pre-processed	||N||POS||brother V||POS||lost N||POS||phone N||POS||room N||POS||mom V||POS||callingV||POS||get V||POS||do V||POS||find N||POS||phone N||POS||shit ||N||

Unigram Feature Bag	{
	wearables, apple, music, reportedly, strikes , popular, social, media, application, love, brother, lost, phone, room, mom, calling, get, do, find, shit
	}

Some of the other feature vectors also add 'bi-grams' in combination with 'unigrams'. For example, 'not good' (bigram) completely changes the sentiment compared to adding 'not' and 'good' individually. Here, for simplicity, we will only consider the unigrams.

TABLE IV. UNIGRAM FEATURE VECTORS

Tweet 1 (Positive)	||H||wearables N||POS||apple N||POS||music R||POS||reportedly V||POS||strikes V||POS||deal A||POS||popular A||POS||social N||POS||media N||POS||application ||P|| ||P|| A||POS||musical N||POS||love ||U|| ||U||		
Tweet 2 (Negative)	||N||POS||brother V||POS||lost N||POS||phone N||POS||room N||POS||mom V||POS||calling V||POS||get V||POS||do V||POS||find N||POS||iphone N||POS||shit ||N||		

Unigram Feature Bag	{
	wearables,
	apple,
	music,
	reportedly,
	strikes ,
	popular,
	social,
	media,
	application,
	love,
	brother,
	lost,
	phone,
	room,
	mom,
	calling,
	get,
	do,
	find,
	shit
	}
		Tweet 1
		1:1,
		2:1,
		3:1,
		4:1,
		5:1,
		6:1,
		7:1,
		8:1,
		9:1,
		10:1
			Tweet 2
			
			11:1,
			12:1,
			13:1,
			14:1,
			15:1,
			16:1,
			17:1,
			18:1,
			19:1,
			20:1

TABLE V. FEATURE VECTORS FILE

```
+1 1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:1          // Tweet 1
-1 11:1, 12:1, 13:1, 14:1, 15:1, 16:1, 17:1, 18:1, 19:1, 20:1 // Tweet 2
```

- **Keyword Analysis: SVM-Classification:** The feature extraction method contains both training and testing. In training, the tweets vectors are labeled with '+1' for positive and '-1' for negative as in Table V. The classifier will use labeled vectors to learn from them and builds its learning model. In testing, however, each new un-labeled tweet will be compared to the bag of word generated from labeled tweets to create the new vector in the same way, however, with no labels given. The classifier will take the model and the new un-labeled vectors to predict the new classification results. The used classifier is based on SVM algorithm and provided by LibSVM library as two main executable applications: *svm-predict.exe* and *svm-train.exe* [6]. Fig. 5 illustrates the SVM classification result.

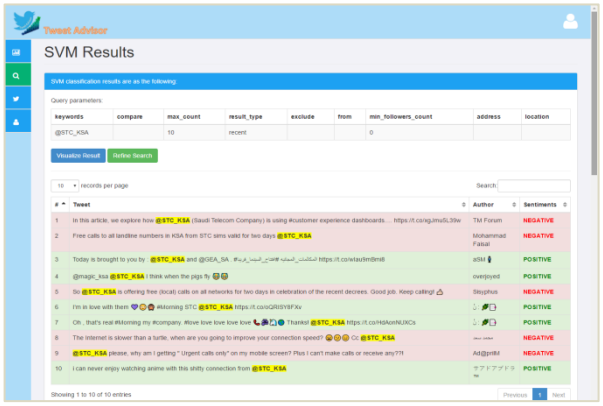


Fig. 5. SVM classification results screen.

- **Keyword Analysis: Compute Metrics:** After SVM classification completes, the following list of metrics will be calculated:

a) **Positive vs. Negative** counts pie chart: It indicates the percentage of passivity and negativity of public opinion about this search term.

b) **Strength:** It is the percentage of tweets count in last 24 hours on the total count of tweets in the result. It indicates how recent this search term is.

c) **Reach:** It is the percentage of different authors' count on the total count of tweets in the result. It indicates the percentage of authors interested in and talking about this search term.

d) **Top Keyword:** It is the list of the most frequented six keywords in the results. It indicated what other topics are related to this search term.

e) **Top Hashtags:** It is the list of the most frequented six hashtags in the results. It indicated what hashtags are related to this search term.

f) **Top Authors:** It is the list of the most Twitter accounts talked about the search term. If reach is 100% then each author has exactly one tweet in total result.

Fig. 6 visualizes a graphical result that illustrates the previous calculated metrics.

- **Keyword Analysis: Comparative Analysis:** We implemented this function by maintaining an array of

keywords entered by the user in the search query. Then, for each step, we run the code in a loop of the size of this array, and store each associated data results list in the corresponding array element. In this way, we will end up with multiple results each stored in its own array as appears in Fig. 7.

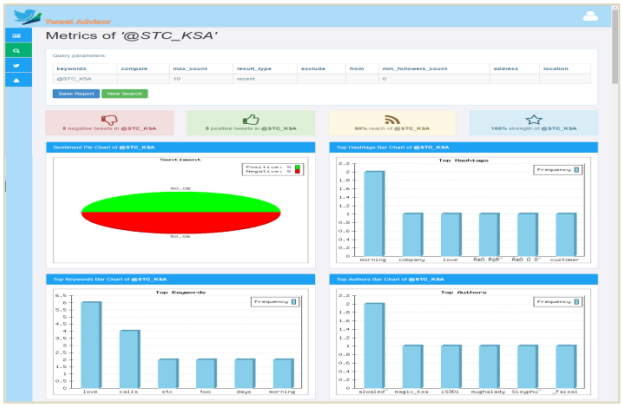


Fig. 6. Visualize metrics screen.

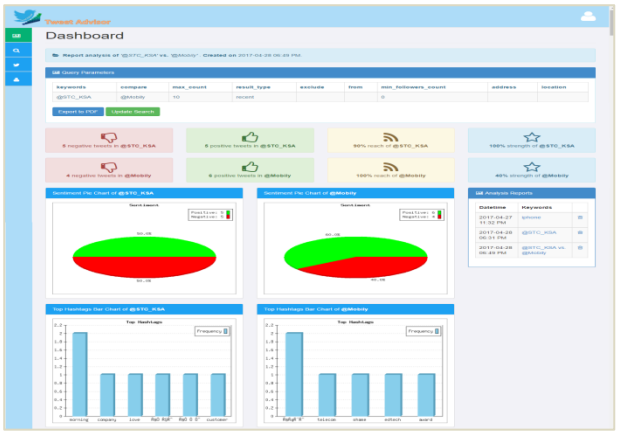


Fig. 7. Comparative keywords analysis.

- **Account Analysis:** User can search for specific Twitter account and analyze its author's activity rate in addition to followers' engagement with this account for the last ten days. For example, user can monitor his own product's account or a competitor's public account. Fig. 8 displays an account analysis for STC company. To implement this function, we need to send queries to *GET users/show* to retrieve user account's information such as *followers_count*. Also, we need to request *GET statuses/user_timeline* to extract tweets posted for last 10 days and get *favourite_count* and *retweet_count* for each tweet. Out of this data, we can calculate the following analysis metrics: Followers count so far, Tweets count last 10 days, Daily tweeting average, Daily interactions with followers, i.e. reply, average, Total likes by followers last 10 days, Average likes by followers per tweet, Total retweets by followers last 10 days, and average retweets by followers per tweet. Moreover, we plot the following graphs:

- **Followers Engagement Graph**: It represents the total likes and retweets by followers for each day during the interval of last ten days. User can notice at what day the engagement was higher or lower and analyze this reading.

- **Author's Activity Graph**: It represents the total tweets, retweets and interactions, i.e. replies, made by the owner of the account for each day during the interval of last ten days. User can notice at what day the activity was higher or lower and analyze this reading.

Fig. 8. Account analysis screen.

IV. Experimental Evaluation Standard

In order to evaluate the efficiency of TweetAdvisor, we conduct a testing process to evaluate the system and its components with the intent to find whether it satisfies the specified requirements or not. In this work, we have considered and performed the following testing types: we first test the used classification algorithm SVM which is a Machine learning method that usually have to deal with big and uncertain data, and the output of the system which is not like traditional system having a good sign of right or wrong.

Therefore, to test a machine learning algorithm accuracy, we need a training dataset, a testing dataset and an independent piece of code as a benchmark to run the algorithm and check the accuracy results. Sometimes, it is better to use different datasets and analyze the properties of the dataset that improved the accuracy of the algorithm.

A. Test Results

In this section, we will present the testing results of SVM classification process in detail as it is the core functionality of our website. We fetched 6 tweets about search term "Google Chrome" and result type = "Both: recent and popular", see Tables VI, VII, VIII and IX.

TABLE VI. RAW TWEETS RESULT

1	Save some valuable time with 20 of the best Google Chrome extensions for social media marketers: https://t.co/94L8Fs7IzV
2	Alert! Google Chrome Is Listening In To Your Room, Sending Data Without Consent - YouTube https://t.co/7QFCfoQYug April 29, 2017 at 06:00AM
3	Big UX step forward: starting in Chrome 59, web notifications will be shown directly as Mac OS X notifications. https://t.co/DyIHjoyjn1
4	Finally! Chrome will soon start using native notifications on macOS. https://t.co/pMCiu94jAj
5	13 Google Chrome Life Hacks that will Blow Your Mind! https://t.co/emjwu6VEuM
6	Google gets more aggressive in alerting users when web pages are not secure https://t.co/VaqjPYR5ID

TABLE VII. PREPROCESSED TWEETS RESULT

1	V||POS||save A||POS||valuable N||POS||time A||POS||best N||POS||google N||POS||chrome N||POS||extensions A||POS||social N||POS||media N||POS||marketers ||U||
2	V||POS||alert ||EXC|| N||POS||google N||POS||chrome V||POS||listening N||POS||room V||POS||sending N||POS||Data N||POS||consent youtube ||U||
3	A||POS||big N||POS||user N||POS||experience N||POS||step A||POS||forward A||POS||starting N||POS||chrome N||POS||web N||POS||notifications R||POS||directly N||POS||mac A||POS||operating N||POS||system N||POS||notifications ||U||
4	R||POS||finally ||EXC|| N||POS||chrome V||POS||start V||POS||using A||POS||native N||POS||notifications macOS ||U||
5	N||POS||google N||POS||chrome N||POS||life N||POS||hacks V||POS||blow N||POS||mind ||EXC|| ||U||
6	N||POS||google A||POS||aggressive N||POS||alerting N||POS||users N||POS||web ||NOT|| A||POS||secure ||U||

Finally, the testing results show that the accuracy given by testing the test cases reached 84%, i.e. one error out of six correct answers.

B. Training/Testing Dataset Collections

The second test was performed using the publicly available data sets of Twitter messages with sentiment analysis. We have used a combination of two datasets to train the SVM machine learning classifier. For the test dataset, we randomly choose 4000 tweets which were not used to train the classifier. The details of the training and test data are explained in Table X.

TABLE VIII. UNIGRAM FEATURE VECTORS RESULT

1	284:1 318:1 756:1 993:1 1004:1 1088:1 1381:1 1496:1 1554:1 2457:1 2766:1 3050:1 3903:1 4069:1 4636:1 5089:1 6553:1 7618:1 8136:1 8825:1 9017:1 9042:1 9090:1 9384:1 9564:1 9707:1 9729:1 10616:1 10783:1 10934:1 11134:1 11926:1 13071:1 14282:1 15093:1 15094:1 15691:1 17041:1 17567:1 17570:1 17629:1 17630:1 17631:1
2	121:1 404:1 711:1 1381:1 1496:1 2457:1 3659:1 3903:1 4274:1 4701:1 5610:1 5994:1 6021:1 6119:1 6744:1 7618:1 8136:1 8186:1 9017:1 9042:1 9115:1 9564:1 9729:1 10029:1 10783:1 10933:1 11363:1 11554:1 11704:1 12477:1 12929:1 13042:1 15093:1 15094:1 15691:1 15913:1 16384:1 16421:1 16489:1 16557:1 17041:1 17629:1 17630:1 17741:1
3	45:1 86:1 317:1 885:1 1381:1 1496:1 1717:1 1770:1 2242:1 2457:1 2854:1 3543:1 3616:1 3747:1 3903:1 4992:1 5014:1 5483:1 5727:1 5993:1 6017:1 6744:1 6846:1 7105:1 7618:1 7929:1 8100:1 8136:1 9042:1 9159:1 9405:1 9564:1 9729:1 10110:1 10354:1 11554:1 11576:1 11774:1 12339:1 15093:1 15094:1 15691:1 17147:1 17629:1 17630:1 17741:1 17758:1
4	1661:1 2457:1 3478:1 3659:1 3747:1 3823:1 5727:1 6017:1 6569:1 6744:1 6974:1 7403:1 7618:1 8136:1 8507:1 8825:1 9042:1 9405:1 9564:1 9729:1 9730:1 10934:1 11554:1 11774:1 13052:1 14017:1 15093:1 15691:1 17629:1 17630:1 17741:1 17758:1
5	98:1 107:1 630:1 1381:1 1496:1 2457:1 3903:1 5427:1 7167:1 7618:1 8136:1 9017:1 9042:1 9564:1 11554:1 11774:1 15093:1 15094:1 15691:1 17041:1 17629:1 17630:1 17758:1
6	1381:1 1496:1 2216:1 2263:1 2457:1 5993:1 8186:1 8825:1 9017:1 9042:1 9152:1 9273:1 9564:1 9729:1 10110:1 10403:1 10426:1 10934:1 11554:1 12339:1 15093:1 15094:1 17041:1 17567:1 17629:1 17741:1

TABLE IX. SVM CLASSIFICATION RESULT

#	SVM Classification	Meaning	Human Classification	Meaning
1	-1	negative	1	positive
2	-1	negative	-1	negative
3	1	positive	1	positive
4	1	positive	1	positive
5	1	positive	1	positive
6	-1	negative	-1	negative

Sanders corpus is designed for training and testing Twitter sentiment analysis algorithms. It consists of 5513 hand-classified tweets. These tweets were classified with respect to one of four different topics. Each entry contains: Tweet id, Tweet text, Tweet creation date, Topic used for sentiment, and Sentiment label i.e. 'positive', 'neutral', 'negative', or 'irrelevant'. We used only the positive and negative tweets out of this dataset for training. To fetch random testing tweets, we used our website interface which searches the Twitter API for a given keyword with recent results. Tweets were downloaded, manually labeled and then subjected to both preprocessing and feature extraction as specified in Section 3. These filtered tweets are fed into the trained classifiers and the resulting output is then saved in a file. The results file was read and compared with the correct classes of chosen tweets. The testing results show that the accuracy given by testing the 4000-tweets dataset reached 87%.

TABLE X. DATASETS USED FOR TRAINING AND TESTING

Dataset	Positive	Negative	Total
Training	9666 (Sanders)	9666 (Sanders)	19,332
Testing	2000 (random)	2000 (random)	4,000

V. Conclusion and Future Work

The social media becomes a reality in people's lives, enabling the growth of many online services. However, the companies maintain and assess the quality of their products or services by analyzing customers' satisfaction through social media platforms. The objective of this work is to propose a system that measures customer's satisfaction using sentiment analysis. Hence, the sentiment analysis is an important phase in the decision making process. We used the SVM as a classification algorithm beside the unigram as a feature extraction method and applied them to measure sentiment in Twitter data. The experimental result indicates that the unigram feature extraction method with SVM classification together bring high score reaches 87%. However, this percentage needs improvement either by using different dataset or different classification algorithm. As a future work, we can test other different classification algorithms and implement different feature extraction in addition to unigram. Moreover, we plan to specialize preprocessing and classification on medical or technology industries as they have definite glossary so the accuracy of the classification will be increased and become more focused. Finally, the algorithm will be applied on the other social media platforms such as Facebook, Instagram and Youtube.

References

[1] M. Rouse, "Social Media Analytics," TechTarget, November 2012.

[2] F. K. Gohar, "Social Media Network Analytics," in Seven Layers of Social Media Analytics: Mining Business Insights from Social Media, 2015.

[3] L. Bing, "Sentiment Analysis and Subjectivity," in Handbook of Natural Language Processing, Illinois - Chicago, 2010, pp. 1-38.

[4] F. Weiguo and G. Michael, "The Power of Social Media Analytics," Communications of the ACM, pp. Vol. 57 Issue 6, PP. 74-81, 2014.

[5] O. Chong, S. Sheila and S. Almahmoud, "Social Media Analytics Framework: The Case Of Twitter And Super Bowl Ads," Journal of Information Technology Management, pp. 1-18, 2015.

[6] E. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study," International Journal of Computer Applications, pp. Vol. 112 , No. 5, PP. 44-48, February 2015.

[7] Z. Thabit, "Data Mining in Social Media," International Journal of Scientific and Engineering Research, vol. 6, no. 7, pp. Vol. 6, Issue 7,152-154, July 2015.

[8] P. Priyanka and M. Khushali, "A Review: Text Classification on Social Media Data," IOSR Journal of Computer Engineering, vol. 17, no. 1, 2015.

[9] K. Faris and K. Jugal, "Classifying Short Text in Social Media: Twitter as Case Study," International Journal of Computer Applications, February 2015.

[10] M. Isaac, "Mining Social Media For Predictive Analytics," Kampala, Uganda: School Of Computing and Engineering, Uganda Technology and Management University, 2015.

[11] toshiakit, "AnalyzeTwitter," 2014. [Online]. Available: https://github.com/toshiakit/AnalyzeTwitter. [Accessed 31 October 2016].

[12] J. Kunal, "Mining YouTube using Python and performing social media analysis," ALS ice bucket challenge, 2014.

[13] P. Bo and L. Lillian, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, pp. Vol. 2, Nos. 1–2 (1–135), 2008.

[14] J. La, "Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata," 23 February 2009. [Online]. Available: https://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/. [Accessed 31 October 2016].

[15] R. Matthew, Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More, O'reilly, 2014.

[16] K. Shamanth, M. Fred and L. Huan, Twitter Data Analytics, Springer, 2013.

[17] G. Chakraborty, M. Pagolu and S. Garla, "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Institute," in Text Mining and Analysis, 2013.

[18] B. Liu, Sentiment Analysis Opinion Mining, Morgan and Claypool Publishers, 2012.

[19] "Subjectivity Lexicon," MPQA , 2016. [Online]. Available: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/. [Accessed 30 october 2016].

[20] L. Naji, "Twitter Sentiment Analysis Training Corpus (Dataset)," 22 september 2012.

[21] K. P. Kishori and R. R. Deshmukh, "Twitter Sentiment Classification on Sanders Data using Hybrid approach," IOSR Journal of Computer Engineering , pp. Volume 17, Issue 4,PP. 118-123, 2015.

[22] T. AzureML , Microsoft, 2 September 2014. [Online]. Available: https://gallery.cortanaintelligence.com/Experiment/Binary-Classification-Twitter-sentiment-analysis-4. [Accessed 31 october 2016].

[23] B. Joel, J. Fredrik, J. Carl and W. Anders, "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises," Security Informatics, vol. 3, no. 1, 28 august 2014.

[24] S. Hridoy, M. Ekram, M. Islam and R. M. Rahman, "Localized twitter opinion mining using sentiment analysis," Decision Analytics, vol. 2, no. 1, 22 october 2015.

[25] G. Pulkit and D. Sapan, Data Mining and Analysis on Twitter, 1st ed., ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2011.

[26] H. Julian, "Build your own Twitter Archive and Analyzing Infrastructure with MongoDB, Java and R [Part 2] [Update] - ThinkToStart," 2013.

[27] J. Hillebrand, "Cluster Twitter Data with R and k-means - ThinkToStart," 13 may 2014. [Online]. Available: http://thinktostart.com/cluster-twitter-data-with-r-and-k-means/. [Accessed 31 october 2016].

[28] A. Bailey and C. Sufandha, "Twitter Analytics Using R Part 3: Compare Sentiments - www.credera.com," 4 June 2014. [Online]. Available: https://www.credera.com/blog/technology-insights/open-source-technology-insights/twitter-analytics-using-r-part-3-compare-sentiments/. [Accessed 31 october 2016].

[29] M. Bonzanini, "Mining Twitter Data with Python Part 1: Collecting Data," Kdnuggets.com, 2015. [Online]. Available: http://www.kdnuggets.com/2016/06/mining-twitter-data-python-part-1.html.. [Accessed 31 october 2016].

[30] L. Shure, "Analyzing Twitter with MATLAB » Loren on the Art of MATLAB," Blogs.mathworks.com, 2014. [Online]. Available: http://blogs.mathworks.com/loren/2014/06/04/analyzing-twitter-with-matlab/. [Accessed 31 october 2016].

[31] M. Zaikin, "Common Architectures: Explain the advantages and disadvantages of multi-tier architectures," in Java Design: Objects, UML, and Process, Robert C. Martin Series, 2007.

[32] A. Agarwal, "Sentiment Analysis of Twitter Data," in In: Proc. ACL 2011 Workshop on Languages in Social Media, 2010 .

[33] B. Carver, "5 Reasons to Invest in Dedicated Social Media Analytics Tools," 2016. [Online]. Available: https://www.linkedin.com/pulse/5-reasons-invest-dedicated-social-media-analytics-tools-bob-carver. [Accessed 31 October 2016].

Toward Exascale Computing Systems: An Energy Efficient Massive Parallel Computational Model

Muhammad Usman Ashraf, Fathy Alburaei Eassa, Aiiad Ahmad Albeshri, Abdullah Algarni
Department of Computer Science
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

***Abstract*—The emerging Exascale supercomputing system expected till 2020 will unravel many scientific mysteries. This extreme computing system will achieve a thousand-fold increase in computing power compared to the current petascale computing system. The forthcoming system will assist system designers and development communities in navigating from traditional homogeneous to the heterogeneous systems that will be incorporated into powerful accelerated GPU devices beside traditional CPUs. For achieving ExaFlops (10^{18} calculations per second) performance through the ultrascale and energy-efficient system, the current technologies are facing several challenges. Massive parallelism is one of these challenges, which requires a novel energy-efficient parallel programming (PP) model for providing the massively parallel performance. In the current study, a new parallel programming model has been proposed, which is capable of achieving massively parallel performance through coarse-grained and fine-grained parallelism over inter-node and intra-node architectural-based processing. The suggested model is a tri-level hybrid of MPI, OpenMP and CUDA that is computable over a heterogeneous system with the collaboration of traditional CPUs and energy-efficient GPU devices. Furthermore, the developed model has been demonstrated by implementing dense matrix multiplication (DMM). The proposed model is considered an initial and leading model for obtaining massively parallel performance in an Exascale computing system.**

Keywords*—*Exascale computing; high-performance computing (HPC); massive parallelism; super computing; energy efficiency; hybrid programming; CUDA; OpenMP; MPI

I. INTRODUCTION

The high-performance computing (HPC) community anticipates that a new supercomputing technology called the exascale computing system will be available at the end of the current decade. This powerful supercomputer system will provide a thousand-fold computing power increase over the current petascale computing system and will enable the unscrambling of many scientific mysteries by computing 1 ExaFlops (10^{18} calculations per second) [1], [22], [23]. This ultrascale computing system will be composed of millions of heterogeneous nodes, which will contain multiple traditional CPUs and many-core General Purpose Graphics Processing Units (GPGPU) devices. In the current petascale computing system, the power consumption is approximately 25-60 MW, by using up to 10 M cores. According to this ratio, the power consumption demand of the exascale computing system will be more than 130 Megawatts. On the way towards the exascale supercomputing system, the United States Department of Energy (US DoE) and other HPC pioneers defined some primary constraints, including power consumption (PC) ≈ 25-30 MW, system development cost (DC) ≈ 200 million USD, system time to delivery (DT) ≈ 2020 and number of cores (NC) ≈ 100 million [25]. The primary limitation for the exascale system is that it does not exist yet. However, in trying to achieve ExaFlops-level performance under these strict limitations, current technologies are facing several fundamental challenges [24]. At a broad level, these challenges can be categorized according to the themes that are listed in Table I.

TABLE I. EXASCALE COMPUTING CHALLENGES

Challenge	Description
Power consumption management	Managing power consumption through new energy-efficient algorithms and devices
Programming models	New programming models are required for programming CPU + GPU-based heterogeneous systems
Novel architectures	New architectures and frameworks that can be implemented with non-traditional processors are required
Massive Parallelism	New parallel programming approaches are required that can provide massive parallelism using new accelerated devices
Resiliency	The system should be able to provide correct computation in the face of faults in the system
Memory management mechanisms	To improve data diversity and bandwidth

One traditional way to enhance the system performance at the exascale level is to improve clock speed. However, in the future, the clock speed will be limited to 1 GHz. An alternative approach is to increase the number of cores in the system. According to the defined limitations for the exascale computing system, the number of cores should not exceed 100 million. Generally, if we increase the number of resources (cores) to enhance the performance, it ultimately will increase the power consumption for computation. Another option is to achieve 'massive parallelism' in the system to improve system performance at the exascale level. Parallelization through different PP models has already been explored and examined, with the aim of exploiting a future exascale computing system. From the start of the current decade, in consideration of the many HPC applications, which include climate and environmental modeling, computation fluid dynamics (CFD) [2], molecular nanotechnology and intelligent planetary

spacecraft [3], new versions of PP models such as High-Performance FORTRAN (HPF) [4], [7] and an explicit message passing interface (MPI) were introduced to attain petaflop-level performance in the system.

To overcome the architectural challenges of petascale systems, many new approaches were introduced, including pure parallelism, *in situ* processing [5], and out-of-core and multi-resolution techniques; however, pure parallelism was conceived as a suitable paradigm. These suggested models were not able to address the challenges of the higher-order CFD applications that are required for computing thread-level parallelism in a cluster system. A new hybrid PP model was required for localizing the work from the distributed system in the spectral element method and performing efficient computations using multiple threads. Therefore, a hybrid model of MPI (to parallelize data at the inter-node level) and OpenMP (to parallelize at the intra-node level) was proposed by Dong et al. [6]. The hybrid model of MPI and OpenMP [21] for coarse-grained parallelism shows good scalability compared to single-hierarchy-level parallelism (pure MPI and pure OpenMP 3.0) with respect to both the problem size and the number of processors for a fixed problem size. However, the use of multiple threading in a hybrid paradigm increases the thread management overhead in thread creation/destruction and synchronization considerably with the increase in the number of threads [9]. To update the thread-level parallelism and address the overhead in thread creation/destruction and synchronization, OpenMP 4.0 was released in 2013 [8]. This new version was equipped with new features for error handling, tasking extensions, atomics and support for accelerated computation.

Recently, a dramatic change occurred in hardware technology development and new powerful computational devices were introduced, such as the General-Purpose Graphical Processing Unit (GPGPU) by NIVIDIA [10], AMD [48], ARM [49] and Many Integrated Cores (MIC) by Intel [11], [12]. These devices are thousands-fold more powerful than the traditional CPU devices. These Single-Instruction Multiple-Data (SIMD)-architecture-based many-core devices contain thousands of cores and are capable of performing thread-level execution. The old GPU models were only used for graphics processing, whereas the latest devices are able to perform general-purpose processing as well. To program GPUs, many PP models have been introduced, including OpenCL [20], OpenACC [50], CUDA and OpenMP [16], which are also available for GPU programming. So far, CUDA is considered the most capable model for performing thread-level optimization. Nevertheless, parallelized thread execution has been transformed from conventional CPU cores to GPU-accelerated devices. A detailed comparative study has been conducted by Ashraf et al. [19].

II. Navigation in the Hierarchy Level

Parallelism has brought about a great revolution in system performance enhancement. Parallelism was introduced in the 90s. The Terascale computing systems were based on coarse-grained parallelism, which was accomplished at the inter-node level through single-hierarchy models such as MPI [31]. To enhance the parallelism, a dual-hierarchy model was introduced for petascale supercomputing systems [32]. The objective of the petascale system was to achieve both coarse-grained and fine-grained parallelism through inter-node and intra-node processing. Many dual-hierarchy-level approaches were proposed to achieve both types of parallelism, including Hybrid MPI + OpenMP. In this dual-level hybrid model, MPI was used to achieve coarse-grained parallelism and OpenMP was used to achieve fine-grained parallelism at the thread level. The major problem with this model was massive power consumption while transferring data over CPU cores [33]. To overcome the power consumption challenge, new energy-efficient devices are introduced, such as GPGPU and MIC. From the software perspective, new programming approaches and models are required that can utilize these energy-efficient accelerated devices with traditional CPU cores through massive parallelism [23]. To achieve massive parallelism in the system, the hierarchy level in PP models is shifted from dual to tri-level, which is considered a promising level for future exascale computing systems. To add a third level of parallelism to the current homogeneous MPI + OpenMP model, a new tri-level model has been considered, which will be a hybrid MPI + OpenMP + X model [34].

Leading to a hybrid approach for massive parallelism, a new tri-level hybrid PP model was proposed for symmetric multiprocessor (SMP) cluster architectures in [12]. This model was based on message passing for inter-SMP node communication, loop directives by OpenMP for intra-SMP node parallelization and vectorization for each processing element (PE). The fundamental objective of this method was to combine coarse-grained and fine-grained parallelism. MPI was used to achieve coarse-grained parallelism and OpenMP was used to achieve fine-grained parallelism by parallelizing loops inside each SMP node. The hybrid approach is advantageous over flat MPI as it does not allow the passage of messages in all SMP nodes. This tri-level hybrid model was implemented to solve 3D linear elastic problems [35] by achieving a performance of 3.80 TFLOPS. In addition, tri-level hybrid and flat MPI programming models achieve similar performance. However, the hybrid model outperforms flat MPI in problems with large numbers of SMP nodes. Due to its monolithic power consumption, this model is not applicable to the exascale computing system. However, according to Amarasinghe et al. [36], unanimous implementation of existing models and powerful GPU devices for better performance of the system should be reinvestigated. For the future exascale system, the tri-level 'X' model will be considered as an additional model that will be responsible for the programming of accelerated GPU devices. To determine the X factor in the tri-level hybrid model, critical studies were conducted, where several models were proposed and compared with respect to performance, computation, optimization and many other metrics [26]-[29]. Evaluations showed that the current compiler of over-simplified OpenACC exceeded the performance of the Compute Unified Device Architecture (CUDA) by approximately 50%; moreover, it exceeded CUDA's performance by up to 98%. Conversely, metrics such as optimization and program flexibility, thread synchronization and other advanced features are attainable in CUDA but not in OpenACC. These metrics prevent full utilization of available resources for HPC heterogeneous computing systems.

Eventually, we finalized the X model as CUDA to compute accelerated GPU devices. Fig. 1 shows the fundamental navigational model for massive parallel programming.

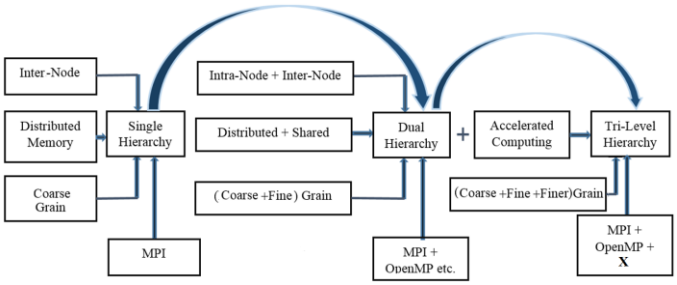


Fig. 1. Hierarchy navigation in the programming model.

This tri-level hybrid model is capable of achieving both coarse-grained and fine-grained parallelism through inter-node and intra-node processing over a heterogeneous cluster system. Leading to this architecture, we have proposed an initiative PP hybrid (MPI, OpenMP and CUDA) model with an optimized approach, which will be a promising framework for achieving the desired performance for exascale computing systems through massive parallelism.

A. MPI

MPI is a well-known traditional independent library that has been used for communication among the explicit processes in a distributed computing system. Historically, the standard version of MPI is considered the MPI-1.0 version from 1994. Many modifications, additions and clarifications have been made in different versions. Recently, in 2015, a new, mature version of MPI, namely, 'MPI 3.1', was released, to which many new features had been added, including environment management, point-to-point message passing, process creation and management, and collective communications [15]. Throughout HPC revolutionary development, MPI has been a prominent model for message passing in distributed nodes and multi-processor systems. In the future, it has been predicted that MPI will remain the best option for message passing among heterogamous devices over the cluster system, even though the original MPI designer did not focus on the exascale computing system, which requires some MPI specifications such as the maintenance of global state per process, memory management during communication within MPI processes, and process synchronization [37]. These MPI specifications must be adapted for the exascale computing system.

B. OpenMP

Open Specification for Multi-Processing (OpenMP) is one of the most frequently used models for SIMD thread-level parallel execution, which determines the set of directives, environment variables and multiple library routines. These specifications are supported in FORTRAN and C/C++ for using shared memory parallelism. The most recent version, namely, OpenMP 4.5, contains various new features, including error handling, tasking extensions, atomics and accelerated computation [38]. A new synchronization strategy has been introduced, where multiple tasks are grouped and synchronized using the 'taskgroup' construct [13]. In this way, many new constructs are added into the OpenMP 4.5 version that manages the threads efficiently. Similarly, loop parallelization with unbalanced amounts of work is also optimized as 'taskloop' using new directives [14]. One shortcoming of OpenMP is that it can be applied only for shared memory platforms on a single node, and not for cluster systems, which limits the use of the MPI option for cluster computing. However, it is anticipated that OpenMP will be promising model for exascale application, to achieve massive parallelism at the thread level.

C. CUDA

Recently, NIVIDA introduced CUDA (Compute Unified Device Architecture), which is a unique thread-level parallel computing platform for programming massive parallel computing accelerated GPUs. CUDA is supported by FORTRAN and C/C++ for programming accelerated GPGPUs. The current CUDA release, namely, CUDA 8.0, which is the most feature-packed and powerful, is available with novel profiling capabilities. In addition, it supports the Pascal GPU architecture and lambda heterogeneous compilers [17], [18]. In CUDA parallel programming, an application that contains the sequential program 'CUDA Kernel' is available, which executes programs in parallel on GPU devices. The Single Program Multiple Data (SPMD)-based kernel is initialized by passing multiple parameters, including grid size and block size. Based on modern GPU architecture, the GPU Block dispatcher schedules the grid by assigning each thread to one of the computational cores, and these threads are synchronized by self-cooperation. Each block has its own shared memory, which is accessible to every core inside it. Threads process data using this shared memory within that block and return the results to the scheduler. This processed data is stored in GPU global memory, which is accessible to host CPU cores. CPU cores read data from GPU global memory and transfer data from GPU to CPU cores and memory. In this way, we can achieve massive parallelism through heterogeneous CPU + GPU computation using CUDA.

III. Tri-Level Hybrid Parallel Programming Model

In this section, we present the proposed tri-level hybrid PP model for the exascale computing system. Based on the hierarchy navigation in previous parallel programming models, the proposed approach is a hybrid of MPI, OpenMP and CUDA.

A. Inter-Node Computation

In the proposed model, initially, some fundamental specifications, such as the number of nodes, number of CPUs per node, number of CPU cores, number of accelerated GPU devices, and memory levels, are the requirements of the system on which the model is to be implemented. After obtaining these fundamental specifications of the system, the parallel computing process is initiated. The top-level inter-node parallelism was achieved through the standard-specification MPI library to parallelize the distributed nodes. Immediately after MPI initialization, some necessary statements were executed to define the MPI communication size and the ranks of the available processes in MPI communication. Usually, the process with rank '0' is considered the master process, while the rest of the processes are considered slave processes. Before

broadcasting begins, data and many other necessary parameters are distributed over connected nodes in the system. For task mapping, the master process communicates with all slave processes to distribute/gather data. To maintain synchronization while sending and receiving data, blocking methods 'MPI_Send' and 'MPI_Recv' were respectively used, instead of non-blocking methods 'MPI_Isend' and 'MPI_Irecv'. These communication methods are better synchronized and more reliable for producing pure error-free parallelism.

B. Intra-Node Computation

Once the data have been shared over all distributed nodes, the second level of multi-threaded intra-node parallel processing is initiated through OpenMP, which uses shared memory among multiple CPU cores of the system. At this stage, multiple OpenMP pragmas were used to achieve fine-grained parallelism by defining all looping and independent parallel computing statements within the OpenMP parallel region. As this is middle-level parallelism, the resources of the current and next levels of parallelism are correlated. Before entering the third step, the number of available CPU threads in the system is determined, followed by the estimation of the number of accelerated GPU devices that are installed in the system. For the optimization of resources and results, the numbers of CPU threads and GPU devices should be same. Consequently, determination of the numbers of CPU threads and GPU devices can facilitate the adjustment of their strengths by using the following pre-defined functions:

*cudaGetDeviceCount (*numGPU*)*; //get number of GPUs

*omp_set_num_threads(*numGPU*);*// Set number of Threads

cudaThreadSynchronize(); // synchronize CUDA threads

C. Accelerated GPU Computation

Within the outer scope of OpenMP, another thread level of parallelism was created through the shared memory system over accelerated GPU devices, which provide finer granularity using GPU cores. This complicated heterogeneous CPU+GPU computation is supported by different programming models using FORTRAN and C/C++. In our proposed model, we used CUDA to perform this heterogeneous computation, where the SIMD-based data segment was transferred from Host to GPU core using built-in CUDA methods. Fig. 2 presents the workflow of tri-hybrid parallel programming as follows.

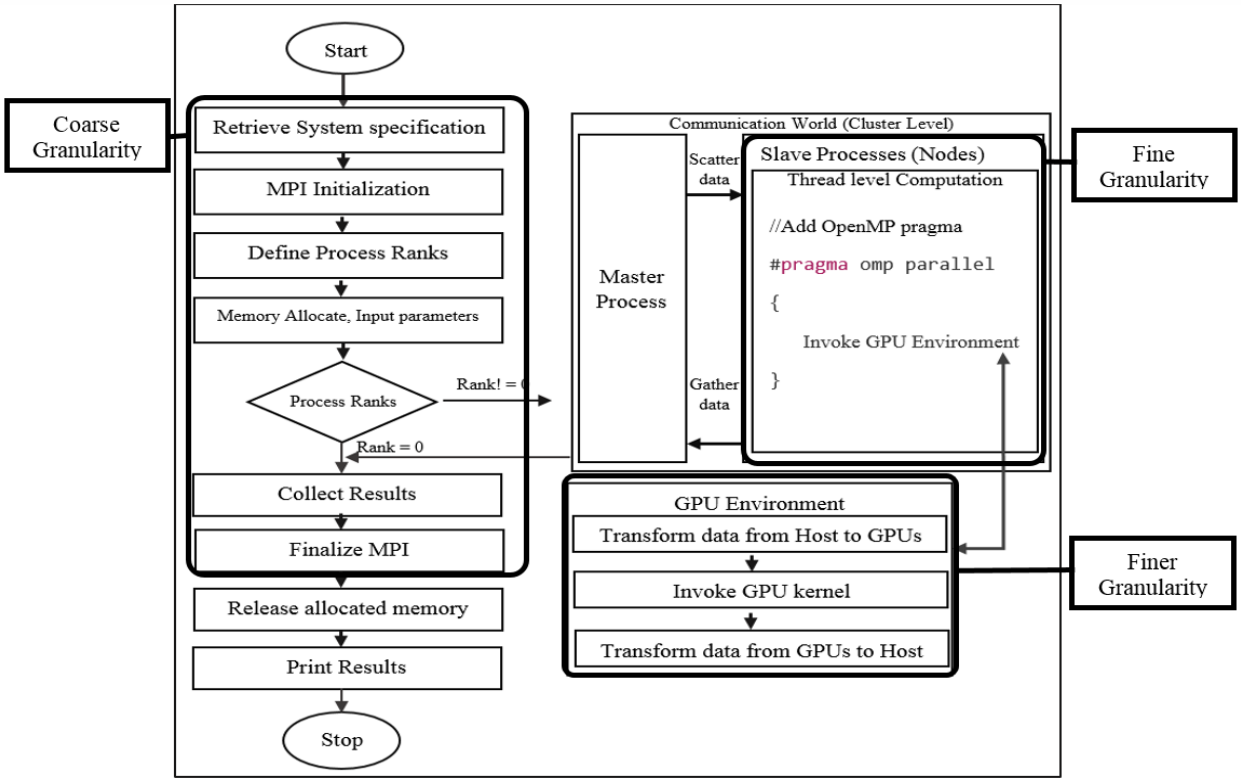


Fig. 2. Workflow of the hybrid parallel programming model.

At the same time, some fundamental information, including grid size and block size, were broadcasted with the CUDA kernel to restrict computation according to given specifications. To create a generic kernel, we defined template datatypes, which were provided by C++, that accept any datatypes as parameters and perform computations accordingly. Once parallel data computation was completed through GPU cores, it used a similar datatype from GPU to Host cores that entered again in OpenMP region. After finishing this complicated heterogeneous computation among CPUs and GPUs, the MPI master process collected all processed data and exited the parallel zone. The detailed sequence of these three levels of parallelism is illustrated in Fig. 2.

IV. EXPERIMENTAL SETUP

This section describes the experimental setup that was used to implement the proposed model. Moreover, we quantified different HPC-related metrics, including performance (number of GFlops/s) and energy efficiency (GFlops/Watt) in the system. A detailed description of these metrics is presented in this section.

A. Experimental Platform

The proposed tri-level hybrid model was implemented on an Aziz-Fujistu Primergy CX400 Intel Xeon Truescale QDR supercomputer, which was manufactured by Fujistu, at the HPC center of King Abdul-Aziz University, Jeddah, Saudi Arabia [39]. In 2015, Aziz was ranked 360th in the list of the top-500 HPC supercomputers [40]. The Aziz supercomputer is comprised of 380 regular (thin) and 112 large (fat) compute nodes. Recently, Aziz was upgraded with two SIMD-architecture-based accelerated GPU compute nodes (NVIDIA Tesla K20 GPU, 2496 CUDA Cores). Moreover, 2 MIC nodes, each with an Intel Xeon Phi Coprocessor with 60 cores, were installed. Aziz consists of a total of 11904 cores. The memories that are offered by regular and large nodes are 96 GB and 256 GB, respectively. Each node that contains an Intel E5-2695v2 processor with 2.4 GHz and 12 Cores is run by the Cent 6.4 operating system. Aziz is linked using three different networks: the InfiniBand network, the User network and the Management network. Moreover, all nodes are interconnected with one another. The file system is parallelized through the InfiniBand network. In addition, the login system and job submission are handled through the user network, while the management network is used for management purposes only. Aziz is capable of achieving 211.3 TFlops/s Linpack performance and 228.5 TFlops/s theoretical peak performance [41].

B. Performance Measurement

Performance is the first and fundamental metric of HPC systems, which is measured in Flops (number of floating-point operations per second) in the current experiments. Usually, in a parallel programing system, Flops are calculated at the peak performance of the system and for implementing algorithms. Let F_p denote the Flops at peak performance and F_m denote the Flops for implementing algorithms. F_c can be calculated as:

$$F_c = \frac{F_p}{F_m} \quad (1)$$

Using the peak performance of 211.3 TFlops/s of the Aziz supercomputer, we measured the performance range by executing target-dense MM with different datasets.

C. Power Measurement

Limiting power consumption is one of the vital challenges for current and future supercomputing technologies. The primary objective of future research for the exascale computing system is the optimal selection of hardware and software for achieving high performance under the power consumption limitations [42]. Many HPC pioneers have initiated and developed energy-efficient devices, such as NVIDIA GPGPU [43], AMD GPU [44], and Intel MIC [45]. Similarly, software development communities are trying to develop new programming models that can provide outstanding performance under energy constraints.

Generally, a system is evaluated according to its energy consumption, which indicates the power rate at which processing was executed, as described in (2).

$$E\ (kWh) = \int_0^t V.I\ \ (dt) \quad (2)$$

From the above equation, we can calculate the total energy consumption of a system by integrating the energy consumption, which is composed of the bandwidth, memory contention, parallelism and behavior of the application in the HPC parallel system, as described in (3).

$$E_{system} = \int_0^t BandW\ (dt) + \ MemC\ (dt) + \ Prll\ (dt) + \ Bhv\ (dt) \quad (3)$$

On the basis of the dictated factors and the fundamental energy evaluation (2), we quantified these factors in the current study with respect to system performance and power consumption. The power consumption is the sum of the products of the power of each component and the corresponding duration [28]. The measurement of power consumption is divided into two categories:

1. System Specification.
2. Application Specification.

Since the system specification has GPU devices installed in it, the power consumption is calculated by (4):

$$\boldsymbol{P_{system}(w)} = \sum_{i=1}^{N} \boldsymbol{P^i_{GPU}}\ (\boldsymbol{w^i}) + \ \boldsymbol{P}_{CPU}\ \sum_j^M\ (\boldsymbol{w^j}) + \ \boldsymbol{P_{mainboard}(w)} \quad (4)$$

From (4), it can be speculated that the approximate power consumption of a system is the sum of the products of the installed GPUs, CPUs and motherboard. The power consumption varies with the workload; however, on the application side, it can be quantified using (5):

$$\boldsymbol{P_{app}(w)} = \sum_{i=1}^{Napp} \boldsymbol{P^i_{GPU}}\ (\boldsymbol{w^i}) + \ \boldsymbol{P}_{CPU}\ \sum_j^M\ (\boldsymbol{w^j}) + \ \boldsymbol{P_{mainboard}}\left(\boldsymbol{w_{app}}\right) \quad (5)$$

According to (4) and (5), the power consumption in watts was measured at the idle state of the system, where only 5 watts of power were consumed by the motherboard and the remaining power was consumed by the cores of system.

V. EXPERIMENTAL RESULTS

In this section, we investigate the proposed tri-level hybrid parallel programming model via implementation of linear algebraic Dense Matrix Multiplication (DMM) [46]. The purpose of this study was to execute the DMM in the proposed model on a heterogeneous-architecture-based Aziz supercomputer and to determine the performance and power consumption, which are vital metrics for emerging exascale computing systems. We recorded different datasets of DMM through multiple CUDA kernels, which demonstrated that multiple kernels could produce energy-efficient results simultaneously. Moreover, during execution, the parallel performance of multiple kernels and the power consumption were evaluated, which indicated that the best performance was attained using a small and optimized number of kernels in an

energy-efficient way. This is due to the optimized computation over heterogeneous CPU and GPU cores using the CUDA platform. In contrast, using many kernels provided lower performance due to unnecessary communication among non-optimized CUDA kernels. A simple implementation of DMM, along with the defined parameters, is presented in Table II.

TABLE II. A NAÏVE CODE AND PARAMETERS OF IMPLEMENTED DMM

Kernel	Naive Code	Parameters & Domains
DMM	Do i = 1; n Do j = 1; n Do k = 1; n z(i, k)=z(i , k) + x(i, j) * y(j, k)	t_i, t_j, t_k *(i,j,k tiles)* *ui,uj(i,j,* unrolls*)* *matrix-Size (msize)* *msize* ∈ [1000, 2000, 3000… 10000]

However, we were unable to find a detailed optimization strategy for DMM due to space limitations, as explained by Tiwari et al. [30]. To explore the implementation strategy for DMM, we reused the z array in the buffer registers and the x and y arrays in the caches. These kernel configurations were obtained by varying parameters. In our implementations, the achieved performance ranged from 200 to 1100 GFlops for all implemented kernels for datasets of sizes 1000 to 10000, and the average was 716 GFlops, as shown in Fig. 3.

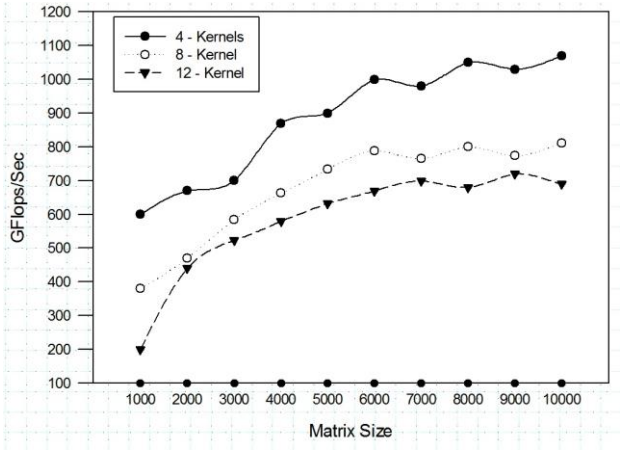


Fig. 3. Performance in DMM through multiple Kernel configurations.

During DMM computation, 4 CPU threads per node with 4 kernels achieved the best performance compared to all other configured kernels and achieved 68% of the peak performance with 1086 Gflops. Using 12 kernels produced efficient performance, but increased the energy efficiency due to unneeded communication in data processing.

Along with performance, we quantified another primary metric, namely, energy consumption, which was 28 Joules. At maximum DMM for a dataset of size 10000 through an optimized 4-kernel configuration, the quantified energy efficiency was 8.3 Gflops/W. The increment of resources affected energy efficiency dramatically and reduced it to 5.6 Gflops/W, as shown in Fig. 4.

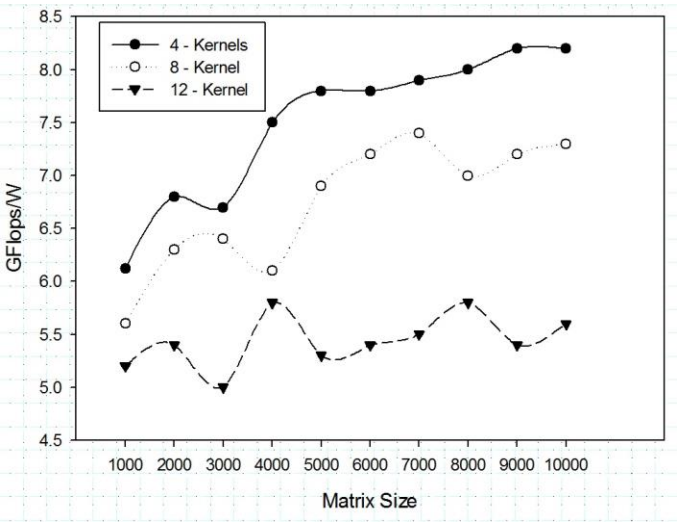


Fig. 4. Energy efficiency in DMM for different multiple-kernel configurations.

Based on performance and energy efficiency, a tradeoff between the two metrics [47] can be determined as follows:

$$\frac{Performance}{Power} = \frac{Execution\ within\ the\ time\ unit}{Energy\ during\ the\ execution\ time\ unit} = \frac{work}{energy}$$

Following this tradeoff, we calculated the ratio between performance and energy efficiency, which describes the performance that is achievable for a given energy efficiency, as shown in Fig. 5. Each vertical and horizontal line represents information about performance and energy efficiency, respectively. We can fix the configuration and parameters at any intersecting point to provide maximum performance and energy efficiency. These evaluations determined that the best performance-energy efficiency that can be achieved using the proposed model on the Aziz supercomputer reached 1086 GFlops, which corresponds to an energy efficiency of 8.3 GFlops/W.

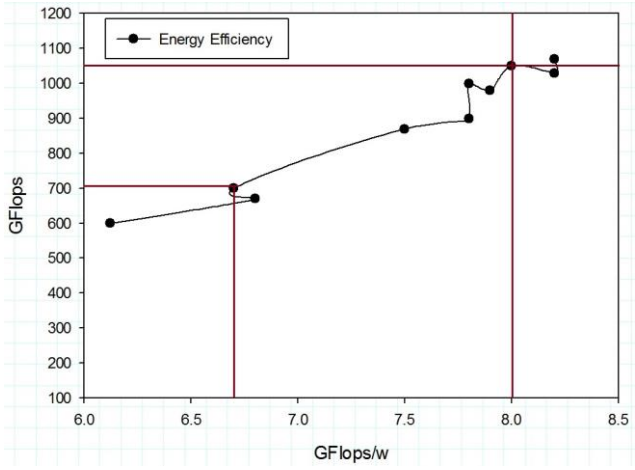


Fig. 5. Performance-energy efficiency tradeoff.

VI. EXASCALE COMPUTING SYSTEM DEMAND

The biggest challenge for the study of the emerging exascale computing system is that such a system does not exist yet. Therefore, predictive exascale-level data can be obtained using current computing systems (Terascale, petascale). In this section, a critical statistical analysis of experimental results that were obtained from the Aziz supercomputer is conducted. This statistical analysis is based on the metrics, including performance and energy efficiency that are required to satisfy the demands of the exascale computing system. The architecture of our experimental platform is heterogeneous (CPU + GPU) and is based on a cluster system that contains 11904 cores, which are integrated over 494 connected nodes. Using processing devices and memory structure, this system can provide 211.3 Tflops/s Linpack performance and 228.5 Tflops/s theoretical peak performance. In our experiments, we implemented DMM using different kernel sizes and obtained 68% of the peak performance with 1086 Gflops by consuming 28 joules of energy, which yielded 8.3 Gflops/Sec energy efficiency. This energy efficiency was determined using the fundamental formula that is given as follows:

$$P(w) = \frac{E_{(j)}}{t_{(s)}}$$

However,

$$watt = \frac{Joule}{\text{Second}} \quad \text{or} \quad W = J / S \tag{6}$$

TABLE III. EXASCALE SYSTEM CONFIGURATION

Feature	Specification
Number of Cabinets	200
Nodes per Cabinet	384
Number of Nodes	76800
Number of Network slice	4
Total router count	19200
Peak PFlops	1258
Max Power Consumption of Processors	230 W
Max Power Consumption / Node	300 W
Max Power Consumption / System	25 MW

According to (6) and the system configurations, our system consumed 130 watts at the best performance and energy efficiency. According to the exascale system constraints and predictive configurations, as listed in Table III, our system required a thousand-fold increase in current resources to perform exaFlops computations.

Based on the ratio of current computation and required resources, the predictive performance and power consumption were calculated, which are presented in Table IV.

TABLE IV. METRIC ANALYSIS FOR DIFFERENT PLATFORMS

Metric	Platforms		
	Aziz Supercomputer		Exascale
	Achieved	Predictive	
Performance	1086 Gflops	≈ 230 Pflops	1 ExaFlops
P. Consumption	130 Watt	≈ 27 M.W	≈ 25 M.W
Energy Efficiency	8.3 Gflops/W	≈ 8.5 Pflops/M.W	≈ 40 Pflops/M.W

Table IV describes a statistical analysis of current and future platforms for massive computation. The currently available Aziz platforms are categorized into achieved and predictive domains. Both platforms are analyzed on the basis of the metrics that were used in DMM. In the predictive platform, the scalability of the Aziz supercomputer was considered according to the configurations of the exascale computing system, which facilitated the determination of a predictive benchmark against each metric. The predictive benchmark does not depend on the demand of the exascale system. Therefore, it can be considered an initial step in achieving the required computational level for the exascale system. In the current study, our evaluations have raised numerous challenging questions, which will open new avenues of research for scientific communities, developers and vendors in the future:

- Which programming layer is responsible for managing the dynamic behavior of resources and code irregularity, and how?
- Sometime algorithms provide better performance with less energy efficiency. What optimized method should be adopted to satisfy the trade-off between the metrics?
- Memory management plays a vital role in enhancing system performance. To increase the efficiency of memory management, what additional hooks are required?
- How can data be managed to reduce power consumption when GPU cores occupy complete the warp for small executions?

These questions suggest new challenges regarding the satisfaction of HPC metrics through massive parallelism. However, we should reconsider our implemented algorithms, frameworks, benchmarks, energy management algorithms, communication mechanisms, memory management mechanisms and load balancing mechanisms, since these factors are paramount concerns for exascale systems.

VII. CONCLUSIONS

HPC technology is being shifted from the petascale to the extreme "exascale" computing system. On the road to the exascale system, due to some strict limitations on energy consumption, system cost, number of cores and time to delivery, there are many vital challenges for vendors and development communities. One of these major challenges is to achieve massive parallelism through energy-efficient mechanisms. In this study, we have proposed a new Tri-Level hybrid (MPI + OpenMP + CUDA) parallel programming model. The proposed model is applicable for heterogeneous (CPU + GPU) distributed systems, to achieve massive parallelism with coarse, fine and finer granularity. To evaluate the proposed model, we implemented DMM with different datasets through multiple kernels. All implementations were performed on an Aziz - Fujitsu PRIMERGY CX400, Intel Xeon E5-2695v2 12C 2.4 GHz, Intel TrueScale QDR supercomputer. We evaluated our model using HPC metrics, including performance, power consumption and energy efficiency. Moreover, we provided some predictive results as

to the performance that will be achievable through exascale-level scalability in the current system. Based on the results of our implementations, the proposed model can be considered a pioneering model for HPC applications. As the exascale system does not yet exist and all the implementations and results are predictive, we must reconsider the generic challenges, including the implementation of algorithms, frameworks, benchmarks, energy management algorithms, and communication mechanisms.

By future perspectives, we have specified some additional questions that are open challenges, while achieving extreme performance with energy efficiency through massive parallelism in the HPC system. Moreover, fixed optimization in a heterogeneous computing system is not possible. Nevertheless, an adaptive framework is required for adjusting the model to the system configuration and the application requirements.

REFERENCES

[1] Perarnau, Swann, Rinku Gupta, and Pete Beckman. "Argo: An Exascale Operating System and Runtime." (2015).

[2] Zhou, Min. Petascale adaptive computational fluid dynamics. Diss. RENSSELAER POLYTECHNIC INSTITUTE, 2009.

[3] Dongarra, Jack J., and David W. Walker. "The quest for petascale computing." Computing in Science & Engineering 3.3 (2001): 32-39.

[4] Jin, Haoqiang, et al. "High performance computing using MPI and OpenMP on multi-core parallel systems." Parallel Computing 37.9 (2011): 562-575.

[5] [5] Ma, Kwan-Liu, et al. "In-situ processing and visualization for ultrascale simulations." Journal of Physics: Conference Series. Vol. 78. No. 1. IOP Publishing, 2007.

[6] Dong, Suchuan, and George Em Karniadakis. "Dual-level parallelism for high-order CFD methods." Parallel Computing 30.1 (2004): 1-20.

[7] Shafto, Mike, et al. "Modeling, simulation, information technology & processing roadmap." NASA, Washington, DC, USA, Tech. Rep 11 (2012).

[8] Martineau, Matt, Simon McIntosh-Smith, and Wayne Gaudin. "Evaluating OpenMP 4.0's Effectiveness as a Heterogeneous PP Model." Parallel and Distributed Processing Symposium Workshops, 2016 IEEE International. IEEE, 2016.

[9] Jin, Shuangshuang, and David P. Chassin. "Thread Group Multithreading: Accelerating the Computation of an Agent-Based Power System Modeling and Simulation Tool--C GridLAB-D." 2014 47th Hawaii International Conference on System Sciences. IEEE, 2014.

[10] Hoegg, Thomas, et al. "Flow Driven GPGPU Programming combining Textual and Graphical Programming." Proceedings of the 7th International Workshop on Programming Models and Applications for Multicores and Manycores. ACM, 2016.

[11] Shan, Hongzhang, et al. "Thread-level parallelization and optimization of NWChem for the Intel MIC architecture." Proceedings of the Sixth International Workshop on Programming Models and Applications for Multicores and Manycores. ACM, 2015.

[12] Nakajima, Kengo. "Three-level hybrid vs. flat mpi on the earth simulator: Parallel iterative solvers for finite-element method." Applied Numerical Mathematics 54.2 (2005): 237-255.

[13] Terboven, C., Hahnfeld, J., Teruel, X., Mateo, S., Duran, A., Klemm, M., Olivier, S.L. and de Supinski, B.R., 2016, October. Approaches for Task Affinity in OpenMP. In International Workshop on OpenMP (pp. 102-115). Springer International Publishing.

[14] Podobas, Artur, and Sven Karlsson. "Towards Unifying OpenMP Under the Task-Parallel Paradigm." International Workshop on OpenMP. Springer International Publishing, 2016.

[15] Dinan, James, et al. "An implementation and evaluation of the MPI 3.0 onesided communication interface." Concurrency and Computation: Practice and Experience (2016).

[16] Concise Comparison Adds OpenMP Versus OpenACC To CUDA Versus OpenCL Debates "techenablement.com/concise-comparision-adds-openmp-versus-openacc-to-cuda-versus-opencl-debates/", 12 Nov 2016.

[17] Fleuret, François. "Predicting the dynamics of 2d objects with a deep residual network." arXiv preprint arXiv:1610.04032 (2016).

[18] NVIDIA Accelerated Computing "developer.nvidia.com/cuda-downloads", 02 Nov 2016.

[19] Ashraf, Muhammad Usman, Fadi Fouz, and Fathy Alboraei Eassa. "Empirical Analysis of HPC Using Different Programming Models." (2016).

[20] Ashraf, Muhammad Usman, and Fathy Elbouraey Eassa. "OpenGL Based Testing Tool Architecture for Exascale Computing." International Journal of Computer Science and Security (IJCSS) 9.5: 238.

[21] Ashraf, Muhammad Usman, and Fathy Elbouraey Eassa. "Hybrid Model Based Testing Tool Architecture for Exascale Computing System."International Journal of Computer Science and Security (IJCSS) 9.5 (2015): 245.

[22] Shalf, John, Sudip Dosanjh, and John Morrison. "Exascale computing technology challenges." International Conference on High Performance Computing for Computational Science. Springer Berlin Heidelberg, 2010.

[23] Reed, Daniel A., and Jack Dongarra. "Exascale computing and big data."Communications of the ACM 58.7 (2015): 56-68. Cappello, Franck, et al. "Toward exascale resilience." International Journal of High Performance Computing Applications (2009).

[24] Marc Snir, Robert W Wisniewski, Jacob A Abraham, Sarita V Adve, Saurabh Bagchi, Pavan Balaji, Jim Belak, Pradip Bose, Franck Cappello, Bill Carlson, Andrew A Chien, Paul Coteus, Nathan A DeBardeleben, Pedro C Diniz, Christian Engelmann, Mattan Erez, Saverio Fazzari, Al Geist, Rinku Gupta, Fred Johnson, Sriram Krishnamoorthy, Sven Leyer, Dean Liberty, Subhasish Mitra, Todd Munson, Rob Schreiber, Jon Stearley, and Eric Van Hensbergen. Addressing failures in exascale computing. International Journal of High Performance Computing Applications, 28(2):129{173, May 2014.

[25] Reed, Daniel, et al. DOE Advanced Scientific Computing Advisory Committee (ASCAC) Report: Exascale Computing Initiative Review. USDOE Office of Science (SC)(United States), 2015.

[26] Hoshino, Tetsuya, et al. "CUDA vs OpenACC: Performance case studies with kernel benchmarks and a memory-bound CFD application." Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on. IEEE, 2013.

[27] Herdman, J. A., et al. "Accelerating hydrocodes with OpenACC, OpenCL and CUDA." High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:. IEEE, 2012.

[28] Lashgar, Ahmad, Alireza Majidi, and Amirali Baniasadi. "IPMACC: Open source OpenACC to CUDA/OpenCL translator." arXiv preprint arXiv:1412.1127 (2014).

[29] [29] Christgau, Steffen, et al. "A comparison of CUDA and OpenACC: accelerating the tsunami simulation easywave." Architecture of Computing Systems (ARCS), 2014 Workshop Proceedings. VDE, 2014.

[30] A. Tiwari, C. Chen, J. Chame, M. Hall, and J. Hollingsworth. A Scalable Auto-Tuning Framework for Compiler Optimization. In IPDPS'09, Rome, Italy, May 2009.

[31] Gabriel, Edgar, et al. "Open MPI: Goals, concept, and design of a next generation MPI implementation." European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting. Springer Berlin Heidelberg, 2004.

[32] Mininni, Pablo D., et al. "A hybrid MPI–OpenMP scheme for scalable parallel pseudospectral computations for fluid turbulence." Parallel Computing 37.6 (2011): 316-326.

[33] Hennecke, Michael, et al. "Measuring power consumption on IBM Blue Gene/P." Computer Science-Research and Development 27.4 (2012): 329-336.

[34] Jacobsen, Dana A., and Inanc Senocak. "Multi-level parallelism for incompressible flow computations on GPU clusters." Parallel Computing 39.1 (2013): 1-20.

[35] Nguyen-Thoi, T., et al. "A face-based smoothed finite element method (FS-FEM) for 3D linear and geometrically non-linear solid mechanics

problems using 4-node tetrahedral elements." International journal for numerical methods in Engineering 78.3 (2009): 324-353.

[36] Amarasinghe, Saman, et al. "ASCR programming challenges for exascale computing." Report of the 2011 Workshop on Exascale Programming Challenges. 2011.

[37] Message passing Interface, https://computing.llnl.gov/tutorials/mpi/ , 20 June, 2017 [03 Aug, 2017]

[38] Royuela, Sara, et al. "OpenMP Tasking Model for Ada: Safety and Correctness." Ada-Europe International Conference on Reliable Software Technologies. Springer, Cham, 2017.

[39] Fujitsu to Provide High-Performance Computing and Services Solution to King Abdulaziz University, http://www.fujitsu.com/global/about/resources/news/press-releases/2014/0922-01.html, 22 Sep, 2014 [06 July, 2017]

[40] King Abdulaziz University, https://www.top500.org/site/50585, June 2015 [03 Aug, 2017]

[41] Aziz - Fujitsu PRIMERGY CX400, Intel Xeon E5-2695v2 12C 2.4GHz, Intel TrueScale QDR, https://www.top500.org/system/178571, June 2015 [03 Aug, 2017]

[42] L. A. Barroso. The price of performance. Queue, 3(7):48–53, September 2005.

[43] Foley, Denis, and John Danskin. "Ultra-Performance Pascal GPU and NVLink Interconnect." IEEE Micro 37.2 (2017): 7-17.

[44] Rohr, David, et al. "An energy-efficient multi-GPU supercomputer." High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst (HPCC, CSS, ICESS), 2014 IEEE Intl Conf on. IEEE, 2014.

[45] Chrysos, George. "Intel® Xeon Phi™ coprocessor-the architecture." Intel Whitepaper 176 (2014).

[46] Gallivan, Kyle A., Robert J. Plemmons, and Ahmed H. Sameh. "Parallel algorithms for dense linear algebra computations." SIAM review 32.1 (1990): 54-135.

[47] Anzt, Hartwig, et al. "Experiences in autotuning matrix multiplication for energy minimization on GPUs." Concurrency and Computation: Practice and Experience 27.17 (2015): 5096-5113.

[48] Rajovic, Nikola, et al. "The low power architecture approach towards exascale computing." Journal of Computational Science4.6 (2013): 439-443.

[49] Rajovic, Nikola, et al. "Tibidabo: Making the case for an ARM-based HPC system." Future Generation Computer Systems 36 (2014): 322-334.

[50] Wolfe, Michael, et al. "Implementing the OpenACC Data Model." Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017 IEEE International. IEEE, 2017.

Image Contrast Enhancement by Scaling Reconstructed Approximation Coefficients using SVD Combined Masking Technique

[1] Sandeepa K S, [2]Basavaraj N Jagadale
Department of Electronics
Kuvempu University
Karnataka, India

[3]J S Bhat
Department of Physics
Karnataka University
Karnataka, India

[4]Mukund N Naragund
Department of Electronics
Christ University
Karnataka, India

[4]Panchaxri
Department of Electronics
SSA Govt First Grade College
Ballari, Karnataka, India

***Abstract*—The proposed method addresses the general issues of image contrast enhancement. The input image is enhanced by incorporating discrete wavelet transform, singular value decomposition, standard intensity deviation based clipped sub image histogram equalization and masking technique. In this method, low pass filtered coefficients of wavelet and its scaled version undergoes masking approach. The scale value is obtained using singular value decomposition between reconstructed approximation coefficients and standard intensity deviation based clipped sub image histogram equalization image. The masking image is added to the original image to produce a maximum contrast-enhanced image. The supremacy of the proposed method tested over other methods. The qualitative and quantitative analysis is used to justify the performance of the proposed method.**

Keywords*—Standard intensity deviation clipped sub image histogram equalization; discrete wavelet transform; singular value decomposition; masking technique

I. INTRODUCTION

The image enhancement has many practical applications. There has been continuous research on developing new algorithms for different applications. In an image analysis, intensity-based enhancement techniques like contrast enhancement help to improve the clarity of the image [1]. The intensity enhancement in an image has been realized through techniques based on the histogram, transform domain filtering and masking approaches. Histogram equalization (HE) or generalized histogram equalization (GHE) [2], are simple approaches as these techniques stretch the gray values effectively but give over enhancement. Some of the methods employed to preserve the brightness are, Bi-histogram, Brightness preserving bi-histogram equalization (BBHE) [3] and dualistic sub image histogram equalization (DSIHE) [4], but these methods fail to suppress noise artifacts. Exposure-based sub image histogram equalization (ESIHE) [5] is best suited for low exposure image but less fit to contrast enhancement of medical images.

The discrete cosine transform (DCT) domain, provide spectral separation [6]. However, there are few drawbacks in transforming the image using block DCT. Therefore, discrete wavelet transform (DWT) has found application in contrast enhancement [7]. Wavelet coefficients have inherent qualities such as sparsity and decomposition [8]. The DCT-SVD and DWT-SVD based illumination enhancement presented by updating singular value matrix of singular value decomposition (SVD). In these methods, the low-low subband of the input image is considered by applying DWT [9], [10].

Recent algorithms for contrast enhancement are based on masking techniques [11]-[14]. In these methods, filtered version of original image have internal scaling processes, and the drawback of these approaches is the use of fixed scale value, irrespective of the input image. The scale values are selected randomly.

In this paper, we tried to address over enhancement issue using effective sub image histogram equalization process. We tried careful rescaling of the reconstructed approximation coefficients using SVD approach to achieve better results. Also, the minute details of the image and sensitive edges are preserved with the help of masking approach.

This work partially motivated by Un-sharp masking technique [11], an excellent result of SVD in contrast enhancement [15] and standard intensity deviation based clipped sub image histogram equalization (SIDCSIHE) [16]. We have used SVD to select intensity information of the SIDCSIHE and reconstructed approximation coefficients. In this paper, we tried to improve upon our previous work [17], the paper is structured as follows: Section 2 contains the proposed method and in Section 3 we discuss experimental results. The conclusions are given in Section 4.

II. PROPOSED METHODOLOGY

The medical image analysis depends upon quality; therefore contrast enhancement is often desirable for

interpretation and diagnosis. This paper presents an improved contrast and enhancement in image quality by using DWT, SIDCSIHE, and SVD along with masking approach.

A. Discrete Wavelet Transformation

The low contrast image is subjected to DWT. The DWT decomposes the image into frequency sub-bands, namely, low low (LL), low high (LH), high low ((HL) and high high (HH) subbands as shown in Fig. 1. The LL subband is approximation coefficients and contains illumination information and image features. The decomposed approximation coefficients are obtained from (1) and its reconstruction is achieved by using inverse discrete wavelet transformation (IDWT) as given by (2).

$$W_{\varphi}(j_0,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\varphi_{j_0,m,n}(x,y) \quad (1)$$

$$f_{(x,y)}^{A}=\frac{1}{\sqrt{MN}}\sum_{m}\sum_{n}W_{\varphi}(j_0,m,n)\varphi_{j_0,m,n}(x,y) \quad (2)$$

where, $\boldsymbol{W_{\varphi}(j_0,m,n)}$ is the approximation coefficient, $\boldsymbol{f(x,y)}$ is the input time domain image of size M x N with discrete variable (x, y). $\boldsymbol{\varphi_{j_0,m,n}(x,y)}$, is the scale function and $\boldsymbol{f_{(x,y)}^{A}}$ is the reconstructed approximation coefficient [2]. The wavelet function (db1) is used for wavelet decomposition of image.

Fig. 2 shows the LL, LH, HL and HH subbands of the brain_MRI image. Here, the reconstructed LL subband intensity information is used to find the best scale with respect to the input image. The other subbands contain edge information and details of the image as it contains high frequency. The illumination improvement can be achieved by scaling the coefficients of the LL subband.

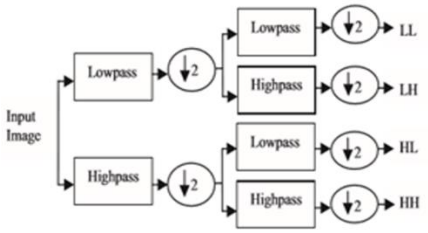


Fig. 1. Block diagram of DWT filter banks of level 1.

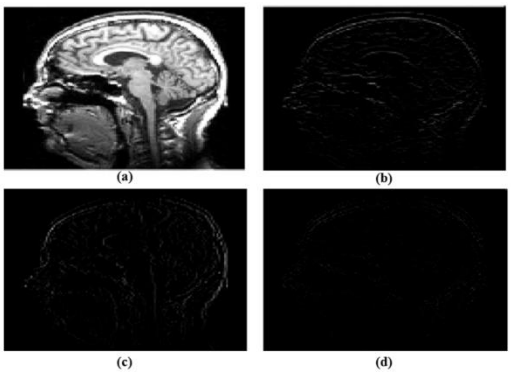


Fig. 2. Results of DWT decomposition of Brain_MRI image (a) Low Low subband; (b) Low High Sub band; (c) High Low subband; (d) High High subband.

B. Standard Intensity Deviation based Clipped Sub Image Histogram Equalization (SIDCSIHE)

Standard intensity deviation value is defined in (3) by using standard deviation function σ as shown in (4) [18].

$$X_{SIDV}=L*\left(1-\left(\frac{\sigma}{L}\right)\right) \quad (3)$$

$$\sigma=\left(\frac{\sum_{i=1}^{L}(i-H_{\mu})^{2}x\,h(i)}{\sum_{i=1}^{L}h(i)}\right)^{1/2} \quad (4)$$

The $H_{\mu,}$ is a mean of the input image histogram is given by (5).

$$H_{\mu}=\frac{\sum_{i=1}^{L}h(i)\,i}{\sum_{i=1}^{L}h(i)} \quad (5)$$

where $h(i)$, is image histogram with its corresponding intensity i and L is its total number of gray levels.

To prevent over enhancement clipping histogram T_c is calculated as in (6) and (7).

$$T_c=\frac{1}{L}\sum_{K=1}^{L}h(k) \quad (6)$$

$$h_c(k)=T_c \text{ for } h(k)\geq T_c \quad (7)$$

$h_{c(k)}$, is the clipped histogram and it is computationally efficient [19]. Based on standard intensity deviation value X_{SIDV}, the clipped histogram is divided into two sub-images I_{low} and I_{up} with ranges varying from 0 to X_{SIDV} and $X_{SIDV}+1$ to $L-1$ respectively. The cumulative distribution function $C_{low}(i)$, $C_{up}(i)$ of each sub-image can be defined as

$$C_{low}(i)=\sum_{i=0}^{X_{SIDV}}h_c(i)\Big/_{N_{low}}\,for\;\;0\leq i\leq X_{SIDV} \quad (8)$$

$$C_{up}(i)=\sum_{i=X_{SIDV}+1}^{L-1}h_c(i)\Big/_{N_{up}}\,for\;X_{SIDV}\leq i\leq L-1 \quad (9)$$

where N_{low} and N_{up} are the total numbers of pixels in each sub-image. The histogram equalization is done individually for two sub-images using the transfer function $F(i)$ as expressed in (10) and sub-images are combined into one final image by the transfer function $F(i)$ [16].

$$F(i)=\begin{cases}X_{SIDV}*\;\;C_{low} & for\;0\;\leq i\leq X_{SIDV}\\ (X_{SIDV}+1)+\;(L-X_{SIDV}+1)*C_{up} & for\;X_{SIDV+1}\leq i\leq L-1\end{cases} \quad (10)$$

The input image is preprocessed by SIDCSIHE that provides the equalized image and its intensity information helps to enhance the image contrast.

C. Masking Technique

In the masking approach, the reconstruction approximation coefficients and SIDCSIHE processed images are considered for masking formulation. The contrast of the reconstructed

low pass signal is enhanced by using SVD and it helps to avoid information loss and better visual result. The SVD decompositions of the reconstruction approximation (I_1) and SIDCSIHE images (I_2) i.e. :

$$I_1 = U1 \times S1 \times V1 \quad (11)$$

$$I_2 = U2 \times S2 \times V2 \quad (12)$$

Here U1, V1, U2, and V2 are the orthogonal matrices identifies as left, right singular matrices of I_1 and I_2 . S1 and S2 are sorted singular value of diagonal matrices, contain intensity information about the images I_1 and I_2. Any changes in the singular matrix will affect the image intensity. The most promising approach of the image contrast enhancement is altering the image intensity information stored in the singular value [20]. To modify the intensity information of the reconstructed approximation coefficients, the S1 matrix is altered by using proper weighting (w) value. The highest singular value of I_1 and I_2 is used to calculate weighting value as shown in (13).

$$w = \frac{\max(S1)+\max(S2)}{2\,Xmax(S2)} \quad (13)$$

The new reconstruction approximation RA is obtained by modifying the S1 matrix as given in (14).

$$RA = U1 \times (w \times S1) \times V2 \quad (14)$$

The formulated mask, which gives the residual intensity information between RA and reconstructed low pass signal. This mask image added to original image and output will be the contrast-enhanced image. The entire process of the method is shown in Fig. 3. The following steps the computational process of the proposed algorithm.

Step 1. Low contrast image was taken for processing.

Step 2. Equalize the image using standard intensity deviation based clipped sub image histogram equalization.

Step 3. Compute 1 level DWT to decompose the image into four sub-band.

Step 4. Perform reconstruction of the approximation coefficients.

Step 5. The SVD is applied to SIDCSIHE image and reconstructed approximation for getting $U1 \times S1 \times V1$ and $U2 \times S2 \times V2$ then max($S1$) and max($S2$) obtained.

Step 6. Calculate weighting value w by using (13).

Step 7. New reconstruction approximation coefficient is obtained by $RA = U1 \times (w \times S1) \times V2$.

Step 8. Subtract the RA from reconstruction approximate matrix [mask].

Step 9. Add the mask with original image to get enhanced output image.

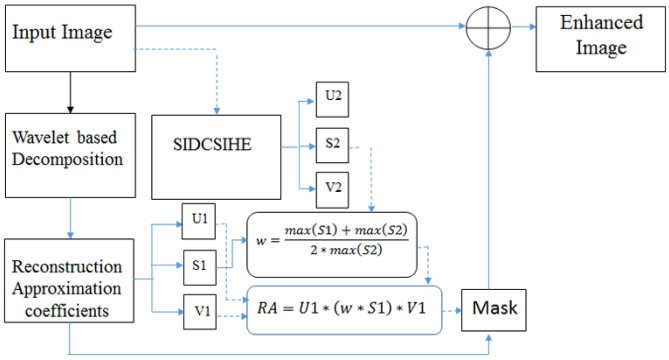


Fig. 3. Block diagram of the proposed method.

III. Experimental Results

The supremacy of the proposed method is illustrated by comparing both qualitative and quantitative analysis with well-known existing methods like HE, BBHE, DSIHE, ESIHE. As far as qualitative analysis is concerned, the performance of the proposed method is judged by visual quality inspection. As far as quantitative analysis of the proposed method is measured in terms of AMBE, entropy, PSNR and SSIM values and are compared with other methods.

The proposed algorithm performance is evaluated using (15)-(18). Given that MN is the size of the image, I_{in} is input image, I_e the enhanced image, Max_{in} the maximum possible pixel value of the input image. The PSNR is representing a measure of the peak error between the input and enhanced image.

$$MSE = \frac{1}{(MN)} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(I_{in}(i,j) - I_e(i,j)\right)^2 \quad (15)$$

$$PSNR = 10\ log_{10}\left(\frac{{Max_{in}}^2}{MSE}\right) \quad (16)$$

$$AMBE = |I_{in} - I_e| \quad (17)$$

Entropy means average information content and is defined in (18).

$$Entropy(p) = -\sum_{k=0}^{L-1} p(k) \log\ p(k) \quad (18)$$

Where $p(k)$, is probability density function at the intensity level k and L is the total number of gray levels of the image.

A. Qualitative Analysis

The performance of the proposed method is well analyzed with medical images (Brain_MRI, Face X-ray, MRI, Ribs X-Ray) [1] and mammogram images (mdb011, mdb013, mdb209, mdb211, mdb213)[2].

Fig. 4(a-f), (g-l), (m-r), (s-x) are the results of the proposed method steps from the input image to the contrast-enhanced

[1] https://medpix.nlm.nih.gov/home
[2] http://www.mammoimage.org/databases

image of (brain_MRI, Ribs X-Ray, mdb013, mdb212). The enhancement provided by SIDCSIHE method as shown in Fig. 4(b, h, n, t). Fig. 4(c, i, o, u) are reconstructed approximation coefficients, it contains necessary illumination information. The scaled reconstructed LL subband based SVD method can be seen in Fig. 4(d, j, p, v). The mask image contains residual intensity information clearly observed in Fig. 4(e, k, q, w). In Fig. 4(q), (w) shows the edge information and extracted intensity residual information. Fig. 4(f, l, r, x) are resultant, contrast-enhanced image with better visual quality than the original image.

The pleasant effect in the appearance of the contrast-enhanced image can be seen in the face X-ray image as in Fig. 5. The HE, BBHE, DSIHE are over enhanced. The proposed image has good contrast enhancement result over ESIHE image.

The MRI image in Fig. 6 is the proposed method that yields contrast-enhanced image. The HE, DSIHE shows over-enhancement and BBHE is dark in nature.

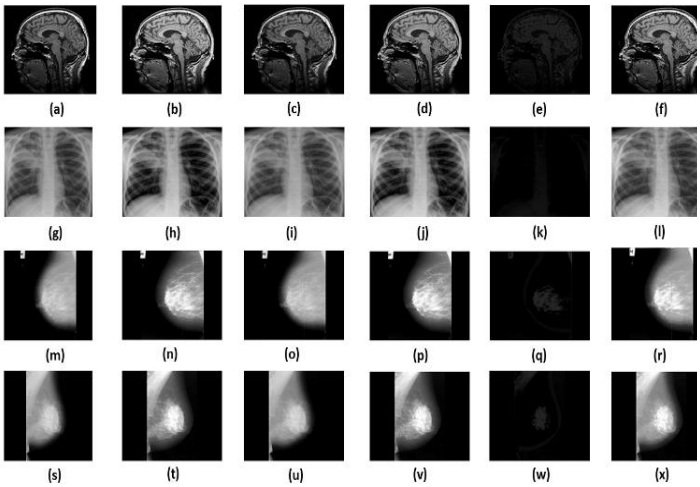


Fig. 4. Results of different stages of proposed method (a,g,m,s) input images; (b,h,n,t) SIDCSIHE ; (c, i, o,u) Reconstructed Approximation coefficient; (d,j,p,s) RA image (e, k,q,w) Mask image; (f,l,r,x) Output image.

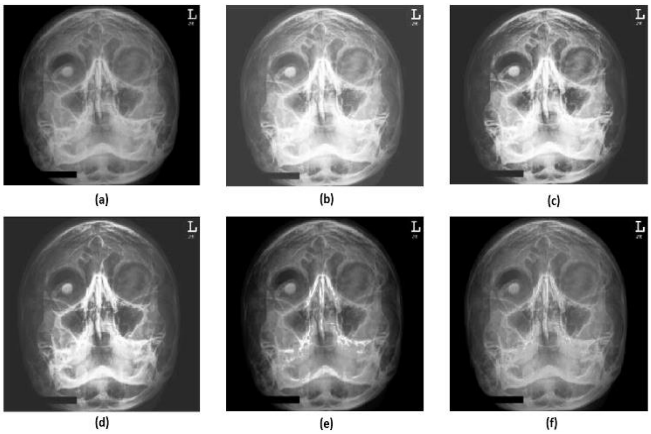


Fig. 5. Results of different methods for the Face X-Ray image (a) Low contrast input image; (b) HE; (c) BBHE (d) DSIHE; (e) ESIHE; (f) proposed method.

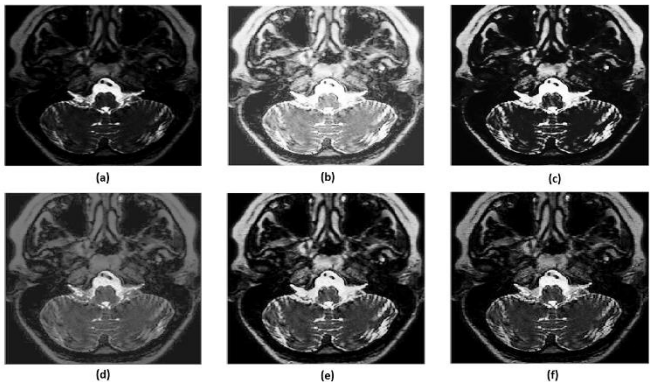


Fig. 6. Results of different methods for the MRI image (a) Low contrast input image; (b) HE; (c) BBHE (d) DSIHE; (e) ESIHE; (f) proposed method.

The analysis of the mammogram image (Fig. 7 and 8) supremacy of the proposed method is to show normal fatty tissue, dense breast tissue and infected area clearly.

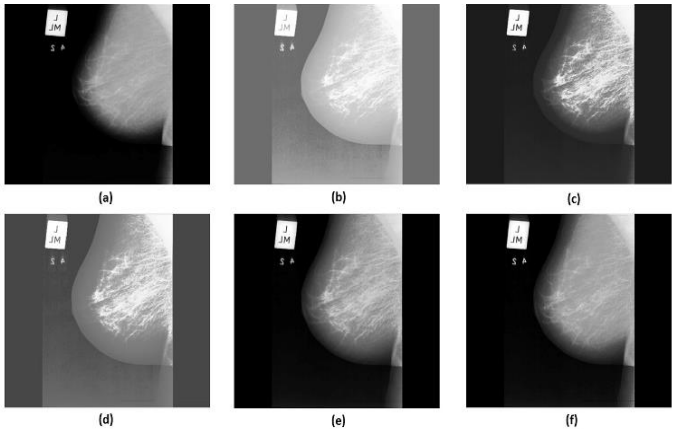


Fig. 7. Results of different methods for the mammogram (mdb011) (a) Low contrast input image; (b) HE; (c) BBHE (d) DSIHE; (e) ESIHE; (f) proposed method.

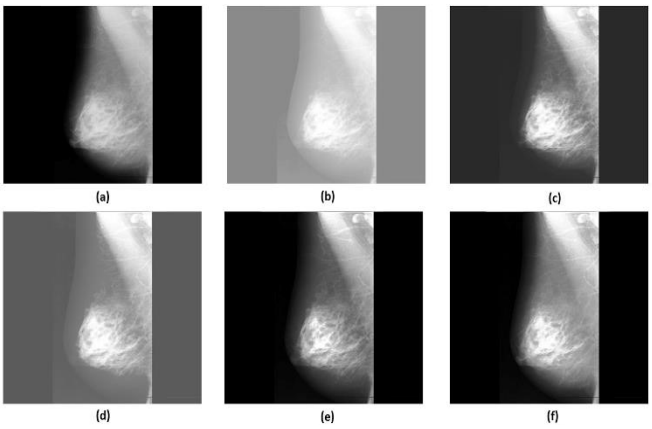


Fig. 8. Results of different methods for the mammogram (mdb211) (a) Low contrast input image; (b) HE; (c) BBHE (d) DSIHE; (e) ESIHE; (f) proposed method.

As compared to other methods in Fig. 7, proposed method clearly highlights the required region. The HE and DSIHE images are over enhanced. The dark image obtained by BBHE method. In Fig. 8, HE, BBHE and DSIHE are over enhanced

and proposed method image provided acceptable visual information than the input image.

B. Quantitative Analysis

To evaluate the performance of the proposed method, quantitative analysis has studied in terms of AMBE, entropy, PSNR, and SSIM.

The quality of the proposed method is provided by AMBE values, which is tabulated in Table I. The average AMBE value lesser than other contrast enhancement technique shows the quality of the contrast improvement. The entropy values of all the input image are tabulated in Table II. The highest entropy value indicates more information content of the image. As compared to HE, BBHE, DSIHE, ESIHE, the proposed has greater entropy value.

The SSIM values also encourage the performance of the proposed method is shown in Fig. 9. To check noise artifacts produced and over enhancement, PSNR values of the enhancement results are compared and shown in Fig. 10. The average PSNR value produced by the proposed method for all images is greater than HE, BBHE, DSIHE, and ESIHE.

TABLE I. AMBE VALUES OF THE COMPARED METHODS FOR CONTRAST ENHANCEMENT

Test Images	HE	BBHE	DSIHE	ESIHE	PROPOSED
Brain_MRI	89.529	29.611	31.566	24.019	**13.600**
Face X-Ray	57.059	36.221	22.554	5.242	**1.987**
MRI	103.587	27.551	42.799	32.576	**18.754**
Ribs X-Ray	11.601	9.941	4.390	2.860	4.498
mdb011	103.552	21.866	62.995	4.279	4.307
mdb013	118.487	27.789	74.376	6.045	**4.316**
mdb209	102.731	25.904	50.772	3.883	4.375
mdb211	112.495	27.558	68.075	4.377	4.402
mdb212	114.276	24.981	69.593	4.136	4.309
Average	90.369	25.713	47.458	9.713	**6.728**

TABLE II. ENTROPY VALUES OF THE COMPARED METHODS FOR CONTRAST ENHANCEMENT

Test Images	HE	BBHE	DSIHE	ESIHE	PROPOSED
Brain_MRI1	5.021	5.859	5.886	6.161	**6.656**
Face X-Ray	4.991	6.113	6.095	6.298	**6.522**
MRI	5.045	5.551	5.613	5.817	**6.205**
Ribs X-Ray	5.980	7.157	7.173	7.216	**7.449**
mdb011	3.742	4.371	4.318	4.587	**4.736**
mdb013	3.001	3.780	3.722	3.970	**4.125**
mdb209	3.842	4.758	4.667	4.943	**5.110**
mdb211	3.107	4.146	4.098	4.329	**4.648**
mdb212	3.223	4.038	4.022	4.240	**4.450**
Average	4.217	5.086	5.066	5.285	**5.545**

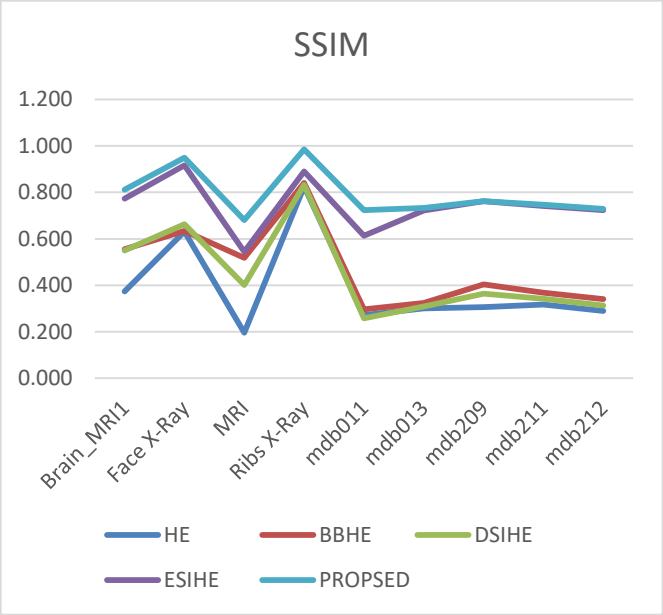


Fig. 9. SSIM values of the compared methods for contrast enhancement.

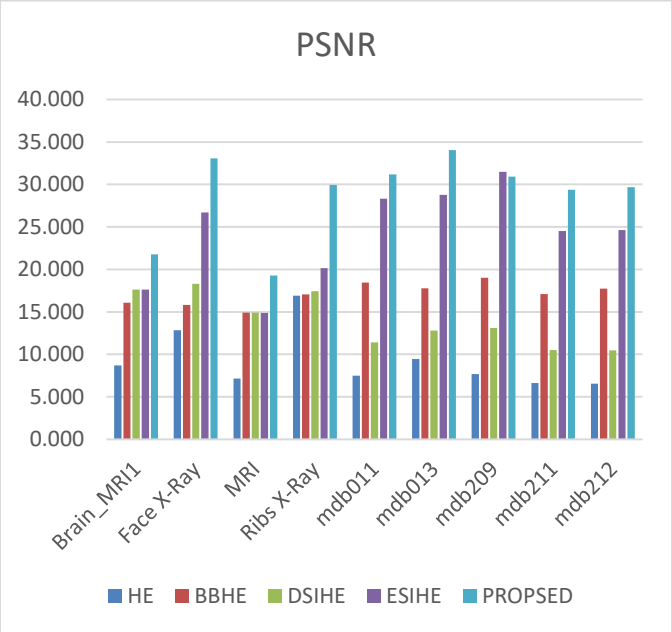


Fig. 10. PSNR values of the compared methods for contrast enhancement.

C. Summary of Analysis and Discussion

The objective of this paper is to obtain improved contrast enhancement by incorporating SIDCSIHE, DWT, SVD and masking technique. The main advantage of this paper is applying SVD technique effectively on SIDCSIHE image and reconstruction of LL image. The masking helps in providing residual information, which can be added to the original image for achieving maximum contrast enhancement. The proposed algorithm has been implemented in MatlabR2010, running on Windows PC with Intel i3 at 1.90GHZ and 4GB RAM.

After qualitative and quantitative analysis it can be concluded that

- Proposed method improves image contrast in comparison to other methods.
- The proposed method provides minimum AMBE vale, it shows the quality of contrast improvement.
- The proposed method provides the highest entropy among other methods, it shows the richness of details.
- Proposed method provide better PSNR and SSIM value, it shows over enhancement control and structural similarity.
- Proposed method produces an image with good contrast enhancement.

IV. CONCLUSION

In this paper, we tried to develop a method to address general issues of the image contrast enhancement. The proposed method effectively used masking approach to collect the sensitive edge information as residual intensity information. The method uses standard intensity deviation based clipped sub image histogram equalization to enhance contrast by controlling over enhancement. The SVD method helps to modify the intensity information. The experimental result shows the robustness of the proposed method as compared to the existing algorithm for the variety of images.

However, the proposed method does not consider the problem of noise enhancement, as the DWT method is shift variant. The work can be improved by using shift invariance transform along with the noise eliminating filter, to suppress the noise during enhancement. We have considered SVD to calculate scale value. The optimized algorithm can incorporate to find scale value. In this work contrast enhancement is achieved, this can be extended into resolution enhancement by using better interpolation algorithm.

ACKNOWLEDGMENT

This research work is supported by UGC-MRP, New Delhi, India.

REFERENCES

[1] Araujo, A.F.D. New artificial life model for image enhancement. *Expertsyst* **2014**, *41*, 5892-5906.

[2] Gonzalez, R.C.; Woods, R. *Digital Image Processing*, 3rd ed.; Pearson, INDIA, 2014, pp. 144–166.

[3] Y. T.Kim, "Contrast Enhancement using Brightness preserving bi-histogram equalization", IEEE Trans. Consumer Electron, **(1997**), 43: 1-8.

[4] Y.Wan, Q. Chen and B. M. Zhang. “Image enhancement based on equal area dualistic sub image histogram equalization method”, IEEE Trans. Consumer Electron, **1999**, 68-75.

[5] Kuldeep sigh, Rajiv Kapoor, "Image enhancement using Exposure-based Sub Image Histogram Equalization", pattern recognition letters 26 (2014) 10-14.

[6] Weeks A R, Sartor LJ, Myler H R. Histogram specification of 24-bit color images in the color difference(C–Y) color space. Proc.SPIE1999; 3646:319–29.

[7] Mukherjee J, Mitra S K.Enhancement of color images by scaling the DCT coefficients. IEEE Trans ImageProcess 2008; 17(10):1783–94.

[8] Mallat G.Theory for multi-resolution signal decomposition: the wavelet representation.IEEE Trans Pattern AnalMach Intell1989;2(7):674–94.

[9] Bhandari A K, Gadde M, Kumar A, Singh, G K.Comparative analysis of different wavelet filters for low contrast and brightness enhancement of multispectral remote sensing images. In: Proceedings of the IEEE international conference on machine vision and image processing(MVIP),p.81–6; 2012.

[10] Bhandari A K, Kumar A, Padhy P K.Enhancement of lowcontrast satellite images using discrete cosine transform and singular value decomposition. WorldAcad Sci EngTechnol2011; 79:35–41.

[11] Guang Deng. A Generalized Unsharp Masking Algorithm. IEEE TRANSACTIONS ON IMAGE PROCESSING**May 2011**, *20*, 123–126.

[12] Ching Chung Yang. A modification for the mask-filtering approach by superposing anisotropic derivatives in an image. Optik - International Journal for Light and Electron Optics**Sep 2011**, *122*, 1684–1687, issue 18.

[13] Polesel, A.; Ramponi,g.; Mathews, V.J. Image enhancement via adaptive unsharp masking. ”, IEEE Transactions on Image Processing **Mar 2000**, *9*, issue 3.

[14] V.S. Hari, V.P.J.Raj, R.Gopikakumari, Unsharp masking using quadratic filter for the enhancement of fingerprints in noisy background, PatternRecognit.46 (2013)3198–3207.

[15] Bhandaria A k, Sonia V. H, Kumaran A, Singhb G K.Cuckoo search algorithm based satellite image contrast and brightness enhancement using DWT–SVD. ISA Transactions **2014**, *53*, 1286-1296.

[16] Sandeepa K S, Basavaraj N Jagadale and J S Bhat, “Standard Intensity Deviation Approach based Clipped Sub Image Histogram Equalization Algorithm for Image Enhancement” International Journal of Advanced Computer Science and Applications(IJACSA), 9(1), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090116

[17] Sandeepa K S, Basavaraj N Jagadale, J S Bhat, Naveen Kumar R, Mukund n naragund, panchaxri, "Image contrast enhancement using DWT-SVD based masking technique", IEEE explore, (ICCES 2017), ISBN:978-1-5090-5013-0.

[18] Shin-Chia Huang and Chien-Hui Yeh. “Image contrast enhancement for preserving mean brightness without losing image features”.Engineering Application of Artificial Intelligence,(2013), 26:1487-1492.

[19] C.H. Ooi and N.S.P. Kong, “Ibrahmin, H. Bi-histogram equalization with plateau limit for digital image enhancement”, IEEE Trans. Consumer Electron ,**(2009),**55 (4), 2072-2080.

[20] H. Demirel, C. Ozcinar, G. Anbarjafari, Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition, IEEE Geosci. Remote Sens. Lett. 7 (April (2)) (2010).

Arijo: Location-Specific Data Crowdsourcing Web Application as a Curriculum Supplement

Justin Banusing, Cedrick Jason Cruz, Peter John Flores, Eisen Ed Briones, Gerald Salazar, Rhydd Balinas, Serafin Farinas

Philippine Science High School – Western Visayas
Iloilo City, Philippines

***Abstract*—Smart devices are quickly becoming more accessible to the general public. With the proper tools, they can be used to supplement the work of educators. According to studies by Beeland Jr. and Roussou, learning through interaction has been considered to be effective by both students and teachers. This study aimed to develop an interactive curriculum supplement for smart devices in the form of a Location-specific Data Crowdsourcing Web Application (Arijo) which teaches students how to conduct experiments and upload their results to the internet for archival purposes. Arijo was developed with a combination of the Appsheet framework, Adobe Photoshop, and Google Maps. Three core functionalities were programmed: data input/output, data interpretation, and information dissemination. Arijo was able to perform its intended features, such as recording and displaying data within specific locations, along with displaying guides on how to conduct an experiment. Arijo was able to fulfill its main objective, to be a curriculum supplement, through the aforementioned features. In the future, Arijo may be expanded to support more year levels and multiple curriculums because of its modular nature.**

Keywords—Web application; smart devices; data crowdsourcing; curriculum supplement

I. Introduction

Science, Technology, Engineering and Math or STEM, is an educational discipline focusing on the four aforementioned fields. Apart from promoting and exposing students to those fields, teaching STEM has been shown to have a positive impact on both students and educators alike. This includes teaching the efficiency and inquiry skills required to succeed in STEM-related endeavors [1].

A study by Estonanto [2] showed that there was a low acceptability of the new Philippine STEM curriculum among stakeholders. It revealed that the major problems were on the areas of the Facility and Instructional Materials, and the difficulty of coursework. He concluded that there was much to be improved about existing learning materials.

New ways can be added to the area of the Facility and Instructional Materials to promote STEM, one of which is interactive learning. There exist studies that prove that interactive learning benefits both students and teachers. Beeland Jr. [3] studied the effects of interactive whiteboards in classrooms, the purpose of which was to determine how student engagement was affected by using interactive tools throughout the learning process. The results showed that the aforementioned tool resonated with learners and lecturers alike while also increasing student engagement.

Nowadays, the technology present in the interactive whiteboards of old can be found in everyday smart devices such as phones, laptops, and tablets. Smart devices are hardware that possess the ability to perform "ubiquitous computing", which gives them the ability to run complex programs known as applications. These programs have multiple uses, one of which is data input and output. Using sensors commonly found in contemporary smart devices, some applications can find location-specific data such as GPS coordinates and ambient air temperature. Using the internet, some applications are able to disseminate information like news, weather forecasts and stock market statistics.

Other applications even involve data crowdsourcing, the act of gathering data from the public. Data gathered via crowdsourcing is valuable for researchers, businesses, and the public alike, as they can help predict trends and identify potential solutions to problems. Location-specific data are among these, it being information related to a specific locale. An example of an application that primarily makes use of crowdsourced location-specific data is Waze, which uses traffic information sourced from public repositories to determine optimal travel routes.

However, with the existence of multiple smart devices and platforms of differing specifications to develop for, the feasibility of developing a platform-specific native application is low when faced with limited time and resources.

This is why an alternative exists in the form of the web application. Utilizing a user's web browser to run the program from a remote source, little to no data is tied to the device itself. Web applications have the same functionality as their offline counterpart but have the ability to work across different platforms so as long as there is internet access. Examples of popular web applications are Yahoo and Google's mail clients [4].

There are a number of ways to develop web applications. Just like native applications, they can be developed using tools such as application programming interfaces (API) and frameworks. These tools vary in complexity, difficulty and capability [5]. Appsheet is one such framework. It is designed to allow developers to create spreadsheet-based web applications for the primary purpose of data crowdsourcing.

Using Appsheet and data crowdsourcing, Arijo (based on the Hiligaynon phrase "Ari oh", which means "here it is") was programmed to be an application that will supplement the STEM discipline by integrating smart devices into activities already done with the current Philippines curriculum.

II. Review of Literature

A. Learning through Interactive Content

Several studies have proven that interactive learning is highly preferred by students and teachers. One such study by Beeland Jr [3] measured students' engagement when their teachers used interactive whiteboards. The study's purpose was to determine the effect of the use of interactive whiteboards to see if the student engagement in the learning process increases while using the interactive whiteboard as an instructional tool. The majority of participating students and teachers indicated that they had a strong preference for the use of the interactive whiteboards in classrooms.

Roussou [6] developed an interactive, participatory, and multi-sensory environment where the combination of physical and virtual realities made for a more exciting learning experience for children. Those who tested her free flowing, interactive learning environment were found to have an increased interest in learning. Roussou stated that her tests turned out the way they did because learners in an activity gain their knowledge by testing ideas and concepts based on their stock knowledge and experience and applying them to new situations, something that interactive environments encourage.

B. Smart Devices

Smart devices are a form of hardware defined as having the ability to perform "ubiquitous computing". These include smartphones, personal computers, tablet computers, wearable computers and the like. The definition encompasses devices that have certain components. First are the power components, the devices' power sources. Second are the memory components that give devices the ability to read and write data. Third are the processing components, allowing devices to execute operations. Lastly, communication interfaces embedded within enable them to communicate with other devices of varying distance [7].

C. Appsheet

Appsheet is an application programing interface (API) that allows users to create hybrid applications on both Android and iOS devices. The API creates form based applications and allows users to store the data to Google Drive, Dropbox, Smartsheet, OneDrive, Microsoft Office, Box, and SQL Server. It also allows users to utilize different phone functions such as GPS, camera, barcode reader, and signature capture.

D. Web Application

Web applications are applications invoked through an internet connection. Within a decade since its mass adoption in 1994, the Web has evolved from a repository of pages for static information into a platform capable of application development and deployment. Creating dynamic applications with the help of new Web technologies, languages, and methodologies represent a new model of cooperation and collaboration among large number of users. Web application development quickly adopts software engineering techniques of component orientation and standard components. For instance, search, syndication, and tagging have become standard components for the new generation of collaborative applications and processes.

Web application development in the future will be driven by browser technology advances, Web infrastructure, protocol standards, software engineering methods, and application trends [8].

E. Use of Smart Devices as Data Gathering Tools

There is an approach in data gathering called participatory sensing where the data is collected by the users and sent through their mobile smart devices. With the increasing commonality of mobile smart devices, such an approach has become viable. Participatory sensing as a concept utilizes a mobile smart device's GPS, camera, microphone, clock, magnetic sensors and other features to collect and send data [9]. However, not all smartphones have the required apparatus to collect complex forms of data; as such, external apparatus is needed in certain cases.

With the rise of the number of people using smartphones today, it is a great to use this opportunity to promote STEM to people and students.

F. Related Studies

The GLOBE Project similarly crowdsources scientific data through use of a web-based form for personal computers and a web application for mobile smart devices. It allows students, teachers and the general public to participate in data collection at predefined and verified points of interest such as schools. [10].

Epicollect is a similar data crowdsourcing application that collects data, scientific and otherwise, from users. Instead of having pre-defined data types however, all data entries are self-defined and associated with a user-made repository [11].

1) Comparison of Similar Studies

Arijo is similar to the other applications at the base level (see Table I). Just like the GLOBE Project and Epicollect, it is to be a data crowdsourcing application.

Arijo, however, is not meant to be a competitor to the applications above; its purpose is entirely different. However, similar to the GLOBE Project, Arijo uses GPS to assign user-submitted data to the appropriate location. They share several function and types of data. Arijo's differences however, lie in two areas: account verification and data collection. While the GLOBE Project requires prospective data collectors to have verified accounts (endorsed by the school) and restricts them to pre-defined data collecting sites, Arijo is to maintain a free-for-all system so as to avoid entry barriers and increase the possible user base. However, Arijo features verification in the form of teacher/supervisor codes.

TABLE I. COMPARISON OF ARIJO TO SIMILAR APPLICATIONS.

Feature	Arijo	GLOBE Project	Epicollect
Mobile Application	✓	✓	✓
Automatic GPS Location	✓	✓	✓
Device Clock	✓	✓	✓
Open Source Spreadsheet	✓	✓	✓
Device Camera	✓	✓	✓
Online Maps Integration	✓	✓	
PC Accessibility	✓	✓	✓
Requires Verification	✓	✓	
Self-defined Data Collecting Sites	✓		✓
Pre-defined Types of Data	✓	✓	

Epicollect shares a number of features with both the Arijo and the GLOBE Project but serves as a repository for user-defined data that fall under user-defined data types. Rather than having pre-defined types of data, it allows users to do so. As such, nothing is specific and everything must be handled on a case-per-case basis. Arijo is to be similar to it in that it allows users to define their own data gathering points and gather data but unlike it allows data gathering in any point and is not limited to predefined data collection sites.

G. Elements of an Accessible User Interface

When designing an application's user interface, there are techniques and patterns to be followed in order for it to be accessible [12].

1) Clarity

Clarity is the most important element of user interface design. The purpose of user interface design is to enable people to interact with the system by communicating meaning and function. If people cannot figure out how the application works, it will confuse people and will fail as a user interface.

2) Concision

Clarity in a user interface is important. However, one must not over-clarify. The developer needs to be aware of the size of the interface. An excessive amount of explanations will cost users more time in order to go through them.

3) Familiarity

Making a user interface familiar is making it appear to look like something that people have encountered before, sparing the users the time to find out how it works.

4) Responsiveness

Making a user interface responsive means making it fast enough that users will not have to wait too long for a response. Animations are a part of an interface's responsiveness.

5) Consistency

Consistency is the similarity in appearance of different parts of an application. Making a user interface consistent makes it easier for users to navigate the application.

6) Attractiveness

Attractiveness is making the application look pleasing. But one must not overdo the attractiveness of an interface for it can make it less pleasing instead. Attractiveness increases the appeal of the application.

7) Clarity

An application's efficiency makes it easier and friendlier to the users. An efficient application is one that performs tasks in the fastest and shortest way possible.

8) Forgiving

A forgiving application is one that makes the users redo tasks when executed wrong. A great example of an application's forgivingness is the undo feature that lets the user redo the task that was performed.

H. Parameters Related to the Philippine School Curriculum

The data types and experiments supported by Arijo were based on the Philippine Department of Education's curriculum guide for grades three, four, and five science subjects (see Table II). There are learning objectives to be accomplished at the end of each grade level according to the aforementioned guide, all of which are satisfied by the data types and experiments supported by Arijo [13].

TABLE II. TABLE SHOWING THE THREE YEAR LEVELS' ASSOCIATED DATA TYPES

Year Level	Data Types
Grade 3	• Types of Water (Salt Water, Fresh Water) • Salinity
Grade 4	• Water Parameters (Temperature, Acidity) • Soil Type • Dissolved Oxygen • Weather (Cloud Type)
Grade 5	• Rock Type

1) Grade 3

According to the DepEd, learners should be able to describe the functions of the different parts of their environment by the end of grade 3, specifically water. Arijo's first featured data type in the grade 3 level is water type, which is part of that year level's goal of teaching students how to classify matter within the first quarter grading period.

a) Water Types

Freshwater is water that is collected from freshwater bodies such as lakes, rivers, falls and others. Lakes, rivers, and wetlands are examples of bodies of freshwater [14].

Saltwater is water that is collected from the sea and ocean. They are known for having a salinity of 35,000 ppt and higher.

2) Grade 4

By the end of grade 4, DepEd states that learners should be able to investigate the observable changing properties of materials when mixed with other materials or when there is an

applied force to it. Learners should be able to classify the different types of soil and which is best for certain plants and infer the importance of water in daily activities. Learners should also apply their knowledge of what makes up the weather and weather conditions.

Arijo's featured data types in the grade 4 level are water temperature, water acidity, dissolved oxygen in water, weather, soil type, and cloud type. These data types are part of that year level's fourth quarter grading period, where learners are taught about the importance of water, soil and weather parameters.

a) Water Parameters Affecting Plants

Temperature is the amount or intensity of heat present in a substance or object. Monitoring temperature is important because it influences water chemistry. The rate of chemical reactions generally increases at higher temperature. Water with higher temperatures can dissolve more minerals from rocks and will therefore have a higher electrical conductivity. It is the opposite when considering a gas, such as oxygen, dissolved in water. Water temperature is measured using a thermometer [15].

Acidity or pH is the measurement on how acidic or basic water is. It is the number of free hydrogen (H) and hydroxyl (−OH) in the water. It is measured by using a water parameter sensor, using reagents or litmus paper.

Dissolved oxygen is the level of free, non-compound oxygen present in water. It is measured by using modern dissolved oxygen sensors, colorimetric method, or titration [16].

b) Types of Clouds

Clouds are formed by tiny ice crystals. There are four basic types of clouds. They are: cirrus, cumulus, nimbus, and stratus. They usually represent the incoming weather of a certain place.

c) Soil Investigation

Soil investigation is investigating and observing of the soil to know important details about the soil.

Soil type is the general classification of soil depending on its consistency, location, color and other factors. The type of soil can affect what plants can grow in an area as well as how plants grow in an area. Soil type is identified by referring to a guide and comparing the characteristics of a soil sample to established archetypes. Examples of soil types are loam, clay, silt and sand

3) Grade 5

By the end of the 5th grade, students under the DepEd curriculum are expected to be able to infer that properties of materials may form new materials due to certain conditions. Arijo's featured data type in the grade 5 level is rock type, the determination of which is part of that year level's 4th quarter activities on changes to the earth [13].

a) Rock Types

Main rock types that are commonly used as basis for identification are these three types namely igneous, metamorphic, and sedimentary. These types may only be found in specific locations and possess unique characteristics; as such, can help scientists determine the nature of a location.

III. Methodology

This study aimed to develop a Location-specific Data Crowdsourcing Web Application (Arijo) using *Appsheet* as an API and Google Sheets as the data repository.

When developing software, the programmer must ensure that all of an *application*'s components work correctly. In order to do so, the development of the *web application* was divided into multiple phases.

Starting from a barebones data input and output application, features were added phase-by-phase. At the end of each phase, a form of evaluation was done to determine if the implementation of new features was successful (see Fig. 1).

Arijo's data flow and operation was aimed to be straightforward. Once a user begins the data submission process, the system simply follows the program flow (see Fig. 2).

A. Materials

- Personal Computer
- Internet Access
- Android OS Mobile Phone
- Apple iOS Mobile Phone
- Appsheet Account
- Google Account
- Google Sheets

B. Diagrams

1) Timeline of Developmental Phases

Development was done in phases (see Fig. 1) with evaluation done after each phase.

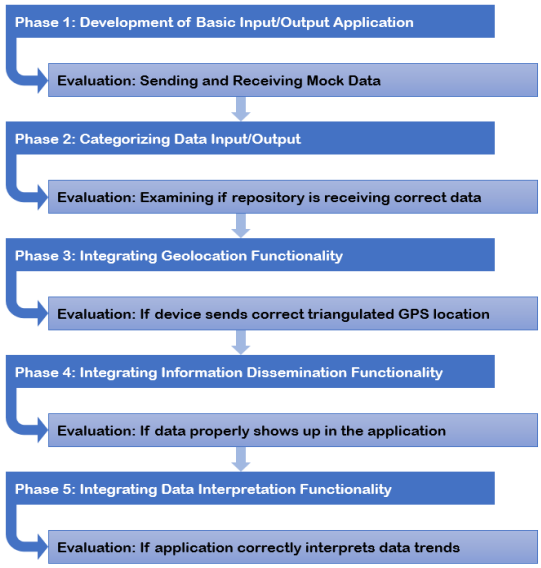


Fig. 1. Timeline of developmental phases for Arijo.

2) Application Operation Flowchart

The flowchart (see Fig. 2) shows Arijo's operation and data flow.

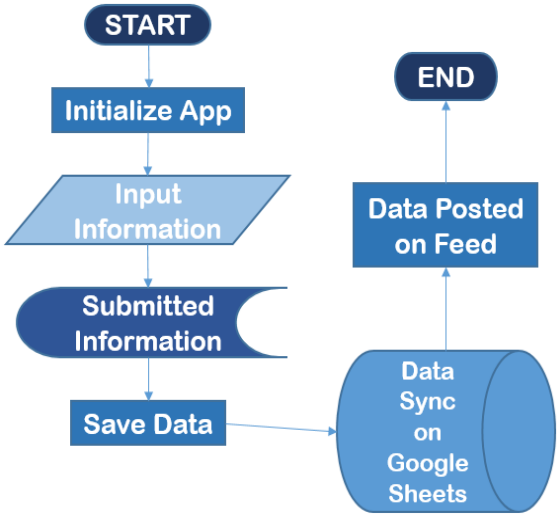


Fig. 2. Application operation flowchart of Arijo.

C. Procedures

1) User Interface Design

User interface is a key concept when developing an *application*. Regardless of an *application*'s function, its *user interface* must follow a set of standards in order to be accessible. Before beginning work on Arijo proper, a *user interface* following such standards was designed so as to streamline development by providing a mockup of what the final product would look like (see Fig. 3).

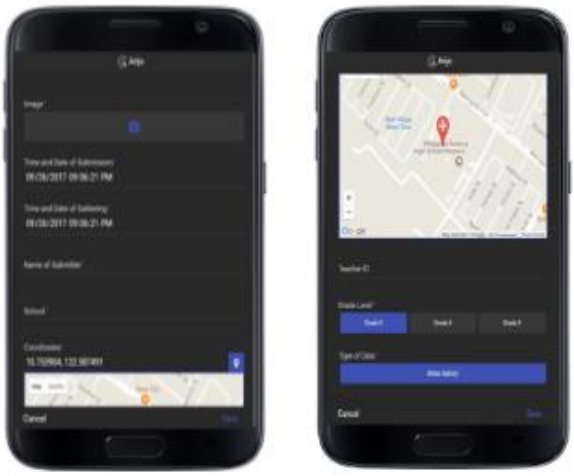


Fig. 3. Mock-up of the input (worksheet) page's user interface.

2) Development

a) Phase 1 – Development of a Basic Input/Output Application

Arijo's initial phase was a basic input/output *web application* using the *Appsheet* API. It was developed to send data from a user's device to the destination data repository, a Google Sheet (see Fig. 4 and 5).

Fig. 4. Input portion of Arijo

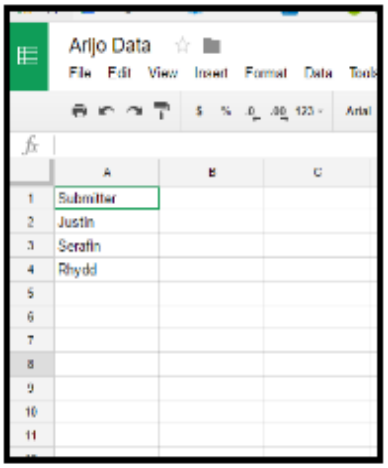


Fig. 5. Output portion of Arijo.

b) Phase 2 – Categorizing Data Input/Output

The second phase of Arijo's development was about presenting the data gathering procedures and categorizing data input and output.

Phase 1's basic input/output *web application* was overhauled to introduce selectors and input fields for the data to be collected as well as other important details like photographs and the submitter's name. These new forms were programmed by adding new fields to the Google Sheet and reconstructing the database in *Appsheet*. As a result, submitted inputs on the data fields are received by the Google Sheet.

User experience behavior was programmed so that data types can only be selected if a previous data field contained a specific input. When a grade level is selected, the only data types able to be chosen for entry are those associated with said grade level. After a data type is selected, it will be the only field present; all other data types will not appear (see Fig. 6 and 7).

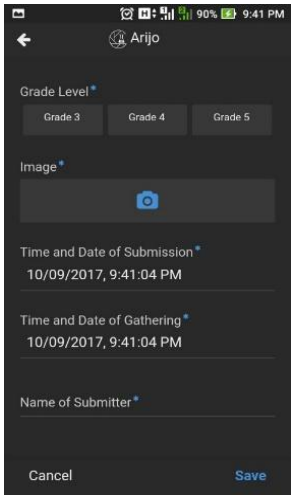


Fig. 6. A portion of the finalized input screen.

	A	B	C	D	E	F	G	H	I
1	Image	Time and Date of	Time and Date of	Name of Submitt	School	Coordinates	Teacher ID	Grade Level	Type of Data
2	Arijo Data_Image	1/15/2017	1/15/2017	Justin Banusing	PSHS-WVC	10.315699, 123.8	N/A	Grade 5	Rock Type
3	Arijo Data_Image	1/17/2017	1/17/2017	Rhydd Jay Balin	PSHS-WVC	10.315699, 123.8	N/A	Grade 4	Soil Type
4	Arijo Data_Image	2/11/2017	2/11/2017	Serafin Farinas	PSHS-WVC	10.315699, 123.8	N/A	Grade 3	Water Type
5	Arijo Data_Image	2/11/2017	2/11/2017	Serafin Farinas	PSHS-WVC	10.315699, 123.8	N/A	Grade 4	Water Temperatu
6	Arijo Data_Image	2/11/2017	2/11/2017	Serafin Farinas	PSHS-WVC	10.315699, 123.8	N/A	Grade 4	Water Acidity
7	Arijo Data_Image	2/11/2017	2/11/2017	Serafin Farinas	PSHS-WVC	10.315699, 123.8	N/A	Grade 4	Dissolved Oxyge
8	Arijo Data_Image	2/11/2017	2/11/2017	Serafin Farinas	PSHS-WVC	10.315699, 123.8	N/A	Grade 4	Weather
9	Arijo Data_Image	3/7/2017	3/7/2017	Serafin Farinas	PSHS-WVC	14.041000, 15.4	N/A	Grade 3	Water Salinity
10	Arijo Data_Image	3/7/2017	3/7/2017	Serafin Farinas	PSHS-WVC	10.753634, 122.8	N/A	Grade 5	Rock Type
11	Arijo Data_Image	4/16/2017	4/16/2017	Farinas	PSHS-WVC	10.753327, 122.8	N/A	Grade 5	Rock Type

Fig. 7. A portion of the finalized output screen.

c) Phase 3 – Integrating Geolocation Functionality

In Arijo's third phase, geolocation or the ability to detect a user's location was added. Appsheet's GPS module was enabled, thus allowing the web application to interface with a phone's GPS sensor when queried by the input form. Arijo is able to automatically add GPS coordinates to a user's data submission, though a user can manually select the location should they be submitting data retroactively (see Fig. 8).

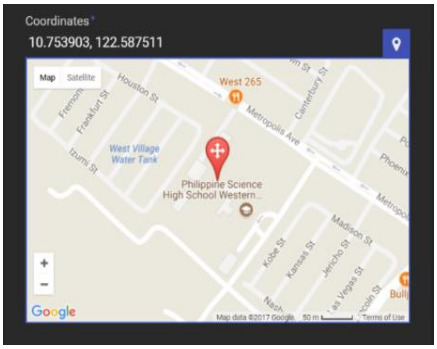


Fig. 8. Coordinate input are of Arijo.

Using *Appsheet*'s Maps view module, which itself is powered by Google Maps, a new portion was added to the *web application* wherein the submitted GPS coordinates will be shown as points on the world map. This creates create a graphic representation of where data was submitted from (see Fig. 9).

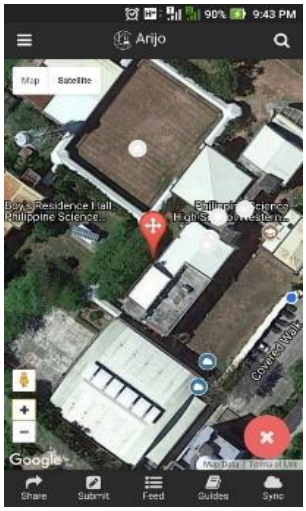


Fig. 9. Data map of Arijo.

d) Phase 4 – Integrating Information Dissemination Functionality

An encyclopedia-like section in the application was added using *Appsheet*'s gallery tool. Each experiment's procedure was conveyed through images designed using Adobe Photoshop (see Fig. 10).

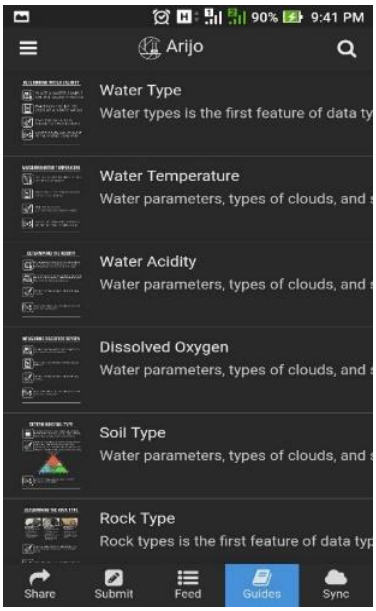


Fig. 10. Procedure selection menu.

e) Phase 5 – Integrating Data Interpretation Functionality

Data interpretation was added as Arijo's fifth phase. Every type of data in a location is shown on a page dedicated to said location. These can be selected from the application's 'data feed' page (see Fig. 11). Each page contains the location of input, name of submitter, data type and data value (see Fig. 12). Data is sourced from Arijo's Google Sheet (see Fig. 13).

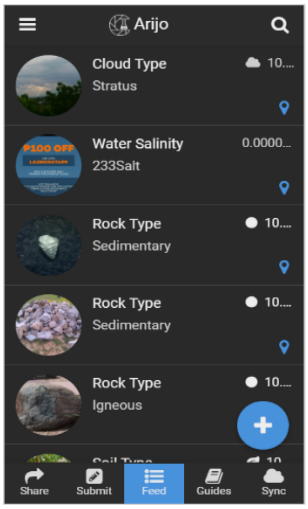


Fig. 11. Live data feed tab.

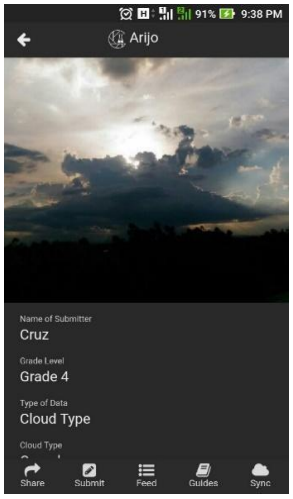


Fig. 12. Example of an interpreted data page.

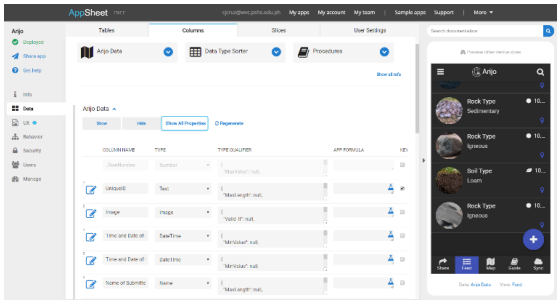


Fig. 13. Appsheet programming interface of Arijo after implementing all phases.

IV. Results And Discussion

This study aimed to develop a Location-specific Data Crowdsourcing Web Application (Arijo) using *Appsheet* to serve as a curriculum supplement by teaching students how to conduct experiments and making them upload their results to the internet for archival purposes.

It also aimed to design an accessible user interface for the web application, describe and record the developmental processes of the web application's data input/output, data interpretation and information dissemination functionalities, and test the web application in an open environment upon completion of development.

Using Arijo, users are able to submit data types supported by the application to a designated Google Sheet for archival purposes. Users are then able to view their submitted data as well as that of other users in a graphical form using the web application's data interpretation functionality. Data types supported by the application are based on experiments conducted in schools following the Philippine curriculum for grades three, four, and five. Users who do not know how to collect the aforementioned data types may learn to do so using the web application's information dissemination functionality.

Due to Arijo's nature as a web application, it is usable on any platform that supports the latest versions of the Microsoft Edge, Google Chrome, Apple Safari, and Mozilla Firefox web browsers. It is also installable and usable on mobile phones that run the latest versions of the Apple iOS and Google Android operating systems.

1) User Interface Design

Using the criteria defined in the review of literature, a user interface was designed. Adobe Photoshop was used to make mock-ups of the user interface; these mock-ups were then successfully implemented during the programming phase using *Appsheet's* built-in design tools. The interface allows users to input data, see submitted data by them or other users through the live feed or world map, and learn procedures. It also has a sync option that enables users to manually update their database should they lack internet access for an extended amount of time.

2) Information Dissemination

A section showing all the procedures needed to gather the data types that Arijo supports was programmed. Users may choose the procedure they aim to learn from the section's gallery-style selection menu. These procedures are taught through minimalistic graphics, which were designed using Adobe Photoshop and inserted into the application in the Portable Network Graphic (.png) image format. The section was programmed to be expandable should more data types be added.

3) Data Flow

Arijo collects data from user submissions and stores said data in a Google Sheet. Upon submission, pre-existing data in the Google Sheet is automatically re-sorted to accommodate the new addition. Arijo can also fetch data from the Google Sheet for data output purposes.

a) Data Input/Output

Users can input data through Arijo's form section. Multiple fields such as the value of the data, time of data gathering, location, name of the collector and other specific information will be filled up by the collector. Through Geolocation and interfacing with the phone respectively, location of data gathering and time of data submission will be filled up by the web application automatically. All submitted data are stored in the Google Sheet.

b) Data Interpretation

Users are able to view data submitted by them and other users. In the data feed section, users can see the data in descending chronological order of submission. They may also conduct a filtered search in order to find data of specific categories.

Another way of viewing data is through the map section. On a world map parsed from Google Maps, users can view data according to where it was gathered and submitted from. Each submission is given a marker on the map; by selecting a marker, details about its corresponding submission can be seen.

4) Testing

Arijo was tested to see if it installs properly, if the data input and output functionality works as intended, and if data from the Google Sheet is displayed and interpreted properly.

a) Field Test

Arijo was tested on both Android and Apple devices. The application was also tested in a desktop computer. Sample mock data was sent to see if the data is properly uploaded and downloaded. The GPS function was in line with a user's phone's reported locations on Google Maps. Likewise, the date and time parsing was in line with the time displayed on a user's phone.

B. Discussion

Arijo had all of its aimed core functionality successfully implemented. This included:

- Data Input/Output: Allowing users to submit and store supported data types on Arijo's Google Spreadsheet through its in-app form, as well as automatically fetching where and when users submitted their data via their phone's geolocation and internal clock functionality respectively.
- Data Interpretation: Allowing users to view data submitted by them and other users in either live feed or map form.
- Information Dissemination: Teaching users how to gather the supported data types.

Area restriction code that limited the data collection to users from the Philippines was not implemented. This was because defining a specific GPS coordinate range for the Philippines is infeasible due to its archipelagic nature; an implementation of this feature would have been inaccurate at best.

Google Maps' points of interests were not removed. This is because the Appsheet API cannot filter what Google Maps feeds the application. This is not considered to be a significant problem as the points of interest may make the data points easier to locate.

Instead of the initial plan of creating an individual form for every data type, the data input forms were organized by using dropdown menus. It gave the application a compact aesthetic compared to the concept.

Only a select amount data types were included in this iteration of the web application. These were taken from the Philippine Department of Education's official curriculum. This does not limit the application's potential however, as additional data types can be added in the future.

V. Summary, Conclusion and Recommendations

This study aimed to develop a Location-specific Data Crowdsourcing Web Application (Arijo) to serve as a curriculum supplement by teaching students curriculum-based science experiments and giving said students a way to archive the results online in a public repository. It also aimed to document the development of the web application's various functionalities.

A. Procedures

Arijo had the following features implemented:

- Data input/output with geolocation functionality through an in-app form and Google Sheets.
- Data interpretation through the data feed and world map view.
- Information dissemination through the web application's procedures page.

B. Conclusion

The development of a Location-specific Data Crowdsourcing Web Application (Arijo) as a Curriculum Supplement is possible using a combination of Appsheet, Google Sheets, and Adobe Photoshop. All of its intended core functionalities were successfully implemented.

C. Recommendations

Procuring a faster and stable internet connection is essential as it is needed in order for the system to display its data regularly and effectively refresh every time new data is sent to the online document.

A more complex API is also recommended. While Appsheet is a suitable solution for medium-scale projects such as Arijo's current state, features such as increased customizability, complex animations, advanced data interpretation functionalities are not possible.

Lastly, future researchers are encouraged to think of methods to make the user interface and experience engaging. As students are the target audience, the aesthetics of applications like Arijo should cater to them.

Acknowledgment

We would like to thank our parents and teachers for supporting us while we developed Arijo. Specifically, we thank our advisors Eisen Ed Briones and Gerald Salazar for lending us their time and expertise for the past year and a half. We also thank the Philippine Science High School – Western Visayas Campus Research Unit, including but not limited to Dr. Aris Larroder and Jarold Mediodia, for guiding us throughout our specialization years. Lastly, we thank Serafin Farinas and Rhydd Balinas for their contributions towards the development of Arijo's prototype.

REFERENCES

[1] Yoon SY, Dyehouse M, Lucietto AM, Diefes-Dux HA, Capobianco BM. 2014. The Effects of Integrated Science, Technology, and Engineering Education on Elementary Students Knowledge and Identity Development. School Science and Mathematics [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/ssm.12090/abstract.

[2] Estonanto A. 2017. Acceptability and Difficulty of the STEM Track Implementation in Senior High School [Internet]. Available from: http://www.apjmr.com/wp-content/uploads/2017/04/APJMR-2017.5.2.05.pdf.

[3] Beeland W. Jr. 2002. Student Engagement, Visual Learning and Technology: Can Interactive Whiteboards Help? [Internet]. Available from: https://s3.amazonaws.com/academia.edu.documents/38455890/COOOOL.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1507301802&Signature=rhnKR2sRncQfYzmDUNmnMO8FjdQ%3D&response-content-disposition=inline%3B%20filename%3DStudent_Engagement_Visual_Learning_and_T.pdf.

[4] Nations, D. 2015. What is a Web Application? About [Internet]. Available from: http://webtrends.about.com/od/webapplications/a/web_application.html.

[5] Christensson, P. 2016. API Definition. TechTerms. Sharpened Productions [Internet]. Available from: http://techterms.com/definition/api.

[6] Roussou M. 2004. Learning by Doing and Learning Through Play: An Exploration of Interactivity in Virtual Environments for Children [Internet]. Available from: http://nexus.hs-bremerhaven.de/Library.nsf/a249ddae15ac617ac12573460029d00b/3a857d47de4a673bc125733e006d1061/$FILE/p10-roussou_Learn_through_Play.pdf.

[7] Davy A. n.d. Components of a smart device and smart device interactions [Internet]. Available from: http://www.m-zones.org/deliverables/d234_1/papers/davy-components-of-a-smart-device.pdf.

[8] Jazayeri M. 2007. Some Trends in Web Application Development [Internet]. Available from: https://dl.acm.org/citation.cfm?id=1254719.

[9] Kanhere SS. 2011. Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces. 2011 IEEE 12th International Conference on Mobile Data Management. 2011:3–6.

[10] Grove C. 2018. Introduction to the GLOBE Research Project on Leadership Worldwide [Internet]. Available from: https://www.grovewell.com/wp-content/uploads/pub-GLOBE-intro.pdf.

[11] Aanensen DM, Huntley DM, Feil EJ, Al-Own F, Spratt BG. 2009. EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection. PLOS ONE.

[12] Fadeyev, D. 2009. 8 Characteristics Of Successful User Interfaces [Internet]. Available from: http://usabilitypost.com/2009/04/15/8-characteristics-of-successful-user-interfaces/.

[13] Department of Education. 2016. K to 12 Curriculum Guide Science [Internet]. Available from: http://www.deped.gov.ph/sites/default/files/page/2017/Science%20CG_with%20tagged%20sci%20equipment_revised.pdf.

[14] Environment Guide. 2015. Freshwater Bodies. [Internet]. Available from: http://www.environmentguide.org.nz/issues/freshwater/the-freshwater-environment/.

[15] Perlman, H. 2014. Water Properties and Measurements. USGS Water Science School. Available from: http://water.usgs.gov/edu/characteristics.html.

[16] Dissolved Oxygen. 2016. Environmental Measurement Systems. [Internet]. Available from: http://www.fondriest.com/environmental-measurements/parameters/water-quality/dissolved-oxygen

A Portable Virtual LAB for Informatics Education using Open Source Software

MiLAB

Ali H. Alharbi, PhD.
The Department of Health Informatics, College of Public Health and Health Informatics
Qassim University
Albukairiyah, Qassim

***Abstract*—The need for students to have hands-on experience is very important in many disciplines to match the requirements of today's dynamic job market. Informatics, which is the science of information engineering, has been recently integrated into many academic programs. Teaching students the main skills in modern software and web development is essential for them to be successful informatics professionals. For any informatics program, students engage in working on projects as essential parts for some courses in their academic programs. This paper presents the development and evaluation of MiLAB (My Mobile Informatics Lab), a portable virtual lab environment for the teaching and learning of modern web development skills. MiLAB has been integrated into an undergraduate health informatics academic program to improve the teaching and learning of essential web development skills, such as databases management and customization of modern content management systems. The evaluation of MiLAB indicated that it served as an interactive personal environment for students to implement, collaborate, and present their web development projects. Strengths, weaknesses and possible improvements are also discussed.**

Keywords—Virtual labs; open source software; e-learning

I. Introduction

It is essential for undergraduate informatics academic programs to include hands-on experience in order to prepare students for the real world jobs. Traditionally, academic institutions rely on physical computer labs that are equipped with essential software to teach students various skills on computing and informatics. However, with limited educational resources and dramatically increasing number of students, efficient and timely access to physical computer labs becomes more difficult. The use of information and communication technology to support education has led to the emergence of innovative paradigms for teaching and learning.

Simulation and virtualization tools provide effective and alternative approaches to help students gain essential informatics skills without the need to be in a physical computer lab [1]. However, this technology is not widely implemented by educators due to the difficulties associated with the integration into the learning environment [2]. One of the key success factors for the use of simulations in education is the integration of these simulations into the traditional teaching and learning environment. Virtual Labs can be defined as "E-learning platforms where learners can gain the experience of practical experimentation without any direct physical involvement on real bench work" [3]. Also, some studies have used the term virtual labs to refer to the concept of providing distance access to physical labs to support the learning in some disciplines, such as physics and engineering [4]. In computer science and informatics education, virtual labs can provide safe and customizable learning environments for students to gain hands-on experience on various skills, such as networking, programming and web development, at any time and place [5]-[7]. However, when it comes to cost, the current technology used in the design and implementation of virtualization and simulation does not come without a high cost [8].

Open Source Software (OSS) has achieved wide adoption in various domains, such as server technologies, networking, databases and enterprise tools. OSS refers to software that complies with the following criteria [9]:

- Free to redistribute.
- The source code is freely available for modification.

OSS leading products such as Apache, MySQL, along with programming languages, such as PHP, Java and Perl are among the technologies that have contributed to the success of the internet.

OSS plays important roles to improve teaching and learning in various disciplines, such as computer science, information systems and science [10], [11].

This paper gets insights into the educational effectiveness of open-source software in creating constructivist educational environments for teaching modern web development. To achieve this, the paper proposes a portable virtual learning lab, MiLAB (My Mobile Informatics Lab), that can be used as a simulated environment for informatics education. As a prove-of-concept, MiLAB was used to teach modern techniques in web development for undergraduate health informatics students. Health informatics is an emerging discipline that combines computer and information science along with health and biomedical science to support healthcare systems. Teaching health informatics requires that candidates gain practical experience on different information systems.

II. Related Work

A. Virtual Labs in Education

Different approaches have been proposed in the literature to design virtual labs for educational purposes in various academic disciplines. For example, in engineering education, various applications and methods have been proposed to use virtual and remote labs to enhance teaching and learning about different engineering topics, such as digital signal processing [12].

In computer science education, virtual labs have been used to augment teaching and learning in various courses. For example, Wu, et al. [13] presented a virtual computer lab that was designed to allow students to experiment with various information security approaches via a computer network constructed using virtual machines. The environment gives students the opportunity to apply various concepts in a simulated environment without compromising any computer network. Another approach was presented by Kumar, et al. [14] , in which a cloud-based virtual lab solution has been used to teach students various computer science concepts.

B. Open Source Software in Education

The open source movement has affected many aspects of our life today including education. As academic institutions move towards reducing cost and increasing efficiency, it is critical for educators to find alternative and efficient tools to supplement education. Open Source Software (OSS) is one of the viable alternative that academic institutions and educators should think about to achieve this vision. In higher education, based on open and free source frameworks, a number of notable educational software has emerged. Moodle [15] , the most well-known open source learning management system, has been extensively adopted in higher education and has an active and large community. It provides different pedagogical tools to help improve the quality of e-learning. Atutor [16] is another open source learning management system, which also gains increased interest and provides the basic functions required to design and manage e-learning contents.

Open source software plays important role to improve education [17]. For example, in computer science education, there are different computer-supported tools and educational resources available for students and educators. However, despite being perceived as educationally effective, these tools and resource are not widely adopted by computer science educators. Previous research, such as [18], [19], relates this problem to the fact that teachers are too busy to find and integrate these tools into the teaching environment, as well as they do not have time to teach students to learn how to use such tools.

In engineering education, Froyd, et al. [20] identified five major shifts that affected the discipline in the last few years. In the fifth shift, the authors pointed out that “technologies (e.g., the Internet, intelligent tutors, personal computers, and simulations) have been predicted to transform education for over 50 years” [19].

III. Portable Virtual LAB for Informatics Education (MiLAB)

This section describes in detail MiLAB, a proposed portable virtual lab for teaching and learning of informatics skills. MiLAB utilizes the state-of-the-art in open source software platforms to provide a portable, virtual, adaptive and personalized environment that can be used for teaching and learning of important web development skills. These skills include web development, content management, and designing and maintaining databases.

A lightweight server was deployed into a USB flash drive, with essential functionality so each student can experiment with a range of OSS platforms in a fully simulated environment. This USB flash drive serves as a personalized lab that is customized for each student's needs and preferences.

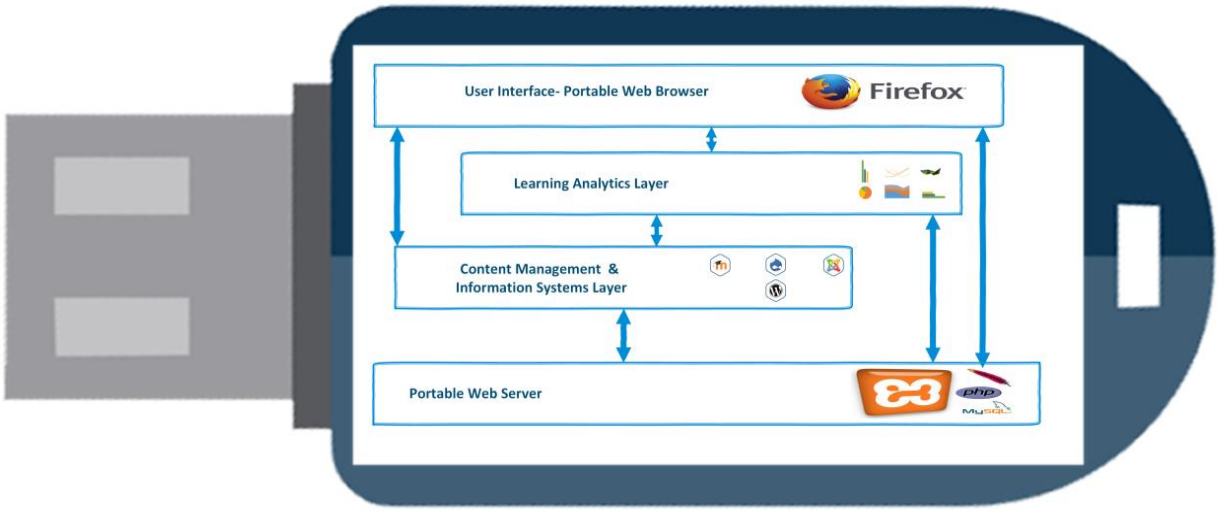


Fig. 1. Architecture of the Portable Virtual Informatics LAB (MiLAB).

The architecture of MiLAB, as depicted in Fig. 1, consists of the following components:

A. Portable Lightweight Web Development Server

The underlying server of this virtual lab is based on XAMPP [21] , a simple, lightweight, free and open source Apache web server distribution that makes it easy to setup and run local web development servers. XAMPP stands for Cross-Platform, Apache, MariaDB, PHP and Perl. It is a cross-platform in the sense that it can work under any operating system. XAMPP serves as a container that hosts all components of the virtual lab. It also provides a control panel for administrators to manage and monitor the lab.

B. Content Management and Information Systems Platforms

This layer serves as the underlying infrastructure to deploy and manage various open source content management systems, as well as other open source information systems. All these platforms share the same underlying server on which different databases can be created and linked to the open source platforms. Students can customize these platforms to meet needs and preferences. Plug-ins and additional modules can be installed on each system to enhance the functionality when needed.

C. Learning Analytics Module

One of the most important components of this virtual lab is a learning analytics module designed to provide insight into the learning process. This module provides a wide range of features to support monitoring of the educational process. The module allows educators and learners to generate different kinds of reports and statistics.

Analytics is a general term that focuses on the use of data and statistical analysis techniques to gain insight into specific issues. Learning Analytics (LA) is an emerging research area that focuses on the retrieval and analysis of data to support teaching and learning [22]. There has been an increasing interest in the analysis of data generated from students' interaction with various learning tools. One of the reasons behind this interest is the great potential for this analysis to discover students' behavior patterns that helps design and deliver adaptive learning contents.

IV. Evaluation of MiLAB In Practice: A Portable Virtual Platform for Health Informatics Students' Projects

A. Methodology

MiLAB was used to improve the teaching and learning of modern web development skills in an undergraduate academic program of Health Informatics. In this program, it is essential for students to be familiar with modern web development techniques, including the customization and management of modern information system applications. MiLAB was used as a portable virtual infrastructure for students to work on their projects individually and in groups.

In the first few weeks of the course, the focus was put on the installation and customization of the portable server. A number of lab sessions have introduced students to the technology of web servers and how to deploy a simple, local and portable web server in a USB flash memory. At the end of the sessions, students successfully installed the portable server and they were familiar with its control panel. In addition to this, students also learnt how to create and manage MySQL databases.

After that, students were introduced to a number of well-known content management frameworks to be used later for students' health informatics projects. In this phase, students learnt how to install and setup each content management system and customize it to support health informatics projects. This study adopted WordPress and Drupal, both of which are well-established and sophisticated content management systems.

The instructor guided students through the process of setting up a database for each content management system, and deploying each system to work as a portable application on the server. This phase covers different functionality of the content management system, including but not limited to, themes, menus, content editing, user management, and plugins installation.

The system was used from 2016 to early 2018 at the College of Public Health and Health Informatics, Qassim University in Saudi Arabia. The main objective of this study is to get insight into the potential benefits for the proposed portable virtual lab as a pedagogical tool to improve teaching and learning of modern web development, particularly in health informatics education. The main research questions that this study is intended to answer can be summarized in the following:

1) To what extent do students accept MiLAB as a pedagogical tool to improve learning and teaching?

2) What are the strengths, weaknesses and possible improvements for MiLAB from the students' point of view?

B. Study Participants

The subjects of the study were students from the undergraduate program of Health Informatics at College of Public Health and Health Informatics, Qassim University, Saudi Arabia. The study was conducted using patches of students in the period from 2016 to early 2018. Voluntarily, 64 students participated to complete the questionnaire, which is described in the section below.

C. Data Collection Instrument

A questionnaire, with both closed and open-ended questions, was administered to students at the end of the study.

The quantitative part of the questionnaire was adapted from the Unified Theory of Acceptance and Use of Technology (UTAUT) [23]. UTAUT is an extension to the original Technology Acceptance Model (TAM), an instrument used to predict users' acceptance of a technology on the basis of a number of constructs. UTAUT, which was originally developed by Venkatesh, et al. [23] , has been described as a cost-effective and easy-to-use questionnaire for predicting user's acceptance of different systems.

UTAUT assumes that user's acceptance of a technology is affected by a number of constructs. Six constructs were

adopted in this study. Table I describes these constructs as reported in [23].

TABLE I. DEFINITIONS OF THE CONSTRUCTS ADOPTED FROM THE UTAUT MODEL

Construct	Definition (in the context of MiLAB)
Performance Expectancy (PE)	The degree to which an individual believes that using MiLAB would enhance his/her learning.
Effort Expectancy (EE)	The degree to which an individual believes that MiLAB is easy to use.
Facilitating Conditions (FC)	The degree to which the user believes that technical infrastructure and resources are available to support the use of MiLAB.
Self-Efficacy (SE)	Judgment of one's ability to use MiLAB to accomplish a particular task.
Attitudes towards using the Technology (A)	The motivation and willingness to use MiLAB.
Behavioral Intention to Use the System (BI)	The user positive or negative intention to use MiLAB.

V. RESULTS AND DISCUSSION

This section outlines and discusses the results of students' responses to the questionnaire. Descriptive statistics techniques were applied to get insight into the extent to which students' support using MiLAB to enhance the teaching and learning environment.

A. User's Acceptance of MiLAB

Fig. 2 to 7 show the distribution of students' responses to each construct in the questionnaire using a 7-point Likert scale ranging from 1(strongly disagree) to 7 (strongly agree).

In the Performance Expectancy construct, as appears in Fig. 2, the vast majority of students (79%) strongly believed that using MiLAB would enhance their learning. They believed the system would enhance their productivity and help them accomplish tasks more quickly and efficiently.

Fig. 3 shows the distribution of students' responses to the Effort Expectancy construct. Half of students believed that MiLAB was easy to use for them. They believed they would be skillful at using the system. However, 13% of the students slightly disagreed and expected they will face difficulties interacting with MiLAB smoothly.

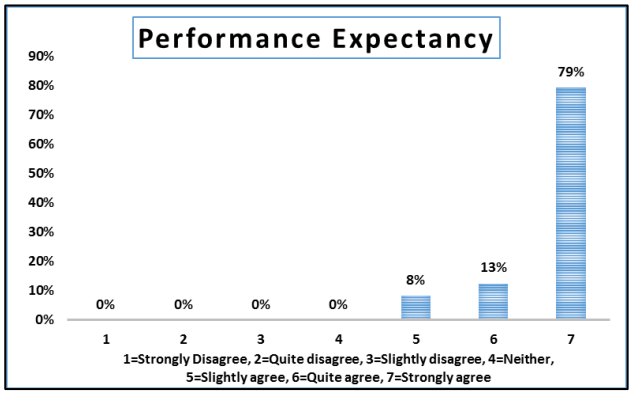


Fig. 2. The distribution of students' responses to the Performance Expectancy construct.

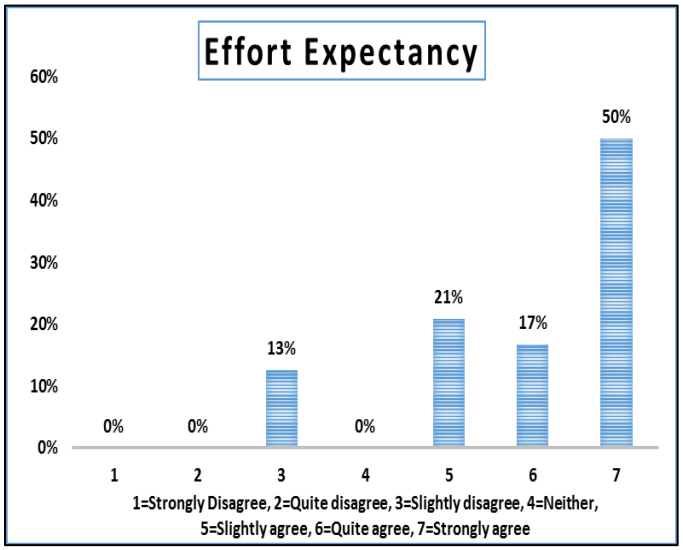


Fig. 3. The distribution of students' responses to the Effort Expectancy construct.

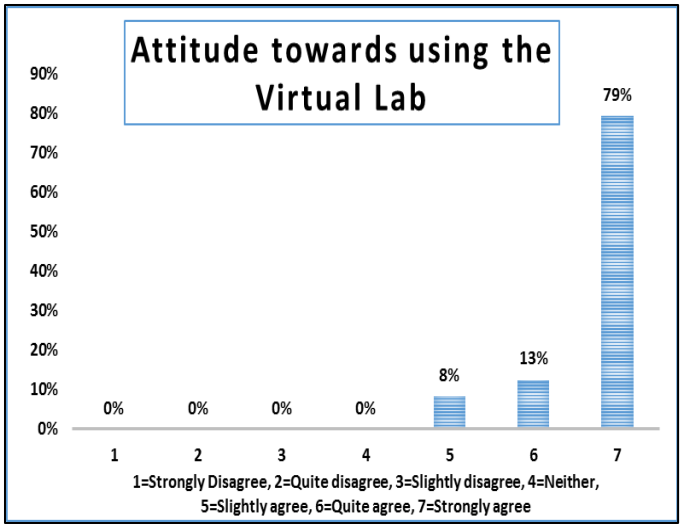


Fig. 4. The distribution of students' responses to the Attitude towards using MiLAB

Fig. 4 summarizes students' attitudes towards using MiLAB. The result reflects a positive attitude towards using the virtual lab, with (79%) of the students reported that using the system was interesting and motivated them to engage with the learning material and tasks.

Fig. 5 shows the result of students' responses regarding the availability of adequate resources and support to use the virtual lab efficiently. (42%) of the students strongly believed that enough technical support and resources were available for them. However, (16%) of the students believed that they do not have the required resources available to support them when they face difficulties. It is also worth noting that (13%) of the students were not sure about the availability of resources or technical support to improve the use of the virtual lab.

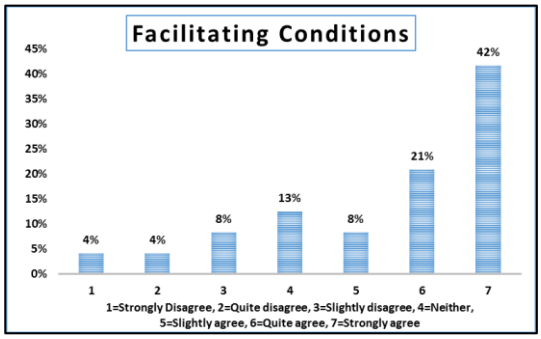


Fig. 5. The distribution of students' responses to the facilitating conditions construct.

Self-Efficacy construct measure the students' judgment of their ability in using the virtual lab. As shown in Fig. 6, the vast majority of students were confident about their ability to use MiLAB to accomplish learning tasks (54% strongly agree, 13% quite agree and 8% slightly agree).

Fig. 7 shows the distribution of students' responses to the questions measuring their behavioral intention to use MiLAB. Almost all students reported that they have a positive intention to use the virtual lab during their study.

Overall, MiLAB received high ratings in almost all aspects as measured by the UTAUT questionnaire. Responses to the questionnaire show that students have the motivation and intention to use the portable virtual lab in their study and self-regulated learning. However, some responses indicated that students do not have enough resources to help them in learning how to use the system efficiently.

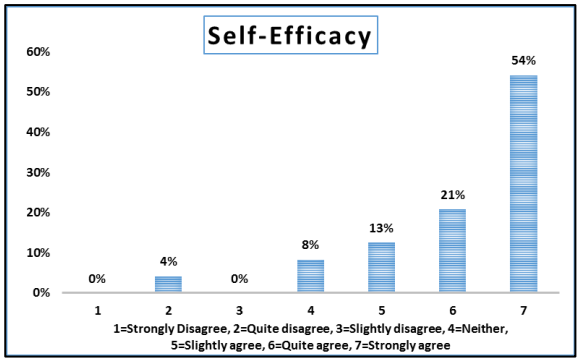


Fig. 6. The distribution of students' responses to the Self-Efficacy construct.

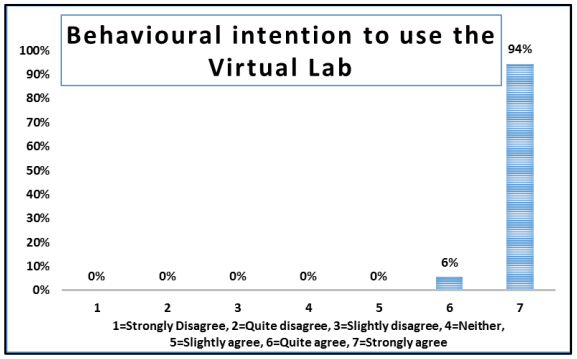


Fig. 7. The distribution of students' responses to the Behavioral intention to use MiLAB.

B. Strengths and Weaknesses

The analysis of the open-ended questions revealed a number of factors that affected the adoption and acceptance of MiLAB from students' point of view. This section outlines the strengths, weaknesses and areas of possible improvement for MiLAB elicited from students' responses to the open-ended questions.

As summarized in Table II, students believed that one of the most important features of MiLAB is that it was easy to use, learn and navigate. However, some students believed that MiLAB needs more functions, and in its current version, only few are available. It is also stressed by students that MiLAB can serve as a reference for their projects and work at other courses, which in turn has the potential to increase their productivity. Moreover, from the students' point of view, MiLAB was interactive and it supported group projects. This enhances students' ability and motivation to collaborate with other group members and serves as a platform for generating and testing new ideas.

Life-long learning is one the most important characteristic of learners in the new era. Based on students' responses, MiLAB had the ability to help students learn individually on their own base to improve web development skills. Students also think that the system will help them improve their career in the future. However, to do this, some weaknesses of the system need to be considered. These weaknesses, as descried in Table III, include stability problems and some hardware compatibility issues.

TABLE II. STRENGTHS OF MILAB FROM THE STUDENTS' POINT OF VIEW

Strength	Examples
Easy to Use	• Easy to setup • Navigation was clear • Easy to learn
Improve Productivity	• The system serves as a reference in many courses • Less save time and effort
Interesting and Interactive	• Transformation to traditional lectures. • Increasing motivation to learn
Support Collaboration	• Supporting group projects • A platform to generate and test ideas
Support Life-Long Learning	• Helping students learn outside the course • Supporting students in their future career

TABLE III. WEAKNESSES OF MILAB FROM THE STUDENTS' POINT OF VIEW

Weakness	Examples
Limited Functionality	• Only few functions were available. • Adding new functions requires plug-ins.
Hardware problems	• The system is based completely on an external USB memory drive. • Some PCs block the server.
Stability	• The server sometimes does not start properly. • Some programs may affect how the server works.

TABLE IV. AREAS OF IMPROVEMENT TO THE PORTABLE VIRTUAL LAB EXTRACTED FROM STUDENTS' RESPONSES

Area of improvement	Examples
Localization	• Easy language switch • Translations
Availability	• Cloud-based • Online and offline modes
Technical Support	• Dedicated system that is fully supported by the IT department
Customization and Upgrade	• Usable user interface • Easy to update

C. Areas of Improvement

Students were given the opportunity to reflect on their experience using MiLAB as a learning tool. Free text responses by students on this question have yielded a number of potential areas of improvement to the design of MiLAB. These are summarized in Table IV.

From the students' point of view, the system needs to support easy customization and localization to meet students' preferences. For some students, it was not easy to add new language support and it was difficult to author multilingual contents in the content management system. It is also noted by students that adding new features is not easy, and there is no instructions to guide them on how to install new plug-ins and modules. From students' responses, it was clear that MiLAB should be accompanied with a dedicated technical support section to provide necessary assistance to students all the time. Finally, as MiLAB is completely based on an external USB hard drive, compatibility and hardware issues arise. To use the system, students need to carry their own external USB drive all the time to work offline. It is recommended to provide an online version of MiLAB as a cloud-based service.

VI. CONCLUSION

Open source software plays significant roles in education in many disciplines. This paper discussed the design, implementation and evaluation of MiLAB (My Mobile Informatics LAB), a proposed portable virtual LAB for the teaching and learning of web development skills using Open Source Software technologies. The proposed virtual lab was then used to teach web development skills for health informatics students. The results of the study showed that MiLAB could serve as an easy and flexible tool for students to gain hands-on experience in informatics skills. However, the study pointed out that there are several challenges that face the adoption of portable virtual labs as educational tools, including compatibility and stability issues as well as lack of technical support.

REFERENCES

[1] I. Al Saeed and M. Eugene Dawson, "Use of Open Source Software and Virtualization in Academia to Enhance Higher Education Everywhere," in Increasing Student Engagement and Retention Using Immersive Interfaces: Virtual Worlds, Gaming, and Simulation, ed, pp. 283-313.

[2] F. Lateef, "Simulation-based learning: Just like the real thing," Journal of Emergencies, Trauma and Shock, vol. 3, p. 348, 2010.

[3] S. Ray, N. R. Koshy, P. J. Reddy, and S. Srivastava, "Virtual Labs in proteomics: New E-learning tools," Journal of proteomics, vol. 75, pp. 2515-2525, 2012.

[4] J. Ma and J. V. Nickerson, "Hands-on, simulated, and remote laboratories: A comparative literature review," ACM Computing Surveys (CSUR), vol. 38, p. 7, 2006.

[5] I. Branovic, D. Markovic, R. Popovic, V. Tomasevic, and D. Zivkovic, "Development of modular virtual lab for introductory computing courses," in Global Engineering Education Conference (EDUCON), 2013 IEEE, 2013, pp. 1027-1031.

[6] V. Baljak and S. Honiden, "Discovery of configurations for indoor wireless sensor networks through use of simulation in virtual worlds," in Sensor Technologies and Applications (SENSORCOMM), 2010 Fourth International Conference on, 2010, pp. 323-328.

[7] Y. Li, L. Xiao, and Y. Sheng, "Virtual laboratory platform for computer science curricula," in Frontiers in Education Conference (FIE), 2015 IEEE, 2015, pp. 1-7.

[8] M. E. Ahmed and S. Hasegawa, "An Instructional Design Model and Criteria for Designing and Developing Online Virtual Labs," International Journal of Digital Information and Wireless Communications (IJDIWC), vol. 4, pp. 355-371, 2014.

[9] Open Source Initiative. (2012). The Open Source Definition. Available: https://opensource.org/osd

[10] W. H. Hsu, "Creating Open Source Lecture Materials: A Guide to Trends, Technologies, and," STEM Education: Concepts, Methodologies, Tools, and Applications, p. 68, 2014.

[11] U. Ruhi, "An experiential learning pedagogical framework for enterprise systems education in business schools," The International Journal of Management Education, vol. 14, pp. 198-211, 2016.

[12] K. Shah, A. Ghosh, M. Hossain, and Y. Lee, "Enhancing Engineering Educational Using Virtual Lab Technology," Retrieved December 04th, 2014.

[13] D. Wu, J. Fulmer, and S. Johnson, "Teaching information security with virtual laboratories," in Innovative Practices in Teaching Information Sciences and Technology, ed: Springer, 2014, pp. 179-192.

[14] P. Kumar, P. Devi, and H. Rohil, "Cloud Computing based Computer Science Lab: Laboratory-as-a-Service," 2015.

[15] moodle.org.(2017). About Moodle. Available: https://docs.moodle.org /34/en/About_Moodle

[16] Autor. ATutor LMS Learning Management System. Available: http://www.atutor.ca/index.php

[17] M. R. Blake and C. Morse, "Keeping your options open: A review of open source and free technologies for instructional use in higher education," Reference Services Review, vol. 44, pp. 375-389, 2016.

[18] P. Brusilovsky, S. Edwards, A. Kumar, L. Malmi, L. Benotti, D. Buck, et al., "Increasing adoption of smart learning content for computer science education," in Proceedings of the Working Group Reports of the 2014 on Innovation & Technology in Computer Science Education Conference, 2014, pp. 31-57.

[19] M. Ivanović, S. Xinogalos, T. Pitner, and M. Savić, "Technology enhanced learning in programming courses–international perspective," Education and Information Technologies, vol. 22, pp. 2981-3003, 2017.

[20] J. E. Froyd, P. C. Wankat, and K. A. Smith, "Five major shifts in 100 years of engineering education," Proceedings of the IEEE, vol. 100, pp. 1344-1360, 2012.

[21] Apache Friends. XAMPP. Available: https://www.apachefriends.org /index.html

[22] S. MacNeill, L. M. Campbell, and M. Hawksey, "Analytics for Education," Reusing Online Resources: Learning in Open Networks for Work, Life and Education, p. 154, 2014.

[23] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS quarterly, pp. 425-478, 2003.

LeafPopDown: Leaf Popular Down Caching Strategy for Information-Centric Networking

*[1]Hizbullah Khattak, [2]Noor Ul Amin, [3]Ikram ud Din, [4]Insafullah, [5]Jawaid Iqbal
Department of Information Technology
[1,2,5]Hazara University Mansehra,
[3]University of Haripur,
[4]Abbotabad University of Sciences & Technology
K-P, Pakistan

***Abstract*—Information-Centric Networking is a name based internet architecture and is considered as an alternate of IP base internet architecture. The in-network caching feature used in ICN has attracted research interests as it reduces network traffic, server overload and minimizes latency experienced by end users. Researchers have proposed different caching policies for ICN aiming to optimize performance metrics, such as cache hits, diversity and eviction operations. In this paper, we propose a novel caching strategy of LeafPopDown for ICN that significantly reduces eviction operation and enhances cache hits and diversity in ICN.**

Keywords—Component; information-centric networking; caching; popularity; least recently used

I. INTRODUCTION

The data traffic and number of users of internet are growing rapidly during the last few years. Global IP video traffic will reach to 82% of all internet users' traffic by 2021 [1].

The ICN is an alternate network paradigm of traditional host based network communication model of internet [2]. The information in ICN is retrieved by name instead of its host locality identifier to provide data to the users with minimum delay. In order to achieve this aim, in-network caching is used in ICN to store the contents for easily access by the end users. The other features of multicast, routing by name and encryption support are included in ICN. The researchers have presented some ICN architectures such as Content Centric Networks [3], NetInf [4] and PURSUIT [5]. However, CCN have received more attention in the research community. Most of the research community focuses on designing efficient caching strategy for ICN as caching is the main characteristic of ICN.

The in-network caches are used in CCN storing several replicas of information in the network. The requests made in the future for these data can be served from these storages reducing access delay to the users and load on server and helps in minimizing the congestion in network. Researchers have proposed novel caching strategies for improving performance of CCN. Recent research work in ICN have identified that content popularity is an important factor in improving performance of ICN [6]. However, the existing policies fail to cache the content effectively so that these caches could be efficiently utilized in order to increase cache hit, diversity and reduce cache eviction operations. It is therefore very important to design caching policy that can enhance the cache utilization and avoid the content redundancy.

We proposed a LeafPopDown caching policy for ICN that cache the unpopular content near the end users and popular content on downward node and on leaf node. LeafPopDown calculate the popularity of content at each node and when content popularity increases from a specific threshold, it then caches the content on downward node and leaf node otherwise, only on leaf node. In this way, load on server or any specific node does not increase from a threshold level and redundancy in the networks is avoided.

This paper is organized as such: Section II discuss related work to our proposed caching policy. Section III discusses our proposed caching policy of LeafPopDown and its algorithm. In Section IV, we discuss the analysis and evaluation of LeafPopDown caching policy and Section V, we conclude our paper.

II. RELATED WORK

Here, we discuss caching strategies related to our proposed caching strategy.

The simplest and default caching strategy for all the architectures of ICN are Leave Copy Everywhere [7], which cache object on each node of a data delivery path. Though this caching policy has an advantage of faster data dissemination however, it causes a huge redundancy and resource consumption to its alternatives.

In MPC [8], the authors calculate popularity of content in content popularity table locally at each node. When content popularity increases from specific threshold; content is cached on the neighbors' node. This caching policy has a drawback of storing content on the neighbor nodes of a serving node away from the nodes near the users.

The authors in [9] proposed a progressive caching policy, in which object is stored on one downstream node of hit node and on intermediate node of incoming links greater than threshold. This caching policy avoids storing the unpopular content. However, it shares the shortcoming of Leave Copy Down and fixed popularity caching because of its reliance on their functionality.

The caching policy of Breadcrumbs [10] is proposed to efficiently utilize off-path caches. After the arrival of content requests to the server, each router stores a pointer called breadcrumbs along the downloading path. This pointer shows the direction of the sent contents. When requests arrive for content, it encounters a breadcrumb and that breadcrumb redirects the request in that direction.

Cho et al. [11] proposed to segment the content and caching the chunks exponentially based on popularity of the content. The idea is to store the content progressively near users with increasing requests. In WAVE, the upstream node recommends its downward node to store the number of chunks by using caching suggestion flag bit in content reply's packet. If the flag bit is 1, chunk is cached otherwise not. WAVE has some limitation. It focuses on accessing the object request and hence it does not enhance the performances of network if the users are requesting a part of an object rather than full objects.

Badov et al. [12] proposed the caching-awareness to in-network caching. The aim of this caching policy is to reduce the download times to the users. CAC avoids using the congested links and storing the object on downstream end of congested link. This caching strategy is based on two factors, i.e., download time to the users and the content popularity and it is performed on every node of a delivery path. The caching capacity of network is considered 5% of the total content population and the Zipf popularity distribution (α = 0.8) is used. This caching policy outperforms in terms of average retrieval delay as compared to other caching policies. However, in case of average hit rate metric; it does not.

III. Proposed Caching Strategy

We assume ICN is a graph of $G = (V, E)$. In this graph, $V = (v_1 \ldots \ldots v_n)$ is a group of nodes where each node is having limited storing capability and $E = (e_1 \ldots \ldots e_m)$ represent links between these nodes.

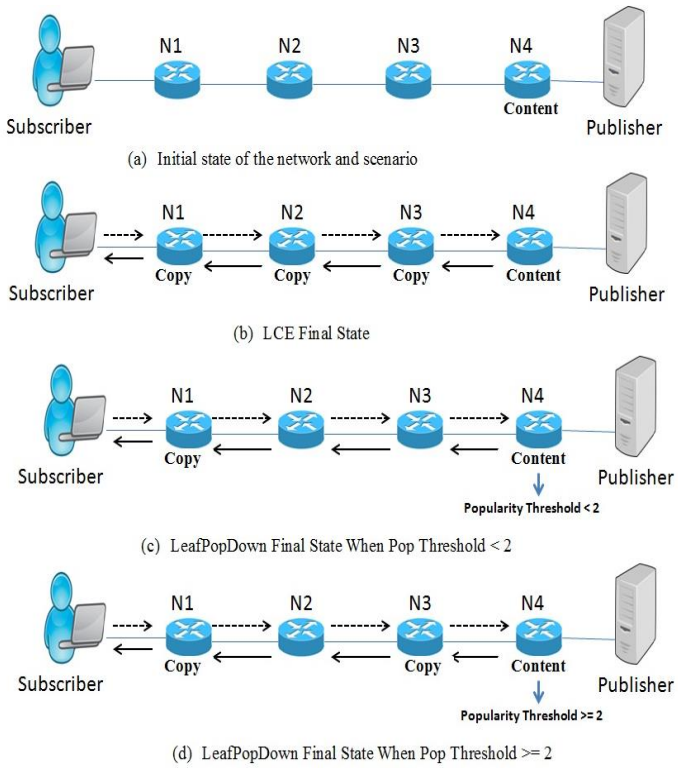
We further assume that request for data follows design of Name Data Networking [2]. An INTEREST packet is forwarded for the desired content and that request is forwarded towards server till it finds the copy of the required content. The routing table is created in Forwarding Information Base (FIB) by OSPFN protocol. We further consider that routing nodes advertise fair information. In response of an INTEREST packet, data packet is delivered on the request traversing path by using the Pending Interest Table (PIT). For simplicity we assume here that the node have same cache size. For calculation of content popularity, we consider that each router count the number of request for content in particular time T.

We illustrate and compare workflow of our proposed caching strategy of LeafPopDown and LCE through an example.

The example is explained in Fig. 1.

Fig. 1(a) represents the general scenario of networks. The content in the network is stored in node N4. Fig. 1(b) indicates the working of Leave Copy Everywhere (LCE) where content replica is caches on each node of a requesting path. The same content is cached on three nodes N1, N2 and N3 causing redundancy in the network.

Fig. 1(c) represents the first part of LeafPopDown caching strategy. When an INTEREST packet is received from the users for the desired content at node N4, it first checks the popularity of that content in its popularity table. If this content is requested for the first time it is cached on the leaf node near the subscriber. In the given Fig. 1(c), copy of the content is cached on node N1. We can clearly see that copy of the content is cached only on single node N1 as compared to LCE that cached the content on three nodes.



(a) Initial state of the network and scenario

(b) LCE Final State

(c) LeafPopDown Final State When Pop Threshold < 2

(d) LeafPopDown Final State When Pop Threshold >= 2

Fig. 1. LeafPopDown caching strategy.

Fig. 1(d) illustrates second part of LeafPopDown. When an INTEREST packet is received for a desired content; popularity of that content is checked if its popularity is greater than or equal to 2, it is cached on downward node of a hit node which is node N3 and on a leaf node that is N1 in the given Fig. 1(d). By comparing it with LCE, number of stored copies is less in our proposed LeafPoPDown caching policy i.e., 2 as compared to LCE that are 3. Similarly, when number of nodes increases in the request path of networks, LCE caches more copies of content causing huge redundancy in the network while LeafPopDown caching strategy cache less copies of content in the networks.

In order to conclude, we proved in the given example that our proposed caching strategy of LeafPopDown creates less redundancy as compared to LCE.

We use the following notations in algorithm of LeafPopDown caching strategy:

Symbol Notations	Description
V	The set of nodes $V = (v_1 \dots \dots . v_n)$, where v_n is the number of nodes
E	Links set $E = (e_1 \dots \dots . e_m)$, Here e_m is the number of links
C_i	Total contents at cache i
$C_{j,i}$	Content j at cache i
int^j	Interest for content j
r_v^j	Number of INTERESTS at node v for content j
$U(k)$	User k

Algorithm

for each ($V_n from\ i\ to\ n$)
if
$C_{j,i} == int^j$
then
if
$r_v^j >= 2$
then
Cache $C_{j,i}$ at v_{i-1} & $U(k+1)$
else
Cache $U(k+1)$
else
forward int^j to v_{i+1}

IV. Evaluation and Analysis

Here, we discuss simulation environment and performance evaluation of LeafPopDown, LCE and MAGIC caching strategies.

A. Simulation Environment

We use SocialCCNSim [8] simulator for the evaluation of LeafPopDown, LCE and MAGIC caching strategies. This simulator is used to evaluate performance metrics of caching strategies for CCN. We conducted the simulation in chosen simulator with its inherited parameters and network topologies.

Table I shows the configured parameters for our simulations. The popularity of files has been formed following MZipf distribution in SocialCCNSim. For simulation and evaluation of LeafPopDown with LCE and MAGIC, we have selected Abilene and Tiger topologies. We set the cache size of 1 GB and catalog size to 10^6. The simulations are conducted for 86400 s. We have chosen the LRU replacement policy in the simulation. The facebook is used as a social graph for simulation. SONETOR is used as a network traffic generator.

TABLE I. Simulation Parameters

Parameters	
Popularity Model	MZipf (α = 0.88, 1.1, 1.5, 2.0)
Cache Size	1 GB
Catalog Size	10^6
Topologies	Abilene, Tiger
Repacement Policy	LRU
Traffic	SONETOR

B. Performance Evaluation

In order to have fair evaluation result, we have simulated LeafPopDown, LCE and MAGIC caching strategies in the same simulation environments for time period of one day. For evaluating performance metrics of cache hits, diversity and eviction operations, these caching strategies are simulated on two topologies of Abilene and Tiger topologies. We have taken MZipf (α = 0.88, 1.1, 1.5 and 2.0).

To summarize simulation parameters, we have taken two topologies of Abilene and Tiger, cache size of 1 GB, popularity distribution values (α = 0.88, 1.1, 1.5 and 2.0).

The simulation results of cache hits of LeafPopDown, LCE and MAGIC caching strategies are shown in Fig. 2 and 3, respectively. The results of diversity of these caching strategies are shown in Fig. 4 and 5 while results of eviction operations of these three caching strategies are shown in Fig. 6 and 7.

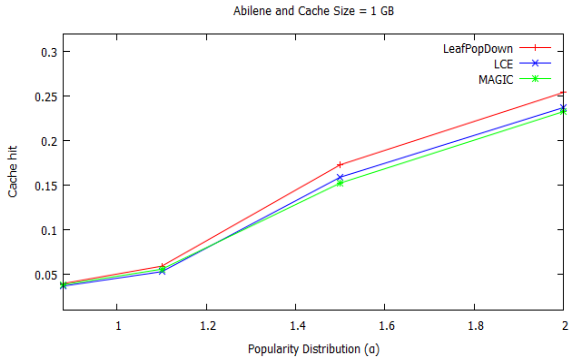


Fig. 2. Cache hits on Abilene Topology.

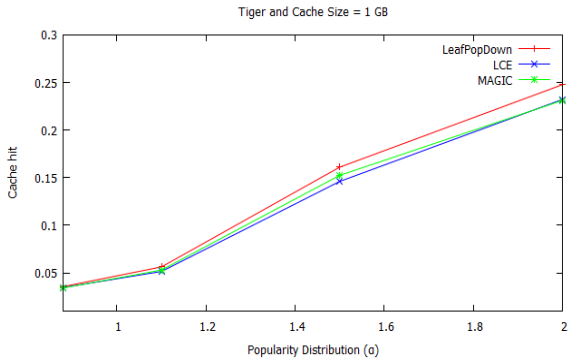


Fig. 3. Cache hits on Tiger Topology.

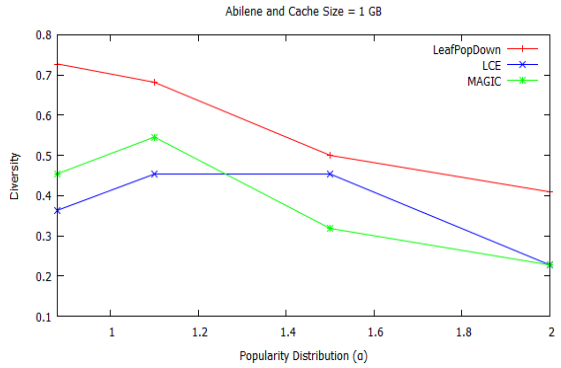


Fig. 4. Diversity on Abilene Topology.

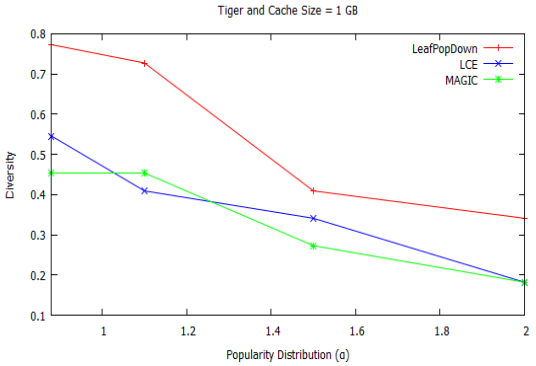


Fig. 5. Diversity on Tiger Topology.

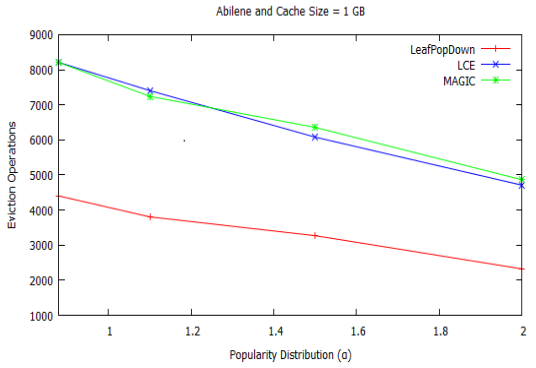


Fig. 6. Eviction operations on Abilene Topology.

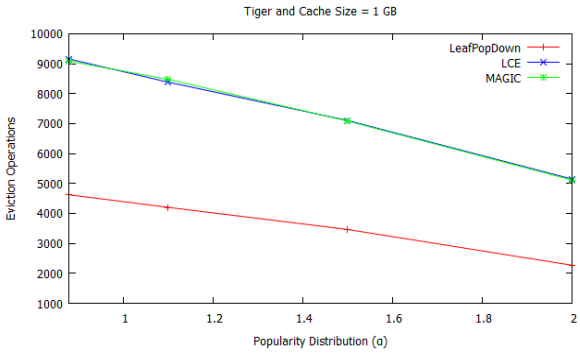


Fig. 7. Eviction operations on Tiger Topology.

By comparing the proposed caching strategy of LeafPopDown with LCE, MAGIC, we can conclude that LeafPopDown receives more cache hits as compared to other two and diversity increase significantly on our designed caching strategy. Moreover, LeafPopDown decreases eviction operations significantly as compared to LCE and MAGIC caching strategies.

V. Conclusion

In this paper, we have proposed a LeafPopDown caching strategy for ICN. LeafPopDown caches content on the leaf node near the user when it is requested in the networks. When its popularity increases to 2 or more in the popularity table and if it is requested again it leave copy of content on downward node of hit node and on leaf node near user. The simulations results show that LeafPopDown performs better than LCE and MAGIC caching policies in terms of cache hits, diversity and eviction operations. Our proposed caching strategy decrease redundancy and eviction operations while enhance the cache hits.

References

[1] Cisco, Visual networking index: Forecast and methodology, 2016-2021, Jun. 2017, White Paper

[2] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos, "A Survey of Information-Centric Networking Research," *IEEE Communications Surveys & Tutorials*, vol. 16, Iss. 12, pp. 1024-1049.

[3] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," ACM CoNEXT 2009.

[4] "SAIL NetInf," http://www.netinf.org.

[5] N. Fotiou, P. Nikander, D. Trossen, and G. C. Polyzos, "Developing Information Networking Further: From PSIRP to PURSUIT," Oct. 2010.

[6] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in Information-Centric Networking," IEEE Communications Surveys & Tutorials, vol. 17, no. 3, pp. 1473 – 1499, 2015.

[7] V. Jacobson *et al.*, "Networking named content," in *Proc. 5th Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Rome, Italy, Dec. 2009, pp. 1–12.

[8] C. Bernardini, T. Silverston, and O. Festor, "MPC: Popularity based caching strategy for content centric networks," in IEEE ICC, Jun. 2013, pp.3619–3623.

[9] J. M. Wang and B. Bensaou, "Progressive caching in CCN," in *Proc. 31st IEEE Glob. Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 2727–2732.

[10] E.J. Rosensweig and J. Kurose, "Breadcrumbs: efficient, best-effort contnet location in cache networks" IEEE INFOCOM, 2009, pp. 2631-2635.

[11] K, Lee M, Park K, Kwon T. T, Choi Y, Pack S, "WAVE: Popularity-based and Collaborative In-network caching for content-oriented networks", In Proc. *IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Orlando, FL, USA, Mar. 2012, pp. 316-321.

[12] Badov M, Seetharam A, Kurose J, Firoiu V, Nanda S, "Congestion-aware caching and search in Information-Centric Networks", *in Proc. 1st ACM Int. Conf. Inf. Centric Netw. (ICN), Paris, France*, 2014; 37–46.

Face Age Estimation Approach based on Deep Learning and Principle Component Analysis

Noor Mualla
Faculty of Computers and Informatics,
Benha University, Egypt

Essam H. Houssein
Faculty of Computers and Information,
Minia University, Egypt

Hala H. Zayed
Faculty of Computers and Informatics,
Benha University, Egypt

***Abstract*—This paper presents an approach for age estimation based on faces through classifying facial images into predefined age-groups. However, a task such as the one at hand faces several difficulties because of the different aspects of every single person. Factors like exposure, weather, gender and lifestyle all come into play. While some trends are similar for faces from a similar age group, it is problematic to distinguish the aging aspects for every age group. This paper's concentration is in four chosen age groups where the estimation takes place. We employed a fast and effective machine learning method: deep learning so that it could solve the age categorization issue. Principal component analysis (PCA) was used for extracting features and reducing face image. Age estimation was applied to three different aging datasets from Morph and experimental results are reported to validate its efficiency and robustness. Eventually, it is evident from the results that the current approach has achieved high classification results compared with support vector machine (SVM) and k-nearest neighbors (K-NN).**

Keywords*—Deep learning; principal component analysis; support vector machine; K-NN; age estimation

I. Introduction

Many applications rely on techniques for age. An example of that is found in vending machines that provide alcoholic drinks or cigarettes as they run the age prediction program to stop underage customers from illegally purchasing the products. Those processes are done through a facial recognition through the computer vision system. It worth noting that the human age estimation is a difficult task due to many a factor; primarily, identifying or labeling people within individual age classes is one hectic mission. While there have been several ways to overcome the problems facing age-estimation through facial recognition, there have been a couple of steps for the past decade that stood out the most. The algorithms designed for the task we accompanied with two components, namely, feature extraction [1]-[3] and age determination [4], [5].

In those systems, the age is recognizable by some unique face features [6]. However, it should be pointed out that age progression seen on faces cannot be accurately estimated due to the different and diverse information that can be seen in human faces. Information obtained from the human face cannot be limited for there are numerous dimensions that can be seen in the race and age. The human eye can identify known faces and the mind can estimate the age range of a human being; however, it is never always accurate despite it being "human". When it comes to the computer, there is a lot that can be read from a person's face regarding their age. Identity, gender, emotion, and age can all be deciphered in the daily communications through developing an automated method for the age estimation of an individual from the face image.

The reason why this work started was that of the need for a fast and efficient system that could provide a successful facial recognition. This work is motivated by the capacity that a system such as that could provide if it was developed successfully which could, in turn, enable numerous applications. If the outcome of any environmental setting changes and it is known, then the research work will be able to improve faster as decisions could be made up faster regardless of whether it is needed to proceed in certain directions or not. The aim of this paper is to present an automatic classification approach for age estimation. Therefore, this paper pays to focus on the classification of age estimation. The problems are analyzed on a set of records taken from Morph database. In fact, the quality of automatic identification systems performance is determined by signal representation algorithm and feature extraction schemes. In this work, a classification approach was proposed that involve PCA for extracting features as well as deep learning for the classification purpose. It is crucial for any developing approach to be tested and compared to other works in order to evaluate where it stands and whether or not it was successful. This work finishes up with a comparison to previous methods that have been used ensure the quality measures and efficacy of the proposed work such as SVM and K-NN.

This research follows the following organization: in Section II, the introduction of the work is presented; Section III follows up on the techniques and methods used within the methodology; Section IV brings forth a discussion on the deep learning approach presented; Section V exposes the results of the work from the experiments developed; and finally, Section VI contains the conclusions and the recommendations of future works proposed by the author.

II. Literature Review

In order to comprehend the steps needed to move forward with the work and bring forth advancements in the field, the work had to lay basis on previous works that relate to the topic. The works presented here vary from experiments to methodologies and even surveys developed on the matter of facial recognition and related topics, and deep learning along with topics similar to them. Several recent studies on age

estimation have been presented, in the following, we will introduce some of them such as, in this work [7], a survey was done on age estimation. Another work [8] was inspired by another problem related to classification. The problem that the work focused on was make-up and plastic surgeries that divided the programs in finding suitable age groups. One more that was studied used the Active Appearance Model (AAM) [9]. A combination of shapes and models was presented in [10]. That approach was evaluated in another work [11] through employing three qualifiers.

Further, Fu et al. [12] have proposed an age estimation method developed to analyze pictures through the employment of the state-of-the-art manifold learning methods that can aid in discovering sufficient low dimension embedding area. The results of the simulation have shown that the linear manifold learning could provide sufficient aging feature extraction for regression ways of age estimation. Also, a similar approach was used by Guo *et al.* [13] that followed an age manifold learning scheme for extracting the aging features. The method also planned a locally adjusted robust repressor for the learning process and to expect the human age.

Eventually, there were various ways that attempted to reach novel approaches for facial recognition with the emphasis of the facial expression characterization and face modeling [14]. The geometric invariants that are important aspects of strain transformation have been observed through previous studies that provided information necessary for the work at hand [15]. Another work dealt with the uncertainty through the formation of nearly definite programming problem [16] or an EM-based algorithm [17]. Other works relied on the Local Binary Pattern (LBP) [18] features to identify series of local features that were also observed by the combination of a strong classifier that enabled a successful method of age classification [19].

III. PRELIMINARIES

A. PCA for Feature Extraction

One of the most popular methods in the field of extracting features is Principle Component Analysis (PCA). The method linearly transforms the high-dimensional input vectors into low-dimensional ones in which the components would be uncorrelated. It computes the eigenvectors of the covariance matrix of the original inputs for processing. The process has been used in developing various algorithms in the past [20]. It is also considered as one of the best means in this area of work along with a reduction in dimensionality.

Through PCA, this work aims to showcase the d-dimensional data in a low-dimensional space which would decrease the degrees of freedom and space and time complexions. The aim of using the method is to present data in a space which expresses the variations in a sum-squared error sense as well as possible. PCA is also beneficial as it divides signals and images alike from several sources. Knowing the number of independent components exist beforehand helps in getting the best out of the method, just like standard clustering methods.

The basic approach in principal components is theoretically rather simple. First, the *d*-dimensional mean vector λ and $m \times m$ covariance matrix *R* is calculated for the full data set. Then, those eigenvalues and eigenvectors are calculated and organized based on the decreasing eigenvalue. Given a group of centered vectors of input $x_t (t=1,...,l \text{ and} \sum_{t=1}^{l} x_t = 0)$, each of which is of m dimension $x_t = (x_t(1), x_t(2), ..., x_t(m))^T$ (usually $m \prec l$), PCA linearly transforms each vector x_t into a new one s_t by:

$$s_t = U^T x_t \quad (1)$$

Where *U* is the $m \times m$ orthogonal matrix whose *i*th column u_i is the *i*th eigenvector of the sample covariance matrix $C = \frac{1}{l} \sum_{t=1}^{l} x_t x_t^T$. Simply put, PCA first solves the eigenvalue problem defined in (2).

$$\lambda_i u_i = C u_i, \quad i = 1, ..., m \quad (2)$$

Where, an eigenvalue of C and the matching eigenvector is u_i. Based on the predicted u_i, the components of s_t are then computed as the orthogonal alterations of x_t:

$$s_t(i) = u_i^T x_t, \quad i = 1, ..., m \quad (3)$$

The new components are called the principal or main components. Through employing solely the first several eigenvectors sorted in descendant order of the eigenvalues, the number of principal components in s_t can be reduced, meaning that PCA has the dimensional decrease feature. The main components of PCA also have the following properties [21].

- $s_t(i), i=1,...,m$ are uncorrelated.
- $s_t(i), i=1,...,m$ have sequentially maximum variances.
- The mean-squared approximation error in the demonstration of the original inputs by the first several principal components is minimal.

B. Support Vector Machines (SVM)

The classification technique helps to discriminate the unknown testing set of observations into their appropriate classes based on the training group of known annotations. A classification technique used a mathematical function named as a classifier to predict the right class of unknown observation of testing data set. SVM was a method that Vapnik and Cortes introduced [22]. SVM is a powerful classifier in biomedical science, image processing and data mining for the detection and classification purposes. SVM is an efficient classifier to classify two different sets of observations into their relevant class. It has the means to handle high-dimensional and non-linear data excellently. Base on the foundation of training data sets, it helps to guess the important characteristics of unknown testing data. SVM mechanisms are based on finding the best hyperplane that divides the data of two different classes of the

category. Accordingly, the best hyperplane is figured by being the one that maximizes the margin.

The design of SVM is based on the regularization parameter, C, which is used to control the relationship between margin maximization and some misclassifications; and kernel functions of nonlinear SVMs which are used for the mapping of training data from an input space to a higher dimensional feature space. It should be noted that all the kernel roles such as linear, polynomial, radial basis function and sigmoid having some free parameters are called hyperparameters. Until the present day, the well-known kernel commonly used research was the Gaussian or radial basis function (RBF) kernel with width σ [23].

$$k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) \quad (4)$$

Where, K(x, y) is termed as the kernel function, which is built upon the dot product of two invariant x and y. suitable trade-off parameter C and the kernel parameter σ are essential to training SVM classifier and usually found by the K-fold cross-validation technique.

C. K-Nearest Neighbors Algorithm (k-NN)

Classification in K-nearest neighbor is based on a majority vote of the nearest k neighbors in the training dataset [24]. Euclidean distances between the unclassified record and the classified records are calculated and sorted. One single observation from the original dataset is selected and used accordingly as a dataset for validation whereas the remainder of observations is selected as training data. That means that each observation is only used one time as the validation data. Leave-one-out cross-validation (LOOCV) of one nearest neighbor (1-NN) is what the pre-mentioned method is known as [24].

IV. Proposed Deep Learning Approach for Classification

A wide array of missions has been accomplished through the use of deep learning using neural networks. The interest herein is in Deep Learning Architectures (DLA) that has the ability of effectively performing FE for hyperspectral data. While DLAs are newly founded developments in neural networks, they have already obtained high classification accuracies in many fields and remain to be at the peak of performance levels, over all of the other methods [25], [26]. DLAs ability is perceived to be closer to AI than other state-of-the-art learning machines [27].

SVMs were shown in the work and in literature to be outperformed through Deep Belief Network and the stacked auto-encoder; both of which are types of DLAs. They outperformed SVMs in classifying the hyperspectral data [28], [29]. The experiments were carried out to predict the effect that the depth of the networks and the principal component numbers of the data on the accuracy of the classification and the time needed to run the experiments. This work was prepared to understand the effects of a DBN on its performance through monitoring the accuracy level and impact by the width of the network and on the time required for the training purpose. Accordingly, it was found in the study that the proposed DBN structure allows for higher accuracy in classification on hyperspectral data that is remotely-sensed via an SVM and K-NN as a reference to aid in the comparison.

The process of deep learning includes a class of models that attempt to hierarchically learn deep features of the input data with very deep neural networks which are often perceived as more than three layers deep. Through unsupervised training, the network is built in the first layer. In the scheme considered for the work, high-level features can be learned through low-level features while the proper ones can be shaped for pattern classification at the end. Accordingly, deep models have the potential to contain more abstracts and complex features in higher layers. Abstract features are known to be invariant to most of the local changes of the input. Some papers [30] argue that deep models can achieve a more accurate approximation to nonlinear functions that shallower models. Orthodox deep neural network architectures contain deep belief networks (DBNs) [31], deep Boltzmann machines (DBMs) [32], restricted Boltzmann machines (RBMs) [33], pooling units, convolutional neural networks (CNNs) [34], etc. This work implements one of the deep learning models mentioned above: namely DBN, for age classification as depicted in Fig. 1.

A. Deep Belief Networks

Deep Learning (DL) can be stacked through feeding a DL the hidden victor of another as input. Layer, in this case, refers to one of the ideal parameters between two vectors of visible or two vectors of hidden units. In the layers beyond the first one, the conditional probabilities are given the following equation:

$$P(h_j^{l+1} \mid h^l) = sigmoid(c_i^{l+1} + \sum_{j \in J} W_{ij}^{l+1} h_j^l) \quad (5)$$

Where h^l denotes the input vector of the RBM at the *l*-th layer of the network and h^0 signifies the input vector of the DBN [31].

When the network is built this way and trained, layer-wise, in an unsupervised style by means of (5), it is known as a Deep Belief Network (DBN). Though a DBN can be used as a generative model because of the RBM's ability to rebuild data, the work's interest here is in the discriminative case. The idea here focuses on classifying data rather than to reconstruct it. To use a DBN discriminatively, the output of the top-layer DL is fed as input to a classifier such as a logistic regression, at which point, the model is then trained in a supervised fashion [35], as shown in Fig. 2.

The pre-training stage is known as the unsupervised training where it initializes the model in order to improve its efficiency during the supervised training stages. Said stages are known as fine-training; where the classifier's prediction is adjusted so that it would match the ground truth of the data. Iterations in either training stages, fine or pre, are known as epoch [36]. Accordingly, DBNs must learn to produce "good" representations.

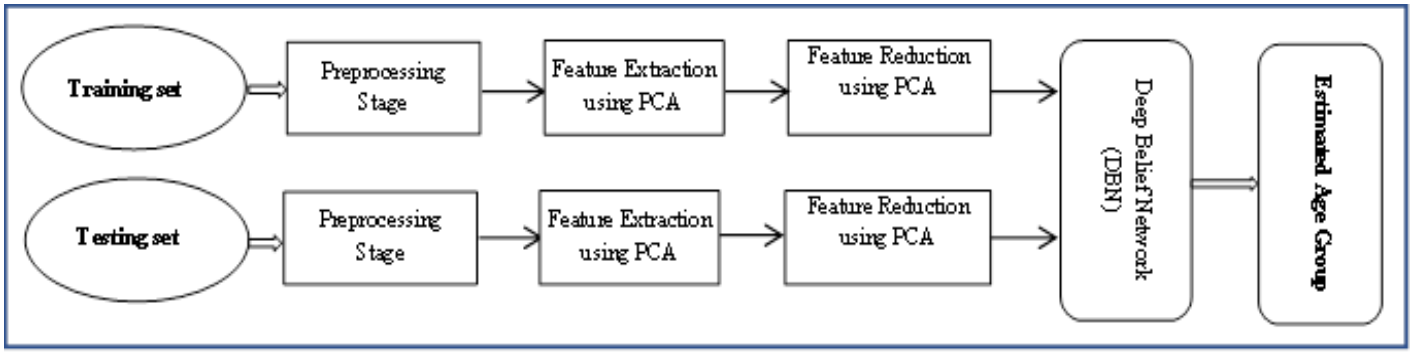


Fig. 1. Diagram of the proposed approach.

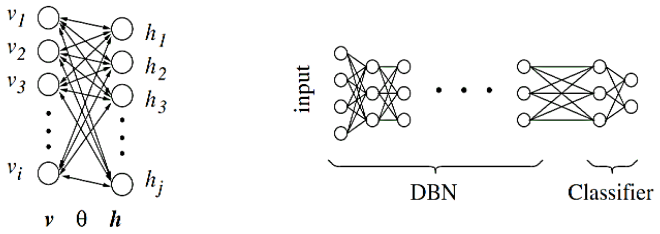


Fig. 2. Left: An RBM with i visible units, j hidden units, and model parameters θ. Right: A DBN with arbitrary width and a classifier.

A good representation is a representation that carries the most crucial underlying patterns of the data while the rest of the data is discarded. The best representation, however, is one that also catches the disentanglements of the patterns from one another, after capturing them. The best representation is similar to the human brain's ability to learn to recognize objects through disentangling them [37]. Those aforementioned definitions explain the term "representation learning" whereas the term "deep learning" refers to the implementation of this theory through a learning machine that contains many layers, such as a DBN. In this paper, the DBN is chosen over other deep constructions as the interest here is in studying one specific technique used by the DBN, pre-training, which remains partly unknown [38].

V. Experimental Results and Discussion

In order to estimate the performance of each model, input facial image dataset is split into training and test sets, then k-fold cross-validation (leave one out) was subsequently run [39]. In this paper, the value of k is set to 3; hence, the facial dataset was divided into 3 distinct parts. Two parts are used to train the classifier, while the third part is used to test classification except for K-NN because this is a non-parametric lazy learning algorithm.

The most popular kernel functions RBF, polynomial, linear and sigmoid kernel functions in this paper for SVM, Gaussian radial basis function (RBF) is applied. Further, in this paper, PCA is used to extract the significant feature and those features are used as input to the different classifier used in this paper. DBN, SVM, and K-NN classifiers are employed to classify dataset. With PCA applied to extract 750 features. Then DBN and SVM classifier was repeatedly trained in order to produce a model that is more precise for age estimation classification. In this section, a number of experiments have been done on a standard data set to measure the accuracy of the facial age approximation method suggested.

A. Dataset Description

The used dataset is built using images from Morph database II [40]. Morph database II the largest publicly available aging face database. It contains facial images of males and females with ages ranging from 16 to 77 years. It consists of 55,134 images of 13,000 persons and it took four years to be collected. Sample images are shown in Fig. 3.

The used dataset consists of 2494 images of males and females that are classified as into four age groups. The details of each age group including the age ranges and the number of images in each group are shown in Table I.

Fig. 3. Sample images from the used Morph dataset.

TABLE I. The Description of the Used Dataset

Age Group	Age Range	Number of Images
Group I	31-40	624
Group II	41-50	1424
Group III	51-60	450
Group IV	61-70	46

A. Experimental Results

The comparison of the performance was done over 3-k fold cross validation as every group of images was divided into a 3-fold where each fold is always saved for testing and the remaining nine were used for training.

Fig. 4 shows the graphical representation of the performance measures for SVM classifier.

Fig. 5 shows the graphical representation of the performance measures for K-NN classifier.

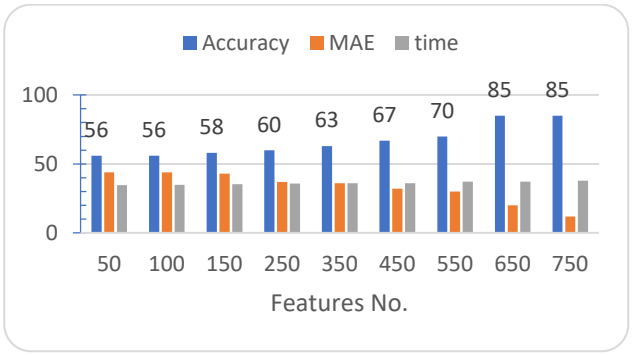


Fig. 4. Performance measurements for SVM.

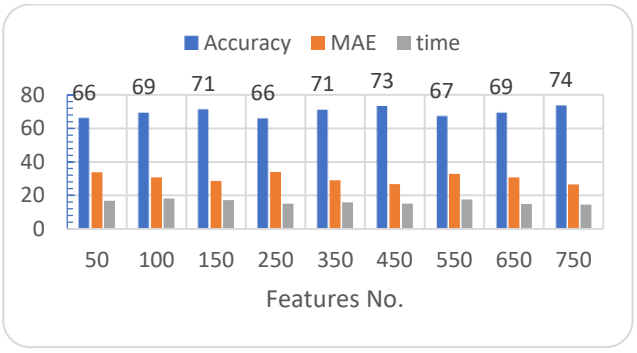


Fig. 5. Performance measurements for K-NN.

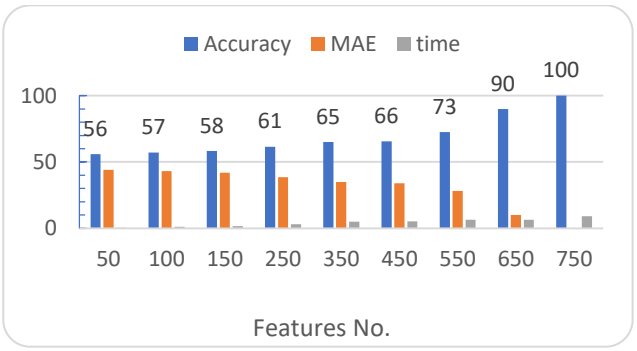


Fig. 6. Performance measurements for DBN.

Fig. 6 shows the graphical representation of the performance measures for DBN classifier.

In summary, Fig. 4, 5 and 6 shows a graphical comparison among all output accuracy measures for DBN, SVM, and K-NN classifiers. The results provided evidence that the DBN outperforms the comparison classifier. Comparing results, it is shown that best results are obtained by DBN classifier was 100% compared to best accuracies by SVM and K-NN classifier.

VI. Conclusion and Future Works

The research area of automated facial age estimation has gained an increasing attention from the researchers due to its relevance to several daily life applications. In this paper, we have proposed an automated facial age estimation method based on several well-known classifiers famous in the machine learning domain such as deep belief network, SVM and K-NN which able to estimate the human age based on the face image. The proposed approach consists of four steps: image preprocessing, feature extraction using PCA, feature reduction using PCA, and DBN based classification process. The proposed approach has been evaluated via a dataset that includes Morph II database images. The experimental results have shown that the proposed approach has a promising performance with achieved classification accuracy up to 100% compared with SVM and K-NN. For future works, different modifications are to be added to the proposed approach such as employing wavelet transform and linear discriminant analysis for feature extraction and nature-inspired algorithms for feature selection and classifier's parameter optimization.

References

[1] Y. Fu and T.S. Huang, "Human age estimation with regression on a discriminative aging manifold", IEEE Transactions on Multimedia (TMM) 10 (4) (2008), pp. 578–584.

[2] X. Geng, Z.H. Zhou and K. Smith-Miles, "Automatic age estimation based on facial aging patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (12) (2007), pp. 2234–2240.

[3] Noor Mualla, Essam H. Houssein, Hala H. Zayed, "Automated Facial Age Estimation Using Deep Belief Network", International Journal of Advancements in Computing Technology(IJACT), Vol. 9, No.3, 2017.

[4] K.Y. Chang, C.S. Chen and Y.P. Hung, "A ranking approach for human age estimation based on face images", in Proceedings of the International Conference on Pattern Recognition (ICPR), 2010, pp. 3396–3399.

[5] S. Yan, H. Wang, X. Tang and T.S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels", in Proceedings of the International Conference on Computer Vision (ICCV), 2007, pp. 1–8.

[6] X. Geng, Z.H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation", in Proceedings of the Fourteenth Annual ACM International Conference on Multimedia, 2006, pp. 307–316.

[7] N. Ramanathan, R. Chellapa and S. Biswas, "Age progression in human faces: a survey", J. Vis. Lang. Comput. 15 (2009) 3349–3361.

[8] [8] Y.H. Kwon and N.D.V. Lobo, "Age classification from facial images", in Proceedings of the 1994 IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 762–767.

[9] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models", IEEE Trans. Pattern Anal. Mach. Intell. (1998), pp. 484–498.

[10] A. Lanitis, C. Taylor and T. Cootes, "Toward automatic simulation of aging effects on face images", IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002), pp. 442–455.

[11] A. Lanitis, C. Draganova and C. Christodoulou, "Comparing different classifiers for automatic age estimation", IEEE Trans. Syst. Man Cybern. 34 (1) (2004), pp. 621–628.

[12] Y. Fu, Y. Xu, and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features," in IEEE Conf. ICME'07, 2007, pp. 1383–1386

[13] G. Guo, Y. Fu, C. Dyer, and T. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression", IEEE Trans. Image Process. 17 (7) (2008), pp. 1178–1188.

[14] S. Z. Li and A. K. Jain. "Handbook of Face Recognition". Springer, New York.

[15] L.S. Mark, J.T. Todd, and R.E. Shaw. "Perception of growth: A geometric analysis of how different styles of change are distinguished", Journal of Experimental Psychology: Human Perception and Performance, pp.855–868, 1981.

[16] S. Yan, H. Wang, X. Tang and T. Huang, "Learning auto-structured regressor from uncertain nonnegative labels", in Proceedings of the Eleventh IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[17] S. Yan, H. Wang, T. Huang, Q. Yang and X. Tang, "Ranking with uncertain labels", in Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 96–99.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002), pp. 971–987.

[19] Z. Yang and H. Ai, "Demographic classification with local binary patterns", in S.W. Lee, S. Li (Eds.), Advances in Biometrics, Lecture Notes in Computer Science, vol. 4642, 2007, pp. 464–473.

[20] Diamantaras, Konstantinos I., and Sun-Yuan Kung. "Principal component neural networks: theory and applications". John Wiley & Sons, Inc., 1996.

[21] Cao, L. J., et al. "A comparison of PCA, KPCA, and ICA for dimensionality reduction in support vector machine". Neurocomputing 55.1 (2003), pp. 321-336.

[22] Cortes, C., and Vapnik, V., "Support-vector networks", Machine learning, (1995), 20(3), pp. 273-297.

[23] Andrew, A.M., "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods" by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000.

[24] Cover, T., and Hart, P, "Nearest neighbor pattern classification", IEEE transactions on information theory, (1967) 13(1), pp. 21-27.

[25] Le, J.H., Yazdanpanah, A.P., Regentova, E.E., and Muthukumar, V. "A Deep Belief Network for Classifying Remotely-Sensed Hyperspectral Data. In International Symposium on Visual Computing, (2015), pp. 682-692. Springer, Cham.

[26] Makantasis, K., Karantzalos, K., Doulamis, A. and Doulamis, N., "Deep supervised learning for hyperspectral data classification through convolutional neural networks", In: IGARSS, pp. 1771–1800 (2015).

[27] Makantasis, K., Karantzalos, K., Doulamis, A., and Loupos, K., "Deep learning-based man-made object detection from hyperspectral data", In International Symposium on Visual Computing (2015), pp. 717-727, Springer, Cham.

[28] Chen, Y., Zhao, X. and Jia, X. "Spectral-spatial classification of hyperspectral data based on deep belief network". IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, pp. 2381–2392 (2015)

[29] Chen, Y., Lin, Z., Zhao, X., Wang, G., and Gu, Y. "Deep learning-based classification of hyperspectral data". IEEE Journal of Selected topics in applied earth observations and remote sensing, (2014), 7(6), pp. 2094-2107.

[30] N. LeRoux and Y. Bengio, "Deep belief networks are compact universal approximators," Neural Comput., vol. 22, no. 8, pp. 2192–2207, Aug. 2010.

[31] G.E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[32] R. Salakhutdinov and G.E. Hinton, "Deep Boltzmann machines," in Proc. Int. Conf. Artif. Intell. Statist. Clearwater Beach, FL, USA, 2009, pp.448–455.

[33] G. E. Hinton, "Apractical guide to training restricted Boltzmann machines," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. UTML TR2010-003, 2010.

[34] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," Neural Comput., vol. 1, no. 4, pp. 541–551, Apr. 1989.

[35] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.

[36] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., Montral, U.D. and Qubec, M., "Greedy layer-wise training of deep networks". In: Sch¨olkopf, B., Platt, J., Hoffman, T. (eds.) NIPS. MIT Press, Cambridge (2007)

[37] Hinton, G.E., Osindero, S. and Teh, Y.W., "A fast learning algorithm for deep belief nets", Neural Comput. 18, pp. 1527–1554 (2006)

[38] Bengio, Y., Courville, A.C. and Vincent, P., "Unsupervised feature learning and deep learning: a review and new perspectives", CoRR abs, pp. 1206.5538 (2012)

[39] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann.

[40] K. Ricanek and T. Tesafaye, "Morph: a longitudinal image database of normal adult age-progression", in: Proceedings of the Seventh International Conference on Automatic Face and Gesture Recognition, 2006, pp. 341–345.

Development and Validation of a Cooling Load Prediction Model

Abir Khabthani[1], Leila Châabane[2]
Université de Tunis El Manar
Ecole Nationale d'Ingénieurs de Tunis
Laboratoire Analyse, Conception
et Commande des Systèmes (LR11ES20)
1002, Tunis, Tunisie

***Abstract*—In smart buildings, cooling load prediction is important and essential in the sense of energy efficiency especially in hot countries. Indeed, prediction is required in order to provide the occupant by his consumption and incite him to take right decisions that would potentially decrease his energy demand. In some existing models, prediction is based on a selected reference day. This selection depends on several conditions similarity. Such model needs deep analysis of big past data. Instead of a deep study to well select the reference day; this paper is focusing on a short sampling-rate for predicting the next state. So, this method requires less inputs and less stored data. Prediction results will be more close to the real state. In first phase, an hourly cooling load model is implemented. This model has as input current cooling load, current outside temperature and weather forecast to predict the next hour cooling consumption. To enhance model's performance and reliability, the sampling period is decreasing to 30 minutes with respect to system dynamic. Lastly, prediction's accuracy is improved by using previous errors between actual cooling load and prediction results. Simulations are realized in nodes located at a campus showing good adequacy with measurements.**

Keywords*—Smart building; energy efficiency; prediction; short sampling-rate; less stored data

I. INTRODUCTION

Research and innovation in the design of technologies to make buildings smart and intelligent have increased enormously during past decade [1]-[6]. Integration of information technologies and computation in Building Management System is an inevitable option. Such system makes easy and possible the on line environmental monitoring [7] and control of many activities and services associated to the building. The role of decision support systems is significant since it contributes to the continuous energy management of a typical building's daily operations, aiming to preserve occupants comfort conditions and minimize energy consumption and cost [8]. Indeed, energy and comfort trade-off in buildings is absorbing interest of many researches around the world. Various aspects of this research domain have been investigated so far [9]-[12]. Enormous attention should be paid to occupants' comfort which is a main concern for building's intelligence. Indeed, human beings spend most of their life in indoor areas (homes, schools, offices). Energy in building, in hot countries, is considered as among the main consumers. According to the Arab Union of Electricity, electricity consumption in residential sector represents 30% of total electricity consumption in Tunisia. The dominant energy consumption in hot countries is the cooling system especially during hot season. In this framework, the World Bank launched a study [13] in the context of a multi-donor trust fund for addressing climate change in the Middle East and the North Africa Region. According to this study, every year, about 2 million air conditioners are added in the Maghreb countries. In Tunisia, in residential sector, the installed capacity of cooling systems is equal to 2156 MW. It represents 69% of maximum power demand equal to 3144 MW. This significant increase of air conditioners causes the appearance of electric consumption peak and a structural change of the load curve during summer. The peak demand has increased by 5.1% annually. Therefore, any saving applied to such equipment has a great potential of energy reductions.

A large number of papers proposed various methods for cooling system's energy management [14]-[16]. The main approach highlighted in these papers is the cooling load prediction. Generally, these methods are physical model [17] [18], [19], black box model [20], [21] and grey box model [22], [23]. The physical model demonstrates good results but requires big data set, various types of weather data, a deep analysis of past data and extensive complex model. For instance, the model, developed in [24], needs several physical parameters and a considerable amount of details (type of internal mass according to its thermal mass and type of radiation it absorbs, detailed parameters of layers, etc.) and site data for identification of parameters. Sometimes, in such physical model, there are data which are difficult to obtain or even missing. Indeed, the authors in [25] proposed a cooling energy prediction model using an enthalpy-based cooling degree days method. Enthalpy calculation is a complex issue and needs some parameters hard to be available. In black box model, several methods achieved acceptable prediction results. However, its major problem is that prediction's reliability and accuracy depend on selected training data. Also, some black box models such as [26], [27] use sophisticated methods. The grey box model is known by its satisfactory reliability, low requirement of training data. However, this model should be improved in terms of computing power and error checking. Among these prediction models, cooling consumption is predicted by developing a method based on the selection of a reference day according to occupancy similarity principle. This prediction method depends also on a weather data which is the

most correlated to current measurements. Then prediction results are calibrated by using the average error of past two hours. This model necessities a deep study to well select the reference day. Also, it requires as input a considerable amount of historical data. This paper therefore proposes a simplified cooling prediction model based on using current hour data as a reference for predicting the next hour consumption. So, the deep analysis of big past daily data and the issue of selecting a reference day are not required. Other enhancement is to decrease sampling period while respecting system response.

The content of this paper is organized as follows. Section 2 elaborates the computing of actual cooling load. Section 3 describes the analysis and modeling processes in this work. This section also discusses the simulations and the results. Section 4 presents the conclusion and the future work.

II. Computing of Actual Cooling Load

Nowadays, several strategies and technologies aim to respect the trade-off between energy and comfort and introduce this concept inside buildings. This approach brings intelligence to buildings. To ensure occupant's comfort, considered as a complicated problem, a new approach is highlighted in this paper. This approach is based on human decisions since a change in his behavior is a key in achieving sustained reductions in energy consumption. For these reasons, cooling consumption, considered as the dominant energy use especially in hot countries, must be predicted and provided to human with useful awareness tips. Indeed, if occupant will be aware of his predicted consumption, he will be able to take right decisions that respect energy management and his preference. Several papers developed various methods for cooling consumption prediction. Among these models, proposed in literature, we cite a load prediction method based on a selected reference day. Cooling consumption of this reference day is taken as the targeted day's initial load prediction result. With respect to smart buildings, this paper proposes a predicted cooling model based on a short-sampling rate during the same day. This prediction method gives occupant fast analyses and drives his behavior in favor of his comfort and his consumption.

The building, proposed as a case of study, is composed of identical rooms. These nodes are illustrated by Fig. 1. T_1 and T_2 are adjacent rooms' temperatures. T_0 is the concerned node's temperature. They are given in degrees Celsius.

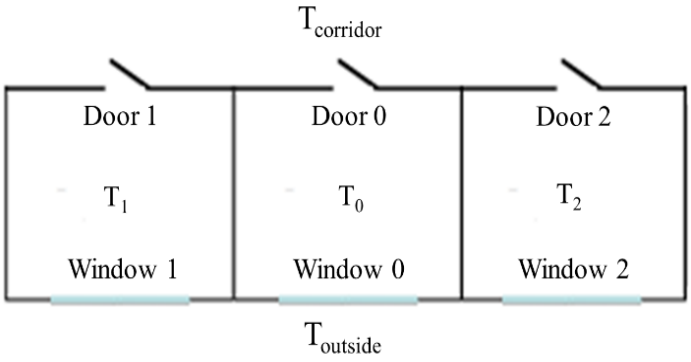


Fig. 1. Node's top view.

Model's simulations have been tested in one node located between two other similar rooms, exposed to solar radiation through windows' glasses and equipped with an air conditioner and temperature sensors. Measurements are done during a significant day in hot season.

A data acquisition system (DAQ) has been realized to measure temperature exchanges: inside temperature, outside temperature, corridor's temperature and adjacent rooms' temperature. This DAQ is composed of:

- microcontroller board based on the ATmega328P,
- memory card reader, and
- temperature sensor DS1821.

The choice of this sensor is justified by its accuracy and its measurement range which is between -55°C and +125°C. Five DAQs are positioned in different measurement points (concerned node, adjacent rooms, corridor and outside) to measure temperatures in a short sampling period. Indeed, a data logging solution is implemented in the microcontroller ATmega328P. This application serves to measure temperatures by sensors DS1821 and store these data in memory card reader.

The air conditioner worked with a set temperature equal to 21°C according to the standard ISO 7730 (Ergonomics of the thermal environment-Analytical determination and interpretation of thermal comfort using calculation of the Predicted Mean Vote (PMV) and the Predicted Percentage Dissatisfied (PPD) indices and local thermal comfort effects). Indoor temperature's sensor was added to calibrate the model. Indeed, to validate predicted cooling consumption, the first step was to determine actual consumption during a day.

To do so, a thermal model is proposed and detailed in the following. This model consists of resistors, capacitors, current source and voltage sources. Its input are data provided from temperature sensors. Its output is the computing of actual cooling consumption. Temperature is modelled by a voltage source, conduction and convection through walls, door and windows are modelled by resistors and capacitors.

Solar radiation is modelled by a current source. Its model is based on the difference between maximum and minimum outside temperature during a small sampling period in order that the model will be closed to the real state. Instead of computing a daily estimation of solar radiation (HS model) [28], [29] that may yield inaccurate results, this paper is focusing on a short-sampling rate. Indeed, in hot seasons, temperature gap is important during a short period. Solar radiation's model depends in turn on temperature variation.

This model is given by (1). Since windows are uncoated single glazing, the solar heat gain coefficient *SHGC* is equal to 0.7. Also, the building, proposed as a case of study, is located in coastal region. Hence, the coefficient K is equal to 0.19.

$$Q_{rad} = SHGC \times S_f \times K \times G_0 \times \left(T_{\max} - T_{\min}\right)^{0.5} \quad (1)$$

Where *SHGC* is the Solar Heat Gain Coefficient equal to 0.7 [30], S_f is the window area equal to 7.4 m^2, K is an

empirical coefficient (0.16 for interior regions and 0.19 for coastal ones), G_0 is the hourly extra-terrestrial radiation (W/m^2), T_{max} is the maximum temperature and T_{min} is the minimum temperature.

The hourly extra-terrestrial radiation G_0 is computed by (2) [28]:

$$G_0 = \frac{12}{\pi} \times SC \times f \times (\cos\varphi \times \cos\delta \times (\sin\omega_2 - \sin\omega_1) + \frac{\pi}{180} \times (\omega_2 - \omega_1) \times \sin\varphi \times \sin\delta) \quad (2)$$

Where SC is the solar constant equal to 1367, f is the eccentricity correction factor, φ is the latitude equal to 36.82°, ω_1 is the solar hour angle for the beginning time, ω_2 is the solar hour angle for the ending time and δ is the solar declination. Angles are given in degrees.

Eccentricity correction factor is described as follows:

$$f = 1 + 0.033 \times \cos\frac{360 \times n}{365} \quad (3)$$

Where n is day of the year (for example 1st January corresponds to 1).

Solar declination is calculated by (4):

$$\delta = 23.45 \times \sin\left(360 \times \frac{284 + n}{365}\right) \quad (4)$$

Solar hour angle is the sun's angular deviation from south. It is expressed as follows:

$$\omega = 15° \times (Solar\_Time - 12) \quad (5)$$

$-180° \le \omega \le 180°$, negative before Solar Noon.

Since measurements are done between 6:00 and 14:00, solar hour angles ω are given by Table I.

TABLE I. SOLAR HOUR ANGLES

6h	7h	8h	9h	10h	11h	12h	13h	14h
-90	-75	-60	-45	-30	-15	0	15	30

After defining analysis data (building parameters and data provided by temperature sensors) and modelling solar radiation during a short sampling rate, a thermal model is proposed in Fig. 2. Where Rwij, Cwi, Rhij and Ti are the parameters of adjacent rooms having same conditions (wall resistor, wall capacitor, convection resistor and node's temperature), Rhci, Rwci and Cwc are the corridor wall's parameters, Rhoi, Rwoi, Cwo are the outside wall's parameters, Rwin is the window's resistor, Qrad is the solar radiation's source and T_0 is the indoor temperature.

Thermal model's parameters are given by Table II.

TABLE II. THERMAL MODEL'S PARAMETERS

Parameters	Values
Resistances (K/W)	
Rwin	0.310
Rwij	0.07
Rwci	0.113
Rhij	0.0106
Rhci	0.0155
Rwoi	0.520
Rho2	0.0518
Rho1	0.029
Capacitance (J/K)	
Cwi	266463.28
Cwc	182746.1
Cwo	934209.45

Variation of indoor temperature depends on outside temperature, corridor temperature and adjacent nodes temperature. Since thermal model's output is the computing of actual cooling consumption, this last depends in turn on indoor temperature variation.

Actual cooling load is given by Fig. 3. Where, 0s is the beginning time corresponding to 6:00. The ending time is 28800s which corresponds to 14:00. This period corresponds to the work hours (7:00 to 14:00). The air conditioner starts running one hour before.

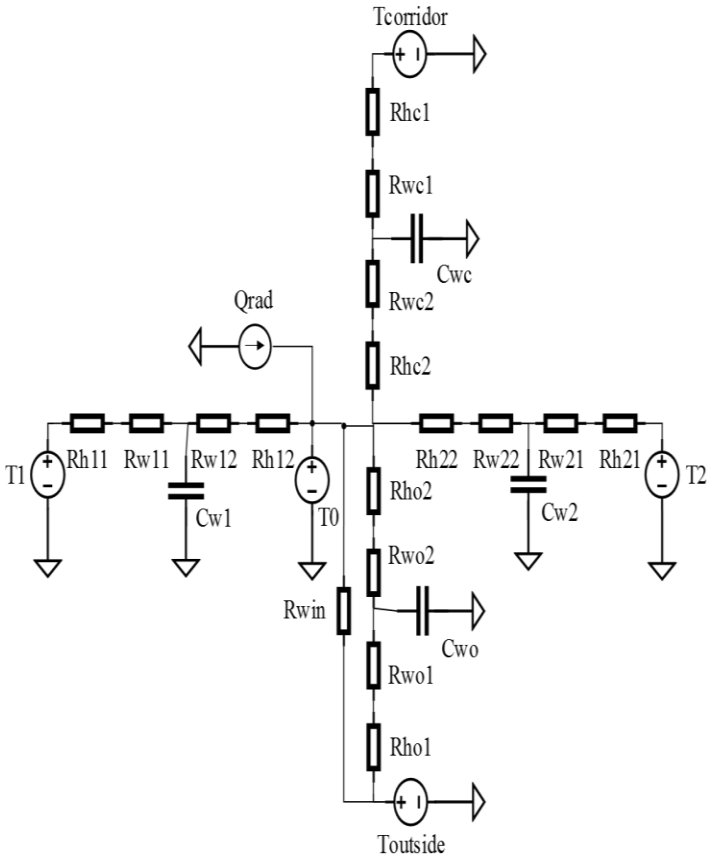


Fig. 2. Thermal model.

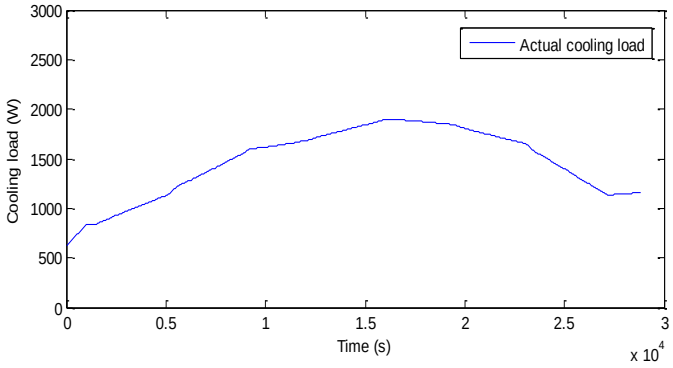


Fig. 3. Actual cooling load between 6:00 and 14:00.

III. Cooling Load Predicton Model

Load prediction method, based on the selection of a reference day according to similarities in occupancy schedule, used an enormous amount of historical data (cooling load and weather data). Then, the model is calibrated by using errors of previous two hours between the predicted cooling load and the actual one.

Modifications were attributed to this model. Certainly, current model requires a deep analysis of great past data. Hence, the model isn't close to reality due to big historical data. Another model's difficulty is how to select the reference day. Indeed, this selection is a complicated issue for some residential buildings because it depends on several parameters essentially environmental conditions. However, in the proposed model, it's enough to deal with current hourly data to predict the next hour cooling load. So, paper's value-added is reducing the amount of data, which presents model's inputs, and discarding the issue of reference day. To enhance model's performance, previous errors are used for calibration. Details will be explained in the following:

In order to predict hourly cooling load during a day, the model requires following inputs:

- $Q^{t}_{cooling/act}$: reference actual cooling load at time t
- T^{t}_{act} : reference actual outside temperature at time t
- T^{t+1h}_{pre} : predicted outside temperature at time t plus one hour
- cte : coefficient

Predicted cooling load for the next hour $Q^{t+1h}_{cooling/pre}$ is given by (6). Predicted outside temperature T^{t+1h}_{pre} are deduced from weather forecast. The coefficient cte is determined by the least square regression algorithm and equal to 0.81. Since measurements are done between 6:00 and 14:00, actual cooling load consumption during the first hour (6:00 – 7:00) is used as prediction's initial vector.

$$Q^{t+1h}_{cooling/pre} = Q^{t}_{cooling/act} \times \left(1 + cte \times \frac{T^{t+1h}_{pre} - T^{t}_{act}}{T^{t}_{act}}\right) \quad (6)$$

The errors of previous two hours between prediction results and actual cooling load are used to enhance model's accuracy. To do so, the average error of past two hours $\Delta Q_{cooling/t}$ is added to the predicted cooling load at time t plus one hour $Q^{t+1h}_{cooling/pre}$. This calibration method is given as follows:

$$Q^{t+1h}_{cooling/final} = Q^{t+1h}_{cooling/pre} + \Delta Q_{cooling/t} \quad (7)$$

$$\Delta Q_{cooling/t} = \frac{1}{2} \times \left(error(t) + error(t-1h)\right) \quad (8)$$

$$\Delta Q_{cooling/t} = \frac{1}{2} \times ((Q^{t}_{cooling/act} - Q^{t}_{cooling/pre}) + (Q^{t-1h}_{cooling/act} - Q^{t-1h}_{cooling/pre})) \quad (9)$$

For the second predicted hour, only the one hour previous error is used. To resume the proposed prediction model, current hourly consumption is sufficient for predicting the next hour cooling load. The second step is to enhance prediction by using average errors of past two hours between actual cooling load and prediction results. That's why, historical data of the previous hour (t minus one hour) and current cooling consumption at time t must be available. Load prediction model's data are illustrated in Fig. 4. Where, t is a reference time for predicting the next hour cooling load at time t plus one hour. So, this prediction is not an average hourly prediction but each instant t is used for computing the next hour consumption (t plus one hour).

Model's simulations are illustrated by Fig. 5. The first hour is used as prediction's initial vector. This justifies that the starting point of predicted cooling load is 7:00. There is a gap between the actual cooling load and the predicted one. To evaluate model's accuracy, the mean relative error MRE between predicted hourly cooling load and actual one is computed.

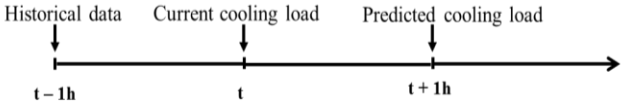


Fig. 4. Load prediction model's data.

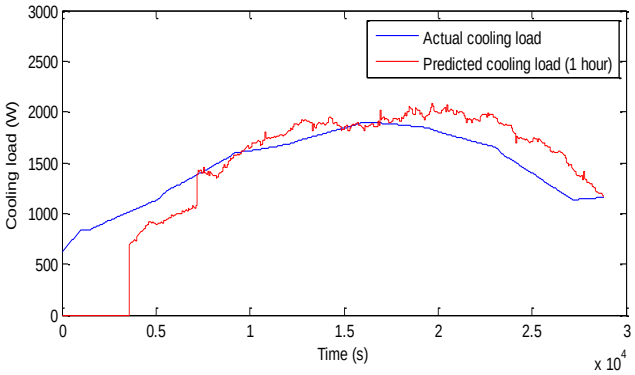


Fig. 5. Cooling load prediction based on a short sampling period equal to one hour.

This MRE is expressed by (10):

$$MRE = \frac{\sum_{k=1}^{N} \left| Q^{k}_{cooling/act} - Q^{k}_{cooling/pre} \right|}{\sum_{k=1}^{N} \left| Q^{k}_{cooling/act} \right|} \quad (10)$$

$$MRE_{1H} = 11.11\% \quad (11)$$

Where, N is the number of data points.

According to MRE, there is a remaining gap essentially during consumption peak. In fact, the node, proposed as a case of study, is located in the east. So, in the morning, the node is too sunny especially during the period (11:00 - 12:00) consequently cooling load consumption is significant. This justifies the presence of consumption peak. Due to the peak's important derivative, the sampling period should be decreased to minimize the gap between predicted results and actual cooling load.

Hence, the model has undergone some refinements that will be detailed in the following. In order to minimize this error, modifications were attributed to the hourly model. Instead of using hourly data, the sampling period was decreased to thirty minutes while respecting system response.

Indeed, this study deals with smart buildings targeted applications and shows consideration for real time constraints so that this paper is focusing on a short-sampling rate to be close to the real state. This justifies the sampling period's choice. Cooling load prediction model becomes:

$$Q^{t+30\min}_{cooling/pre} = Q^{t}_{cooling/act} \times \left(1 + cte \times \frac{T^{t+30\min}_{pre} - T^{t}_{act}}{T^{t}_{act}} \right) \quad (12)$$

Where $Q^{t+30\min}_{cooling/pre}$ is the predicted cooling load for the next thirty minutes, and $T^{t+30\min}_{pre}$ is the predicted outside temperature for the next thirty minutes.

This proposed method is about to deal with current data to predict next cooling consumption after thirty minutes. Model's accuracy was enhanced by adding the average error of two previous 30 minutes.

Equations (8) and (9) are rewritten as follows:

$$\Delta Q_{cooling/t} = \frac{1}{2} \times \left(error(t) + error(t - 30\min) \right) \quad (13)$$

$$\Delta Q_{cooling/t} = \frac{1}{2} \times ((Q^{t}_{cooling/act} - Q^{t}_{cooling/pre}) + (Q^{t-30\min}_{cooling/act} - Q^{t-30\min}_{cooling/pre})) \quad (14)$$

Through simulations given by Fig. 6, the 30 minutes model has largely improved cooling load prediction showing good adequacy between actual consumption and prediction results. This enhancement is substantiated by computing MRE.

$$MRE_{30\min} = 4.52\% \quad (15)$$

$$\frac{MRE_{1H}}{MRE_{30\min}} = 2.46 \quad (16)$$

This criterion highlights that prediction based on a short sampling period, fixed according to system response, has calibrated the model and enhanced its performance. In order to evaluate the proposed model's performance, it's compared with cooling load prediction model developed in [24]. This existing model's MRE is equal to 9.50%. Whereas, through (15), MRE is equal to 4.52%. Hence, the prediction model based on a short sampling period equal to thirty minutes is more accurate than the existing model. The proposed prediction model's methodology is also compared to other cooling load prediction based on physical model [19]. This latter approach may be sophisticated since it requires several thermal parameters, a deep physical representation of the building and extensive complex model. The development and validation of such approach are difficult due to several details involved and to a considerable amount of data hard to be measured or obtained. However, the prediction model, developed in this paper, is simplified and only depends on current data which are easily available.

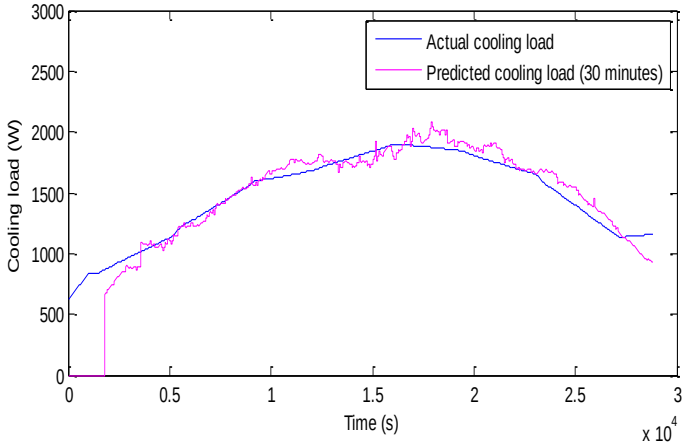


Fig. 6. Cooling load prediction model based on a short sampling period equal to 30 minutes (first significant day).

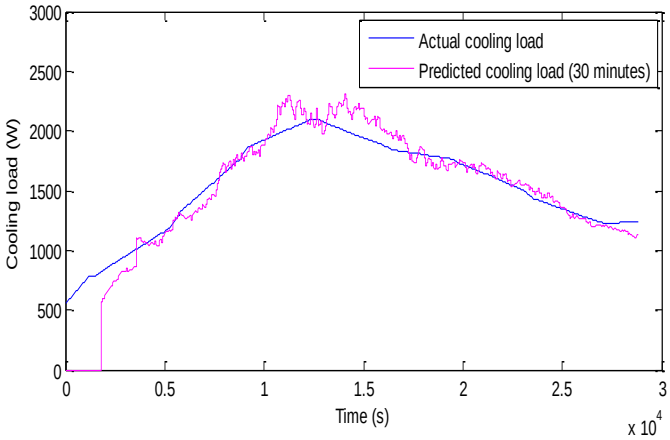


Fig. 7. Cooling load prediction model based on a short sampling period equal to 30 minutes (second significant day).

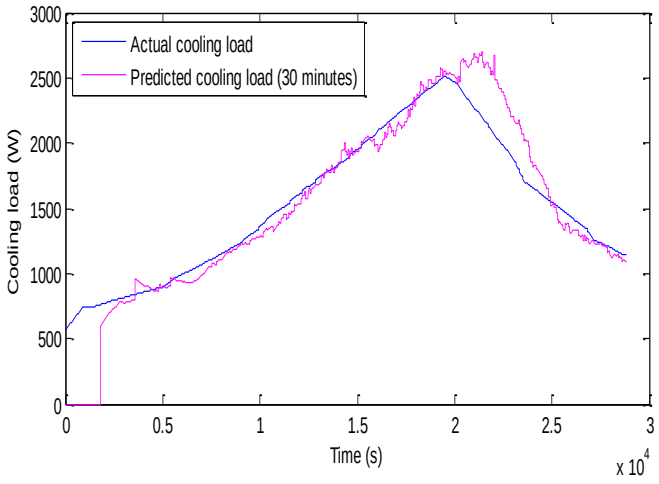


Fig. 8. Cooling load prediction model based on a short sampling period equal to 30 minutes (third significant day)

To generalize this proposed prediction, simulations, illustrated by Fig. 7 and 8, are applied to other significant days. These figures underscore that this prediction cooling model is valid even in other circumstances since meteorological parameters may vary enormously from a day to another due to their random fluctuation.

IV. Conclusion

A cooling load prediction model, based on a short sampling rate, is developed in this paper. In first phase, it's about an hourly prediction. So, the model depends on current hourly data used as a reference to predict the next hour consumption. To calibrate this model, previous errors between prediction results and actual cooling load are helpful to enhance prediction's accuracy. Other enhancement is to decrease the sampling period to 30 minutes with respect to system response. Prediction has largely enhanced with MRE equal to 4.52%.

To validate the proposed method, actual cooling consumption is computed. Since this consumption mainly depends on solar radiation, this last is estimated during a small sampling period with respect to smart buildings. Solar radiation model is based on the difference between maximum and minimum outside temperature during a short-sampling rate.

Simulations have been done in rooms located at National Engineering School of Tunis in real conditions and have shown good adequacy with measurements.

In perspective, this model can be generalized in the whole building while taking account several details such as building structure, orientation according the sun, windows' sizes and number and wall material of each local... . This prediction model can be also enhanced by involving other parameters such as natural ventilation and blinds' position.

As far as further work on the prediction issue, heating systems, considered as the dominant energy use in cold countries, deserves attention to be studied and analyzed in the sense of energy reductions. So, the proposed prediction model will be expanded and applied to the sector of heating load.

References

[1] Quoc-Dung, N., Yanis, H. S., Stéphane, P., Ujjwal, M., Md., "Automation generation of model for building energy management", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, No. 9, 2016.

[2] Yasser, M. A., Muhammad, N. K., "Simulation of building evacuation: performance analysis and simpliied model", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7, No. 8, 2016.

[3] Eunji, L. and Hyokyung, B., "Electricity usage scheduling in smart building environments using smart devices", The Scientific World Journal, Vol. 2013.

[4] Fazli, W. and Do, H. K., "An efficient approach for energy consumption optimization and management in residential building using artificial bee colony and fuzzy logic", Mathematical Problems in Engineering, 2016.

[5] Thomas, M. L., Richard, T. W., Marie-Claude, B. and Javad, M., "Data flow requirements for integrating smart buildings and a smart grid through model predictive control", Procedia Engineering, Vol. 180, pp. 1402-1412, 2017.

[6] Rigo-Mariani, R., Roccuzzo, V., Sareni, B., Repetto, M. and Robaum, X., "Power flow optimization in a microgrid with two kinds of energy storage", COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, Vol. 35, pp. 860-870, 2016.

[7] Jonathan, R. T., Nicholas J. K., Travis, M. A. and Tat, S. F., "Wireless sensor system for detection of occupants to increase building energy efficiency", Smart Cities Conference (ISC2), 2016 IEEE International, Italy, 2016.

[8] Bin, Z., Wentao, L., Ka, W. C., Yijia, C., Yonghong, K., Xi, L. and Xiong W., "Smart home energy management systems: concept, configurations, and scheduling strategies", Renewable and Sustainable Energy Reviews, Vol. 61, pp. 30-40, 2016.

[9] Althaher, S., Mancarella, P. and Mutale, J., "Automated Demand Response From Home Energy Management System Under Dynamic Pricing and Power and Comfort Constraints", IEEE Transactions on Smart Grid, Vol. 6, pp. 1874-1883, 2015.

[10] Yaser, I. A., Maria, d. M. C., José, D. A. and Antonio, R., "An economic model-based predictive control to manage the users' thermal comfort in a building", Energies, 10(3), 321, 2017.

[11] Fakeha, S., Manisa, P. and Saifur, R., "Integrated automation for optimal demand management in commercial buildings considering occupant comfort", Sustainable Cities and Society, Vol. 28, pp. 16-29, 2017.

[12] Kan, E. M., Kan, S. L., Ling, N. H., Soh, Y., Lai, M., "Multi-zone building control system for energy and comfort management", Hybrid Intelligent Systems, Advances in Intelligent Systems and Computing, Vol. 420. pp. 41-51, 2015.

[13] Khalfallah, E., Missaoui, R. El Khamlichi, S. and Ben Hassine, H., "Energy-efficient air conditioning: a case study of the Maghreb", the European Union, ESMAP and the World Bank, Washington, 2016.

[14] Sun, Y., Wang, S. and Xiao, F., "Development and validation of a simplified online cooling load prediction strategy for a super high-rise building in Hong Kong", Energy Conversion and Management, Vol. 68, pp. 20-27, 2013.

[15] Amjad, A. M., Hassan, M. and Ashkan, R. K., "Optimal smart home energy management considering energy saving and a comfortable lifestyle", IEEE Transactions on Smart Grid, Vol. 6, pp. 324-332, 2014.

[16] Riccardo, M. V., Francesco, B., Luigi, P., Martin, S. and Maria, P., "Energy management of a building cooling system with thermal storage: an approximate dynamic programming solution", IEEE Robotics and Automation Society, Vol. 14, pp. 619-633, 2017.

[17] Nada, S. A., Elfeky, K. E. and Attia, A. M. A., "Experimental investigatioons of air conditioning solutions in high power density data centers using a scaled physical model", International Journal of Refrigeration, Vol. 63, pp. 87-99, 2016.

[18] Ma, Y., Matusko, J. and Borrelli, F., "Stochastic model predictive control for building HVAC systems: complexity and conservation", IEEE Transactions on Control Systems Society, Vol. 23, pp. 101-116, 2015.

[19] Sandels, C., Brodén, D., Widén, J., Nordstrom, L. and Andersson, E., "Modeling office building consumer load with a combined physical and behavioral approach: Simulation and validation", Applied Energy, Vol. 162, pp. 472-485, 2016.

[20] Afram, A. and Janabi-Sharifi, F., "Theory and applications of HVAC control systems – A review of model predictive control (MPC)", Building and Environment, Vol. 72, pp. 343-355, 2014.

[21] Coakley, D., Raftery, P. and Keane, M., "A review of methods to match building energy simulation models to measured data", Renewable and Sustainable Energy Reviews, Vol. 37, pp. 123-141, 2014.

[22] Coninck, R. D., Magnusson, F., Akesson, J. and Helsen, L., "Toolbox for development and validation of grey-box building models for forecasting and control", Journal of Building Performance Simulation, Vol. 9, pp. 288-303, 2015.

[23] Maomao, H. and Fu, X., "Investigation of the demand response potentials of residential air conditioners using grey-box room thermal model", Energy Procedia, Vol. 105, pp. 2759-2765, 2017.

[24] Ying, J., Peng, X., Pengfei, D. and Xing, L., "Estimating hourly cooling load in commercial buildings using a thermal network model and electricity submetering data", Applied Energy, Vol. 169, pp. 309-323, 2016.

[25] Minjae, S., and Sung, L. D., "Prediction of cooling energy use in buildings using an enthalpy-based cooling degree days method in a hot and humid climate", Energy and Buildings, Vol. 110, pp. 57-70, 2016.

[26] Chirag, D., Lee, S. E., Junjing, Y. and Mattheos, S., "Forecasting Energy consumptivn of institutional buildings in Singapore", Procedia Engineering, Vol. 121, pp. 1734-1740, 2015.

[27] Radu, P., Vahid, R. D. and Jacques, M., "Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis", Energy and Buildings, Vol. 92, pp. 10-18, 2015.

[28] Yacef, R., Mellit, A., Belaid, S., and Sen, Z., "New combined models for estimating daily global solar radiation from measured air temperature in semi-arid climates: Application in Ghardaia, Algeria", Energy Conversion and Management, Vol. 79, pp. 606-615, 2014.

[29] Daut, I., Irwanto, M., Irwan, Y. M., Gomesh, N. and Ahmad, N. S., "Combination of Hargreaves method and linear regression as a new method to estimate solar radiation in Perlis, Nothern Malaysia", Solar Energy, Vol. 85, pp. 2871-2880, 2011.

[30] Owen, M. S., ASHRAE : Amercian Society of Heating, Refrigerating and Air-Coonditioning Engineers, Tullie Circle, N. E., Atlanta, GA 30329, 2009.

A Multilingual Datasets Repository of the Hadith Content

Ahsan Mahmood
Department of Computer Science
COMSATS Institute of information Technology,
Attock, Pakistan

Fawaz K. Alarfaj
Computer and Information Science Department
Al-Imam Mohammad Ibn Saud Islamic University, Al-Hofuf, Kingdom of Saudi Arabia

Hikmat Ullah Khan*
Department of Computer Science
COMSATS Institute of information Technology,
Wah, Pakistan

Muhammad Ramzan, Mahwish Ilyas
Department of Computer Science and Information Technology, University of Sargodha,
Sargodha, Pakistan

***Abstract*—Knowledge extraction from unstructured data is a challenging research problem in research domain of Natural Language Processing (NLP). It requires complex NLP tasks like entity extraction and Information Extraction (IE), but one of the most challenging tasks is to extract all the required entities of data in the form of structured format so that data analysis can be applied. Our focus is to explain how the data is extracted in the form of datasets or conventional database so that further text and data analysis can be carried out. This paper presents a framework for Hadith data extraction from the Hadith authentic sources. Hadith is the collection of sayings of Holy Prophet Muhammad, who is the last holy prophet according to Islamic teachings. This paper discusses the preparation of the dataset repository and highlights issues in the relevant research domain. The research problem and their solutions of data extraction, pre-processing and data analysis are elaborated. The results have been evaluated using the standard performance evaluation measures. The dataset is available in multiple languages, multiple formats and is available free of cost for research purposes.**

Keywords*—Data extraction; preprocessing; regex; Hadith; text analysis; parsing

I. Introduction

Data mining, Information retrieval and knowledge extraction have become attractive fields for the researchers during the last decade due to the birth of Social media [1]. These fields are getting researchers' interest because textual data over the internet is expanding exponentially during the last decade. The internet users are shifting from conventional methods of communications to online social networks at a rapid rate [2]. Sharing textual data over the internet is common due to the Social media channels [3]. Data mining and knowledge discovery tasks are carried out using the machine learning, statistical and database oriented approaches [4]. For the purpose of Knowledge discovery, researchers have used databases of different languages and domains to meet the requirements [5]. The recent research focuses on diverse techniques to make the unstructured data over the internet to be converted into such structured form so that various text mining and content analysis tasks can be accomplished and the data become more understandable as well machine readable [6].

Hadiths are regarded as one of the major sources of knowledge of the religion of Islam. The Hadith are the sayings of the Holy Prophet Muhammad, who is the last apostle according to Muslims. Analyzing Hadith text results in knowledge discovery from Hadith with the help of natural language processing methods. Although many researchers work in this field, there is no work solely focused on Hadith data. Moreover, it is not possible to compare these works with one another due to the unavailability of the common Hadith data corpus. There are a number of web sources which contains the Hadith contents. However, there is a lack of a repository containing data sets of Hadith for researchers to work in various research domains, such as text mining, data analysis, information retrieval and knowledge extraction. In this paper, we focus on preparation of a repository of the Hadith content data sets. The Hadith content is extracted from the reliable online sources. The volume of the data related to Islamic knowledge is present in a huge amount and is available in two major forms including the Quran and Hadith. Many researchers around the world who have worked on Natural language processing tasks, used Quran and Hadith data for knowledge discovery and Data mining tasks. However, most of the times, researchers have used Quran data for their research and knowledge discovery and have overlooked Hadith data. One of the most important reasons is the unavailability of data corpus of Hadith data [7]. A number of data mining researchers develop their own data corpus. Some of the researchers used the existing datasets, but those datasets are not present in enough amount considering the real data of Hadith. After a Data corpus of Hadith become available, it will become easy for the researchers to achieve Data mining and knowledge discovery tasks on Hadith data and compare performance of different works in the field.

In this research paper, we discuss our research contribution for the Hadith data repository preparation from different websites, processing them through different techniques and preparation of a data set repository of Hadith content that can

further be used for knowledge discovery and Data mining tasks by researchers. The rest of the paper is as follows: Section II reviews earlier studies, Section III discusses online Hadith resources used, Section IV discusses the details of the proposed research methodology and Section V discusses experimental setup and evaluation of our results before concluding the paper.

II. Background

Muslims believe that Muhammad (Peace and Blessings may Allah be upon him (PBUH)) is the last messenger of Allah. In religious terms, Hadith, meaning "tradition", is a report of the actions and sayings of Prophet Muhammad (PBUH). There are a number of Hadith books, but mainly six books, known as Sihah-e-Sita are regarded as the most authentic books.[1] The six most authentic books are Sahih Bukhari, Sahih Muslim, Sunnah Abu Dawood, Sunnah Nasai, Sunnah Tirmidhi, and Sunnah ibn Majah. After the Holy book of Quran, Hadiths are regarded as the second most important source of guidance in the Islam. Each Hadith consists of two things, the chain of narrators called isnad, and Hadith text called meeting. Rawi AL-Hadith, the person who reports a prophetic tradition is called Narrator or Rawi of Hadith. In all the books of Hadith, the content is divided into multiple parts. Usually, a book consists of volumes, each volume contains multiple chapters and each chapter has many Hadith referred in it. Each Volume and chapter has its own name and number while each hadith is assigned a number, list of narrator(s) and its content. While there are many parts of Hadith, the most important parts of Hadith are sanad and Matn that contains the actual textual content. Fig. 1 presents the hierarchy of a Hadith content in Hadith books.

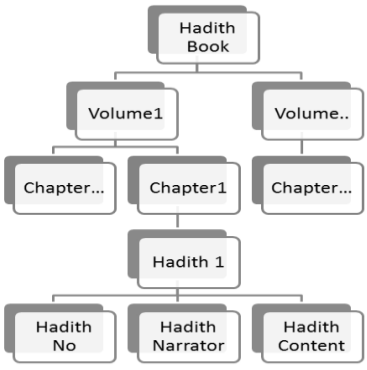


Fig. 1. Structure of the Hadith Data

A Hadith book consists of many volumes. Each volume consists of a number of chapters. The chapters are the topic-wise pre-categorization of Hadith. For data extraction, we extract all the required attributes that are associated with each Hadith including volume number and name, chapter number and name, hadith number, Sanad and Hadith matn.

III. Related Work

In the modern era, the amount of Information available on the web is growing at a very fast pace and have become the largest source of knowledge for the users, however processing and accessing such a large amount of data is not possible due to its vastness. [8]. Although the web traffic is constantly increasing thus the usage of regular expression is also increasing for packet inspections. Usually Regex matching is slow due to backtracking, but some new work has been done in this field using future matching techniques and achieve up to 70% performance boost [9]. Although many researchers have used the Hadith data for research purposes, there is no combined research work done in this regards. None of those researchers are able to compare their work with the others because there is no proper dataset of Hadith. In different hadith books, data has been divided into multiple parts and each book has its own format so each book takes its own time to extract data. Moreover, Hadith data is available in different languages that make it difficult to perform and compare the results of natural language processing tasks or to propose a framework of data extraction of Hadith that can be used with all the books or languages. Many researchers have worked to extract the hadith data from separate books using different kinds of techniques.

Aldhlan et al. [10], presents their understanding of new techniques in data mining to collect the Islamic knowledge from multiple resources, and represent the knowledge to the users in a better way. In their work, they used Hadith as a source of knowledge and proposed an approach for the classification of Hadiths in multiple categories using Supervised learning. They also discussed several ways to extract the knowledge from Hadith related to the goal of knowledge. Harrag et al., [11] perform Named Entity Recognition, extraction of words and entities from Hadith data. Siddiqui et al., [12] Proposed a system to extract isnad from Hadiths using Named entity extraction and classification in a form of network. Another work discusses extraction of the surface information from Sahih Bukhari. Author proposed a system based on Finite State Transducers (FST) to extract the knowledge from hadith text and their work shows 71% and 39% precision and recall respectively [13]. Another work proposed for NER for Hadith documents using Maximum Entropy Classifier, Naïve Bayes Classifier and Support Vector Machine based techniques and achieved an F-measure of 95.3% [14]. Mahmood A et al [15], proposed a framework for knowledge extraction from Sahih Bukhari urdu translation book. Other than named entity extraction, different sorts of other techniques have been used to extract the data from Hadith books. The method of handling missing data while extracting the Hadith data has been presented which is called missing data detector (MDD) [16]. The authors proposed a method based on the Isnad validity methods in Hadith science. Another tool proposed on the basis of Vector Space Model to let the users search for a particular Hadith from the complete Data repository with better precision. [17]. The authors performed Hadith classification according to the similarity between them.

The extraction of chain of narrators is also an important step. The authenticity of Hadith depends on authentication of the chain of narrators' and its content. The first step in the hadith authentication process is to determine chain of narrators' authentication. The chain of narrators' can be represented in the form of network graph. Network graph has

[1] https://en.wikipedia.org/wiki/Hadith. Accessed on August 14, 2016.

an element of the chain that can help in search of chain of narrators' for hadith [18]. The authentication measurement of the Hadith data is also important. Another research work share a reliable method to extract Hadith text from Islamic web pages [19]. In this work, researchers used Shiekh Al-Albani Hadith Database collection and finds out the correctness of each item. Working with different languages of Hadith has different experience due to their structures and diversity in the format of Hadith in each book. Therefore, it is not possible to propose a single algorithm for data extraction from the Hadith that can be applied to all the Hadith books of different languages. A study examined the knowledge discovery from Al-Hadith content by using classification algorithms [20]. It classifies the hadith content into one of the pre-defined books of hadith (classes in terms of classification). According to them Arabic language has a complex morphological structure and orthography variations. An android based application [21] targets Hadith retrieval system. Arabic is a major language in around two dozen countries of the world [22]. It differs from conventional retrieval systems because it allows Hadith retrieval in non-conventional manner. It allows the users to search for Hadiths using root based search. These kinds of sophisticated searches require extensive database, however, we kept the database as simple as possible and utilized regular expression (RE), which is supported by many modern programming languages.

According to another model, which create parts of each unit in *isnad* and *matn* and further process each part. It also creates a graph based on the relation between transmitters using an AraMorph morphological analyzer (RAM) and explains the text content [23]. In addition to hadith data, the work done on Quranic content is also valuable to mention. In the holy Quran the semantic web ontology has been applied for the purpose of search and extraction of semantic knowledge, including Quranic Wordnet, and mapping of domain ontology with higher level ontologies and it can also be applied to Hadith Data. [24]. Different kinds of models have been presented to perform the semantic search in the Holy Quran. A relational WordNet model [25] is presented to perform the semantic search in the Holy Quran that has been carried out in the latest tools and researchers used Surah AlBaqrah as a sample and produced their results on that basis. Quranic Arabic Dependency Treebank (QADT) model [26] reports on the approaches and solutions used in applying NLP to the Challenging Language of Quran. The authors proposed a complex linguistic model based on the Arabic language. It has been argued that memorization and methodologies are important factors that enhance the practices of memorization in the Islamic world [27].

Alqahtani M et al. [28], discussed different search techniques and proposed a model built on those techniques on Quran for searching purposes, including ontologies and semantic search tools for holy Quran. A sub-path mining algorithm built for the Holy Quran content to generate frequent patterns that can also be used for indexes and clusters in Quran Data [29].

Hadith books are not present in the form of Dataset on the internet. All the researchers who work on Hadith data develop their own Hadith dataset and do their experiments. In this research our focus is to collect Hadith data across different websites and develop a central Data corpus where users can download any Hadith book dataset. For this purpose, we use different websites to collect data of Hadith. The source websites are discussed in the next section. During data crawling, we faced some problems like slowness and noise in the web data, but we retrieved data chapter by chapter and volume by volume so data can be easily manageable. For Sahih Muslim data, we used SahihMuslim.com website that has all the Hadith of Sahih Muslim in a better format that can easily be crawled. There are some other websites that we have used for data crawling purpose as mentioned at the start.

IV. Research Methodology

In this research work, we select a number of sources and process textual data of Hadith and develop datasets of different books. Regular expressions [30] are used for data extraction purpose. We discuss the details about our extraction process that we use to extract the Hadith data from different sources.

A. Selection of Hadith Sources

The authentic and reliable sources are selected from the sources of Hadith content. Data from the sources, whether in the form of text on websites or in the form of documents such as in PDF form have been used for data extraction. Although Hadith books are present over the internet in a number of formats and types, we focus on Hadith books available in Unicode format so the data present in the book can be easily processed. The sources from which data is taken are as follows:

1) Hadith Websites

There are a number of websites that contain reliable Hadith content, possesses such a structure which allows us to apply regex and retrieve the desired data by matching the patterns [31]. It is notable that websites with AJAX [32] (asynchronous JavaScript and XML) do not allow its users to scrap the website content. Table I shows the list of books downloaded for extraction along with their source.

TABLE I. Hadith Book Sources

S#	Book Name	Source
1	Sahih Muslim English	http://sunnah.com/muslim
2	Sunan Abu Dawud	http://ahadith.co.uk/
3	Mawta Imam Malik	http://ahadith.co.uk/maliksmuwatta
4	Sahih Al-Bukhari	http://www.sahih-bukhari.com/, http://hadithcollection.com/

B. Hadith Content Extraction

Regular expressions search for a particular pattern and retrieve output based on the pattern. It helps to extract the required tokens from Hadith Data but we face the issues of data variability as each book has its own format and in each book there are variations of length, content and difference in structure.

In case of Sahih Bukhari Data extraction, the proposed regex extracts all the parts of Hadith easily as Sahih Bukhari data on the website is properly managed. Fig. 2 shows the extraction process from Hadith in Sahih Bukhari. Moreover, the entity relationship description diagram (ERD) presented in Fig. 3 that shows the entities as well as the attributes extracted. The diagram shows the structure of Mawta Imam Malik Hadith book. Each book has its own structure and we need to create a separate database design for each book. Fig. 3 shows an ERD diagram of Hadith Book "Mawta Imam Malik".

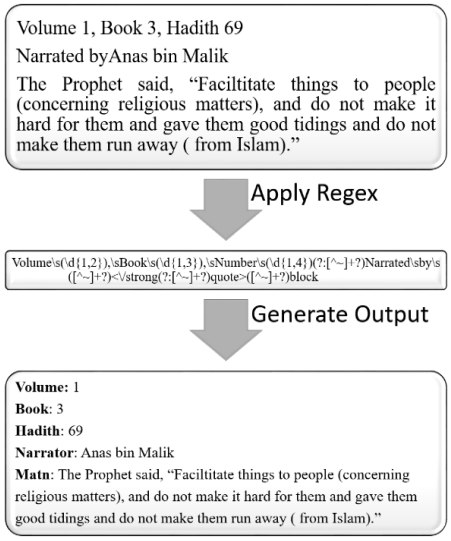


Fig. 2. Sahih Bukhari data extraction process.

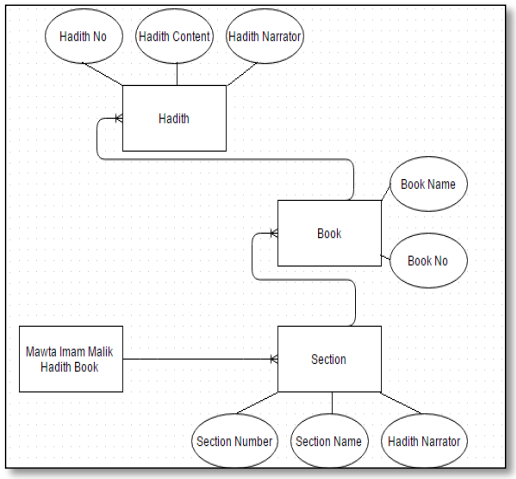


Fig. 3. ERD diagram for database of Mawta Imam Malika Hadith book.

1) Regular Expression

Regular expressions or Regex is used to search and retrieve a particular pattern in a stream of text document. The output generated by Regex can further be used to explore the extracted text. According to the context of the textual data, different types of regular expressions used in this research. Writing a regex is a very sensitive work as for detecting a particular text pattern, different regex can be used. The major priority during regex proposition is to write a regex with lowest possible steps and least backtracking. The higher number of steps involves matching a pattern, the higher system resources such as memory and processing power use of the system. According to Fang Yu et al. [33], firm and efficient ways of regex matching are required.

a) Regex example for recognizing Hadith Sanad and Matn

Below is an example of Hadith text. In order to extract the Sanad and Matn from below example a regex can be proposed. Regex shown in Fig. 4 extracts Sanad and Matn part of the hadith from below example.

Narrated by Al-Hasan: 'Amr bin Taghlib said, "Some property was given to the Prophet and he gave it to some people and withheld it from some others. Then he came to know that they (the latter) were dissatisfied. So the Prophet said, 'I give to one man and leave (do not give) another, and the one to whom I do not give is dearer to me than the one to whom I give. I give to some people because of the impatience and discontent present in their hearts, and leave other people because of the content and goodness Allah has bestowed on them, and one of them is 'Amr bin Taghlib." 'Amr bin Taghlib said, "The sentence which Allah's Apostle said in my favor is dearer to me than the possession of nice red camels."

```
Narrated\s(?:[Bb]y|[Ff]rom)\s(.+):\s?([^+]+)(?<=').+?bin.+?(?=\s)
```

Fig. 4. Regex Example3 for extraction of narrator and content.

b) Regex example for recognizing Hadith Number, Sanad and Matn data from Html Source

In the below example, there are three attributes of Hadith that can be extracted. These attributes are Hadith No, Hadith Sanad and Matn. Regex shown in Fig. 5 can be used to Extract those attributes from the below Hadith text.

<aname=18.1.1> 18.1.1

</td><td class="QuranData" bgcolor="#FFFFFF" valign="top"> Yahya related to me from Malik from Nafi from Abdullah ibn Umar that the Messenger of Allah, may Allah bless him and grant him peace, once mentioned Ramadan and said, "Do not begin the fast until you see the new moon, and do not break the fast (at the end of Ramadan) until you see it. If the new moon is obscured from you, then work out (when it should be)."

</td></tr><td class="QuranData" bgcolor="#FFFFFF" valign="top">

```
(\d{1,3}\.\d{1,3}\.\d{1,3}).+(?=;);(.+)that([^+]+(?=&))
```

Fig. 5. Regex Exanmple 4 for extraction of Narrator and Hadith cContent from Html document.

1) Issues and solutions

While working on text processing some minor issues that arises during the data retrieving and saving to database. At some places, those issues are negligible, but some issues are non-negligible. One of the major issues in Hadith text processing is escape sequences in the textual data when saving or retrieving the data into the databases. These escape

sequences become even more challenging when processing the data that is present on a website. In these cases, we remove escape sequence before retrieving. Therefore, sometimes when we process the Hadith data, we are not able to remove the escape sequence from textual data and backslashes are visible in the text when we retrieve the Hadith text for further processing.

C. Dataset Preparation

After text extraction, we execute other steps to in text processing to save the data into the database and develop a proper data corpus. Data preparation steps are shown in Fig. 6.

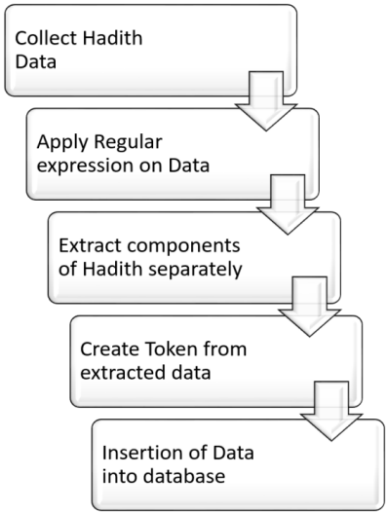


Fig. 6. Dataset preparation steps.

1) Diversity in Dataset preparation

As each of the Hadith books has its own format, we cannot combine all the datasets in one dataset. During dataset preparation and tokenization of the text we faced different kinds of problems.

An example of a webpage is given in Fig. 7. This example shows some Hadiths from Sahih Bukhari website.

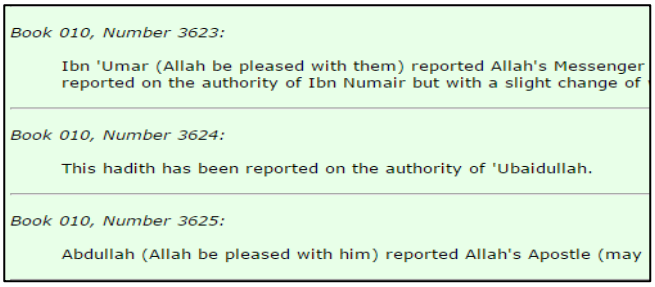


Fig. 7. Hadith website data example.

The problem however with this book is that it does look so simple in the webpage, but a lot of customized fonts etc. are present. Therefore, its source file has a lot of unnecessary things so to cope with these issues our regex became complex. In Fig. 8, it can be seen that the source code has a very huge amount of unnecessary text and small amount of information so proposing and debugging regex in such cases are challenging.

```
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Ibn Umar</strong></td></tr>
same as above Hadith 59. </blockquote></td></tr>
14"><strong><img src"..="" image="" book.gif"="" align="absbottom"> Volume 1,
ed by Anas bin Malik</strong></td></tr>
While we were sitting with the Prophet in the mosque, a man came riding on a
was sitting amongst us (his companions) leaning on his arm. We replied, "This
```

Fig. 8. HTML source of Hadith website data.

In these cases, our regex become complex while in case our data is simply available on a webpage with small amount of noise in the data, the regex become simple and it can extract the hadith text with the smallest number of steps required to extract the Hadith text. Some of our proposed regex are given in Table II.

TABLE II. SAMPLE REGEX FOR DIFFERENT BOOKS

Type	Regular Expression
Sahih Bukhari Arabic	(?+[`^])حَدَّثَنَاReference[^`]+?Vol. (\d{1 ,3}),Book (\d), Hadith (\d{1,3})
Sahih Bukhari English	(\d{1,3}),\sBook\s(\d{1,3}),\sNumber \s(\d{1,4})(?::|;)?,?\s\r\n?([^%]+?)(?::| ;|,)?\s?\r\n?([^%]+?)(?:Volume|Book)
Sahih Muslim English Initial	(?:[^%]+?(?:by|of|narratedon)?([^%]+ ?)(?:that|</font))
Sahih Muslim English Final	(?:(\d{1,4}),(?:N|n)umber\s(\d{1,5}) :.?\r\n)(?:(.+?)(?:said|that|narrates|rep orted:|who|that|by.+?\)|authorityin||sa me.+?\r\n|like|by another chainof|\, while|Apostle|somanyahadith|Messg er|addition|reported,|authority.+?\r\n. +?authority|chai|slight|andhe|change ofwords))([^%]+?)(?:Book\s\d|Capp)
Sunan Abudawud English	(?:\s(\d{1,3}),\sNumber\s?(\d{1,4})[^%]+?Narrated\s(.+?):\s?\r\n([^%] +?)Book)
Sunan Abudawud Urdu	\:حدیث نمبرs(\d{1,3})[^%]+?\.\. (([^%]+?)(?:کہتے|روایت|مروی \?+[%^](:کہا|کہتی|r\n\r\n)

Due to the difference in structure of the Hadith books, we use different kinds of Regex for data extraction. In some books data was present in a particular format so extraction was easy while in other books the data was not available in easy format so Regex become complex. In some books, it is not easy to extract Sanad and Matn separately so proposed Regex are very complex thus show. Table II shows some examples of Regex for different books.

D. Website Creation

Fig. 9 shows this, we plan to launch this webpage in the near future; however, dataset a screenshot of the webpage that we have developed for the end users so that they can download their required dataset in their required format can be accessed by contacting first author.

Fig. 9. Website Hadith database.

Fig. 10 shows an activity diagram about how the user can choose against different options to download their required dataset.

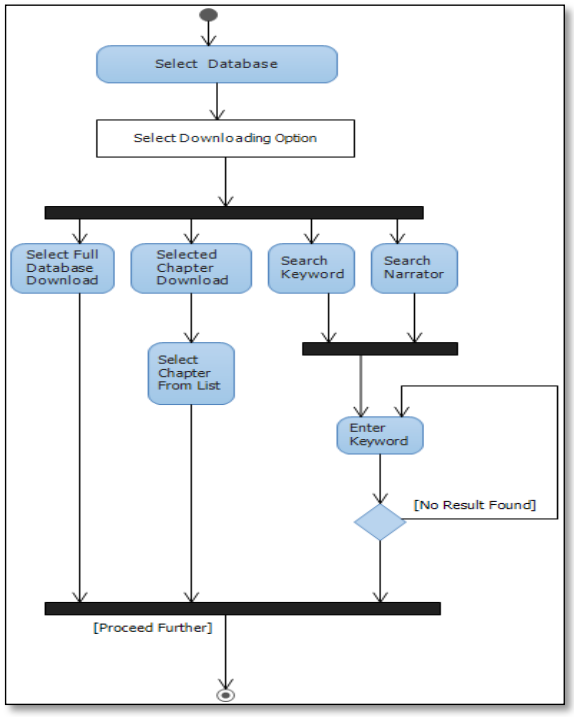


Fig. 10. Activity diagram of downloading the database.

V. Experiments and Results

We use Precision, Recall and F-Measure to measure the performance and accuracy of our methods. Due to the difference in structure of different books, the accuracy level is different. Precision can be given as the ratio of all entities extracted by our system to the correct entities extracted by our system. Its equation can be given as:

$$\text{Precision} = \frac{\text{No of Correct entities Extracted}}{\text{No of all entities Extracted}} \quad (1)$$

Recall can be given as the ratio of total entities extracted by human to the correct entities extracted by our system. Its equation can be given as:

$$\text{Recall} = \frac{\text{No of Correct entities Extracted}}{\text{No of Actual entities extracted}} \quad (2)$$

F1 score can be given as:

$$\text{F1} - \text{Score} = \frac{2 * \text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})} \quad (3)$$

On the basis of all the above calculation methods some of the results obtained by our system are given in Table III.

TABLE III. Results of Different Hadith Books

Book Name	Precision	Recall	F1 Measure
Sahih Muslim English	96%	91%	93%
Sahih Bukhari English	99%	99%	99%
Sunan Abudawud	100%	100%	100%
Mawta Imam Malik	100%	100%	100%

In the Sahih Muslim book, our accuracy is less because the structure of Hadiths is complex and noisy and format changes around almost every chapter. Fig. 11 shows the Precision, Recall and F1 Measure rate across different books.

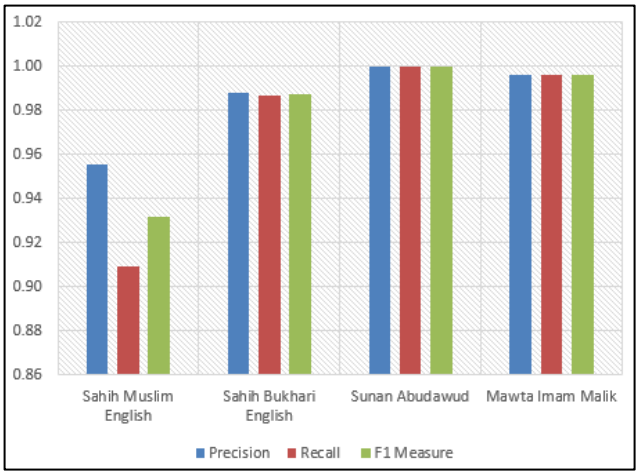
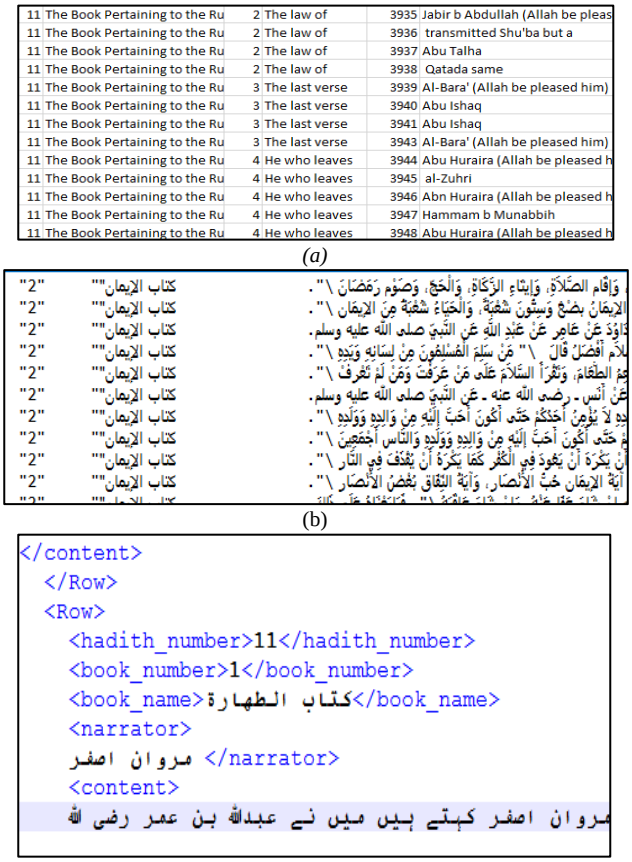


Fig. 11. Precision, recall, F1 measure performance.

Fig. 12(a), (b) and (c) shows Hadith data in different formats. Users can download the dataset in any of these formats. The formats in which user can download dataset are plain Text Format, CSV Format, XLS Format, XML FORMAT.



11	The Book Pertaining to the Ru	2	The law of	3935	Jabir b Abdullah (Allah be pleas
11	The Book Pertaining to the Ru	2	The law of	3936	transmitted Shu'ba but a
11	The Book Pertaining to the Ru	2	The law of	3937	Abu Talha
11	The Book Pertaining to the Ru	2	The law of	3938	Qatada same
11	The Book Pertaining to the Ru	3	The last verse	3939	Al-Bara' (Allah be pleased him)
11	The Book Pertaining to the Ru	3	The last verse	3940	Abu Ishaq
11	The Book Pertaining to the Ru	3	The last verse	3941	Abu Ishaq
11	The Book Pertaining to the Ru	3	The last verse	3943	Al-Bara' (Allah be pleased him)
11	The Book Pertaining to the Ru	4	He who leaves	3944	Abu Huraira (Allah be pleased h
11	The Book Pertaining to the Ru	4	He who leaves	3945	al-Zuhri
11	The Book Pertaining to the Ru	4	He who leaves	3946	Abn Huraira (Allah be pleased h
11	The Book Pertaining to the Ru	4	He who leaves	3947	Hammam b Munabbih
11	The Book Pertaining to the Ru	4	He who leaves	3948	Abu Huraira (Allah be pleased h

(a)

"2"	كتاب الإيمان	وَإِقَامِ الصَّلاَةِ، وَإِيتَاءِ الزَّكَاةِ، وَالْحَجِّ، وَصَوْمِ رَمَضَانَ \".
"2"	كتاب الإيمان	الإِيمَانُ بِضْعٌ وَسِتُّونَ شُعْبَةً، وَالْحَيَاءُ شُعْبَةٌ مِنَ الإِيمَانِ \".
"2"	كتاب الإيمان	دَاوُدَ عَنْ عَامِرٍ عَنْ عَبْدِ اللَّهِ عَنِ النَّبِيِّ صلى الله عليه وسلم.
"2"	كتاب الإيمان	لاَمِ أَفْضَلُ قَالَ \" مَنْ سَلِمَ الْمُسْلِمُونَ مِنْ لِسَانِهِ وَيَدِهِ \".
"2"	كتاب الإيمان	مِ الطَّعَامَ، وَتَقْرَأُ السَّلاَمَ عَلَى مَنْ عَرَفْتَ وَمَنْ لَمْ تَعْرِفْ \".
"2"	كتاب الإيمان	عَنْ أَنَسٍ ـ رضى الله عنه ـ عَنِ النَّبِيِّ صلى الله عليه وسلم.
"2"	كتاب الإيمان	دِهِ لاَ يُؤْمِنُ أَحَدُكُمْ حَتَّى أَكُونَ أَحَبَّ إِلَيْهِ مِنْ وَالِدِهِ وَوَلَدِهِ \".
"2"	كتاب الإيمان	مْ حَتَّى أَكُونَ أَحَبَّ إِلَيْهِ مِنْ وَالِدِهِ وَوَلَدِهِ وَالنَّاسِ أَجْمَعِينَ \".
"2"	كتاب الإيمان	أَنْ يَكْرَهَ أَنْ يَعُودَ فِي الْكُفْرِ كَمَا يَكْرَهُ أَنْ يُقْذَفَ فِي النَّارِ \".
"2"	كتاب الإيمان	آيَةُ الإِيمَانِ حُبُّ الأَنْصَارِ، وَآيَةُ النِّفَاقِ بُغْضُ الأَنْصَارِ \".

(b)

```
</content>
  </Row>
  <Row>
    <hadith_number>11</hadith_number>
    <book_number>1</book_number>
    <book_name>كتاب الطهارة</book_name>
    <narrator>
    مروان اصفر </narrator>
    <content>
    مروان اصفر کہتے ہیں میں نے عبداللہ بن عمر رضی اللہ
```

(c)

Fig. 12. (a) Hadith Database Downloaded in CSV Format, (b) Hadith Database Downloaded in Text Format Arabic Language, (c) Hadith Dataset Downloaded in XML format.

VI. Conclusion

In this paper, we discuss how we prepared hadith repository by applying regular expressions to extract the Hadith data from Multiple Hadith books that are present in different forms, and on both sorts of online & offline data. In the process, we crawled different websites to gather data from the sites directly and have also extracted the data from different sort of files like pdf, doc, etc. we then made a website in WPF to make all the databases downloadable for public. In the future, we plan to analysis data using different data mining and text mining algorithms. In addition, we plan to launch our website to provide online and free access of the dataset repository so that researchers all over the world may download the Hadith data from our website. The output dataset of Hadith can be used in many applications of data mining, text mining and information retrieval. Moreover, there are many fields of NLP which can get benefits from this dataset and can be applied on hadith dataset to extract knowledge in different ways through these datasets.

References

[1] Clinton Cardoza, Rupali Wagh, "Text analysis framework for understanding cyber-crimes," International Journal of Advanced and Applied Sciences, vol. 4, no. 10, pp. 58-63, 2017.

[2] Rehan Khan, Hikmat Ullah Khan, Muhammad Shehzad Faisal, Khalid Iqbal, Muhammad Shahid Iqbal Malik, "An Analysis of Twitter users of Pakistan," International Journal of Computer Science and Information Security, vol. 14, no. 8, 2016.

[3] Altaher, Altyeb, "Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting," International Journal of Advanced and Applied Sciences, vol. 4, no. 8, pp. 43-49, 2017.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, p. 18, 1996.

[5] Puteri N. E. Nohuddin, Zuraini Zainol , Angela S. H. Lee , A. Imran Nordin , Zaharin Yusoff, "A case study in knowledge acquisition for logistic cargo distribution data Mining framework," International Journal of Advanced and Applied Sciences, vol. 5, no. 1, pp. 8-14, 2018.

[6] Zulfiqar Ali, Waseem Shahzad , Syed Khuram Shahzad, "A review on comparative performance analysis of associative classifiers," International Journal of Advanced and Applied Sciences, vol. 4, no. 6, pp. 96-103, 2017.

[7] Saloot, M. A., Idris, N., Mahmud, R., Ja'afar, S., Thorleuchter, D., & Gani, A, "Hadith data mining and classification: a comparative analysis," Artificial Intelligence Review, vol. 46, no. 1, pp. 113-128, 2016.

[8] Crescenzi, Valter, Giansalvatore Mecca, and Paolo Merialdo, "Roadrunner: Towards automatic data extraction from large web sites.," VLDB, p. 1, 2001.

[9] Michela Becchi ,Anat Bremler-Barr ,David Hay ,Omer Kochba ,Yaron Koral, "Accelerating regular expression matching over compressed HTTP," 2015.

[10] Aldhlan, K.A, Zeki, AM, "Datamining and Islamic knowledge extraction: alhadith as a knowledge resource," 2010.

[11] Fouzi Harrag, Eyas El-Qawasmeh, and Abdul Malik Salman Al-Salman, "Extracting named entities from prophetic narration texts (Hadith)," Berlin, 2011.

[12] Siddiqui, Muazzam Ahmed, Mostafa El-Sayed Saleh, and Abobakr Ahmed Bagais, "Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification," 2014.

[13] F. Harrag, "Text mining approach for knowledge extraction in Sahîh Al-Bukhari," Computers in Human Behavior archive, vol. 30, pp. 558-566, 2014.

[14] Mohanad Jasim Jaber, Saidah Saad, "NER in english translation of hadith documents using classifiers combination," Journal of Theoretical and Applied Information Technology, vol. 84, no. 3, pp. 348-354, 2016.

[15] Ahsan Mahmood, Hikmat Ullah Khan, Zahoor-ur-Rehman, Wahab Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," in 13th International Conference on Emerging Technologies (ICET), Islamabad, 2017.

[16] Aldhaln, Kawther A., Akram M. Zeki, and Ahmed M. Zeki, "Knowledge extraction in Hadith using data mining technique," 2012.

[17] Fouzi Harrag, Aboubekeur Hamdi-Cherif, and Eyas El-Qawasmeh, "Vector space model for Arabic information retrieval—application to "Hadith" indexing," 2008.

[18] Nursyahidah Alias, Zulhilmi Mohamed Nor, Nurazzah Abdul Rahman, "Searching Algorithm of Authentic Chain of Narrators' in Shahih Bukhari Book," MALAYSIA, 2016.

[19] Shatnawi MQ, Abuein QQ, Darwish O, "Verification Hadith Correctness in Islamic Web Pages Using Information Retrieval Techniques," 2011.

[20] K. Jbara, ". "Knowledge discovery in Al-Hadith using text classification algorithm," Journal of American Science, vol. 6, no. 11, pp. 409-19, 2010.

[21] Azmi AM,Alkhalifah F, Alsaeed A, Barnawi Y, "Using non-conventional search schemes to retrieve Hadiths," 2014.

[22] Maheen Akhter Ayesha, Sahar Noor, Muhammad Ramzan, Hikmat Ullah Khan, Muhammad Shoaib, "Evaluating Urdu to Arabic Machine Translation Tools," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 90-96, 2017.

[23] Boella M, Romani FR, Al-Raies A, Solimando C, Lancioni G, "The SALAH Project: Segmentation and Linguistic Analysis of Hadith Arabic Texts," 2011.

[24] Khan HU, Saqlain SM, Shoaib M, Sher M, "Ontology Based Semantic Search in Holy Quran," International Journal of Future Computer and Communication, vol. 2, no. 6, 2013.

[25] Muhammad Shoaib M, Yasin MN, Khan HU, Saeed MI, Khiyal MS, "Relational WordNet Model for Semantic Search in Holy Quran," 2009.

[26] Kais Dukes, Tim Buckwalter , "A Dependency Treebank of the Quran using Traditional Arabic Grammar," 2010.

[27] M. Y. Alfi, "An Applied Linguistics Approach to Improving the Memorization of the Holy Quran: Suggestions for Designing Practice Activities for Learning and Teaching.," Journal King Saud Univ, vol. 16, pp. 1-32, 2004.

[28] Mohammad Alqahtani, Eric Atwell, "Arabic Quranic Search Tool Based on Ontology," Natural Language Processing and Information Systems, vol. 9612, pp. 478-485, 2016.

[29] Ali, Imran, "Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of the Arabic," International Journal of Software Engineering and Its Applications, vol. 6, no. 3, 2012.

[30] Chee Yong Chan, Minos Garofalakis, Rajeev Rastogi, "Indexed Regular Expression Matching," Springer US, pp. 1-6, 2014.

[31] Brodie, Benjamin C., David E. Taylor, and Ron K. Cytron, "A scalable architecture for high-throughput regular-expression pattern matching," ACM SIGARCH Computer Architecture News, vol. 34, no. 2, 2006.

[32] Garrett, Jesse James, "Ajax: A new approach to web applications," 2005.

[33] Fang Yu, Zhifeng Chen, Yanlei Diao, T.V. Lakhsman, Randy H.Katz, "Fast and memory-efficient regular expression matching for deep packet inspection," 2006.

A 1NF Data Model for Representing Time-Varying Data in Relational Framework

Nashwan Alromema
Department of Computer Science,
Faculty of Computing and Information Technology
Rabigh, Saudi Arabia

Fahad Alotaibi
Department of Information System,
Faculty of Computing and Information Technology
Jeddah, Saudi Arabia

***Abstract*—Attaching Date and Time to varying data plays a definite role in representing a dynamic domain and resources on the database systems. The conventional database stores current data and can only represent the knowledge in static sense, whereas Time-varying database represents the knowledge in dynamic sense. This paper focuses on incorporating interval-based timestamping in First Normal Form (1NF) data model. 1NF approach has been chosen for the easily implementation in relational framework as well as to provide the temporal data representation with the modeling and querying power of relational data model. Simulation results revealed that the proposed approach substantially improved the performance of temporal data representation in terms of required memory storage and queries processing time.**

***Keywords*—Time-oriented data model; time-varying data, valid-time data; transaction time data; bitemporal data; data model; N1NF; 1NF**

I. INTRODUCTION

Temporal Database (TDB) is database modeling technique that is considered as repositories of time-dependent data. Several research works have been conducting in this research area starting from the 1970s [1]. Some of these works deal with storage structure and temporal DBMS prototype, while others concentrated on query processing temporal time indexing [2]-[6]. The research work by Snodgrass in [7] treats the problems of temporal databases models, integrity constraints, storage structures, and implementation techniques using different DBMS. A debate within the last three decades was on how to model, implement and query temporal database in efficient way [8]. Since conventional relational database is used to store and process the data that refer to the current time [2], commercial DBMS and standards for the query language do not fully support temporal features [3], [21]. There are two fundamental methods of creating temporal database applications. The first one is an integrated method where the time-varying features of the data are supported by an extended or modified internal model in DBMS. The second method is the stratum method, in which the temporal features of the data are implemented in top of standard DBMS by a layer over DMBS which then changes the outcome into its temporal data [9]. The greatest efficiency is offered by the first method however the second method has greater popularity due to its realism.

A number of temporal data models have been proposed since the early 1980s. These data models are based on schema extension approach of relational data model. There are two common approaches for these extensions, tuple timestamping with First Normal Form (1NF), and attribute timestamping with Non-First Normal Form (N1NF). The study in [17] generalized the models under 1NF approach into Tuple Timestamping Single Relation (TTSR), and Tuple Timestamping Multiple Relations (TTMR) according to the way of data representations. TTSR approach is not efficient since it introduces redundancy, where attribute values that change at different time are repeated in multiple tuples. However, TTMR approach have solved the problem of data redundancy in TTSR, the problem with this approach is that the fact about a real world entity is spread over several tuples in several relations, and combining the information for an object a variation of join known as temporal intersection join would be needed which is generally expensive to be implemented. For N1NF, the problem with this approach, as stated in Jensen [6], there are some difficulties of temporal data models capturing an object in a single tuple such that "the models may not be capable of directly using existing relational storage structures or query evaluation techniques that depend on atomic attribute values". The study in [3] shows an approach of partial implementation of temporal database capabilities in top of widely used commercial DBMS, the model in this study is categorized under TTSR. This study also lacks most of temporal features as well as data redundancy of the proposed representational data model. The study in [10], [19] show an approach of temporal database representation in standard SQL under TTMR approach, the study explains number of examples of temporal data and how temporal manipulations of such data can be effected using standard SQL. A Column Level Temporal System (CLTS) proposed by Kvet in [20] is TTMR approach, the main issue of this model is to keep the duplicity of data minimal. As reducing the duplicities of the data is considered one of the important factors which improve processing speed to get a current snapshot and all data during life cycle of the database object [22]. Atay and Tansel in [18] proposed the Nested Bitemporal Relational Data Model (NBRDM) under N1NF approach [18], NBRDM model attached bitemporal data to attributes and defined a bitemporal relational algebra and a bitemporal relational calculus language for the proposed data model.

In this paper, we describe an approach for implementing temporal database in the framework of relational data model over the most widely used commercial DBMSs (Oracle RDBMS). The proposed approach does not significantly

This Research Project is sponsored by the Deanship of Scientific Research (DSR)-King Abdulaziz University –Jeddah-Saudi Arabia

change the procedures of designing and developing information systems. The major contributions of this research project can be formulated as follows:

- Describe the meaning and use of temporal features in the framework of relational data model.
- The approach is restricted to use the existing technology of designing and implementing databases applications.
- Incorporating temporal aspects need to minor modifications without affecting the performance of the parts of the system that do not use temporal data.
- The proposed implementation approach represents the temporal database in a data model that has expressive power, has less storage memory comparing to other works under TTSR approach, and efficient query processing.
- The implementation is easy, does not cost much, and based on relational database not on XML files as in [11].

This paper utilizes the following concepts on temporal database theory: the representation of real world time as a line. Every point in the line is referred to as an instance, a period is the time separating two instances, and an interval is the duration of loose segment of the time-line. Temporal data types in a temporal database can be identified as an instant of time, period and interval [7]. It is conceivable that time extends infinitely into the past of the future, as such when the relational database model has time introduced to it, it should be limited to delineate a particular time. Time-line chronons is the term for the reading of the time-line clock in the time-line. A time instance is delineated by each tick of a clock. To increase familiarity with temporal description times on a time line clock are expressed though a calendar.

The time line clock chronon is defined as day, month, and year on the Gregorian calendar. The date "22nd of June 2009" is an example. Granules are time points and the dividing scheme that splits the time line into a measurable collection of time segments is referred to as granularity and is an aspect of all temporal information [12]. Temporal databases are depicted by the discrete time model because it is easy and comparatively simple to use [9]. Temporal databases have formulated a taxonomy of time which identifies when a particular event happens or when a given statement can be regarded at factual. User-defined time is one interpretation of the time feature employed in temporal databases. It is expressed in the data that is of the date/time kind (the birth date column for example) and does not suggest anything correlated to the validity of the other columns or temporal time, wherein the column(s) that contain date/time information types are employed to mark the related tuple's time aspects. There are three categories of temporal time. Valid time: where in the related time is employed to determine when a particular statement (event-based) happened or when a particular statement (interval based) is regarded as being factual in the real world [13]. Transaction time: the related time is in reference to the period when the data was actually retained inside the database. Bitemporal-time: the related time is connected to the yield of valid-time and transaction time in the model of bitemporal data. Tuples are regarded as valid at instances of that time by rollback databases [7], [8].

II. METHODOLOGY

Designing any database systems usually goes through three phases, namely, (1) Conceptual design; (2) Logical design; and finally (3) Physical design. The temporal aspects of database schemas are complex and difficult; therefore it is an error-prone to design. In designing temporal database, the same steps as the mentioned above can be followed, in addition to that, defining new features concerning the time aspects because both conventional conceptual model and relational data model do not fully support time-varying aspects. The following steps summarize the proposed methodology for designing temporal database in relational data model.

- Designing the conceptual model for the business logic of the system and map it into conventional relational data model using the mapping methodology described in [5], [7], [10], where all temporal aspects that need to be modeled are ignored at this step. All conventional methods which are used to construct good relational database schema by analyzing the design and applying different forms of normalization should take place in this step.
- Adding the temporal aspect for all the database objects that need to keep the historical changes of the entities' data.

To make this process clear, an example (proof of concept) of the conceptual data model shown in Fig. 1 for EMPLOYEE and DEPARTMENT relations are mapped into relational data model shown in Fig. 2. Adding the temporal aspects to these two relations is shown in Fig. 4. These approaches can apply to any other domain of database technology like biomedical domain, business intelligent, metrological and any other domains.

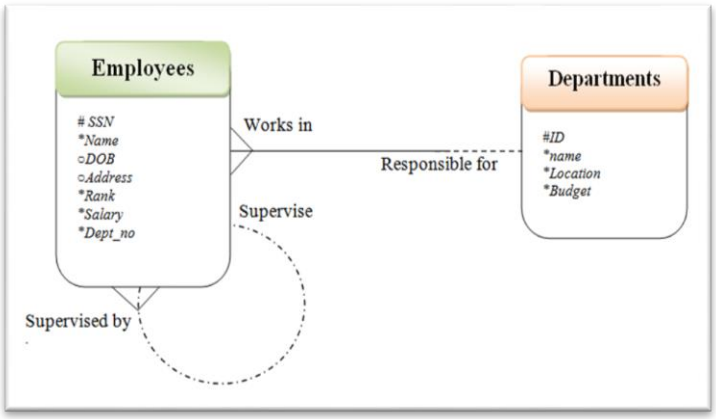


Fig. 1. The conceptual schema of EMPLOYEE and DEPARTMENT database entities.

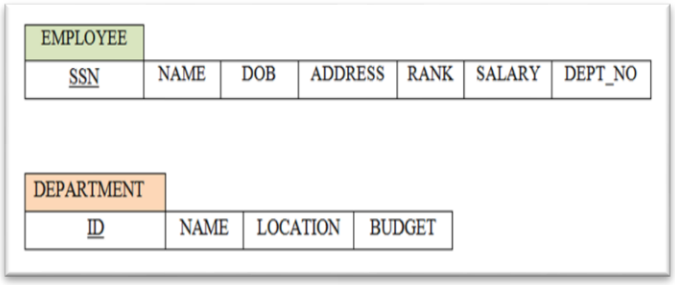


Fig. 2. The relational schema of EMPLOYEE and DEPARTMENT database relations.

III. Representations Methodology

The methodology of representing temporal database in this paper is accomplished by using Tuple Timestamp Historical Relational (TTHR) data model. Fig. 3 shows the conceptual structure of TTHR model. The database applications is directly connected to the main tables which hold the current valid time data, this feature gives the advantages that TTHR can be adapted to any functioning database systems without any changes to the infrastructure. The historical changes of each time-varying attributes in any table are stored in corresponding temporal database table (auxiliary tables) as shown in Fig. 3. The data representation of temporal database in TTHR is accomplished by firstly, defining the database object (entities /relations) for which we want to track the historical changes of the stored data, then we add for each such relations two additional columns Lifespan Start Time (LSST) and Lifespan End Time (LSET), which indicate the beginning and the end of the time interval within which the database object exists in the modeled reality [14], [19]. Secondly, for each such entity /relation, we create an additional relation with the same name as in the basic schema with the suffix VT, we use VT to indicate the valid time model. Example, the relational table EMPLOYEE in Fig. 2, is represented into temporal database (Fig. 4) by adding two additional columns LSST and LSET, after that we create a new table, the schema representation of Table_VT as an example of EMPLOYEE_VT will be: *EMPLOYEE_VT= (SSN, index, Update_A_VST, VET).*

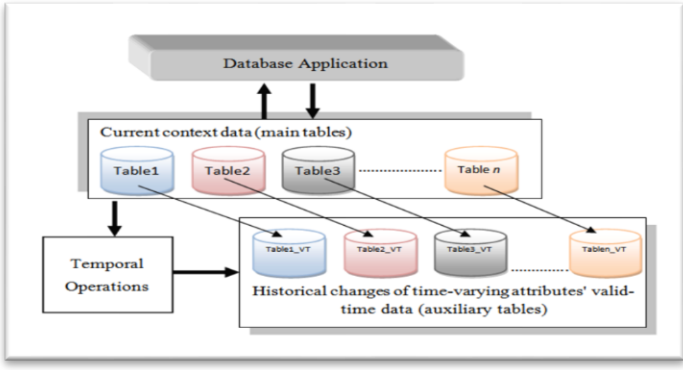
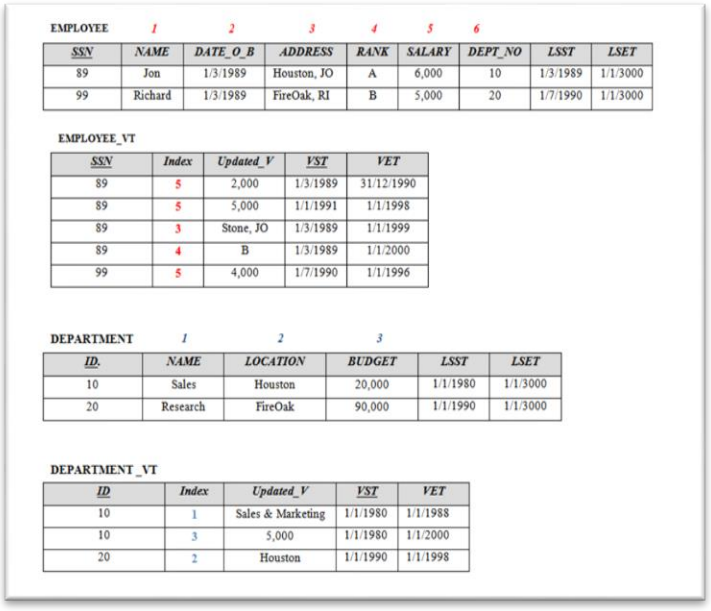


Fig. 3. The conceptual structure of TTSR Model.

The data in the basic table keeps the latest updated data (current data), whereas *Table_VT* stores the historical changes of the validity of the updated attributes in the basic table.



EMPLOYEE

SSN	NAME	DATE_O_B	ADDRESS	RANK	SALARY	DEPT_NO	LSST	LSET
89	Jon	1/3/1989	Houston, JO	A	6,000	10	1/3/1989	1/1/3000
99	Richard	1/3/1989	FireOak, RI	B	5,000	20	1/7/1990	1/1/3000

EMPLOYEE_VT

SSN	Index	Updated_V	VST	VET
89	5	2,000	1/3/1989	31/12/1990
89	5	5,000	1/1/1991	1/1/1998
89	3	Stone, JO	1/3/1989	1/1/1999
89	4	B	1/3/1989	1/1/2000
99	5	4,000	1/7/1990	1/1/1996

DEPARTMENT

ID.	NAME	LOCATION	BUDGET	LSST	LSET
10	Sales	Houston	20,000	1/1/1980	1/1/3000
20	Research	FireOak	90,000	1/1/1990	1/1/3000

DEPARTMENT _VT

ID	Index	Updated_V	VST	VET
10	1	Sales & Marketing	1/1/1980	1/1/1988
10	3	5,000	1/1/1980	1/1/2000
20	2	Houston	1/1/1990	1/1/1998

Fig. 4. The relational schema of EMPLOYEE and DEPARTMENT database relations.

A. Modification Operation

Modify temporal data database is a challenges because of the time dimension [15]. In this representational data model, we consider the insertion, deletion, and update of records in the table of the basic schema, the data in the Table_VT are updated automatically using database triggers or application's function. The following is the rules of data modification operations:

Insertion Operation: Inserting a new record into a table of the basic schema is accomplished as in conventional database, in addition to that the value of LSST field is set to the current date, and the value of the LSST field is set to a very far future time, for example, 1/1/3000. This date is always greater than the current date for the lifespan of the application. Inserting data into Table_VT is accomplished as consequences of updating any attribute in the table of the basic schema as it will be explained in updating operation. Thus, the data in the table of the basic schema always represents the latest current valid data.

Updating Operation: updating a record in a table of the basic schema results into the following actions: 1) If the updated data is an indexed attribute(s) as shown in Fig. 4, then the old value of this attribute and its index with the same value of the primary key and VST and VET fields are inserted into Table_VT, the value of VST and VST can be calculated as follows: (a) if this is the first time to update this attribute (this attribute has not been updated before or no record for this attribute is found in Table_VT), then VST value will have the same value as LSST in the table of the basic schema, and VET will be having the value of the current time. (b) If this attribute has been updated before, then VST will be having the value of VET plus one time granule of the latest update of this attribute. An example of this case is shown in Fig. 3, when the value of the SALARY attribute indexed by 5 has been updated (at time point '1/1/1996') for Richard, and then we look at Table_VT at

that time point, since no record has been found for this attribute and for this object, thus a new record for the updated value of this attribute and corresponding database object has been inserted into Table_VT table with these values:

$$(SSN:99, index:5, Update\_A:4,000, VST:'1/7/1990', VET:'1/1/1996').$$

Another scenario is shown in Fig. 4, for database object Jon when the SALARY attribute indexed by 5 has been updated (at time point '1/1/1998'), then we look for the latest update for SALARY attribute for Jon in Table_VT which is on '31/12/1990', We add one day (assumed in our example the time granularity is one day). The new row is will be having these values:

$$(SSN:89, index:5, Update\_A:5,000, VST:'1/1/1991', VET:'1/1/1998').$$

2) If the updated data is LSET attribute with instance time not equal to 1/1/3000, then this action is considered as logical delete of this record and this record stops to be a life or valid, as if one employee has resigned from the company.

Delete Operation: Delete a record form the basic schema is accomplished by setting the value of LSET to current time as explained in update operation.

In our proposed schema representation Table_VT tables keep the historical changes of the validity of the updated attributes in the basic table. Each record in Table_VT represents the validity of the changed attributes in the basic table during the time interval [VST, VET]. The historical changes of the validity is continuous, the timestamp in VST field coincides with the value of VET field of the preceding record with the same primary key. Fig. 3 shows the schema representation and the update operations on the basic schema tables (EMPLOYEE and DEPARTMENT) and the temporal tables (EMPLOYEE_VT, DEPARTMENT_VT).

The scheme described in Fig. 4 does not address many subtle issues specifically for temporal database [16]. An example of these issues are, constraints on the upper and lower time boundaries of interval-based data model, since the time is discrete, the above schema cannot guarantee that LSST should be less than LSST in the basic schema, and VST should be less than VET in temporal schema. Overlapping of the same fact that belong to the same object, an example the SALARY of an employee is $5000 in the interval [1/1/1990, 1/1/1999] and different salary $7000 is valid in the interval [1/1/1996, 1/1/2005]. Referential integrity constraints, it might have an object in reference relation refers to another object in the referee relation in different time points. As an example, in EMPLOYEE relation (reference) the foreign key Dept_no can have the ID value of DEPARTMENT relation (referee) that is either logically deleted or have interval lifespan time [LSST, LSET] that is not fully cover the interval lifespan time of EMPLOYEE object. Although, these issues can't be verified by conventional DBMS, these problems can be solved by an additional check through triggers of applications functions.

Although the historical changes of data are in temporal schema and the latest current valid data available from the basic schema, our approach is useful for the following reasons:

- Integrity constraints in the basic schema as well as temporal schema can be defined and implemented in DBMS easily without any major update to the existing applications. The purpose of this implementation is to ensure the creation of highly reliable databases.
- The proposed implementation removes data redundancy and satisfied high level of memory storage saving comparing to other implementation techniques discussed in [17], reducing the redundant data will help to facilitate efficient query execution.
- The tables in the temporal schema are updated only by insert operation when specific attribute in the basic schema table updated, thus the growth of this table depends on the frequency of attributes updates.

The current valid data in basic schema table helps in efficient query execution because some queries does not need to have temporal data. Temporal-joins involving data from the temporal schema are less efficient than joins of the tables in the basic schema.

B. Query Operations

Querying temporal databases represented by our approach using standard SQL2 can be classified into current query, sequenced query and non-sequenced query [7], [27]. Current query provide the current valid data which is in the basic schema table, while sequenced query provide the data that were valid during a certain interval of time where this data can be obtained from basic schema, temporal schema, or both depends on the complexity of the query, non-sequenced provide the historical changes of database objects' data. This work presents the following types of queries:

Current Queries: Current query is an ordinary query which provides current values of the data regardless of the time dimension. We project current queries on the basic table schema where the latest current values are stored for example the query that selects the current SALARY and RANK of an employee is

```
SELECT E.SALARY, E.RANK
FROM EMPLOYEE E
WHERE E.SSN = 89;
```

Some current queries involving time predicates for excluding/including valid/not valid lifespan entities, an example, the query that selects the latest SALARY of not valid lifespan employees is

```
SELECT E.SALARY
FROM EMPLOYEE E
WHERE E.LSET <> '1/1/3000';
```

This query selects all employees whom are logically deleted by setting the value of LSET to an instance time not equal to the END_TIME which we consider it in our approach equal to this date '1/1/3000'.

Sequenced Queries: Sequenced query provide the data that were valid during a certain interval of time, and the result of the query is valid-time table unlike current query which returns snapshot state. For example the query that returns the salary of an employee in a certain point of time or in a certain interval of time is

Q1 for point of time t

```
SELECT ES.SSN, ES.upadated_v
FROM EMPLOYEE_VT ES
WHERE ES.index = 5 and
         ES.VST <= t and
         ES.VET >  t and
         ES.SSN = 89;
```

Q2 for interval of time [t_1 ,t_2]

```
SELECT ES.SSN, ES.upadated_v
FROM EMPLOYEE_VT ES
WHERE ES.index = 5 and
      ES.VST < t2 and
      ES.VET >= t1 and
      ES.SSN = 89;
```

Q1 returns exactly one record, whereas Q2 returns one or more records because the time intervals for the salary historical changes of same employee might have an overlap with the input time interval [t1 ,t2]. No duplicated records will be returned for both queries because the data in our model are coalesced [10]. In contrast to other models that need more processing for coalescing function.

Non-sequenced query: provide the historical changes of a database objects' data during their lifespan time, the result of the query is valid-time table like sequenced query. The complexity of Non-sequenced queries depends on number of tables involved because the intervals in which the selected records were valid must be overlap for different tables. For temporal queries we need to define three functions for time interval manipulations as follows:

- Overlap([X,Y], [Z,W]) function takes two time intervals as a parameters, and returns one (1) if the time intervals are overlap and zero (0) otherwise. The following the code in SQL2 for this function.

```
CREATE FUNCTION OVERLAP (X IN NUMBER, Y
   IN NUMBER, Z IN NUMBER, W IN NUMBER)
   RETURN NUMBER    IS
BEGIN
    RETURN
     CASE
      WHEN X < W AND Y >= Z
        THEN 1
        ELSE 0
        END;
  END OVERLAP;
```

- Upper_bound(Y,W) function takes the tow upper boundaries of two time intervals as a parameters, and returns upper boundary of the overlapped time intervals. The following is the code in SQL2 for this function.

```
CREATE FUNCTION UPPER_BONUD (Y IN
   NUMBER, W IN NUMBER) RETURN NUMBER
   IS
   BEGIN
     RETURN
     CASE
      WHEN Y >= W THEN W
      WHEN Y < W THEN Y
      ELSE 0
      END;
  END UPPER_BONUD;
```

- lower_bound(X,Z) function takes the tow lower boundaries of tow time intervals as a parameters, and returns lower boundary of the overlapped time intervals. The following is the code in SQL2 for this function.

```
CREATE FUNCTION LOWER_BONUD (X IN
NUMBER, Z IN NUMBER) RETURN NUMBER
IS
BEGIN
  RETURN
     CASE
    WHEN X >= Z THEN X
    WHEN X <  Z  THEN Z
    ELSE 0
    END;
END LOWER_BONUD;
```

Since the current data are in the basic schema table and the historical changed data are in the temporal schema, then combining these data into one place can be accomplished by database views. We can create view for each time-varying attributes in the basic schema table, for example the SALARY_V view can hold the track log data including the current data for the salaries of all employees. The SALARY_V view is defined as follows:

```
CREATE VIEW SALARY_V AS
   SELECT E.SSN, E.SALARY,
          MAX (CASE
               WHEN ES.VET IS NULL
               THEN E.LSST
               WHEN ES.VET IS NOT NULL
               AND E.LSST > ES.VET
               THEN E.LSST
               WHEN ES.VET IS NOT NULL
               AND E.LSST < ES.VET
               THEN (ES.VET +1 )END)
             AS VST, E.LSET AS VET
FROM EMPLOYEE E LEFT OUTER JOIN
      (SELECT ES.SSN,
      TO_NUMBER(ES.UPADATED_V), ES.VST,
      ES.VET FROM EMPLOYEE _VT ES
      WHERE ES.ATT_INDEX = 5)
      ON E.SSN = ES.SSN
GROUP BY E.SSN, E.SALARY, E.LSET
UNION
```

```
SELECT SSN, TO_NUMBER(UPADATED_V), VST,
VET
FROM EMPLOYEE _VT WHERE INDEX = 5;
```

An example of the query that returns the track log of the salary of an employee for his lifespan time is

```
SELECT * FROM SALARY_V
WHEN SSN =89;
```

Another query that selects the track log information about salary and address (ADDRESS_V is view created by the same way as SALARY_V) of an employee is

```
SELECT S.SSN, AD.ADDRESS, S.SALARY,
  LOWER_BONUD(S.VST, AD.VST)AS VST
  ,UPPER_BONUD(S.VET,AD.VET)AS VET
    FROM ADDRESS_V AD, SALARY_V S
   WHERE SSN =89 AND AD.SSN = S.SSN
  AND OVERLAP (AD.VST,AD.VET, S.VST,
              S.VET) = 1;
```

Above queries can be applied for any other temporal information in employee or department tables. With time, the tracking log query that retains a data for a certain time interval might have a different data in other time interval.

IV. Results and Discussion

The performance evaluation of the proposed model is considered in terms of memory storage efficiency and query processing time. TTHR is compared with the main models in literature namely TTSR and TTMR. The Employees relation in Fig. 4 is represented by the three models, and the size in byte for the attributes in Employees relation is given as in Table I. The cost improvement of the memory storage is considered during one lifespan time and with a frequency of time-varying attributes update equal to 5. The results of memory storage efficiency for the three models are shown in Table I.

Note: Snp Stands for Snapshot and His for History.

TABLE I. Cost Model of Employees Relation Represented by TTSR, TTHR and TTMR

Attribute name	S/ Byte	Cost of data representation where $\delta = 5$								
		TTSR			TTHR			TTMR		
		Snp	His	Total	Snp	His	Total	Snp	His	Total
SSN	9	9	27	36	9	27	36	63	45	108
Name	100	100	300	400	100	0	100	100	0	100
B_date	10	10	30	40	10	0	10	10	0	10
Address	20	20	60	80	20	0	20	9	9	18
Tel_no	9	9	27	36	9	0	9	9	0	9
Spr_SSN	9	9	27	36	9	0	9	9	0	9
Dno	3	3	9	12	3	0	3	3	6	9
Salary	6	6	18	24	6	0	6	6	12	18
Rank	1	1	3	4	1	0	1	1	0	1
VST	10	10	30	40	0	30	30	70	50	120
VET	10	10	30	40	0	30	30	70	50	120
LSST	10	10	30	40	10	0	10	10	0	10
LSET	10	10	30	40	10	0	10	10	0	10
index	1	0	0	0	0	3	3	0	0	0
$\beta = S(\alpha)$	20	0	0	0	0	60	60	0	0	0
Total Cost				1176			371			542

Many parameters affect the cost improvements of TTHR over other models, Fig. 5 shows the cost improvements where all the parameters have been fixed with varying the values of the frequency of time-varying attributes update from 5 to 440 times in a period of time. TTHR has achieved significant saving in storage memory space that ranges between 68%-81% over TTSR approach, and 10%-32% over TTMR that is based on the average change of the time varying attributes. TTHR has achieved some significant saving in storage memory space that is roughly equal or greater than TTMR. The proposed temporal data model is suggested for its simplicity as fewer database objects will be needed to capture the temporal aspects of time-varying data compared to TTMR. Moreover, applying TTHR to an existing database application does not require many changes compared to TTMR. Moreover, the only need is to create the auxiliary relation to capture the historical changes of time-varying attributes but without touching the system itself. This is contrary to TTMR, where the relations need to be decomposed and the integrity constraints need to be redefined.

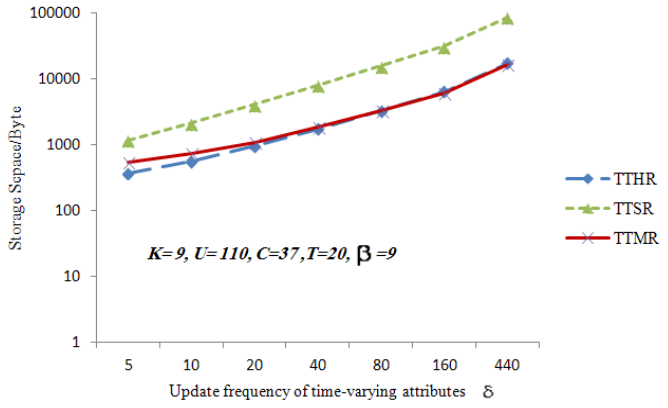


Fig. 5. Cost improvement of Employees relation represented by TTSR, TTHR and TTMR in one lifespan time [0, 10], and variations of δ.

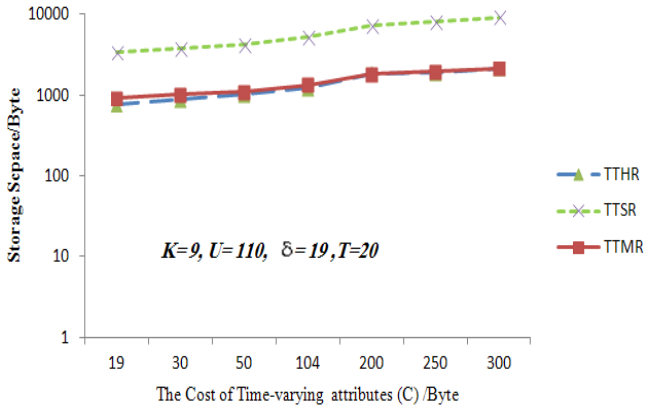


Fig. 6. Cost improvement of Employees relation represented by TTSR, TTHR and TTMR in one lifespan time.

Fig. 6 shows the storage costs of the temporal relational approach after freezing all the parameters and varying the sizes of the time-varying attributes. For these values, TTSR-based approach shows worse storage costs comparing to TTMR-based and TTHR-based approaches. However, the graph shows a positive indication that TTHR can be used as an efficient storage that is better than TTMR-based approach until the value of 150 byte. After this point it seems that both TTHR and TTMR have the same storage efficiency.

Fig. 7 shows the storage efficiency after freezing all the parameters and varying the sizes of key attributes (K) value variations. We increase value from 9 to 300 bytes. As we can see, the TTHR-based approach shows the best storage efficiency than the others. However, it is shown that the difference of storage efficiency is marginal between the TTHR-based approach and the TTMR-based approach.

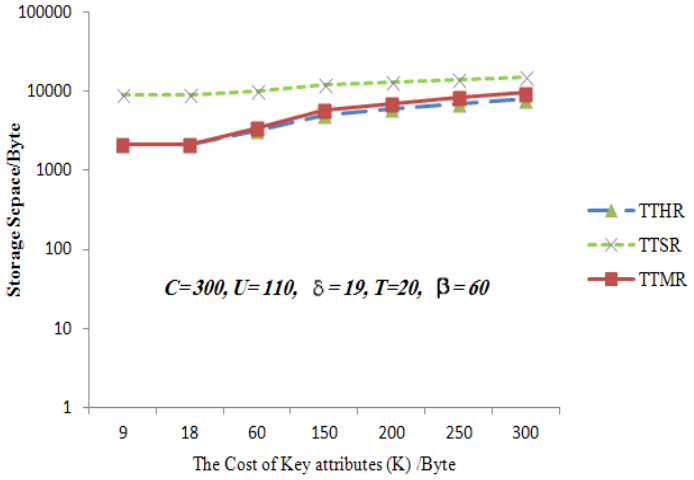


Fig. 7. Cost improvement of Employees relation represented by TTSR, TTHR and TTMR in one lifespan time [0, 10], and variations of Key attributes' size (K).

For query processing time, an experiment has been carried out on the database shown in Fig. 4 with a data set consists of 108,004 instances of Employees. This data set has been randomly generated in the three models to simulate real-world scenarios (the same approach has been taken by Anselma [23]. The SQL Trace facility and TKPROF (Transient Kernel Profiler) are two basic performance diagnostic tools that have been used for queries analysis in the three approaches. TKPROF program outputs the parameters of each query as CPU, Elapsed, Disk, and Query such that:

CPU(C): is time in seconds executing.

Elapsed (E): is the time in seconds executing.

Disk (D): is the number of physical reads of buffers from disk.

Query (Q): is the number of buffers gotten for consistent read.

Queries from 1 to 10 have been run in sequence for each approach. Table II shows the experimental results of executing these queries for each Model.

TABLE II. AN OUTPUT OF QUERY PROCESSING EXPERIMENTAL RESULTS

		TTHR			TTSR			TTMR		
Temporal	**Q**	**C**	**D**	**Q**	**C**	**D**	**Q**	**C**	**D**	**Q**
Current	Q1	0.00	3	3	0.00	1	4	0.00	15	25
	Q2	0.32	1193	8325	0.40	1251	8315	1.10	598	11467
	Q3	0.01	0	1199	0.01	0	1260	0.03	0	356
	Q4	0.00	0	3	0.00	0	7	0.00	0	6
	Q5	0.00	2	5	0.00	2	11	0.00	2	11
Non-sequenced	Q6	0.15	6	8332	2.43	0	351872	0.17	0	7552
	Q7	0.17	0	1206	2.17	0	351896	0.10	0	7672
	Q8	1.31	0	9538	5.84	0	696645	1.70	0	8054
Sequenced	Q9	0.01	12	18	0.01	5	30	0.00	6	12
	Q10	0.29	0	2869	1.03	0	95272	0.28	0	2038

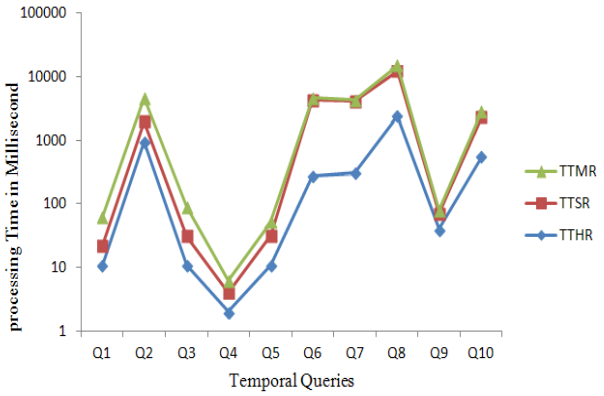


Fig. 8. Query processing time for the 10 queries in the three models.

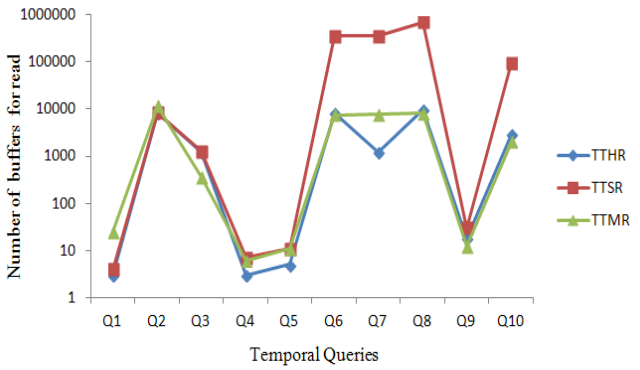


Fig. 9. Number of Buffers read in the three models for the 10 Queries.

From Table II, Fig. 8 and 9 have been plotted to compare the performance of each model in graphical view. It can be shown that TTSR satisfies good query performance in current query (Q1-Q5); the same performance is achieved by TTHR. However, TTMR costs a lot for current queries, but it costs less for both, sequenced (Q6, Q7 and Q8) and non-sequenced (Q9, and Q10) queries and the same performance is achieved by TTHR. TTSR costs a lot for both sequenced and non-sequenced queries due to coalesce function that needs to be applied to the query results to make sure the query result is in snapshot equivalence.

SQL developer suite with TKPROF has been used for these experiments. Measuring the performance of the query by only running the query few times is a pretty bad idea - equivalent to just accepting that the cost of the explanation plan that tells you the best query. Therefore, it is really a need to take into account what resources query is taking up and therefore how it could affect the production system.

V. Conclusion

The 1NF temporal data model proposed in this study uses a novel approach for modeling and implementing interval-based temporal database in relational framework [24]-[27]. In our approach the issues concerning the memory storage and query efficiency, and application development procedures are considered. All of these issues ensure the development of efficient and reliable temporal database over conventional DBMS. In this paper, we proposed an approach for representing temporal data that achieves saving in memory usage range from 68-81% over other temporal representations, and speed up the processing time of current snapshot data. Finally, our approach has better storage representation, reduce query complexities.

Acknowledgment

This paper was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University. The authors, therefore, acknowledge with thanks to DSR's technical and financial support.

References

[1] Findler, N. V., & Chen, D. (1973). On the problems of time retrieval of temporal relations causality, and coexistence. International Journal of Computer & Information Sciences, 2, 3, 161-185.

[2] Date, C. D., Darwen, H., & Lorentzos, N. A. (2003). Temporal data and the relational data model. San Francisco: Morgan Kaufmann.

[3] Novikov, B. A., & Gorshkova, E. A. (2008). Temporal databases: From theory to applications. Programming and Computer Software, 34, 1, 1-6. Pleiades Publishing, Ltd., 2008. Original Russian Text

[4] Tansel, A. U. (2004). On handling time-varying data in the relational data model. Information and Software Technology, 46, 2, 119-126.

[5] Elmasri, R., and Navathe (2000). Fundamentals of Database Systems. 3rd edition. Addison Wesley.

[6] Jensen, C. S., Clifford, J., Gadia, S. K., Segev, A., & Snodgrass, R. T. (1992). A glossary of temporal database concepts. ACM Sigmod Record, 21, 3, 35-43.

[7] Snodgrass, R. T., (2000). Developing Time-Oriented Database Applications in SQL, 1st edition, Morgan Kaufmann Publishers, Inc., San Francisco.

[8] Jensen, C. S., Snodgrass, R. T., & Soo, M. D. (1995). The tsql2 data model (pp. 157-240). Springer US. http://people.cs.aau.dk/~csj/Thesis/pdf/chapter12.pdf

[9] Patel, J. (2003). Temporal Database System Individual Project. Department of Computing, Imperial College, University of London, Individual Project, 18-June-2003, http://www.doc.ic.ac.uk/~pjm/teaching/student_projects/jaymin_patel.pdf

[10] Zimányi, E. (2006). Temporal aggregates and temporal universal quantification in standard SQL. ACM SIGMOD Record, 35, 2, 16-21.

[11] Wang, F., Zhou, X., & Zaniolo, C. (2006, April). Using XML to build efficient transaction-time temporal database systems on relational databases. In Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE'06 (pp. 131-131). IEEE.

[12] A-Qustaishat, M. (2001). A visual temporal object-oriented model embodied as an expert C++ Library. ADVANCES IN MODELLING AND ANALYSIS-D-,6, 3/4, 3-43.

[13] Bohlen, M. H., Busatto, R., & Jensen, C. S. (1998, February). Point-versus interval-based temporal data models. In Proceedings of 14th International Conference on Data Engineering, (pp. 192-200). IEEE.

[14] Dyreson, C., Grandi, F., Käfer, W., Kline, N., Lorentzos, N., Mitsopoulos, Y., ... & Wiederhold, G. (1994). A consensus glossary of temporal database concepts.ACM Sigmod Record, 23, 1, 52-64.

[15] Tansel, A. U. (2006). Modeling and Querying Temporal Data. Idea Group Inc.

[16] Tansel, A. U. (2004). Temporal data modeling and integrity constraints in relational databases. In Computer and Information Sciences-ISCIS 2004 (pp. 459-469). Springer Berlin Heidelberg.

[17] Halawani, S. M., & Romema, N. A. (2010). Memory storage issues of temporal database applications on relational database management systems. Journal of Computer Science, 6, 3, 296.

[18] Atay, C. (2016). An attribute or tuple timestamping in bitemporal relational databases. Turkish Journal of Electrical Engineering & Computer Sciences. (2016) 24: (pp. 4305 – 4321). doi:10.3906/elk-1403-39.

[19] Noh, S.Y., Gadia, S.K. and Jang, H., (2013). Comparisons of three data storage models in parametric temporal databases. Journal of Central South University, 20(7), pp.1919-1927.

[20] Kvet, M., Matiako, K. and Kvet, M., (2014). Transaction management in fully temporal system. In Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on (pp. 148-153). IEEE.

[21] Snodgrass R, Ahn I. Performance evaluation of a temporal database management system. Commun ACM 1986; 15:96-107.

[22] Arora, S. (2015). A comparative study on temporal database models: A survey. In Advanced Computing and Communication (ISACC), 2015 International Symposium on (pp. 161-167). IEEE.

[23] Anselma, L., Stantic, B., Terenziani, P., and Sattar, A. (2013). Querying now-relative data. Journal of Intelligent Information Systems, 41(2), 285-311.

[24] Halawani, S.M., AlBidewi, I., Ahmad, A.R. and Al-Romema, N.A., 2012. Retrieval optimization technique for tuple timestamp historical relation temporal data model. Journal of Computer Science, 8(2), p.243.

[25] Nashwan Alromema, Mohd Shafry Mohd Rahim and Ibrahim Albidewi, "A Mathematical Model for Comparing Memory Storage of Three Interval-Based Parametric Temporal Database Models" International Journal of Advanced Computer Science and Applications (ijacsa), 8(7), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080741

[26] Alromema, N.A., Rahim, M.S.M. and Albidewi, I., 2016. Temporal Database Models Validation and Verification using Mapping Methodology. VFAST Transactions on Software Engineering, 11(2), pp.15-26.

[27] Ab Rahman Ahmad, N.A., Rahim, M.S.M. and Albidewi, I., 2015. Temporal Database: An Approach for Modeling and Implementation in Relational Data Model. Life Science Journal, 12(3).

Sentiment Analysis using SVM: A Systematic Literature Review

Munir Ahmad[1], Shabib Aftab[2], Muhammad Salman Bashir[3], Noureen Hameed[4]
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

***Abstract*—The world has revolutionized and phased into a new era, an era which upholds the true essence of technology and digitalization. As the market has evolved at a staggering scale, it is must to exploit and inherit the advantages and opportunities, it provides. With the advent of web 2.0, considering the scalability and unbounded reach that it provides, it is detrimental for an organization to not to adopt the new techniques in the competitive stakes that this emerging virtual world has set along with its advantages. The transformed and highly intelligent data mining approaches now allow organizations to collect, categorize, and analyze users' reviews and comments from micro-blogging sites regarding their services and products. This type of analysis makes those organizations capable to assess, what the consumers want, what they disapprove of, and what measures can be taken to sustain and improve the performance of products and services. This study focuses on critical analysis of the literature from year 2012 to 2017 on sentiment analysis by using SVM (support vector machine). SVM is one of the widely used supervised machine learning techniques for text classification. This systematic review will serve the scholars and researchers to analyze the latest work of sentiment analysis with SVM as well as provide them a baseline for future trends and comparisons.**

Keywords*—Sentiment analysis; polarity detection; machine learning; support vector machine (SVM); support vector machine; SLR; systematic literature review

I. Introduction

With the rapid development of mobile internet in the recent years, the usage of micro-blogging websites have seen a considerable increment. On the other hand, trend of sharing the views and experience regarding products and services is increasing day by day. Moreover, users rely on the feedback of the previous customers before targeting the new product or service to purchase. In the same way the companies can get the reviews about their products or services from their clients via micro blogging sites (Facebook, twitter, etc.) to explore and analyze the response and ultimately can improve those targeted products or services. However, it is not practically possible to read all the reviews in tweets. Several researchers have been working to develop automated techniques and algorithms for sentiment analysis and text classification. The term sentiment analysis is basically aims to classify the given text into positive, negative and neutral category. Three basic approaches are available in the literature today for sentiment analysis: Lexicon driven, Machine learning based, and Hybrid (integration of lexicon and machine learning). The authors in [1] explored different Lexicon driven sentiment analysis tools and techniques. In [2], different machine learning techniques have been discussed in detail which are used for sentiment analysis. Moreover, in order to take the results up a notch, researchers combined lexicon based techniques and machine learning techniques to formulate a hybrid framework to dig-up even better results as explained in [3]. SVM belongs to the supervised category of machine learning algorithms. Supervised machine learning algorithm is one which has to be trained first with the pre identified output class (training data) and then it makes itself capable to classify the real input data (test data). Several annotated datasets regarding different domains are available which are used by machine learning algorithms for classification and sentiment analysis. Some of these annotated datasets include: the customer review dataset [4], [5], Pros and Cons dataset [6], Amazon product review dataset [7] and gender classification dataset [8]. In this study, published papers regarding sentiment analysis with SVM technique from year 2012 to 2017 are analyzed. Two online libraries are used for this research: ACM and IEEE. Nine hundred and one articles were selected based on the particular query strings. After following the complete systematic framework, 8 papers were finally selected for in-depth and detailed review.

Further organization of this paper is as follows. Section II describes the related work in this domain. Section III defines research methodology used for this SLR. Section IV presents in-depth review of selected papers. Section V discusses the findings of this detailed review. Section VI finally concludes the paper.

II. Related Work

Development and refining the automated techniques of sentiment extraction and analysis is one of the hot research topics today. Many researchers have worked on sentiment analysis techniques via different approaches (Lexical, Machine Learning and Hybrid) however, in-depth analysis and review of latest literature on sentiment analysis with SVM was still required. Some of the related studies on sentiment analysis are as follows.

Authors in [9] conducted a systematic literature review regarding opinion mining from the reviews of mobile app store users. The researchers focused on the importance of mobile applications in now days and further highlighted the increasing demand of user reviews about those apps. Obviously these reviews are crucial for the new users, who are going to buy these apps and also for those who develop or sell these apps.

The authors highlighted the proposed solutions of mining problems, and also identified the remaining unsolved issues and new challenges. In [10], a systematic literature review is conducted to analyze the current state of Arabic text mining. For this review, more than one hundred papers are selected from different reliable sources and then were classified according to their specific domains. A quantitative analysis of selected articles is also conducted with respect to publication type, year, category and contribution. The researchers in [11] conducted a literature review on sentiment analysis and opinion mining of social issues. The selected papers have taken the data from social web sites. According to authors, different types of classification techniques, if combined, can provide the better results. In [12], a literature survey is conducted about opinion and spam mining. For this purpose, most cited research articles from these domains are considered. Authors found the proposed architecture and methods imperfect in those selected researches. They highlighted that the important thing in spam detection is not only the spam identification but also, not to filter the real ones. In [13], a systematic literature review is conducted for the classification of burn care parameters with machine learning techniques. A total of 1503 topic relevant research articles were primarily selected, after screening and extracting the most relevant literature, 15 studies were selected for the analysis. All the selected studies demonstrated the benefits of machine learning techniques in burn care however different research articles reflected different accuracies. The authors in this SLR focused on the benefits of using machine-learning techniques in burn care as well as highlighted the importance of common metrics and goals for effective evaluation and validation of these techniques. In [14], the authors have performed sentiment classification of Arabic tweets by using Naïve Bayes, Decision Tree and Support Vector Machine. In this study, a framework for Arabic tweets classification is followed which consisted of several subtasks such as: Term Frequency Inverse Document Frequency (TF-IDF) and Arabic stemming etc. Moreover three information retrieval metrics are used for performance evaluation: precision, recall, and f-measure. In [15], a literature review is conducted covering the domain of data mining applications in customer relationship management. The study considered the research literature from year 2000-2006, covering 24 journals. For this study, 900 articles were shortlisted and then 87 most relevant papers were selected to classify in four CRM dimensions i.e. customer identification, customer attraction, customer retention and customer development. In [16], the authors have predicted the rainfall in Malaysia by using machine-learning techniques. They have used following classification algorithms: Naïve Bayes, Decision Tree, Support Vector Machine, Neural Network and Random Forest. A comparative analysis was performed to identify the particular technique which can bring good results with little amount of training data. The comparative analysis showed that Decision Tree and Random Forest both can get well trained by using lower amount of training data and can get high F-measure score. However, Support Vector Machine and Naive Bayes both showed lower F-measure score, when trained with lower amount of training data. Neural Network required large amount of training data to predict very little amount of test data. In [17], the authors have focused on the effects of preprocessing feature in sentiment classification process. They have classified the 1000 Arabic tweets and compared their implemented stemmer with light stemmer. They have used two approaches for comparative analysis, Machine Learning and Semantic Orientation. According to authors, the used stemmer achieved 1% of improvement with Machine Learning approach. However, with semantic orientation approach, the improvement was 0.5%. In Machine learning approach, SVM used twice, once before applying the preprocessing phase and then again used after each stage of preprocessing to analyze the system's performance. They claimed the improvement of 4.5 percent in all measures. Same steps were adopted for semantic orientation approach and achieved 2-7% improvement in different measures. In [18], the authors have analyzed the performance of Support Vector Machine for polarity detection of textual data. A sentiment analysis framework is proposed and performance of SVM was evaluated on three datasets. Two datasets were taken from twitter and one from IMDB review. Performance of SVM was compared for each dataset by keeping in view three different ratios of training data and test data: 70:30, 50:50 and 30:70. Precision, recall and f-measure scores were used for performance evaluation. In [19], student's academic performance was predicted by using three data mining techniques: Decision tree (C4.5), Multilayer Perceptron and Naïve Bayes. These techniques were applied on student's data, which was collected from 2 undergraduate courses in two semesters. According to results, Naïve Bayes showed overall accuracy of 86% and outperformed MLP and Decision tree.

III. RESEARCH PROTOCOL

The purpose of this research is to extract the valuable information from most relevant research articles on sentiment analysis/opinion mining, published in last five years.

A Systematic literature review analyzes the gap between different researches, spanning within a particular time period as explained by [20]. Research Protocol defines the structure in which different steps are specified which have to be followed in a particular sequence. For the selection of most relevant research articles with high quality measures, a detailed procedure is adopted in this study along with some specific structure and boundary lines as explained by [21] and [22]. Guidelines for this Systematic Literature Review are also taken from latest review papers in software engineering domain such as [23], [24], [25].

Research protocol/methodology of this study consists of following steps (Fig. 1):

- Identification of research questions
- Selection of keywords for query string
- Identification of search space
- Outlining the selection criteria
- Extraction of literature with selection criteria
- Quality assessment of extracted literature
- Data extraction and synthesis
- Presentation of results

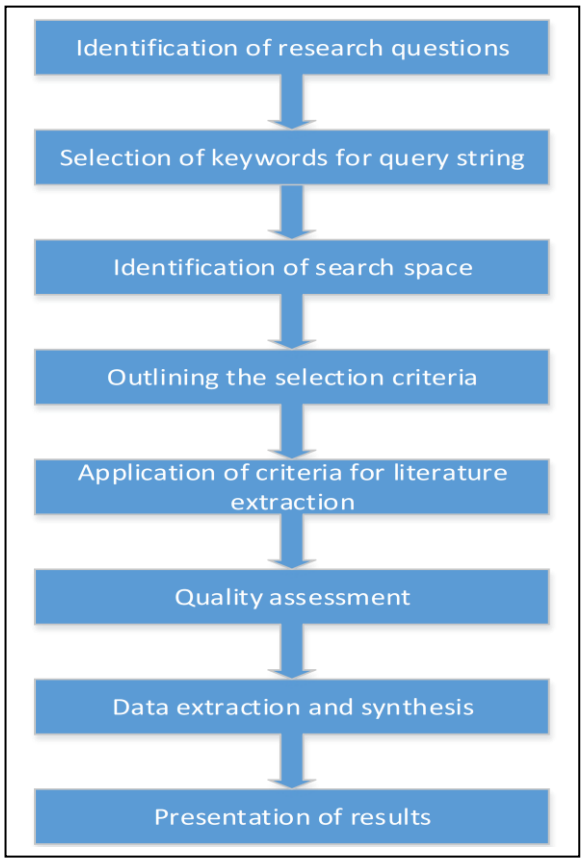


Fig. 1. Steps of SLR.

A. Research Questions

Research questions reflect the objectives of SLR and during the critical review of most relevant extracted articles; those questions have to be answered. Research questions of this SLR are given below.

RQ1: Which are the latest research trends in the domain of sentiment analysis?

RQ2: Which machine learning/lexicon/hybrid technique is considered for comparison with SVM?

RQ3: Which areas of sentiment analysis are considered for investigation by the researchers?

RQ4: Which factors affect the classification results?

RQ5: Which type of dataset is used for performance evaluation?

B. Query String and Search Space

Query String is the combination of selected keywords used to extract the research articles from concerned libraries.

Keywords extracted from research questions are given below:

Sentiment, Polarity, Opinion, Analysis, Extraction, Detection, Mining, Support Vector Machine, SVM

The following search query is finalized with the above key words.

((“Sentiment” OR “Polarity” OR “Opinion”) AND (“Analysis” OR “Extraction” OR “Detection” OR “Mining”) AND (“Support Vector Machine” OR “SVM”))

Two well-known search libraries are selected for the extraction of literature: ACM and IEEE. Both of these libraries have different characteristics and options to search the material. Therefore, slight adjustments are made in query string to obtain more relevant and appropriate literature. The Query had to be searched for multiple times with different combinations of selected keywords. Results of search query along with some significant parameters can be seen in Table I.

TABLE I. SEARCH SPACE

Sr#	Digital Library	Search Scheme	Date Searched	Total Results
1	IEEE Xplore DL	Query Search	2017-11-05	460
2	ACM DL	Query Search	2017-11-06	441

C. Selection Criteria

In this section, most relevant literature is selected with the particular selection criteria. The selection criteria further consists of IC (inclusion criteria) and EC (exclusion criteria).

1) Inclusion Criteria (IC)
Inclusion criteria is formed with the following rules:

IC1: Papers published from year 2012 till 2017.

IC2: Papers that used Support Vector Machine for Sentiment Analysis.

IC3: Papers that used Hybrid Model for sentiment analysis, which includes Support Vector Machine.

IC5: Papers that used other machine learning algorithms in comparison with Support Vector Machine.

IC5: Papers that used other lexical/Hybrid techniques in comparison with Support Vector Machine.

2) Exclusion Criteria (EC)
Exclusion criteria is formed with the following rules:

EC1: Papers which are not in English.

EC2: Papers published before 2012 or after 2017.

EC3: Papers which did not use Support Vector Machine.

EC4: Papers that do not target sentiment/opinion/polarity analysis of textual data.

EC6: Papers that do not contain any results.

EC7: Papers that used Hybrid Model, which does not include Support Vector Machine.

Only those papers are shortlisted which are more relevant to the research questions. After applying IC and EC, 92 most relevant studies are found. All the remaining studies were excluded as defined in EC.

D. Quality Assessment

Quality assessment parameters must be followed in order to provide effective results. Following parameters are considered for this SLR to maintain the quality.

- Top rated scientific libraries are selected to find the relevant research material.
- Most recent research articles were selected to ascertain the best quality
- Selection process is un-biased.
- All the steps of SLR (as discussed above) are followed in the true sense.

E. Data Extraction and Syntesis

After applying the search process (Fig. 2), 8 most relevant research articles were short listed as provided in Table II where CP stands for Conference Paper.

TABLE II. MOST RELEVENT RESEARCH LITERATURE

Sr. #	Digital Library	Type	Selected Papers	No. of Researches
1	IEEE	C.P	[26]–[31]	6
2	ACM	C.P	[32], [33]	2

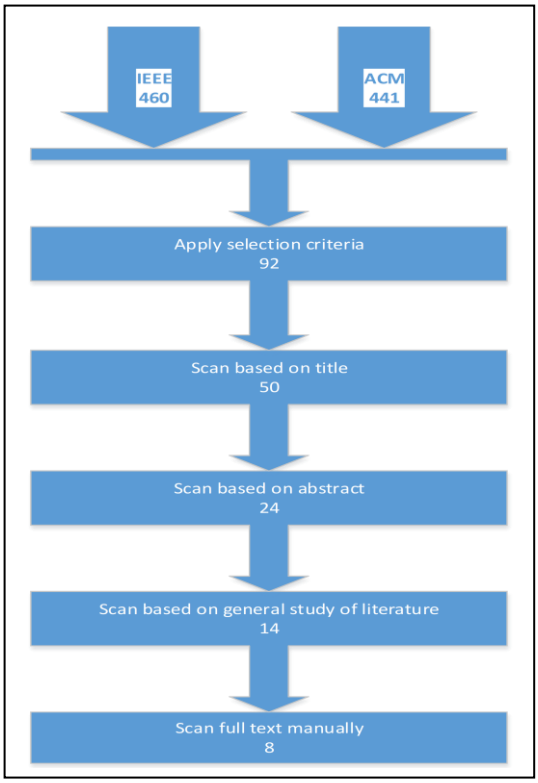


Fig. 2. Search process.

IV. LITERATURE ANALYSIS

A. A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine [Result Needed]

The authors in [26] developed a process for feature based sentiment analysis using Support Vector Machine (SVM). In the proposed model the dataset has to pass through five different phases before the conclusion of final result. First of all, the Sentence level classification is performed. Only reviews which have sentimental meaning are to be stored such as positive, negative or neutral. Questions or comments which aren't actually reviews will be filtered out using POS tagging as keeping them would led to an unnecessary extension of the vocabulary dictionary and unwanted scoring. After sentence level sentiment classification, extraction of aspects is performed which is most important and challenging task. POS tagging is used to extract words with tags like NNS (noun plural), NN (Noun), NNP (Proper noun singular), etc. In the next phase, the opinion words for aspects are extracted using the Stanford parser [34]. After that the dataset are labeled using SentiWordNet [35]. And finally for the opinion regarding whole product, SVM classifier was applied on the labeled dataset. SVM plots vectors in a 3D virtual space and distinctively allocates testing data's points to particular group which it belongs to, e.g. positive, negative, neutral or whatever the predefined groups are. The dataset considered for this research was taken from user reviews about laptops which were from a variety of companies like HP, Apple, Dell, Lenovo, etc.

B. Modeling Sentiment Terminologies: Target Based Polarity Phenomena

In [27], the researchers presented a subject sensitive sentiment analysis approach, which includes the context of tweets. According to authors the text cleansing techniques for input data before classification process can improve the results. Text cleansing includes normalization and vector representation of input data. They have pointed out that the subject aware classification brings the better results as compare to subject un-aware classification. The results can be further improved, if uni-gram approach is used instead of bi-gram or n-gram approach. A twitter dataset about word "Obama" was selected first. Features from tweets of selected dataset were extracted through Alchemy API, Tweet NLP and NTLK. From dataset, 30% of the data was used for training purpose and the rest of 70% as the test data. The collected tweets were scanned for feature extraction and then the features were stored in a separate dictionary - Keyword_Bundle - in conjunction with their specific topics to retain the target and context of the tweets. This technique further helped for the development of input matrix for SVM to classify the tweet with improved accuracy. Then two more datasets were selected "Movie Review", and "Apple" to have a comparative analysis. 85.00 %, 84.00%, and 88.00% accuracies were achieved of "Obama", "Movie Review" and "Apple" datasets respectively, making a cumulative accuracy of 85.60%.

C. Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process

In [28], authors presented an approach that classified the documents into multiple categories by keeping in view the multiple aspects. The existing problems of document level sentiment analysis such as entity identification, subjectivity detection and negation were also taken into consideration in this study. The proposed framework was used for educational data mining. The faculty performance was evaluated using the

comments provided by the students as feedback. The dataset contained 5000 comments about the faculty. The objective reviews which had no polarity inclination were filtered out, such as social comments, replies and questions. Java string tokenizer was used to divide the reviews into two token groups. After this, stopwords removal algorithms were used to remove special characters and some pronouns which would hold no significant value in the actual classification. They used TF-IDF to represent the acquired data in a numerical form, which is further used by the classifiers. Two machine learning classifiers, i.e. Naïve Bayes and Support Vector Machine were applied on the pre-processed dataset. 81.00% and 72.80% accuracy were achieved by the SVM and Naïve Bayes, respectively for aspect based document level sentiment analysis.

D. Tweeple's Microblogs on Illegal Immigration in USA

Authors in [29] presented a process for opinion mining of tweeps (People who use Twitter). The topic that was specifically chosen in comparison with some other political topics was "Illegal Immigration" as it has been under discussion for decades in the US. The dataset used in this research was collected after the US Republican Presidential election debate on Oct 28, 2015. Three major categories of the topic were selected i.e. "reform/give citizenship to illegal immigrants", "deport all immigrants" or "deport only the criminal illegal immigrants". Binary classification of first two and multinomial classification of all three categories was done using the Random Forest, Multinomial Naïve Bayes, Linear SVM and Logistic Regression classifiers. The results obtained for all the four classifiers were promising with 82% of overall average. Linear SVM and ensemble based approach using Random Forest classifiers depicted optimal results and accuracy with the mean score of 90% and 84%, respectively for binomial and multinomial classification, for individual classes with lower error rate.

E. A Proposal of a Method to Automatically Estimate Evaluations of Various topic of Traverler's Reviews

Authors of [30] conducted a study to evaluate the performance of SVM. For performance evaluation, results of SVM were compared with the dependency search tree results. For the SVM based estimation, one-against-one method was used and this parameter was selected by Scikitlearn3, which is a programming package of python. The ease with which SVM can be extended to multiclass classification by one-against-one method played a key role in the performance evaluation. In addition, SVM's RBF kernel was used. The Dependency Tree Search was used for the comparison because the researchers expected that linguistic dependency relationship would prove to be useful for obtaining evaluation data from texts as evaluation words often appear after the evaluated object. In order to obtain evaluation data from scriptures, evaluation-attribute dictionary was used. Three Polarities were defined: positive, negative and neutral. 1000 Reviews from TripAdvisor of 2014 were used as the dataset for the experimental evaluation. Three different experiments were performed A, B, and C with different number of valued scores, 5, 3, 3 respectively. These three targeted different procedures. A was used to determine the basic results of machine learning. B was basically used to assess the estimation in laxer score. C was an estimation of individual topics using the dependency tree search so the results of machine learning and methods used by the dependency tree could be compared. This architecture is incapable of designing completely foolproof feature vectors. The researchers have suggested future work to focus on automatic estimation by machine learning.

F. Sentiment Analysis of Textual Reviews

In [31], the researchers presented an experimental study for performance evaluation of different approaches for document-level sentiment classification of movie reviews. The approaches included two supervised machine learning based classifiers: Support Vector Machine and Naïve Bayes, one unsupervised technique: Semantic Orientation Approach (SO-PMI-IR Algorithm) and one lexical driven approach: SentiWordNet. For Naïve Bayes, the multinomial version of NB was implement using Java with Eclipse IDE and the labeled dataset was fed as k-folds where k was chosen to 3, 5 and 10. For SVM algorithm the dataset was converted to vector space representation using TF-IDF, afterwards same k-fold scheme was used. The Unsupervised SO-PMI-IR algorithm was implemented using Java in accordance with a POS tagger. Firstly, POS tagging was applied on the data and then feature extraction was performed for each review. The SentiWordNet approach was implemented after performing POS tagging and feature extraction. In this approach, the researchers not only used the SentiWord's lexical dictionary but rather used an enhanced procedure to increase the result of classification to a greater degree of accuracy. This was accomplished by scheming out an adjective and adverb correlation in essence with SentiWord's predefined dictionary. In this method SentiWord's scoring and Adjective Priority Scoring (APS) were assigned different weighting and the combined score of the composites was used to compute final results. 35% weightage was given to APS and the rest 65% to SentiWord's scoring. Two existent datasets of "movie reviews" were used along with one created individually for sentiment classification with different amendments in the procedures. Accuracy didn't fall out of the range of 65%-68% for SentiWordNet but SO-PMI-IR method went up to an accuracy of 89.00% but the only drawback is that a lot of PMI values have to be computed. On the other hand Naïve Bayes performed better than SVM.

G. Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain

Authors in [32] presented a novel target-oriented hybrid sentiment analysis system. It consisted of three major modules: preprocessing module, lexicon-based sentiment feature generator module and finally Machine learning module. The pre-processing module performed the optimization process and normalized the data. Sentiment Feature Generation Module started with replacing slangs with English words holding the same meaning using a slang dictionary and then tagging all the words in the dataset either by score or type. A total of 14 feature types were selected by the researchers using this module. After the feature selection phase, the data was forwarded to the machine learning classifier, which was a linear SVM. The dataset used in the evaluation was based on the occurrences of the word "iPhone". It consisted of 940 tweets which were labeled by a group of 22 human annotators. 470 tweets had a positive polarity whereas 470 tweets had negative polarity. The

proposed hybrid model achieved an overall accuracy of 89.13% outperforming the SVM's Baseline accuracy of 86.70%. The researchers concluded that use of sentiment features instead of conventional text processing features can bring the better results.

H. A Boosted SVM based Sentiment Analysis Approach for Online Opinionated Text

Authors in [33] proposed a hybrid sentiment classification model. For evaluation purpose two different datasets were used. A "movie reviews" dataset which was acquired from imdb.com in 2004 and a "hotel review" dataset which was acquired from tripadvisor.com and yatra.com. The authors came up with hybrid architectures like Adaptive Boosting (AdaBoost)) or bagging combined with SVM. This research proposed to use bagging technique to construct the SVM ensemble. In bagging, several SVMs are trained independently via a bootstrap method and then they are aggregated to formulate a strong classifier via an appropriate combination technique. The vector space model (VSM) was utilized in order to generate the bag of words representation for each document. The text documents were pre-processed with basic natural language processing techniques like word tokenization, stop word removal and stemming. The residual tokens were arranged as per their frequencies or occurrences in whole dataset. The average accuracy achieved from both the datasets went up to 93.00%. The study goes on to conclude that SVMs usually suffer from biased decision boundaries (in case of the hyper plane), and their prediction performance drops significantly when the data is highly skewed. The authors concluded that the obtained results are considerably better when multiple technologies are used in correlation instead of using SVM alone.

V. Results and Discussions

Finally, 08 research papers are selected by using systematic framework explained in Section II. These papers have been discussed in detail in Section IV of this research. Following answers are obtained against the identified Research Questions (RQs) while having an in-depth exploration and analysis of the selected papers.

RQ1: Which are the latest research trends in the domain of sentiment analysis?

As per systematic research process, 8 most relevant papers have shown the latest trends in the domain of sentiment analysis. The latest trends included the proposal of new techniques for polarity detection and sentiment analysis, customization of already proposed techniques and introducing the novel ideas to use the hybrid techniques more effectively. Moreover, one of the most important latest trends covered by our shortlisted papers is to target the new domain or area from where significant knowledge can be extracted by using classification techniques.

RQ2: Which machine learning/lexicon/hybrid technique is considered for comparison with SVM?

All selected papers [26]–[33] have used one or more techniques in comparison with SVM. The purpose of comparative analysis is to identify the difference between accuracy of that technique and the accuracy of SVM. The algorithms or techniques which are used in comparison with SVM include supervised machine learning, unsupervised machine learning, lexicon, and the hybrid of supervised and lexicon.

RQ3: Which areas of sentiment analysis are considered for investigation by the researchers?

The selected papers discussed sentence level sentiment analysis as well as document level sentiment analysis. For this purpose, different techniques are used including machine learning, lexicon based and hybrid. However, the focal point of investigation was the performance evaluation and comparative analysis to identify the best technique for sentiment analysis.

RQ4: Which factors affect the classification results?

All the selected papers have investigated the performance of their proposed techniques in terms of accuracy. To check the performance of any classification technique the output result has to be compared with pre classified or pre labeled dataset. It has been seen by analyzing the selected papers that accuracy of results may depend upon the following: the steps and techniques of preprocessing phase, the selection of input dataset along with its subject and ratio of training data & test data (in case of supervised classifier). Moreover, some of researches have claimed that the use of multiple techniques can bring more accurate results instead of using single technique.

RQ5: Which type of data sets are used for performance evaluation?

The selected papers have used the following as input dataset: tweets on different topics, user reviews about product or services and student comments about faculty. It also has been noted from the selected papers that the performance of sentiment classification techniques depends upon the selected dataset as well as the preprocessing techniques.

Limitations of Research:

Following are the limitations of this research:

1) Although all the published literature was obtained through a rigorous and thorough research process that depicts the completeness of this study however there may be still possibilities of missing some important relevant work.

2) The enhanced and optimized algorithms were mostly evaluated by the researchers themselves; therefore, the actual results might not be as accurate as claimed. This may affect the interpretation of this research.

VI. Conclusion and Future Work

Sentiment Analysis is considered as one of the hot research topics in the domain of knowledge discovery. Large amount of online data is being added on daily basis ranging from social media posts and comments to movie and software reviews. By using sentiment analysis techniques, these data sources can be used to fetch the useful information such as: prediction of election results, getting user's feedback about any software, analyzing the market reputation of particular brand and obtaining public opinion before launching a new product etc. Multiple approaches are available for sentiment analysis such

as lexicon based, machine learning based and the hybrid of both. SVM is one of the widely used machine learning techniques for the detection of polarity from text. Now days, along with conventional machine learning classification techniques, many customized and integrated models have been proposed by researchers for sentiment analysis and polarity detection. This study has provided a compact and comprehensive review of latest research by focusing on SVM technique of sentiment analysis. This study has followed a systematic framework for review and provided the answers of identified research questions after critical review of selected papers. For future work it is suggested to perform a comparative analysis of the customized techniques with same dataset.

REFERENCES

[1] M. Ahmad, S. Aftab, S. S. Muhammad, and U. Waheed, "Tools and Techniques for Lexicon Driven Sentiment Analysis : A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 1, pp. 17–23, 2017.

[2] M. Ahmad, S. Aftab, and S. S. Muhammad, "Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.

[3] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, 2017.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, 2004, p. 168.

[5] X. Ding, X. Ding, B. Liu, B. Liu, P. S. Yu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," Proc. Int. Conf. Web search web data Min. - WSDM '08, p. 231, 2008.

[6] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," Proc. 22nd Int. Conf. Comput. Linguist. - COLING '08, vol. 1, no. August, pp. 241–248, 2008.

[7] N. Jindal and B. Liu, "Opinion spam and analysis," Proc. Int. Conf. web search web data Min. 2008, pp. 219–230, 2008.

[8] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," Proc. 2010 Conf. Empir. Methods Nat. Lang. Process., no. October, pp. 158–166, 2010.

[9] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," J. Syst. Softw., vol. 125, no. November, pp. 207–219, 2017.

[10] H. Al-Mahmoud and M. Al-Razgan, "Arabic Text Mining a Systematic Review of the Published Literature 2002-2014," 2015 Int. Conf. Cloud Comput., no. November, pp. 1–7, 2015.

[11] V. Singh and S. K. Dubey, "Opinion Mining and Analysis: A Literature Review," 2014 5Th Int. Conf. Conflu. Next Gener. Inf. Technol. Summit, pp. 232–239, 2014.

[12] A. A. Sheibani, "Opinion mining and opinion spam: A literature review focusing on product reviews," 2012 6th Int. Symp. Telecommun. IST 2012, pp. 1109–1113, 2012.

[13] N. T. Liu and J. Salinas, "Machine learning in burn care and research: A systematic review of the literature," Burns, vol. 41, no. 8, pp. 1636–1641, 2015.

[14] M. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," vol. 6, no. 6, pp. 1067–1073, 2016.

[15] E. Ngai, L. Xiu, and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert Syst. Appl., vol. 36, no. 2, pp. 2592–2602, 2009.

[16] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1148–1153, 2016.

[17] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," Fourth Work. Comput. …, no. November, pp. 47–56, 2012.

[18] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[19] A. Mueen, "Modeling and Predicting Students ' Academic Performance Using Data Mining Techniques," no. November, pp. 36–42, 2016.

[20] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," Inf. Softw. Technol., vol. 51, pp. 7–15, 2008.

[21] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," J. Syst. Softw., vol. 80, no. 4, pp. 571–583, 2007.

[22] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," 2012 Int. Symp. Manag. Technol., pp. 572–576, 2012.

[23] S. Ashraf and S. Aftab, "Latest Transformations in Scrum: A State of the Art Review," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 7, pp. 12–22, 2017.

[24] S. Ashraf and S. Aftab, "Scrum with the Spices of Agile Family : A Systematic Mapping," I.J. Mod. Educ. Comput. Sci.,vol. 9, no. 11, pp. 58–72, 2017.

[25] F. Anwer and S. Aftab, "Latest Customizations of XP : A Systematic Literature Review," vol. 9, no. 12, pp. 26–37, 2017.

[26] D. V. N. Devi, C. K. Kumar, and S. Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine," 2016 Ieee 6Th Int. Conf. Adv. Comput., pp. 3–8, 2016.

[27] Y. Khaliq and M. Khaleeq, "Modeling Sentiment Terminologies : Target Based Polarity Phenomena," pp. 700–705, 2016.

[28] N. D. Valakunde and M. S. Patwardhan, "Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process," Proc. - 2013 Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol. CUBE 2013, pp. 188–192, 2013.

[29] S. M. Altarrazi and S. Sasi, "Tweeple's microblogs on illegal immigration in USA," Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016, pp. 2011–2018, 2016.

[30] K. Kawabata, M. Okada, N. Mori, and K. Hashimoto, "A Proposal of a Method to Automatically Estimate Evaluations of Various Topics of Travelers' Reviews," 2016 5th IIAI Int. Congr. Adv. Appl. Informatics, pp. 262–266, 2016.

[31] V. K. Singh, R. Piryani, A. Uddin, P. Waila, and Marisha, "Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches," 2013 5th Int. Conf. Knowl. Smart Technol., pp. 122–127, 2013.

[32] I. Alfina, D. Sigmawaty, F. Nurhidayati, and A. N. Hidayanto, "Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain," pp. 43–47, 2010.

[33] A. Sharma and S. Dey, "A boosted SVM based sentiment analysis approach for online opinionated text," Proc. 2013 Res. Adapt. Converg. Syst. - RACS '13, pp. 28–34, 2013.

[34] D. Klein and C. D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," in Advances in Neural Information Processing Systems, 2003, vol. 15, pp. 3–10.

[35] A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," Proc. 5th Conf. Lang. Resour. Eval., pp. 417–422, 2006.

Mining Trending Hash Tags for Arabic Sentiment Analysis

Yahya AlMurtadha
Department of Computer Science
Faculty of Computing and Information Technology, University of Tabuk
Tabuk, Kingdom of Saudi Arabia

***Abstract*—People text millions of posts everyday on microblogging social networking especially Twitter which make microblogs a rich source for public opinions, customer's comments and reviews. Companies and public sectors are looking for a way to measure the public response and feedback on particular service or product. Sentiment analysis is an encouraging technique capable to sense the public opinion in a fast and less cost tactic than traditional survey methods like questionnaires and interviews. Various sentiment methods were developed in many languages, such as English and Arabic with much more studies in the first one. Sometime, hash tags are misleading or may have a title that does not really reflects the subject. Tweets in trend hash tags may contain keyword or topics titles better represent the subject of the hash tag. This research aims at proposing an approach to explore Twitter Hash tag trends to retrieve tweets, group retrieved tweets to learn topics' profiles, do sentiment analysis to test the subjectivity of tweets then develop a prediction model using deep learning to classify a new tweet to the appropriate topic profile. Arabic hash tags trends have been used to evaluate the proposed approach. The performance of the proposed approach (clustering topics within hashtag trend to learn topics profiles then do sentiment analysis) shows better accuracy than sentiment analysis without clustering the topics.**

Keywords*—*Arabic sentiment analysis; twitter; opinion mining; trending hashtags; text analysis; deep learning

I. Introduction

With the development of information technology, everyone is interacting with electronic devices of various sizes. Different users from different age groups are dependent on the Internet to collect information, exchange news and jokes. The technological development has strong effect in the transformation and the impact on social life, which prompted all social, cultural and political figures and even government agencies to create accounts in various means of social communication to disseminate what is going on in their minds and to influence the followers.

With the rapid increasing in the number of social networks and the awareness of the crucial role social networks play through the posts published by users, opinion miners are interested in discovering automated ways to retrieve, analyse and report users response so that decision makers can draw their policies to develop services to the community. Previously, public opinion poll methodologies survey relevant people by distributing questionnaires or interviews which are considered as traditional methods. Mentioned previously traditional survey methods are based on small samples and may not be prepared accurately to cover what the public really feel on specific domain, which may result in inaccurate results, not to mention the cost. While the use of automated artificial intelligence methods may result in larger sample scans and actually measure their wishes indirectly by analysing their emotions and expressions that represents their feelings. In addition, the little cost that may be negligible compared to the cost of traditional survey methods mentioned above.

Data mining professionals use artificial intelligence techniques to develop methodologies that can identify the public's response, measure their opinions and interests, and classify them into different categories that can be used to promote products that are related to them. AI based techniques depend on the analysis of written posts according to the natural language in which they are written and the methodologies of natural language processing. Social networks such as Twitter can be used by specialists to sense the pulse of the public and to measure the extent of their impact or interest in a particular subject via opinion mining. Companies use opinion mining technology to measure the interest and admiration of the public to specific products for better marketing and guide users with ads lining to increase profits.

Opinion miners investigate and develop methodologies for better understanding of natural languages to improve the results of automated retrieval and understanding of desired written texts. Proposed techniques face glitches in natural language processing sine each language needs its own methodology and rules which vary from language to language such as Arabic language which is rich in synonyms, morphology and diversity of dialects [1]. Furthermore, when referring to web sites that provide a broadcasting service for Twitter hash tags that reaches trends as in Fig. 1, for example, clarify that most of popular trends may be obsessed with uncommon words that may not reflect the Twitter trend contents. Hence, different topics that might be found in a trend may lead commercial companies to post in unwanted hash tag trend Therefore, knowing the key words of Twitter within the trends may solve raised confusion. In addition, identifying the accurate search query to accurately measure the public response to specific topic is crucial for the interested individuals or companies.

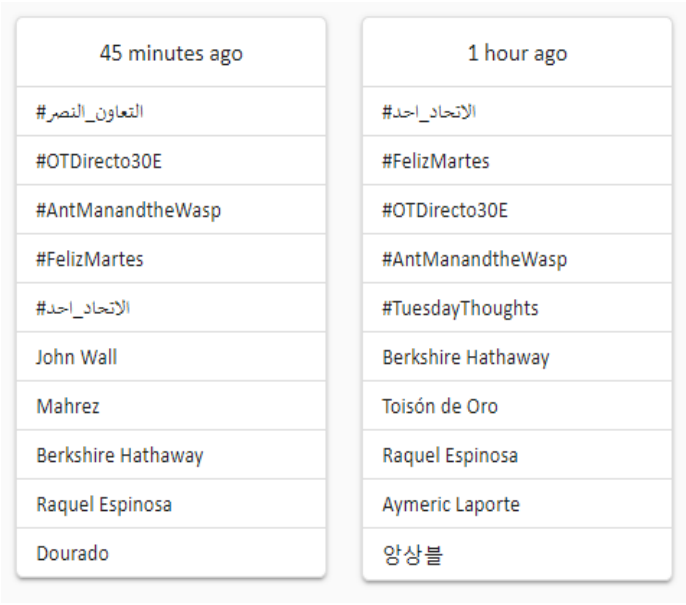


Fig. 1. Twitter worldwide trends (https://trends24.in/) Retrieved: 30-01-2018.

Therefore, a hot area of research is to develop a methodology that identifies key phrases and synonyms in a manner that is adapted to the change in the interests of the public that may retrieve more relevant and relevant texts then develop a mechanism to suit the written language processing.

This research is concerned with finding a mechanism that will quest for better key words and phrases (seed words) that can better retrieve the related tweets in a way that ensures continuous follow-up in the public opinion change. Twitter as the most famous social networking in Saudi Arabia will be used as the platform of analysis and experiment. Experiments of this research will test the proposed approach using seed words from Arabic language then report precision, recall and accuracy of classification.

The rest of the paper is organized as follows: Section II overviews the work related to sentiment analysis in Arabic. Section III describes the data collection and dataset construction. Section IV includes a description of the proposed methodology for mining the trending hashtags. Section V describes the experimental setup. Section VI presents the results of the evaluation of the proposed approach and Section VII concludes the paper.

II. Related Works

Sentiment analysis attracts researchers to focus on mine the tremendous amount of available information, discussions, texts, reviews and opinions in the digital form [2]. Recently, opinion mining or sentiment analysis has been used widely for various purposes due to its encouraging findings. Habitually, sentiment analysis is used in the business sector to measure the public response and points of view [3]. Government agencies, business sectors and research agencies are using sentiment analysis for better and indirect understanding of public opinion in specific service, product or suggestions. Sentiment analysis shows promising results to its efficiency and less-cost compared to traditional survey methods. Various articles investigated the use of sentiment analysis for different social networks such Facebook and Twitter and apply sentiment analysis for different languages such as English and Arabic. Since microblogging such as Twitter has been introduced relatively recently, view researches have been investigated sentiment analysis with increasing attention [4]. Other research highlights the importance to do sentiment analysis for unstructured information in social network such as Facebook [5]. In [6] author develop a sentiment prediction model to investigate the added value of auxiliary data, including leading, lagging information, and traditional post variables on Facebook posts.

Each natural language has its own rules, lexicons and morphology and diversity of dialects. Hence, sentiment analysis should consider the difference in morphology and rules to get the desired results. Sentiment analysis has been investigated sufficiently in English language resulting in two derived methodologies: Corpus-based and lexicon-based [7]. Arabic language includes a huge number of words synonyms resulting of data sparsity [8]. Various methods for subjectivity and sentiment analysis for Arabic language has plotted in [9]

SemEval-2017 Task4 [10] describes the fifth year of the Sentiment Analysis in Twitter task as a competition for 48 teams in both English and Arabic languages. In [11] author presents NileULex, which is an Arabic sentiment lexicon containing thousands of Arabic terms and compound phrases. A semantic approach has been developed to extract the user outlook from social networks in Arabic language investigating both regular and Arabic dialects to announce an Arabic Sentiment Ontology (ASO) includes various terms that describes how robust extracted terms express the feelings [12]. Expressions/proverb phrases lexicon has been used to advance sentiment analysis polarity in Arabic sentences [13].

AdaMC was presented to boost the best accuracy of sentence-level negative/positive classification via adaptive Multi-Compositionality [14]. This research focus in finding an approach to find the best seed words that express the public opinion keeping in mind the frequent change over time.

III. Data Collection

Millions of tweets are posted every day in different languages as a response to popular hashtags. Twitter hashtags attract public attention immediately at particular times becomes trending hashtags. The main aim of the research is to mine the Arabic trending hashtags to sense public response to particular trend hashtag. A trending hashtags dataset is needed to assess the proposed approach. To construct Twitter data set that can be analysed to evaluate the proposed methodology, following steps followed:

1) For the purpose of obtaining a large number of tweets for experiments with the potential to produce clear results, a site https://trends24.in/ was used – as suggested by SemEval-2017 [10] - to identify the trending topics of interest in the search interval of time. Further supporting information such as filtering based on the geographical location will be of extra benefit.

2) Subscribe to Twitter API Streaming service provided by Twitter. Such registration will give you access token that permit you to download the tweets containing the search query.

3) Using Python code, create a session to download tweets texts (along with extra supporting information) then save retrieved tweets to a file. The limit of the number of tweets to be inspected for each search query was set to 100000.

4) Do tweets pre-processing to remove duplications and URLs from the tweets contents.

5) Only topics with tweets more than 100 are retained for the dataset constructions.

IV. Methodology

The research aims to provide a methodology that can help better understand and measure the public response or opinion on a particular topic by analysing their sensation expressed in social networks. Twitter is chosen for its popularity. Twitter streaming API was used to search and retrieve the trending hashtags. The proposed methodology can be described as in Fig. 2 and as follows:

- The search query (mother seed) used to extract the tweets of a hash tag trend to look for candidate seed words representing the public opinions of a desired topic. Additional factors may give a more subtle twist when used, such as the geographic location of Tweets, which may sometimes be mentioned in some tweets and can be added as key words.
- After that, tokenize the tweets into words, exclude the very common words that cannot distinguish the subject from another stop words. Microblogs differs from regular texts since microblogs includes noisy text blocks, therefore, TF-IDF proves ability to extract keywords from microblogs [15]. Stem the remaining words, then weight them using weighting techniques such as TF-IDF.
- Considering that the public feelings change over time and the trend might contains altered topics, cluster the tweets into groups to formulate topics profiles within a trend. Each cluster (topics profile) includes weighted words as mentioned in the previous step.
- Sort the weighted words according to their weights (highest to lowest) as Algorithm 1 directs, then the top five words are picked to be the surrogate seed words that better represent the tweets' topic according to the users' interests and change over time.
- Topic profiles are mined using sentiment analysis technique such as AYLIEN Text Analysis. Sentiment analysis classify the tweets into objective (contains factual information) or subjective (useful for describing the opinion or feelings about specific topic).
- Finally, apply several classification methods to calculate the precision, recall and accuracy. The classification method with the highest accuracy is chosen.

Algorithm 1: Build Surrogate Seed Words

Input: Main Keyword (Mother Seed Word *k*), Seed-List *S*
Output: Adaptive Seed Words

```
Function Build-Adaptive-list (k)
C← Extract candidate words related to k
for (w ∈ C) do
    Weight words
C← Sort words
S ←Choose highest 5 words
Return S
```

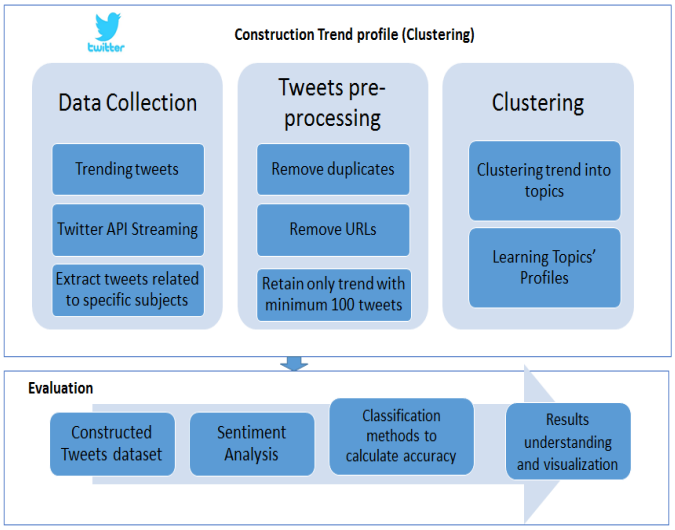


Fig. 2. The proposed approach.

V. Experiments

This section describes the experimental set up and settings to give a clear understanding for the readers.

1) First, as proposed by SemEval-2017 [10] using https://trends24.in/ to explore the recent hash tag trends while setting the region as Saudi Arabia.

2) Second, for each hash tag trend

a) Use Rapidminer studio to retrieve the tweets posted related to that trend using *Twitter search operator* which requires setting the access token sent by Twitter during subscription on Twitter streaming API. Write retrieved tweets to a file for later processing. The saved file contains the tweets in addition to some details *{created at, from user, from user ID, to user Id, language, source, text, IDgeo-location latitude, geo-location longitude, ID, retweet count}.*

b) Remove tweets duplication.

c) For each tweet text, remove retweet symbol and links (-http).

d) Use Support Vector Clustering technique to cluster the collected tweets of a specific trend into groups represent topics of related seed words.

e) Process each cluster: tokenize the texts into words, remove stop list, stem the words and then weight them.

f) Learn the topic profile by sorting the weighted words, then pick the top five words as the surrogate words best represent the topic profile.

g) Do sentiment analysis. The following attributes added to the file {*polarity, subjectivity, polarity_confidence, subjectivity _confidence*}.

h) Apply deep learning as a multi-layer feed-forward Artificial Neural Network (ANN) to build a prediction model helps to classify new tweet into the appropriate topic profile. Deep learning. Deep learning permits computational system represented by many processing layers to pick up data exemplifications with various stages of generalization [16].

i) Use 10 *k* cross validation to test the prediction accuracy of the developed predict model. Cross validation splits the dataset into k subsets whereby each time one subset is used for testing and the remaining k-1 subsets are used for training. The average value for all the k experiments are used as the validation value.

VI. Results and Discussion

Keeping in mind that the hash tag might contains more suitable seed words that better represent the topics, the evaluation of the proposed approach first cluster the tweets then do the sentiment analysis finally develop a prediction model to classify the new tweet to the class belongs to wither subjective or objective. Precision, recall and accuracy have been used to test the proposed approach efficiency. Precision, recall and accuracy are usually used to assess and relate various detection algorithms [17]. For better understanding of the evaluation method in relation to the sentiment analysis of tweets, precision, recall and accuracy has been defined as follows:

Precision measures the portion of relevant tweets among the truly positive and false positive retrieved tweets on a specific class:

Precision = TP/ (TP + FP), where TP is truly positive and FP is false positive. (1)

Recall measures the portion of retrieved relevant tweets over the total amount of relevant tweets:

Recall = TP/ (TP + FN) where TP is truly positive and FN is false negative. (2)

Accuracy is the = TP+TN/ (TP+TN+FP+FN) where TP is truly positive, FP is false positive, TN is truly negative and FN is false negative. (3)

Test analysis classifies the texts into two categories: facts and opinion [18]. Facts are "objective expressions about entities, events and their properties" while opinion "are subjective expressions that describe people's sentiments, appraisals or feelings" [19]. Several experiments were conducted on various trending hashtags but only trends with more than 100 tweets were reported. Tables from I to VIII report the precision and recall for retrieved four highest hash tags trends investigated in Arabic language from the website https://trends24.in/. Sentiment analysis for tweets normally classify the tweet into either objective or subjective. Subjectivity feelings of opinions can be used to tell about two concerns: people's feelings and language expressions used to describe that feelings [19] while objectivity describes some factual information. Tables illustrate the efficiency of the proposed approach (clustering the trending hashtags into profiles then do sentiment analysis for the tweets' texts) against the usual sentiment analysis (do sentiment analysis for tweets' texts without learning trends profiles first).

TABLE I. Precision for Hash Tag1

اي_جهاز_تفضل_ايفون_او_سامسونج **Which one you prefer, iPhone or Samsung**	**Class Precision (Normal Sentiment Analysis)**	**Class Precision (Proposed Approach)**
pred. subjective	0.00%	0.00%
pred. objective	97.10%	91.18%

TABLE II. Recall for Hash Tag1

اي_جهاز_تفضل_ايفون_او_سامسونج **Which one you prefer, iPhone or Samsung**	**True Subjective**	**True Objective**
class recall (Normal Sentiment Analysis)	0.00%	100.00%
class recall (Proposed Approach)	0.00%	98.41%

TABLE III. Precision and Recall for Hash Tag2 (Normal Sentiment Analysis)

يصلحه_ لعالم_ما يفسدة **What the world destroy, name a thing that fix it**	**True Objective**	**True Subjective**	**Class Precision**
pred. objective	149	12	92.55%
pred. subjective	26	6	18.75%
class recall	85.14%	33.33%	

TABLE IV. Precision and Recall for Hash Tag2 (Proposed Approach)

ما يفسدة العالم يصلحه **What the world destroy, name a thing that fix it**	**True Objective**	**True Subjective**	**Class Precision**
pred. objective	154	12	92.77%
pred. subjective	21	6	22.22%
class recall	88.00%	33.33%	

TABLE V. Precision and Recall for Hash Tag3 (Normal Sentiment Analysis)

اسم_اللي_تحبه_بالرموز **Name the one you love with symbols**	**True Objective**	**True Subjective**	**Class Precision**
pred. objective	207	15	93.24%
pred. subjective	67	4	5.63%
class recall	75.55%	21.05%	

TABLE VI. PRECISION AND RECALL FOR HASH TAG3 (PROPOSED APPROACH)

اسم_اللي_تحبه_بالرموز Name the one you love with symbols	True Objective	True Subjective	Class Precision
pred. objective	229	18	92.71%
pred. subjective	45	1	2.17%
class recall	83.58%	5.26%	

TABLE VII. PRECISION AND RECALL FOR HASH TAG4 (NORMAL SENTIMENT ANALYSIS)

الشباب_الاهلي Ahli_Youth	True Objective	True Subjective	True?	Class Precision
pred. objective	344	35	93	72.88%
pred. subjective	18	58	98	33.33%
pred. ?	142	42	269	59.38%
class recall	68.25%	42.96%	58.48%	

TABLE VIII. PRECISION AND RECALL FOR HASH TAG4 (PROPOSED APPROACH)

الشباب_الاهلي Ahli_Youth	True Objective	True Subjective	Class Precision
pred. objective	495	36	93.22%
pred. subjective	9	99	91.67%
class recall	98.21%	73.33%	

Table IX reports the accuracy calculated for the four hash tags illustrating the number of tweets retrieved for each hash tag and the number of topics profiles (clusters) for each hash tag. The following observations obtained from Table IX:

1) The performance of the normal sentiment analysis is higher than the proposed approach only for the first hash tag. The reason beyond that can be explained due to the small number of tweets retrieved for that hash tag (96 tweets) compared to the remaining three hash tags. Hence, results prove the assumption mentioned in the methodology that we retain only hash tag with a minimum of 100 tweets.

2) Notably as illustrated by hash tag from 2 to 4, the more tweets we retrieve in a hash tag the better accuracy we get.

TABLE IX. ACCURACY FOR THE 4 HASH TAGS

#Trend	Normal Sentiment Analysis Accuracy	The proposed Approach Accuracy	No. of Tweets	No. of Clusters
اي_جهاز_تفضل_ايفون_او_سامسونج Which mobile phone you prefer, iPhone or Samsung	97.14 %	90.00 %	69	5
ما يفسدة العالم يصلحه What the world destroy, name a thing that fix it	80.32 %	82.84 %	194	13
اسم_اللي_تحبه_بالرموز Name the one you love with symbols	72.08%	78.57 %	294	14
الشباب_الاهلي Ahli_Youth	61.05 %	92.96 %	1100	42

VII. CONCLUSION

Mining the tremendous amount of text, information, posts and customers' comments is essential to extract the desired knowledge. Companies might survey customers via traditional methods like questionnaires and interviews which is time consuming, costly to attract large sample size and people might respond incorrectly for different reasons. Sentiment analysis is a technique helps to understand and measure the targeted folk's opinions. The proposed approach aims at learning topics profiles helps to better understand the public response to a particular service, product or feedback by analysing recent Twitter hash tags trends. Occasionally, some hash tag titles is not understandable or misleading to tweets that represents different topics. To solve such problem, clustering of tweets was proposed to learn topics profiles. Recent Arabic hash tag trend-as listed by trends website announcer- were retrieved then the proposed approach was tested on the tweets of popular hash tags that becomes trends. Results show that the more tweets retrieved for a hash tag, the more groups or cluster (topics profiles) leading to enhanced sentiment analysis. Applying deep learning, findings show that the accuracy of the proposed approach is better than the normal sentiment analysis. As a future work, further investigation in using the proposed approach to automatically use the learned topics profile in each hashtags to retrieve similar topics.

REFERENCES

[1] Sallab AA Al, Baly R, Hajj H. "Deep Learning Models for Sentiment Analysis in Arabic". ANLP Work. 2015;9–17. Available from: http://www.aclweb.org/anthology/W15-32#page=21

[2] Liu B. "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies.2012;5:1–167. Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016.

[3] Neuendorf KA. "The Content Analysis Guidebook" (Second Edition). 2017. Available from: http://academic.csuohio.edu/kneuendorf/SkalskiVitae/SkalskiNeuendorfCajigas17.pdf

[4] Patodkar VN, I.R S. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". IJARCCE [Internet] 2016;5:320–2. Available from: http://ijarcce.com/upload/2016/december-16/IJARCCE 74.pdf

[5] Dasgupta SS, Natarajan S, Kaipa KK, Bhattacherjee SK, Viswanathan A. "Sentiment analysis of Facebook data using Hadoop based open source technologie"s. In: Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015. 2015.

[6] Meire M, Ballings M, Van den Poel D. "The added value of auxiliary data in sentiment analysis of Facebook posts". Decis. Support Syst. 2016;89:98–112.

[7] Assiri A, Emam A, Aldossari H. "Arabic Sentiment Analysis: A Survey". Int. J. Adv. Comput. Sci. Appl. 2015;6:75–85.

[8] Abdul-Mageed M, Diab M, Kübler S. "SAMAR: Subjectivity and sentiment analysis for Arabic social media". Comput. Speech Lang. 2014;28:20–37.

[9] Korayem M, Crandall D, Abdul-Mageed M. "Subjectivity and Sentiment Analysis of Arabic: A Survey". In: Communications in Computer and Information Science. 2012. page 128–39.

[10] Rosenthal S, Farra N, Nakov P. "SemEval-2017 Task 4: Sentiment Analysis in Twitter". In: Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017). 2017. page 502–18.Available from: http://aclweb.org/anthology/S/S17/S17-2088.pdf

[11] El-Beltagy SR. "NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic". to Appear Proc. Lr. 2016 2016;2900–5.

[12] Tartir S, Abdul-Nabi I. "Semantic Sentiment Analysis in Arabic Social Media". J. King Saud Univ. - Comput. Inf. Sci. 2017;29:229–33.

[13] Ibrahim HS, Abdou SM, Gheith M. "MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis". In: 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS 2015 - Proceedings. 2015. page 353–8.

[14] Li Dong FWMZKX. "Adaptive Multi-Compositionality for Recursive Neural Models with Applications to Sentiment Analysis". 2014;1–7.

[15] Simsek A, Karagoz P. "Sentiment enhanced hybrid TF-IDF for microblogs". In: Proceedings - 4th IEEE International Conference on Big Data and Cloud Computing, BDCloud 2014 with the 7th IEEE International Conference on Social Computing and Networking, SocialCom 2014 and the 4th International Conference on Sustainable Computing and Communications, SustainCom 2014. 2015. page 311–7.

[16] Lecun Y, Bengio Y, Hinton G. "Deep learning". Nature2015;521:436–44.

[17] Zhu M. "Recall, precision and average precision". Dep. Stat. Actuar. Sci. 2004;1–11. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Recall+,+Precision+and+Average+Precision#0

[18] Pang B, Lee L. "Opinion Mining and Sentiment Analysis". Inf. Retr. Boston. 2008;2:271–8. Available from: http://www.aclweb.org/anthology/P/P04/P04-1035.pdf%5Cnhttp://portal.acm.org/citation.cfm?id=1218990%5Cnhttp://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf%5Cnhttp://arxiv.org/abs/cs/0409058v1

[19] Liu B. "Sentiment Analysis and Subjectivity". Handb. Nat. Lang. Process. 2010;1–38. Available from: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf%255Cnhttp://people.sabanciuniv.edu/berrin/proj102/1-BLiu-Sentiment Analysis and Subjectivity-NLPHandbook-2010.pdf%255Cnhttp://www.cs.uic.edu/ ~ liub/FBS/NLP-handbook-sentiment-analysis.pd

A New Task Scheduling Algorithm using Firefly and Simulated Annealing Algorithms in Cloud Computing

Fakhrosadat Fanian
Department of Computer Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran

Vahid Khatibi Bardsiri
Department of Computer Engineering, Bardsir Branch, Islamic Azad University, Kerman, Iran

Mohammad Shokouhifar
Department of Electrical Engineering, Shahid Beheshti University G.C. Tehran, Iran

***Abstract*—Task scheduling is a challenging and important issue, which considering increases in data sizes and large volumes of data, has turned into an NP-hard problem. This has attracted the attention of many researchers throughout the world since cloud environments are in fact homogenous systems for maintaining and processing practical applications needed by users. Thus, task scheduling has become extremely important in order to provide better services to users. In this regard, the present study aims at providing a new task-scheduling algorithm using both firefly and simulated annealing algorithms. This algorithm takes advantage of the merits of both firefly and simulated annealing algorithms. Moreover, efforts have been made in regards to changing the primary population or primary solutions for the firefly algorithm. The presented algorithm uses a better primary solution. Local search was another aspect considered for the new algorithm. The presented algorithm was compared and evaluated against common algorithms. As indicated by the results, compared to other algorithms, the presented method performs effectively better in reducing to make span using different number of tasks and virtual machines.**

Keywords*—Firefly; make span; simulated annealing; task scheduling; cloud

I. Introduction

Cloud computing has recently been introduced as a new technology for users. From a historical perspective, the first computers used were those of the first generation, mainly the mainframes. As time went by, these computers became smaller with higher processing power until personal computers were developed and distributed amongst all users. Next, the technology of networks providing higher processing power emerged by connecting a few small personal computers. However, processing requirements increased exponentially and the need for bigger computing systems became crucially essential.

Thus, smaller networks were privately joined to form bigger networks across the internet. By then, millions of users had access to the internet mostly never using their computers processing power to its full capacity and preferring to give away the idle processing time of their computers to be used for computational tasks. Therefore, many small computational resources were connected; however, it was not possible to completely use these sources within the created network, since these computers were not purposefully created to handle commercial applications. This led to the establishment of a new approach. An approach in which the details were hidden from the user and users did not need to allocate or control infrastructural cloud technologies they were using [1].

In layman's terms, cloud computing was a new user-driven model based on users demands with easy access to flexible and configurable computational sources such as networks, servers, storage areas, practical applications, and services, such that this access is rapidly made with the minimum need for resource management or intervention by the service provider. In general, cloud-computing users are not proprietors of the cloud infrastructure, but rather rent these services from third parties in order to avoid large costs [2]. These users utilize the existing resources in the form of services and only pay for whichever sources they are using [3]. Like any other public service, the costs are based on the amount of service the user requires [4]. Hence, considering that hundreds of people make use of virtual machines, manual allocation of computational sources for different tasks is very troublesome in cloud technology [5]. This highlights the need for an efficient algorithm for task scheduling in cloud environments. This scheduler must be consistent with environmental changes and change in task types [6]. At any moment, millions of users are demanding cloud resources. Scheduling this number of tasks is a serious challenge in cloud processing environments, especially since allocation of optimized resources or task scheduling in clouds must be done in accordance with optimized number and need of systems within the cloud environment so as to maintain the clouds integrity. On the other hand, this scheduling must be done in a way minimizing energy consumption within the cloud. Ergo, this study tries to present an efficient algorithm for task scheduling in clouds using the combination of both firefly and simulated annealing optimization algorithms. This study is organized as follows: Section II reviews related and previous works. Section III discusses and presents a new method. Section IV contains the results of the presented algorithm, and finally Section V gives a conclusion of the entire study.

II. Review of Literature

Cloud computing is currently made up of various aspects, making it a challenging subject. Thus, many researchers have made efforts to investigate the various aspects of cloud computing [7] and have tried to make virtualization and automation technologies focus on improving services in clouds. In this regard, task scheduling and reducing energy consumption in clouds is a very challenging issue for these

environments. Kusic [8] investigated the issue of energy management in virtual heterogeneous environments and used Kalman filters as a method complying with system demands and as a means for prediction and actual implementation.

Kalman filters are used for estimating future demands in order to predict system status and allocate resources accordingly. On the other hand, some researchers focused on the effects of scheduling virtual machines on I/O virtual performance and emphasized on monitoring optimization for better I/O performance. For instance, Ongaro et al. [9] studied the effect of virtual machine observer on performance and presented an idea for arranging processors in an executional queue based on remaining and current value. They ultimately presented an optimization algorithm for scheduling even I/O distribution. However, this scheduling procedure did not take into account the workload and the reallocation of virtual machines. In [10], Kim presented a task-aware scheduler with an emphasis on developing I/O performance.

This scheduler did not consider the heterogeneous workload and variety of weights only focusing on I/O performance. Liao [11] presented a scheduler for scheduling real time applications for supporting respond time, and instead of placing the processor at the end of the executive queue, this method compute the state in which the virtual processor is inserted based on its delay. Goiri [12] presented a task dynamic scheduling policy for allocating informed sources at cloud data centers. The presented scheduler worked to stabilize workload by connecting large tasks of individual devices with necessary hardware, in order to maintain service quality. In other words, these methods reduced energy consumption at data centers turning off servers. Wood [13] presented a virtual machine-driven scheduling policy based on using resources including processor, memory, and subnet components. However, instead of optimizing and scheduling operational energy, his study mainly focused on developing an algorithm for avoiding local traps. Dorigo et al. presented the ACO algorithm [14]. The ACO was a random search algorithm, which used positive feedback and followed actual ant colony behavior. In [15] this algorithm was used to allocate optimized sources for tasks in a dynamic cloud environment in order to minimize make span.

Liu et al. [16] worked on a scheduling algorithm based on genetic and ant colony algorithms. They tried to make use of the advantages of both algorithms. This algorithm uses the global search in genetic algorithm in order to reach the optimized solution faster. It also utilizes initial values for pheromones in the ACO algorithm. Guo et al. [17] used a formulated particle swarm optimization (PSO) model for minimizing process costs. They also tried to use crossover and mutation functions of the genetic algorithm along with the PSO model. Lakro et al. [18] investigated various variables and their optimization in cloud computing environments. They tried to present a multi-variable optimization algorithm for scheduling and improving performance of data centers. Jia et al. [19] investigated scheduling of various tasks of different sizes on a set of parallel batch machine and presented a meta-heuristic algorithm based on max-min and ant system for minimizing make span.

III. Methodology and Suggested Algorithm

Cloud computing is one of the newest technologies today, which allows users to send their requests to clouds and pay a certain amount of fees based on the service provided. On the other hand, cloud environments are in fact homogenous systems suitably storing large applications and data for services. Considering this, scheduling of these data and large applications in these systems is of great importance. The present study tries to present a new algorithm based on firefly and simulated annealing algorithms called FA-SA in order to schedule tasks in clouds. The details of the suggested combination are expressed below. The general framework for this study is shown in Fig. 1.

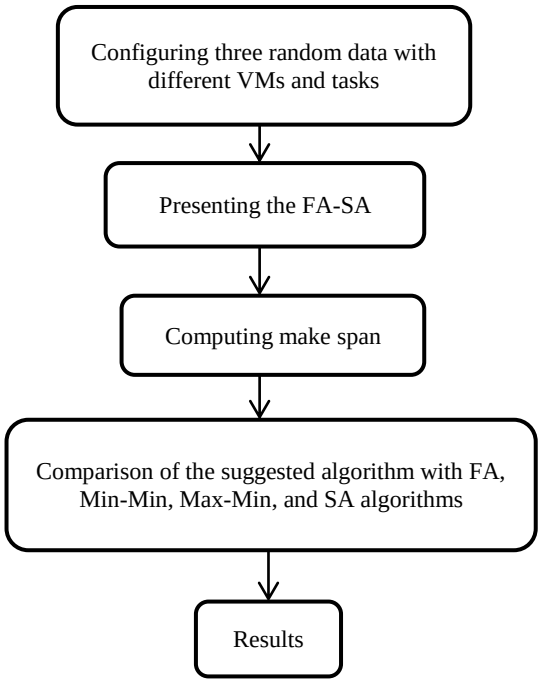


Fig. 1. General framework for the study.

A. Problem Statement

The allocation of tasks to virtual machines in cloud computing systems is a problem, in which m number of tasks, $V= \{t_1, t_2, \ldots, t_m\}$ are to be allocated to certain virtual machines. In this study, the total number of tasks are randomly selected from 10 to 100 tasks and categorized into three different data sets with different number of virtual machines. The tasks are made randomly. Also $P= \{ v_1, v_2, \ldots , v_n \}$ are the n virtual machines used. All systems are the same, meaning tasks are performed in a homogeneous environment.

B. Possible Solutions

This study uses a combination of firefly algorithm (FA) and simulated annealing (SA). The feasible solution in this study is a string of m characters, where m is the total number of tasks. According to (1), if task i is allocated to a virtual machine, j, the i_{th} place in the relative string, has a value of "j". 20 virtual machines are considered for all m tasks. A feasible solution for the problem is shown in Fig. 2.

$$Solution_i = j \qquad \text{if task } i \text{ is allocated to machine } j \qquad (1)$$

	1	2	3	4	5	6	7	....	M
A Solution:	13	1	17	8	2	6	14	....	7

Fig. 2. A feasible solution: for example, the first task is given to machine I3 and the second task is given to machine 1.

C. Objective function in the Suggested Algorithms (FA, SA and FA-SA)

As previously mentioned, an objective function is needed for all algorithms in order to schedule tasks and minimize amount of make span. Task scheduling is an optimization problem in which tasks are to be allocated to sources at certain times. In other words: *n* tasks, j_1, j_2, …, j_n, with different sizes are to be allocated to *m* identical scheduler machines such that make span is minimized. Make span is defined as the total amount of time required to perform all tasks (after all tasks have been done). Recently, this problem has been introduced as a dynamic scheduling problem, in which for every task, the dynamic algorithm must use the existing information to make a decision before the next task comes. This is one of the most famous dynamic problems and the first for which a competitive analysis was presented by Graham in 1966 [20].

D. Overall Stages of Allocating Tasks to Virtual Machines using the Suggested FA-SA Algorithm

Evolutionary algorithms are generally based on population and make use of a very suitable global search strategy. The firefly algorithm [21] was used in this study. This algorithm is a meta-heuristic algorithm inspired by the behavior and motion of fireflies in nature. This algorithm is similar to other population-based algorithms and computes the optimized solution (or near to optimized) in an iterative manner. The algorithm starts by performing a search procedure in a randomly developed population. Each member of the population (location of each firefly in the search space) is a possible solution for the problem, which is shown in Fig. 2 according to (1). Each iteration in the FA algorithm has two main stages: Stage 1, evaluating the suitability of the solutions and Stage 2, updating the population (establishing a new population). These two stages are continuously performed in iteration until the termination criteria of the algorithm is satisfied.

The termination condition in this study is the completion of all tasks. The FA algorithm is a population-based algorithm with the ability to perform a very suitable global search since it has a very high convergence rate and each firefly tries to find the best state individually; thus, it avoids local optimums and searches for the global optimum [22]. On the other hand, the SA algorithm has a very convenient local search procedure. It is for this reason that both of these algorithms were combined in this study to form the FA-SA algorithm in order to benefits from the advantages of both of these algorithms for performing a better scheduling of tasks in clouds.

In the presented method, the FA algorithm initiates first in order to perform a global search in the search space. After the FA algorithm, the SA algorithm is executed to perform a local search near the previous solution provided by the FA algorithm. In other words, the initial population for the SA algorithm is not selected randomly, rather it gets the value provided by the FA algorithm which is in fact the optimum value provided by the FA algorithm. The general flowchart for the suggested method is shown in Fig. 3. The stages of the suggested algorithm will be explained in more detail in the following section.

a) Producing a random initial population for the FA algorithm): As previously mentioned, the first stage for all evolutionary algorithms is producing initial solutions, which are mostly done randomly. The initial solutions for the FA algorithm in this study are produced considering the following regulations:

1) Perform the following stages for m iterations (where m is the total number of tasks): find the virtual machine(s) with the least termination time (since the data are random multiple machines may have the same value).

2) Perform the following stages for m iterations (where m is the total number of tasks): find the virtual machine(s) with the least termination time (since the data are random multiple machines may have the same value).

3) If a virtual machine is found, select the virtual machine, otherwise randomly select a virtual machine with the least termination time (since data are random multiple machines may produce the same value).

4) Search the initial data set (containing tasks and virtual machines) and find the virtual machine selected in stage 2 and choose the task with the least time from the unallocated tasks for that machine.

5) If a task exits with the least amount of time, select that task; otherwise, randomly select a task.

6) All tasks are assigned?

7) no, go to stage 1, otherwise terminate.

In other words, each task is allocated to a virtual machine according to the regulations mentioned above. It is worth noting that since the data sets of this study are random in nature and according to the regulations, random selection is performed two times, the initial population or rather the initial solutions are different for each iteration, though due to the nature of the regulations, these initial solutions are near optimum.

b) Competency assessment for produced solutions: The solutions produced by the FA algorithm are evaluated in each iteration after the population has been updated. This evaluation works on the basis of the objective function. In order to evaluate each member of the population (each firefly), allocated tasks for each machine are considered first. Next, execution time on each machine is computed and finally termination time for all tasks are computed.

c) Updating population in the FA Algorithm: The firefly algorithm was presented by Yang [23] and is inspired by the motion and behavior of fireflies in nature. Fireflies produce short and rhythmic lights. These rhythmic lights, light radiation

rate, and distance are what make two fireflies attract each other. Light intensity at a distance of r from the light source has a relationship with the reverse squared amount of distance. In the firefly algorithm, light can be considered as the objective function to be optimized. In short, the firefly algorithm is based on the following three principles:

1) All fireflies are unisexual and each firefly attracts the other firefly despite their sexuality.

2) Attraction of fireflies is proportional to their radiance such that the firefly with less light intensity is attracted to the one with higher light intensity, and if there is no firefly with a higher light intensity in the locality, fireflies move randomly. 3- The light intensity of fireflies is determined as the objective function [24], [25]. In FA algorithm, the location of each firefly in m-dimension space determines a solution for the optimization problem, where m is the number of optimization variables (total number of virtual machines). Considering that fireflies' location is defined in a continuous space, this study considers the location of each firefly within the (0,n] range, where n is the total number of machines. Therefore, each dimension value for each firefly is a value from 0 to n. In each iteration of the evaluation stage, each dimension for each firefly is rounded up to the nearest natural number that is bigger than the current number.

Therefore, evaluation of fireflies takes place in a discrete space. However, fireflies' motion and attraction are done continuously. After determining the time for the solution of each firefly using relative objective function, radiance of each firefly i is computed using (2) (since radiance in this algorithm denotes higher competency, every firefly with a lower objective function has a higher f_i), where $objective\ function_{\ i}$ and f_i denote error rate (objective function) and radiance for the ith firefly, respectively. Each iteration selects fireflies with the highest radiance. Then, each of the remaining fireflies moves towards the nearest radiant firefly. The distance between firefly i and firefly j is computed by (3):

$$f_i = \frac{1}{\text{objective function}_{\ i}} \quad (2)$$

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \quad (3)$$

where x_i and x_j are the locations of the i_{th} and j_{th} fireflies, respectively. d is the number of optimization variables, which in this case is equal to the total number of tasks. Movement of firefly i towards firefly j is formulated as (4). The second expression in this statement shows the attraction if firefly i towards firefly j and the third expression shows a random movement in the attraction procedure. α and β are two static variables that configure the effect of the two expressions when firefly i moves. η determines the way fireflies move and is usually selected between 0 and infinity.

$$x_i = x_i + \beta * e^{-\eta r_{ij}}(x_j - x_i) + \alpha * (rand - 0.5) \quad (4)$$

d) Addition of local search to the FA algorithm: Three types of local searches were added to the firefly algorithm in this study, where each type is used with a probability of 1/3 for each iteration (generation). These searches include exchange mutation, inverted exchange mutation, and a suggested local search called hybrid max-min to exchange (HHME). These procedures are explained in more detail in the following section.

- Exchange mutation: In this procedure, two machines are randomly selected and their tasks are exchanged [26]. Fig. 4 shows the search procedure used in the proposed algorithm.

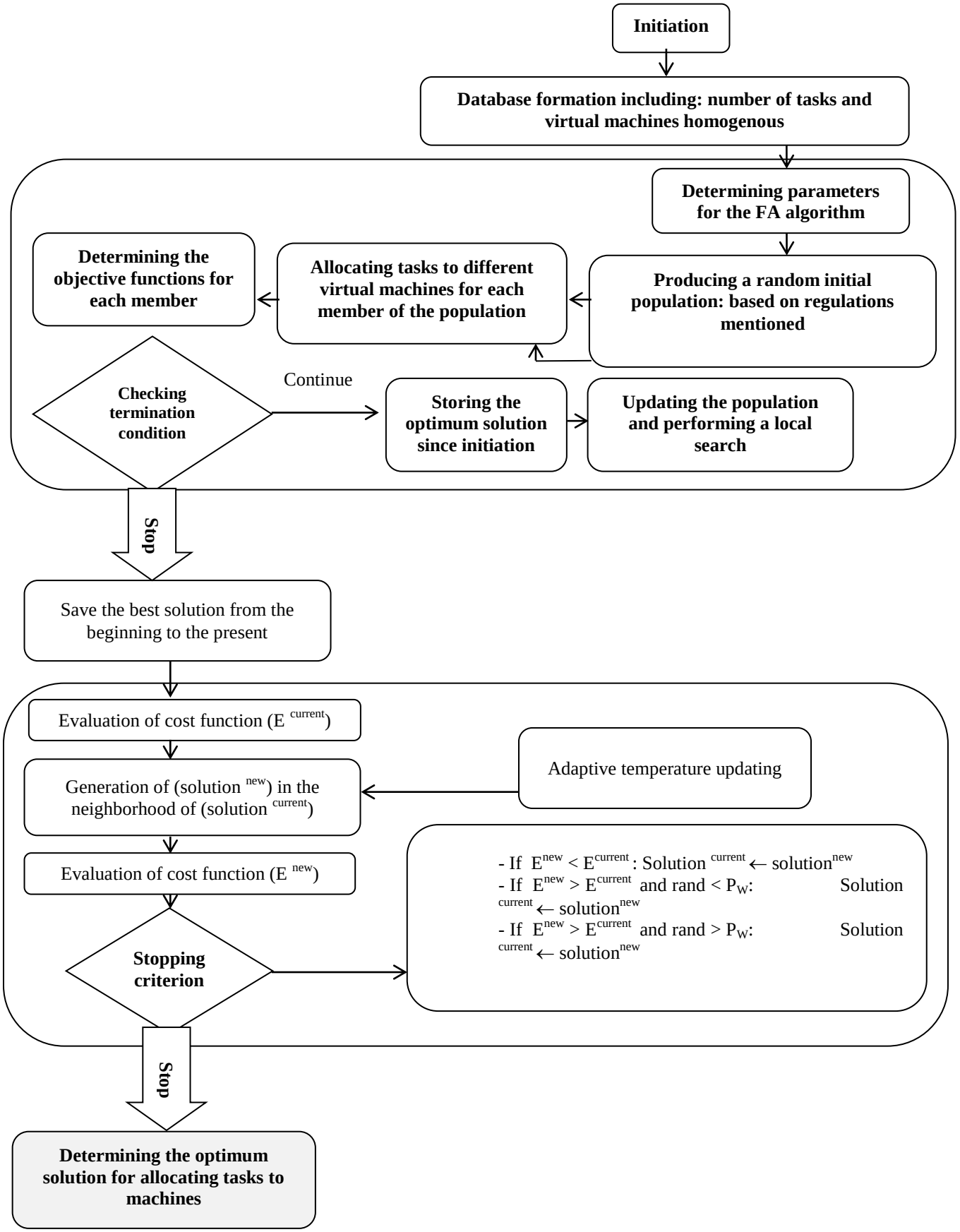


Fig. 3. General flowchart for the FA-SA algorithm.

Fig. 4. Search procedure used in the suggested algorithm, before exchange mutation (current solution) and after exchange mutation (new solution).

- Inverted exchange mutation: two machines are randomly selected and their tasks are inverted [26]. Fig. 5. shows the search procedure used in the proposed algorithm, before inverted exchange mutation (current solution) and after the inverted exchange mutation (new solution).

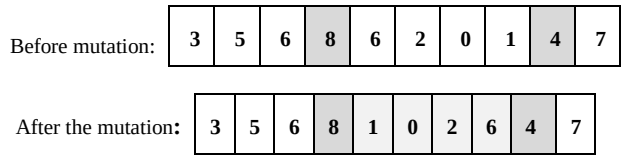


Fig. 5. search procedure used in the suggested algorithm, before inverted exchange mutation (current solution) and after inverted exchange mutation (new solution).

- Hybrid max-min to exchange (HMME): In this procedure, virtual machines with the highest and lowest value of termination time are selected and; 1) a minimum task from the machine with the highest termination time is transferred to the machine with the lowest termination time or 2) a task is randomly selected from the machine with the highest termination time and transferred to the machine with the lowest termination time or 3) a maximum task from the machine with the highest termination time is transferred to the machine with the lowest termination time.

e) Local search using SA algorithm: the simulated annealing, presented in 1983 [27] algorithm, is an optimization algorithm that uses local search. The gradual annealing technique is used by metallurgists in order to reach a state where the solid material is sorted properly with minimized energy. In this technique, the substance is placed at high temperature then cooled down gradually. During this algorithm, each state s in the search space is similar to a state of a physical system and the E(s) function which must be minimized is similar to the internal energy of the system in that specific state. The purpose of this procedure is to transfer the system from its initial random state to a state where the system has the lowest amount of energy.

For an optimization problem, the algorithm starts with a random initial solution and gradually moves towards neighboring solutions in an iterative manner. In each iteration, if the neighbor solution ($solution^{new}$) is better than the current solution ($solution^{current}$), the algorithm selects the former solution as the new current solution. Otherwise, the algorithm selects the new solution with a probability of of $p_w = exp(-\Delta E/T)$, where $\Delta E = E^{new} - E^{current}$ is the difference between the objective function value of the current solution and that of the neighboring solution and T is the temperature variable. This algorithm iterates for each temperature, and gradually decreases the temperature. The temperature is initially high so that the possibility of choosing worse solutions is high. However, with the gradual decrease in temperature, the possibility of choosing worse solutions decreases and better solutions are selected. Therefore, the algorithm converges to a proper solution. As seen in Fig. 6, in this study, random changes in one dimension of the solution have been selected for local search. The possibility for performing this procedure is defined as P_m, where the value of each dimension in the current solution changes with the probability of P_m.

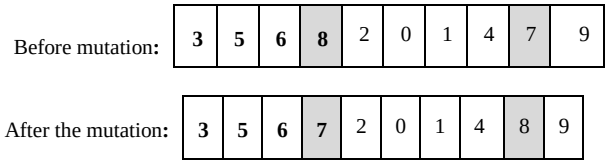


Fig. 6. Local search in the SA algorithm, before exchange mutation (current solution) and after exchange mutation (new solution).

IV. Results

The suggested FA-SA algorithm was simulated in MATLAB and was compared with three data sets and the following algorithms: min-min, max-min, firefly, and simulated annealing. Comparison results are provided below.

A. Datasets

Three different datasets were randomly selected in this study considering that the minimum and maximum numbers of tasks were 10 and 100, respectively. The number of tasks and machines were chosen randomly such that the first dataset, name data1, contained 100 tasks and 8 homogeneous virtual machines randomly selected according to the mentioned criteria. Detailed specification of these three datasets are provided in Table I.

TABLE I. Specifications of Randomly Selected Datasets

Type data	Number of task	Number of VM	Min_task	Max_task
Data1	100	8	10	100
Data2	200	20	10	100
Data3	500	20	10	100

B. Parameter Configuration for Optimization Algorithms

Configuration circumstances for the FA, SA, and FA-SA optimization algorithms are shown in Table II. As can be seen, considering that SA is a single-population algorithm, the number of iterations in SA is more than that of FA and the circumstances for the FA-SA algorithm are a combination of those related to FA and SA, since the final solution of FA algorithm is used as the initial solution for the SA algorithm.

C. Evaluation using Makespan

This section compares the FA-SA algorithm with SA, FA, min-min, and max-min algorithms based on objective function for computing and minimizing make span. Fig. 7 shows the evaluation results of FA-SA and other algorithms on the data1 dataset. As it is observed, since workload and number of virtual machines is lower compared to other datasets, all algorithms, except max-min, showed a similar make span value and the suggested algorithm outperformed other algorithms in reducing make span. It is worth mentioning that all results were

based on 10 iterations of the algorithms and were expressed as mean values.

TABLE II. CONFIGURABLE PARAMETERS FOR FA AND SA ALGORITHMS

Algorithm (Value)			
Parameter	**FA**	**SA**	**FA-SA**
FA	200		200
FA $_{population}$	50		50
FA $_{alpha}$	0.05		0.05
FA $_{beta}$	2		2
FA $_{gama}$	0.001		0.001
SA $_{Max\_iter}$		500	500
SA $_{T\_initial}$		0.001	0.001
SA $_{T\_final}$		0	0

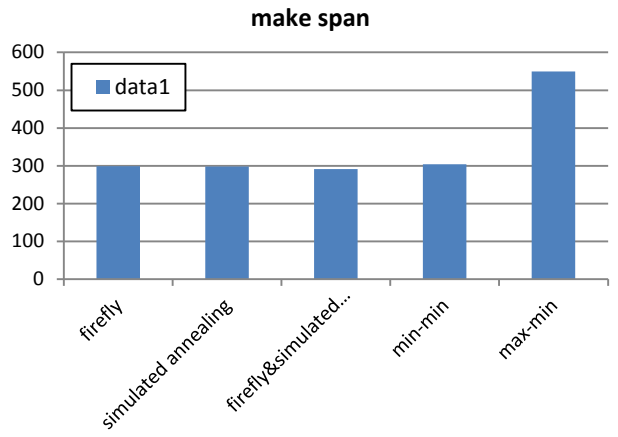


Fig. 7. Comparison of scheduling algorithms on the data1 dataset.

Results of computing make span using the suggested algorithm on the data2 dataset indicate that the FA-SA algorithm was more successful in minimizing make span compared to other algorithms; thus, it can be said that the FA-SA algorithm also creates a good workload balance on virtual machines. Complete results of this comparison are shown in Fig. 8.

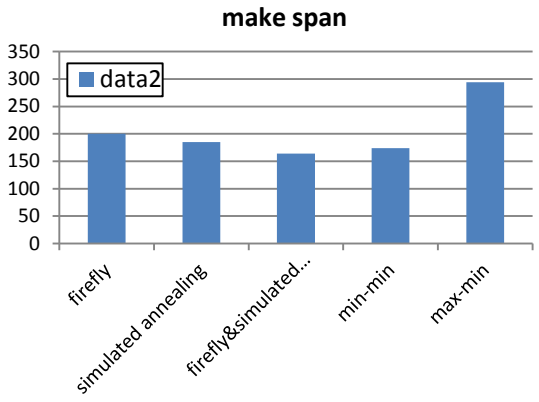


Fig. 8. Comparison of different scheduling algorithms on the data2 dataset.

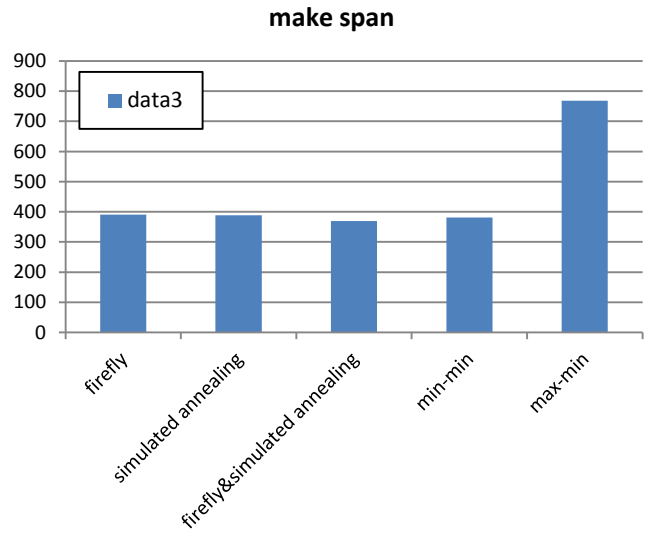


Fig. 9. Comparison of different scheduling algorithms on the data3 dataset.

TABLE III. MAKE SPAN RESULTS FOR EACH OPTIMIZATION ALGORITHM ON ALL THREE DATASETS

Type algorithm	**Data1**	**Data2**	**Data3**
Firefly	299	200	391
Simulated annealing	298	185	388
Firefly& Simulated annealing	291	164	369
Min-min	304	174	381
Max-min	550	294	768

The FA-SA algorithm was further tested on the data3 dataset compared to the previous datasets, had a higher workload. Results of this performance are shown in Fig. 9. These results indicate once again that the FA-SA algorithm is superior to FA, SA, min-min, and max-min algorithms in reducing make span and balancing workload on machines. Overall results of 10 iterations of the algorithms with mean values are shown in Table III.

V. CONCLUSION

Cloud processing, parallel computing and development of distributed computations are all new concepts in computer sciences. One of the major issues in this regard known as a major challenge and an NP-hard problem is the scheduling of tasks in cloud computing. Task scheduling in cloud computing have been discussed in regards to meta-heuristic algorithms such as genetic, ant colony, and other algorithms. However, this study aimed to combine two optimization algorithms, namely the firefly and the simulated annealing algorithms in order to create the new hybrid FA-SA algorithm. Also, a new mechanism for producing initial population and a new method for local search were presented. The suggested algorithm was compared with firefly, simulated annealing, min-min, and max-min algorithms. Results indicated that the FA-SA algorithm can perform much better in reducing make span in different scenarios with different numbers of tasks and virtual machines.

For future works, we will try to focus our attention on energy performance and resource allocation in these systems.

REFERENCES

[1] M. Miller, "Cloud computing: web based applications that change the way you work and collaborate online", Que Publishing, 2008.

[2] Y. Gao, H. Guan and Z. Qi, et al., "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing", Journal of Computer and System Sciences, Vol. 79, pp. 1230 –1242, 2013.

[3] K. Danielson," Distinguishing cloud computing from utility computing", (http:/ / www. ebizq. net/ blogs/ saasweek/ 2008/ 03/ distinguishing_cloud_computing/), 2008.

[4] J. Beliga, R. W. A. Ayre and K. Hinton, et al. "Green cloud computing: balancing energy in processing, storage and transport", Proceedings of the IEEE, Vol. 99, pp.149-167, 2011.

[5] H. Qiyi and H. Tinglei, "An optimistic job scheduling strategy based on QoS for cloud computing", Proceedings of the International Conference on Intelligent Computing and Integrated Systems (ICISS), pp.673-675, 2010.

[6] F. Chang, J. Ren and R. Viswanathan, "Optimal resource allocation for batch testing" ", Proceedings of the International Conference on Software Testing Verification and Validation (ICST), pp.91-100, 2009.

[7] G. Lin, G. Dasmalchi and J. Zhu, "Cloud computing and IT as a service: opportunities and challenges", Proceedings of the International Conference on Web Services (ICWS), pp.1-5, 2008.

[8] D. Kusic, J. O. Kephart and J. E. Hanson, et al. "Power and performance management of virtualized computing environments via lookahead control", Cluster Computing, Vol. 12, pp. 1-15, 2009.

[9] D. Ongaro, A. L. Cox and S. Rixner, "Scheduling I/O in virtual machine monitors", Proceedings of the Fourth ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, pp.1-10, 2008.

[10] H. Kim, H. Lim and J. Jeong, et al. "Task-aware virtual machine scheduling for I/O performance", Proceedings of the ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, pp. 101-110, 2009.

[11] G. Liao, D. Guo and L. Bhuyan, et al. "Software techniques to improve virtualized I/O performance on multi-core systems", Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), pp. 161-170, 2008.

[12] I. Goiri, F. Julia and R. Nou, et al. "Energy-aware scheduling in virtualized datacenters", IEEE International Conference on Cluster Computing (cluster), pp. 58 – 67, 2010.

[13] T. Wood, P. Shenoy and A. Venkataramani, et al. "Black-box and gray-box strategies for virtual machine migration", Proceedings of the 4th USENIX conference on Networked systems design & implementation (NSDl), pp.17-17, 2007.

[14] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey", Theoretical Computer Science Vol.344, pp.243–278, 2005.

[15] M. A. Tawfeek, A.El-Sisi and A. E. keshk, et al.," Cloud task scheduling based on ant colony optimization", Proceedings of the 8th International Conference on Computer Engineering & Systems (ICCES), pp. 64 – 69, 2013.

[16] C.Y. Liu, C. M. Zou and P. Wu, "A task scheduling algorithm based on genetic algorithm and ant colony optimization in cloud computing", Distributed Computing and Applications to Business, Engineering and Science (DCABES), pp. 68-72, 2014.

[17] L. Guo, S. Zhao, and S. Shenet, et al. " Task Scheduling Optimization in Cloud Computing Based on Heuristic Algorithm", Journal of networks, Vol. 7, pp.547-553, 2012.

[18] A. V. Lakra and D. K. Yadav, "Multi-objective tasks scheduling algorithm for cloud computing throughput optimization", Proceedings of the International Intelligent Computing, Communication & Convergence (ICCC), pp. 107 – 113, 2015.

[19] Z. H. Jia, C. Wang and J. Y. T. Leung," An ACO algorithm for makespan minimization in parallel batch machines with non-identical job sizes and incompatible job families", Applied Soft Computing, vol.38, pp.395-404, 2016.

[20] R. L. Graham, "Bounds for certain multiprocessing anomalies", Bell System Technical Journal, Vol.45, pp.1563–1581, 1966.

[21] X. S. Yang, "Firefly algorithms for multimodal optimization". Stochastic Algorithms: Foundations and Applications, Vol. 5792, pp. 169–178, 2009.

[22] O. Jafarzadeh-Shirazi, "Task scheduling with firefly algorithm in cloud computing", Science International, Vol.1, pp-167-171, 2014.

[23] S. Yang, "Nature-Inspired Metaheuristic Algorithms", Luniver Press,2010.

[24] X.S. Yang, "Firefly algorithm, stochastic test functions and design optimization," International Journal of Bio Inspired Computation, Vol. 2, pp. 78–84, 2010.

[25] X.S. Yang, "Firefly algorithm, levy flights and global optimization," Research and Development in Intelligent Systems, pp. 209–218, 2010.

[26] K. Deep, H. Mebrahtu, "Combined mutation operators of genetic algorithm for the travelling salesman problem", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, pp. 1-23, 2011.

[27] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi,"Optimization by Simulated Annealing", Science, Vol.220, pp. 671–680, 1983.

The Proposed Model to Increase Security of Sensitive Data in Cloud Computing

Dhuratë Hyseni
Department of Computer Science
University Ukshin Hoti
Prizren, Kosovo

Artan Luma
Department of Computer Science
South East European University
Tetovo, Macedonia

Besnik Selimi
Department of Computer Science
South East European University
Tetovo, Macedonia

Betim Cico
Department of Computer Engineering
Epoka University
Tirana, Albania

***Abstract*—There is a complex problem regarding security of data in cloud, it becomes more critical when the data in question is highly sensitive. One of the main approaches to overcome this problem is the encryption data at rest, which comes with its own difficulties such as efficient key management, access permissions and similar. In this paper, we propose a new approach to security that is controlled by the IT Security Specialist (ITSS) of the company/organization. The approach is based on multiple strategies of file encryption, partitioning and distribution among multiple storage providers, resulting in increased confidentiality since a supposed attacker will need to first obtain parts of a file from different storage providers, know how to combine them, before any decryption attempt. All details of the strategy used for a particular file are stored on a separate file, which can be considered as a master key for the file contents. Also, we will present each strategy with the results and comments related to the realized measurements.**

***Keywords*—ITSS-IT security specialist; partitioning; confidentiality; cloud service provider; cloud service client; platform as a service; service as a service; third party auditor**

I. Introduction

Cloud computing has brought impressive advantages to the clients interested to use cloud services such as flexibility in managing the space, automatic software update, easier access to needed information and pay per use services etc. The encryption of data at rest is considered to be one of the main issues related to security in the cloud computing and especially cloud storage [1], [12].

People are becoming more interested in cloud computing due to low cost services that it offers [20], [12]. However, two major concerns lie on the security of data: Data confidentiality and audibility, which seem to be one of the main obstacles to the adoption of cloud computing. In addition, security concerns are preventing some organizations from adopting cloud computing to their businesses, others are considering using combination of a secure internal private cloud with less secured public cloud. Moreover, this is an approach where sensitive data can be deployed in private cloud while less sensitive data can be externally deployed in a public cloud. However, this approach seems to have problems when allocation applications in clouds usually operate on an ad-hoc, per-application basis which is not ideal as it lacks rigorosity and audibility.

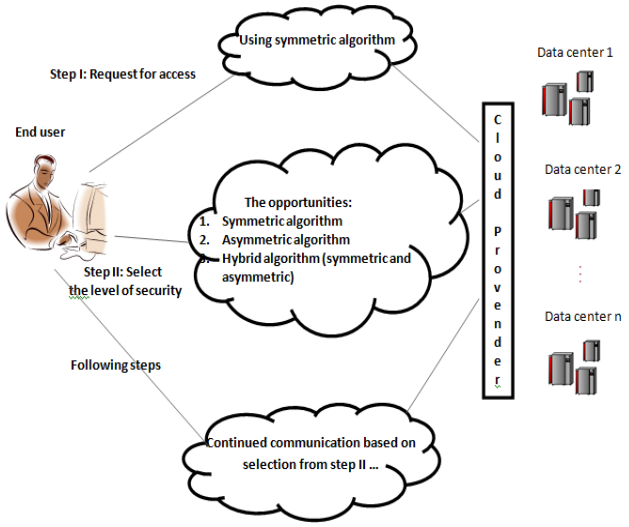


Fig. 1. Scenario of proposed model in the cloud security controlled by the ITSS [2], [3].

To be able to control security from the ITTS of the company/organization, we have proposed a model as an adequate system solution in the cloud computing [2], [3], Fig. 1. In our model, the main actor that manages the security of data in the cloud is the ITSS of the company/ organization. The ITSS selects different security parameters, such as encryption algorithms and keys, as well as partitioning strategies in order to distribute files in the cloud. However, all files transferred to the cloud should follow these company/organization wide rules in the future.

This research starts from our previous works [1], [2] and it continues with proposed scenarios (as new solutions) based on sensitive data.

II. Literature Review

The following research papers claim that there is a large number of researches [12]-[19], even though, the indication is that there is lack of reliability in cloud computing by users.

Based on the literature review, we are going to discuss some security solution in the cloud computing which had great impact in our proposed model.

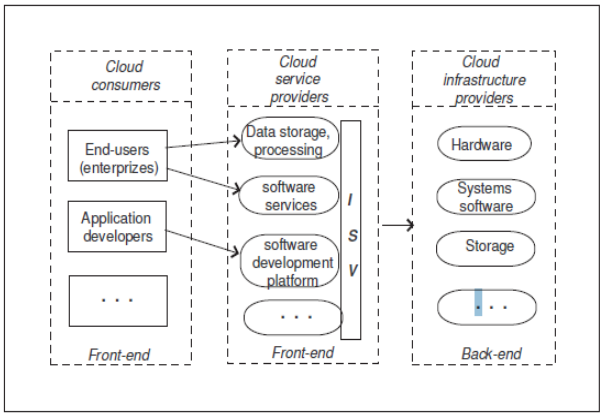


Fig. 2. Levels of abstractions of cloud computing.

The proposed model from [4], consists of three different security scanners with different choices depending on their requests from interested parties for use in the cloud computing, Fig. 2. The researchers began to carry out a thorough analysis of security in the cloud computing modules and then they were focused on the basic requirements to secure a system's cloud protection [4]. Their work orientation is geared towards the Advanced Cloud Protection System (ACPS), which is the result of security for the Linux Kernel Virtual Machine. In this scientific work through the ACPS, it is possible to protect the integrity of virtual machines (VM) and distributed computing middleware, which appear as supporting elements in the cloud computing. Also, different monitoring offered by the VMs in cooperation with the components of the infrastructure is proposed against the various attacks [5].

Authors from [6] have conducted a survey where the main objective is the security on the multi-tenant software platform, offered as part of the PaaS Model. In the PaaS Model, they have discovered the technical weaknesses of the multi-tenancy support platforms, similar to the .Net or Java. In this paper [6], the authors suggested that inside the PaaS, the code isolated by the Cloud Service Client (CSC) and reduces the probability of possible errors in the other applications. Therefore, based on the two weaknesses during the application development, all the rules to ensure this code is developed in the PaaS. As a conclusion, the Cloud Service Provider (CSP) for the PaaS Model should use all the possible mechanisms provided by the security environment, to minimize the potential risks that come to this model. Here the encryption is seen as a choice of the data security during processing and rest of the database, the data has not been tampered with or seen by the other parties in the cloud. Although it has been proposed to encrypt the data for the above mentioned problems, it is not safe for those problems. In the cloud computing, exploited patterns (for management, data processing and storage), security and privacy processes cannot be used with the same encryption techniques as traditional ones [5].

According to [7], authors proposed security for integrity and privacy of the data, based on the efficient audit at low cost. It is also proposed the audit to be carried out by a third party auditor (TPA), which audits the data from time to time and the same data should be available to the client by passing the load and the cost for validation and downloading data at the local level. Practically, this is presented in this way, the data owner has assigned a secret cell used to process the file which is divided into several blocks. Before sending the file and the verification parts to the CSP, one part of the public verification information is already generated and stored in the TPA. Based on the requirements of the data owner, the TPA uses the data retrieval protocol and then enables auditing or controlling data integrity by using public verification information. The perception of this architecture is that it can be implemented in the TPA without including the data owner.

Authors from [8] have requested that the third-party privacy and auditing problems and the data integrity should be resolved by means of the TPA. In this paper, the integrity audit is supported by using the homomorphic encryption. There are solutions in order to increase the efficiency of the TPA which tries to offer both data collection techniques, integrity and the data privacy.

III. Research Methodology

The aim of this study is to propose a model to control data security in the cloud. Recent trends in cloud security have played an important role to attract organizations and companies to deploy sensitive data in cloud. In this context, the proposed model offers different scenarios based on the level of sensitivity of data. In another point of view, it increases reliability of clients in cloud computing. This reliability will be increased by offering controls of data security to the end user- ITSS.

Our model, Fig.1 was proposed in [2], [3], it is the same philosophy for controlling the security in the cloud and it is based on two objectives:

- Control of security depending on the ITSS of a certain organization.
- Possibility to select the security options, based on different algorithms.

The proposed model offers three scenarios based on the data sensitivity:

Scenario I: Security is based on the choice of the ITSS organization, depending on the information proposed by the model.

Scenario II: Based on the features of the file, algorithms which are proposed and the ITSS makes a choice.

Scenario III: Security, based on the file encryption and partition by the ITSS of organization with two possibilities:

A. *The File is Partitioned than Encrypted in Particular Parts (P1, P2,…, Pn)*

It is based on the partitioning of file then encrypts these partitions by algorithm which the ITSS selects, depending on the sensitivity of data in the organization, Fig. 3. This alternative is preferable when it reads the partitions before retrieving completely all the parts of the file. For example a video starts to play (meaning there won't be delays for the client) before retrieving other parts of the file uploaded to cloud.

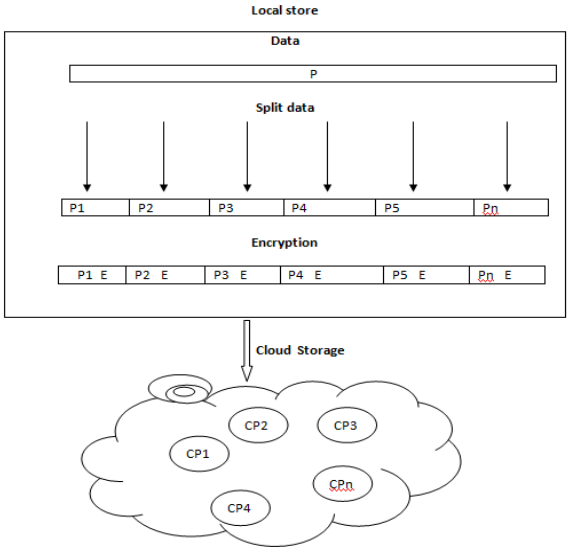


Fig. 3. Partitioned and then encrypted respectively parts of the file - CASE_I.

B. *The File is Encrypted than Partitioned in Particular Parts (P1, P2,…, Pn)*

In the second case of proposed model, we encrypt the file then split it in partitions based on the scenario and algorithm selected. This scenario offers higher security of the data, because we must have all parts of the file in order to be able to read them, Fig. 4.

Each part can be stored in different clouds. A new file *p0*, contains selected algorithms, index and position of the file. File *p0* is significantly smaller, encrypted by DSA algorithm and can be stored anywhere in the cloud or in the local machine.

The steps needed for our proposed model to increase security in the cloud, Fig. 5:

Step I: Access to program using password and user name.

Step II: Selection of scenario of security based on the sensitivity data (Scenario I, Scenario II or Scenario III).

Step III: Selection of algorithms of data encryption.

Step IV: Selection of uploading/downloading files to the cloud.

The proposed model is implemented on the .NET Framework 4.5, which is developed by Microsoft that runs primarily on the Microsoft Windows and was developed in the c# programming language.

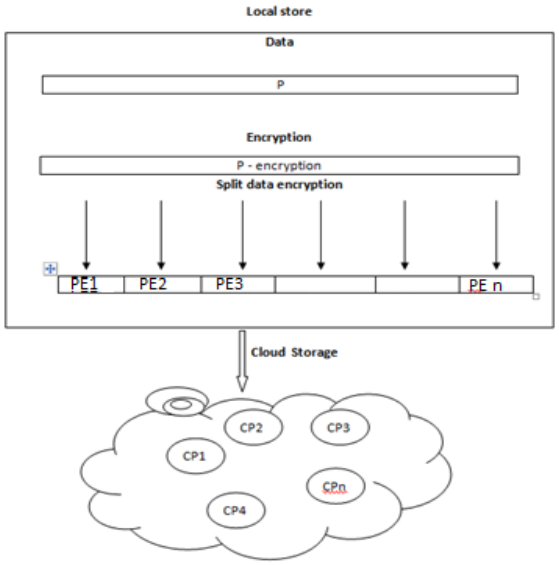


Fig. 4. Encrypted then partitioned file- CASE_II.

IV. Result and Analysis

Considering all measurements, we have the following working conditions: Processor: Initial (R) Celeron (R) CPU 1005 M 1.90 GHz and network details, Ping: 55ms, Download: 15.46 Mbps, Upload: 2.22 Mbps.

For all measurements we have used measuring unit of time execution in milliseconds. Also, for every case of measurements, there have been two ways of measurements for Upload and Download.

These measurements are realized using three different schemas for every type of algorithms (Symmetric, Asymmetric and Hybrid). Moreover, in this part, the execution time of measurement starts after the file is partitioned and special measurement (i.e. t1,t2…tn) are made for every partition. Also as total time in general is taken T, while in partitions t1,t2…tn, Table I (see Appendix). Also, at Table I (see Appendix), obtained measurements for the type of file, are provided in different colors.

For measurements we have used three symmetric algorithms: AES, DES, TripleDES and three asymmetric algorithms: RSA Diff-Hellman and El Gamal, as well as hybrid algorithms (combination of both symmetric and asymmetric algorithms).

As for the measurements, we have used types of files from Table II:

TABLE II. Files Used For Measurement

Type	Size	Comment
.doc	2969KB	Large
.doc	606KB	Medium
.pdf	606KB	Medium
.png	606KB	Medium
.mov	454KB	Medium

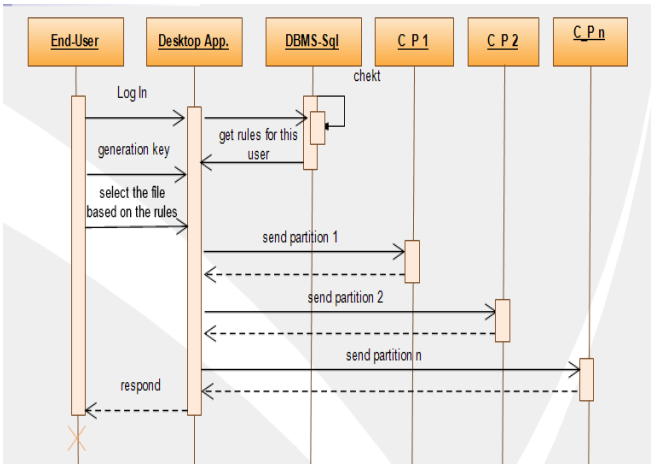


Fig. 5. Communication between user and proposed model.

A. Schema for Symmetric Algorithms

Fig. 6 shows the schema used for symmetric algorithms. It is seen that the file is partitioned then every partition is encrypted with a different symmetric algorithm. The time of measurement for upload starts from partitioning and continues with encryption then sending the file to the cloud and vice versa for download. Measurements are obtained separately for every partition: as in Fig. 9. In Table I (see Appendix), this type of measurements is shown in column: "Type of Algorithm: S".

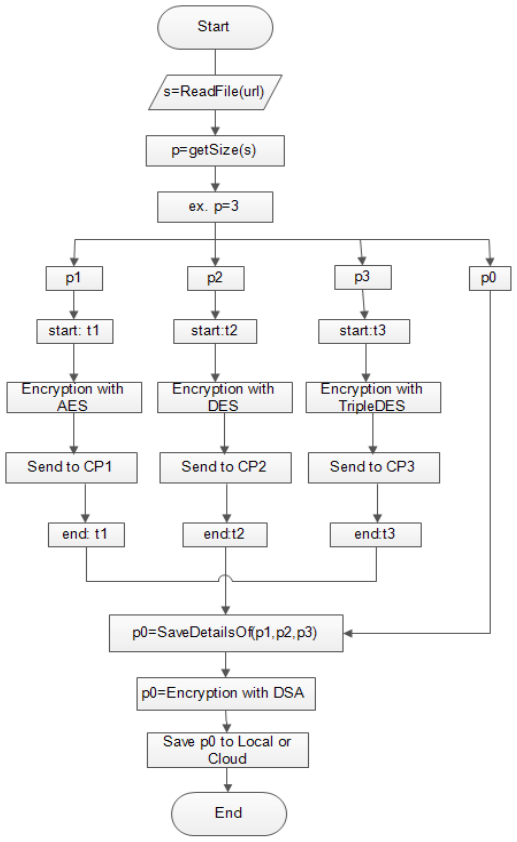


Fig. 6. Schema used of measurements for symmetric algorithms.

B. Schema for Asymmetric Algorithms

Fig. 7 shows the schema used for asymmetric algorithms. It is seen here that the file is partitioned then every partition is encrypted with a different asymmetric algorithm. The time of measurement begins from the separation of the file then it is encrypted and sent to the cloud providers and vice versa for download. Measurements are done separately for each partition as in Fig. 9. In Table I (see Appendix), this group of measurements is shown in column: "Type of Algorithm: A".

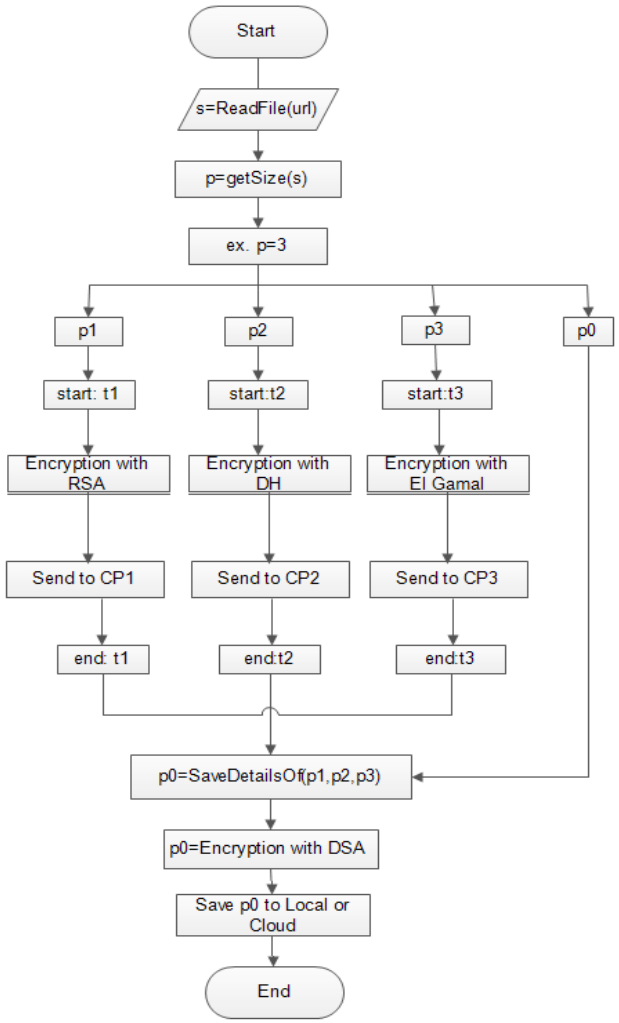


Fig. 7. Schema used of measurements for asymmetric algorithms.

C. Schema for Hybrid Algorithms

Fig. 8 shows the schema used for hybrid algorithms. Here we see that the file is separated in partitions, each partition is encrypted with a different algorithms, symmetric and asymmetric. The time of measurement for upload begins from partitioning, continues with encryption then sending the file to different cloud providers and vice versa for download. Measurements were made separately for every partition, as in schema Fig. 9. In Table I (see Appendix), this group of measurements is shown in the column: "Type of Algorithm: H".

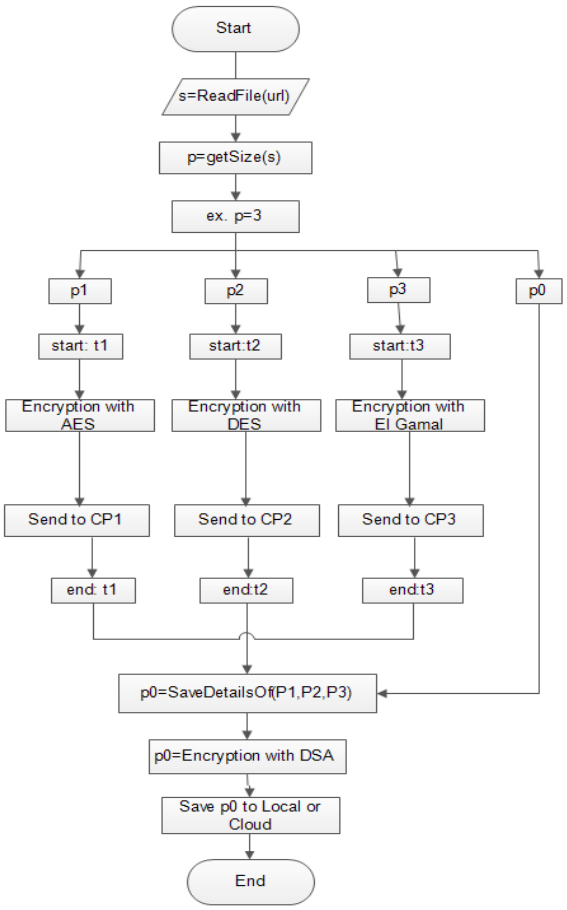


Fig. 8. Schema used of the measurements for hybrid algorithms.

Fig. 8 and Table III (see Appendix), present general results from three types of schemas (Fig. 6, 7 and 8). At this graph we have a better view for every type of file and the execution time.

V. Discussion of the Results

In these measurements we tested types of files from Table II, the focus at this part of measurements were three schemas from Fig. 6, 7 and 8.

The main reason of any measurement was the difference of execution time for three groups of algorithms: symmetric, asymmetric and hybrid. At this point we can conclude that symmetric algorithms are faster than asymmetric algorithms, whereas the hybrid algorithm are in the middle (it is confirmed form Fig. 9 (see Appendix)), therefore the difference of the time of execution for Upload/Download is emphasized when the file is large as in our case with size 2969 KB, Table II. In Fig. 9, we confirm that the type of file does not affect in the time of execution, for both processes (Upload/Download).

The proposed model provides different options for security of data, so ITSS of the organization/company decides on the options:

- High level of security, for very sensitive data. In this part of measurements we can conclude that high level for security of data from the proposed model uses the method of sending partitions in cloud Fig. 3, CASE_I(file is partitioned then encrypted) and scenario from Fig. 7 (use of asymmetric algorithms) that are safer, referring to [10], [11] and [12].
- Moderate level of security for less sensitive data. For data that the level of security is moderate, we still propose that partitions be sent to cloud as in Fig. 3, CASE_I and scenario in Fig. 8 (use of hybrid algorithms), also hybrid algorithms are proposed by [9] as well, as a better solution for data encryption.
- Lower level of security for data that are least sensitive. For data that is not required a high level of security and big data, then we suggest that partitions should be sent to the cloud as in Fig. 4, CASE_II (the file is encrypted then sent to clouds) referring to the results from Fig. 9, we believe that this method is faster for big and less sensitive data. Also for this case we suggest that symmetric algorithms for the encryption of the partitions (referring to Fig. 6 and Table III (see Appendix), tend to be much faster, also referenced on [10]).

VI. Conclusion

Based on the recent trends of cloud computing, security practices in current researches have often overlooked the importance of mutual trust. Therefore, the growth of this trust has been the main motivation for our research.

Despite the fact that different ideas exist for security in cloud, our proposed model offers the possibility of controlling security by the ITSS [2], [3], controlling the security in cloud based on different options.

There are offered different schemas for the strategy it uses for the model based on the sensitivity of the data. In addition, another important issue of this research is the measurements realized in two parts:

- First: Using the encryption of partitions by symmetric and asymmetric algorithms (example FileM(p1-RSA, p2-RSA, p3-RSA…pn-RSA)), for methods of sending the partitions in cloud proposed in our model, Fig. 3 and 4.
- Second: The encryption of partitions of the same file with different algorithms. Symmetric and asymmetric (example FileN(p1-RSA, p2-AES, p3-DES…pn-RSA)), for the methods of sending the partitions to the cloud proposed in our model, Fig. 3 and 4.

The proposed model for security in cloud is possible in different working conditions, especially for those environments that work is based on sensitive data and for those companies that still hesitates to deploy in cloud.

As future work we are going to realize measurements for other types of files, also the possibility of realizing measurements for other scenarios including other algorithms in different working environment. In addition, advance of application in a way that supports users of different professions.

REFERENCES

[1] Chauhan, N. S., & Saxena, A. (2013). Cryptography and Cloud Security Challenges. CSI Communications.

[2] Hyseni, D., Cico, B., & Shabani, I. (2015). The proposed model for security in the cloud, controlled by the end user, In Embedded Computing (MECO), 2015 4th Mediterranean Conference on (pp. 81-84). IEEE.

[3] Hyseni, D., Çiço, B., & Selimi, B. (2016).Conception, design and implementation of an interface for security in cloud controlled by the end user. International Journal on Information Technologies & Security, 8(2).

[4] Lawal, B. O., Ogude, C., & Abdullah, K. K. A. (2013). Security management of infrastructure as a service in cloud computing. *African Journal of Computing & ICT*, *6*(5).

[5] Ouedraogo, M., Mignon, S., Cholez, H., Furnell, S., & Dubois, E. (2015). Security transparency: the next frontier for security research in the cloud. Journal of Cloud Computing, 4(1), 12.

[6] Rodero-Merino L, Vaquero LM, Caron E, Muresan A, Desprez F (2012). Building safe PaaS clouds: a survey on security in multitenant software platforms. Computers & Security 31(1):96–108

[7] Zhu Y, Hu H, Ahn GJ, Yau SS (2012). Efficient audit service outsourcing for data integrity in clouds. J Syst Softw 85(5):108–1095, Elsevier

[8] Wang C, Wang Q, Ren K, Lou W (2010). Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing. In: Proceedings of the 29th conference on Information Communications (INFOCOM 2010). IEEE, San Diego, pp 525–533

[9] Hofheinz, D., & Kiltz, E. (2007). Secure hybrid encryption from weakened key encapsulation. Advances in Cryptology-CRYPTO 2007, 553-571.

[10] Janakiraman, V. S., Ganesan, R., & Gobi, M. (2007, July). Hybrid Cryptographic Algorithm for Robust Network Security. In The International Congress for global Science and Technology (Vol. 17, No. 24, p. 33).

[11] Arockiam, L., & Monikandan, S. (2013). Data security and privacy in cloud storage using hybrid symmetric encryption algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(8), 3064-70.

[12] Khan, M. S. S., & Tuteja, R. R. (2014). Security in cloud computing using cryptographic algorithms. IJCA.

[13] Kamara S, Lauter K (2010) Cryptographic cloud storage. In: Proceedings of Financial Cryptography. Workshop on Real-Life Cryptographic, Protocols and Standardization, Springer, Heidelberg

[14] Juels A, Kaliski BS Jr (2007) Pors: proofs of retrievability for large files. In: Proceedings of the 2007 ACM Conference on Computer and Communications Security (CCS 2007). ACM Digital library, New York, pp 584–597

[15] Ateniese G, Burns R, Curtmola R, Herring J, Kissner L, Peterson NJZ, Song D (2007) Provable Data Possession at Untrusted Stores. In: Proceedings of CCS'07, Alexandria, VA. ACM, New York, pp 598–609

[16] Doelitzscher F, Reich C, Knahl M, Clarke N (2012) An agent based business aware incident detection system for cloud environments. Journal of Cloud Computing:Advances, Systems and Applications 1(9):1–19, Springer-Verlag, Berlin

[17] Mather, T., Kumaraswamy, S., & Latif, S. (2009). Cloud security and privacy: an enterprise perspective on risks and compliance. " O'Reilly Media, Inc.".

[18] Cruz, Z. B., Fernández-Alemán, J. L., & Toval, A. (2015). Security in cloud computing: A mapping study. Computer Science and Information Systems, 12(1), 161-184.

[19] Chandra, J. V., Challa, N., & Hussain, M. A. (2014). Data and information storage security from advanced persistent attack in cloud computing. International Journal of Applied Engineering Research, 9(20), 7755-7768.

[20] Anshel, M., & Boklan, K. D. (2007). Introduction to cryptography with coding theory. The Mathematical Intelligencer, 29(3), 66-69.

APPENDIX

TABLE I. RESULTS OF MEASUREMENTS FROM PROPOSED MODEL

No.	File type	Process	File size	Scenario	Algorithm	Execution Time/ms	Type of Algorithm
t1	.doc	Upload	2969KB	CASE_I	AES	4828	
t1	.doc	Download	2969KB	CASE_I	AES	2230	
t2	.doc	Upload	2969KB	CASE_I	Des	5112	S
t2	.doc	Download	2969KB	CASE_I	Des	4120	
t3	.doc	Upload	2969KB	CASE_I	TripleDES	15575	
t3	.doc	Download	2969KB	CASE_I	TripleDES	11352	
					T=t1+t2+t3	43217	
t1	.doc	Upload	2969KB	CASE_I	RSA	27954	
t1	.doc	Download	2969KB	CASE_I	RSA	21562	
t2	.doc	Upload	2969KB	CASE_I	Diffie-Hellman	28260	
t2	.doc	Download	2969KB	CASE_I	Diffie-Hellman	23323	A
t3	.doc	Upload	2969KB	CASE_I	ElGamal	41837	
t3	.doc	Download	2969KB	CASE_I	ElGamal	32500	
					T=t1+t2+t3	175436	
t1	.doc	Upload	2969KB	CASE_I	RSA	27954	
t1	.doc	Download	2969KB	CASE_I	RSA	21562	
t2	.doc	Upload	2969KB	CASE_I	Des	5112	
t2	.doc	Download	2969KB	CASE_I	Des	4120	H
t3	.doc	Upload	2969KB	CASE_I	ElGamal	41837	
t3	.doc	Download	2969KB	CASE_I	ElGamal	32500	
					T=t1+t2+t3	**133085**	
t1	.doc	Upload	606KB	CASE_I	AES	4476	
t1	.doc	Download	606KB	CASE_I	AES	1203	
t2	.doc	Upload	606KB	CASE_I	Des	6603	S
t2	.doc	Download	606KB	CASE_I	Des	3520	

t3	.doc	Upload	606KB	CASE_I	TripleDES	7,405	
t3	.doc	Download	606KB	CASE_I	TripleDES	3850	
					T=t1+t2+t3	27057	
t1	.doc	Upload	606KB	CASE_I	RSA	7833	A
t1	.doc	Download	606KB	CASE_I	RSA	5422	
t2	.doc	Upload	606KB	CASE_I	Diffie-Hellman	9267	
t2	.doc	Download	606KB	CASE_I	Diffie-Hellman	6055	
t3	.doc	Upload	606KB	CASE_I	ElGamal	8574	
t3	.doc	Download	606KB	CASE_I	ElGamal	5391	
					T=t1+t2+t3	42542	
t1	.doc	Upload	606KB	CASE_I	AES	4476	H
t1	.doc	Download	606KB	CASE_I	AES	1203	
t2	.doc	Upload	606KB	CASE_I	Diffie-Hellman	9267	
t2	.doc	Download	606KB	CASE_I	Diffie-Hellman	6055	
t3	.doc	Upload	606KB	CASE_I	TripleDES	7,405	
t3	.doc	Download	606KB	CASE_I	TripleDES	3850	
					T=t1+t2+t3	32256	
t1	.pdf	Upload	606KB	CASE_I	AES	4922	S
t1	.pdf	Download	606KB	CASE_I	AES	2051	
t2	.pdf	Upload	606KB	CASE_I	Des	5897	
t2	.pdf	Download	606KB	CASE_I	Des	3625	
t3	.pdf	Upload	606KB	CASE_I	TripleDES	6,070	
t3	.pdf	Download	606KB	CASE_I	TripleDES	3986	
					T=t1+t2+t3	26551	
t1	.pdf	Upload	606KB	CASE_I	RSA	6568	A
t1	.pdf	Download	606KB	CASE_I	RSA	4203	
t2	.pdf	Upload	606KB	CASE_I	Diffie-Hellman	10077	
t2	.pdf	Download	606KB	CASE_I	Diffie-Hellman	6950	
t3	.pdf	Upload	606KB	CASE_I	ElGamal	10453	
t3	.pdf	Download	606KB	CASE_I	ElGamal	7106	
					T=t1+t2+t3	45357	
t1	.pdf	Upload	606KB	CASE_I	Des	5897	H
t1	.pdf	Download	606KB	CASE_I	Des	3625	
t2	.pdf	Upload	606KB	CASE_I	TripleDES	6,070	
t2	.pdf	Download	606KB	CASE_I	TripleDES	3986	
t3	.pdf	Upload	606KB	CASE_I	ElGamal	10453	
t3	.pdf	Download	606KB	CASE_I	ElGamal	7106	
					T=t1+t2+t3	37137	
t1	.png	Upload	606KB	CASE_I	AES	4784	S
t1	.png	Download	606KB	CASE_I	AES	2130	
t2	.png	Upload	606KB	CASE_I	Des	5218	
t2	.png	Download	606KB	CASE_I	Des	2962	
t3	.png	Upload	606KB	CASE_I	TripleDES	6,213	
t3	.png	Download	606KB	CASE_I	TripleDES	3908	
					T=t1+t2+t3	25215	
t1	.png	Upload	606KB	CASE_I	RSA	6859	A
t1	.png	Download	606KB	CASE_I	RSA	4799	
t2	.png	Upload	606KB	CASE_I	Diffie-Hellman	8554	
t2	.png	Download	606KB	CASE_I	Diffie-Hellman	5210	
t3	.png	Upload	606KB	CASE_I	ElGamal	9688	
t3	.png	Download	606KB	CASE_I	ElGamal	5865	
					T=t1+t2+t3	40975	
t1	.png	Upload	606KB	CASE_I	RSA	6859	H
t1	.png	Download	606KB	CASE_I	RSA	4799	
t2	.png	Upload	606KB	CASE_I	Diffie-Hellman	8554	
t2	.png	Download	606KB	CASE_I	Diffie-Hellman	5210	
t3	.png	Upload	606KB	CASE_I	TripleDES	6,213	
t3	.png	Download	606KB	CASE_I	TripleDES	3908	
					T=t1+t2+t3	35543	
t1	.mov	Upload	454KB	CASE_I	AES	3601	S
t1	.mov	Download	454KB	CASE_I	AES	1956	
t2	.mov	Upload	454KB	CASE_I	Des	3606	
t2	.mov	Download	454KB	CASE_I	Des	1833	
t3	.mov	Upload	454KB	CASE_I	TripleDES	6359	
t3	.mov	Download	454KB	CASE_I	TripleDES	4662	
					T=t1+t2+t3	15658	
t1	.mov	Upload	454KB	CASE_I	RSA	6923	A
t1	.mov	Download	454KB	CASE_I	RSA	4856	

t2	.mov	Upload	454KB	CASE_I	Diffie-Hellman	7694	
t2	.mov	Download	454KB	CASE_I	Diffie-Hellman	5887	
t3	.mov	Upload	454KB	CASE_I	ElGamal	8711	
t3	.mov	Download	454KB	CASE_I	ElGamal	4985	
					T=t1+t2+t3	39056	
t1	.mov	Upload	454KB	CASE_I	Des	3606	H
t1	.mov	Download	454KB	CASE_I	Des	1833	
t2	.mov	Upload	454KB	CASE_I	Diffie-Hellman	7694	
t2	.mov	Download	454KB	CASE_I	Diffie-Hellman	5887	
t3	.mov	Upload	454KB	CASE_I	ElGamal	8711	
t3	.mov	Download	454KB	CASE_I	ElGamal	4985	
					T=t1+t2+t3	32716	

TABLE III. RESULTS OF MEASUREMENTS PRESENTED IN GENERAL

File type	File size	Algorithm	Execution Time / ms
.doc	2969KB	S	43217
.doc	2969KB	A	175436
.doc	2969KB	H	133085
.doc	606KB	S	27057
.doc	606KB	A	42542
.doc	606KB	H	32256
.pdf	606KB	S	26551
.pdf	606KB	A	45357
.pdf	606KB	H	37137
.png	606KB	S	25215
.png	606KB	A	40975
.png	606KB	H	35543
.mov	454KB	S	15658
.mov	454KB	A	39056
.mov	454KB	H	32716

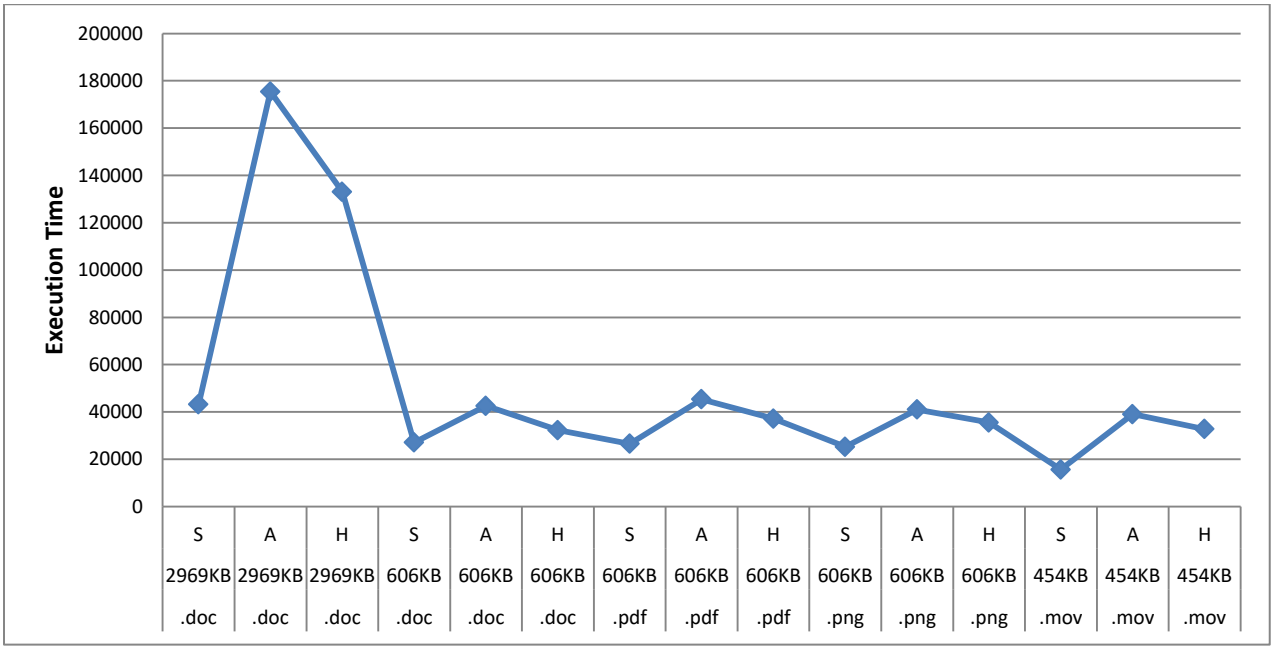


Fig. 9. Graphical presentation of data from the Table III.

Improvement of the Frequency Characteristics for RFID Patch Antenna based on C-Shaped Split Ring Resonator

Mahdi Abdelkarim
Physics Department
Faculty of Science, University of ElManar
Tunis, Tunisia

Lassad Larach
Physics Department
Faculty of Science, University of ElManar
Tunis, Tunisia

Seif Naoui
Physics Department
Higher Institute of Applied Science and Technology
Kasserine, University of Kairouan
Tunis, Tunisia

Ali Gharsallah
Physics Department
Faculty of Science,
University of ElManar
Tunis, Tunisia

***Abstract*—In this paper, we present a new technique for improving frequency characteristics and miniaturizing the geometric dimension of the RFID patch antenna that operates in the SHF band. This technique consists in implementing a periodic network based on a new model of a split C-ring resonator on the square slots of radiating rectangular patch antenna. The results of the simulation have proved that the proposed technique has an excellent radiation efficiency and size reduction compared to the reference patch antenna.**

***Keywords*—Rectangular patch antenna; C-shaped split ring resonator; RFID reader antenna; metamaterials antenna**

I. Introduction

RFID (Radio Frequency Identification) is an identification technology that has been appearing for military applications in 1950 [1], [2]. In 1990, RFID became prevalent in many industry services, such as access control, document tracking, distribution logistics, automotive systems, and animal tracking [3]-[6]. In these applications, the setup of a wireless communication link is crucial. So in such a case, we seek to increase the directivity, gain and minimize the size of the RFID antenna as possible to make the reader more compact and reduce the cost of the antenna.

There are several techniques to improve performance and miniaturize the dimension of UHF RFID antennas. The most well-known technique used in RFID is the inclusion of left-handed metamaterials in the RFID antenna.

The metamaterials were proposed by Victor Viselago in 1968 [7]-[13]. They are artificial materials consisting of the network of periodic split ring resonators for obtaining a medium with negative refractive index. Since their appearance, we found a lot of studies that increasingly numerous on the implementation electromagnetism application which aims to improve performance and minimize RFID antenna.

In [14] the author provided a simple method to improve the performance of the rectangular patch antenna by using a novel Chiral Metamaterial (CM). This method consists of placing the substrate layer with metamaterials above the antenna. The study results show that for the antenna with the CM, the gain increases from 4.84 dB to 7.46 dB, by a rate of 2.6 dB. In [15] a microstrip patch antenna based on an array of Complementary G-Shape Split Ring Resonator (CGSRR) is proposed to improve the gain and bandwidth of patch antenna. The measured results show that by owing to the presence of CGSRRs around the rectangular patch, the gain improves by 2.5 dB compared to the original antenna. Together the bandwidth of the antenna increases from 3.5% to 5.1%. Moreover, in [16] the author developed a novel technique to reduce the size and increase the bandwidth of the rectangular patch antenna. It consists of implementing an array of complementary split ring resonators into the ground plane. The measured results proved that the proposed method able to reduce the size of patch antenna of 11.2% and improve the bandwidth of 202.6%. In [17] the author presents a study on the performance and miniaturization of patch antennas with complementary metamaterials. Based on simulation result, the integration of a homogeneous and periodic array of complementary split ring resonators on the surface of the patch antenna has an interesting consequence on the antenna size and negative effect on the antenna gain. Therefore, according to the developed works, we can conclude if we seek to enhance the performance of the patch antenna, we lose the miniaturization of antenna dimension and it is the same thing in revere case.

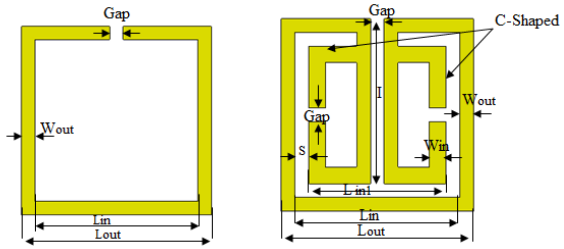
Thus, in this paper, a new method based on the integration of novel C-Shaped split-ring resonator structure on the radiating patch antenna is proposed, studied and evaluated to miniaturize the dimension and improve the performance of the RFID patch antenna. The dimension of the proposed antenna

is calculated according to [17] by using a microstrip line feed to operate at 2.4 GHz RFID Band.

The paper is organized as follows, In Section 2 the design of the proposed SRR with C-Shaped technique is given and the simulation results were compared with simulation results of classic SRR at the same geometric forms in term of S11, surface currents, electric and magnetic fields, frequency band and effective permeability. In Section 3, the new technique based on the integration of an array of CSSR on the radiating surface of the RFID patch antenna is explained. Finally, in Section 4, the proposed structure is compared to the previous works and the traditional microstrip patch antenna in order to select the perfect technique [14]-[17].

II. C-Shaped Split Ring Resonator

In this session, a novel C Shaped Split Ring Resonator (C-Shaped SRR) can give a left-hand medium (LHM) around its resonance frequency is presented. So, Based on the Split Ring (SRR) (Fig. 1(a)) which was introduced by Tang Chun Ming in 2010 [18], we will add in SRR two C-shaped implemented within the external ring for obtaining C-shaped split ring resonator. Details of the C-Shaped SRR unit cell are described in Fig. 1(b). The proposed metamaterials was designed and fixed on Roger 5880 substrate with dielectric constant ε_r =2.2 and thickness h=1.6 mm. The dimensions of C-Shaped SRR are: Lout= 7 mm, Lin= 6.5m, Lin1= 6 mm, Gap= 0.5 mm, Win=0.5 mm, Wout=0.5 mm, S= 0.5 mm, I=5.4 mm.



(a) (b)

Fig. 1. Structure of (a) split ring resonator (b) Proposed C-shaped split ring resonator.

A. Simulation and Results

Two wave ports are set left-right faces of the box of proposed SRR to obtain the reflection and transmission coefficients. The effect of two C -Shaped on the reflection coefficient of the proposed cell is shown in Fig. 2. It is observed that by adding the C shaped in SRR, the resonance frequency shifts toward a lower frequency. So the proposed metamaterials illustrate at a frequency 2.6 GHz with a transmission coefficient S21 of -21 dB. Therefore, this resonance frequency is slightly smaller than the original SRR by a difference arriving up to 2 GHz. This degradation at the resonant frequency due to the inductive and capacitive effect which appears in the metallic conductor lines lc and gab between the C-shaped. The relationship (1) [10] shows that if lc changes, the inductance L also changes and thus, there is a major gap change in resonant frequencies which can be interpreted from Fig. 2. Similarly, when the gap between the C-shaped (2) is decreased, then the capacity C_{gap} also increases and thus according to (3) it is to be noted resonant frequencies must shift towards left.

$$L = \frac{\mu_0 h_c}{W} l_c \quad (1)$$

$$C_{gap} = \varepsilon_r \varepsilon_0 \frac{e}{gap} \quad (2)$$

Where,

hc: Conductor thickness [mm]

μ0: permeability in the void $1.56 \times 10^{-6} [m.Kg.S^{-2}.A^{-2}]$

e: surface area [mm²]

Ɛr: relative permittivity 3.2

Ɛ0: permittivity in the void 8.854*10^{12}A.sV^{-1}.m^{-1}

$$f = \frac{1}{2\pi\sqrt{LC}} \quad (3)$$

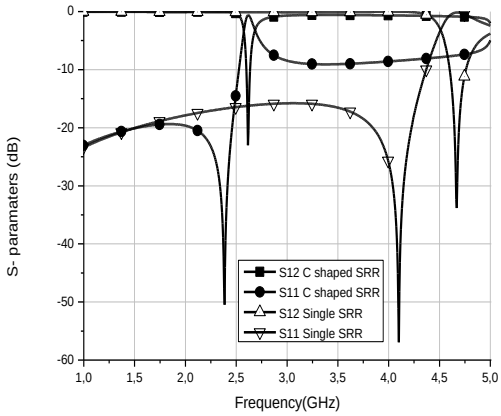


Fig. 2. Curve of the reflection and transmission coefficients produced by SRR and C- Shaped SRR structure.

$$\mu_{eff} = \pm \sqrt{\frac{(1+S_{11})^2 - S_{12}^2}{(1-S_{11})^2 - S_{12}^2}} \quad (4)$$

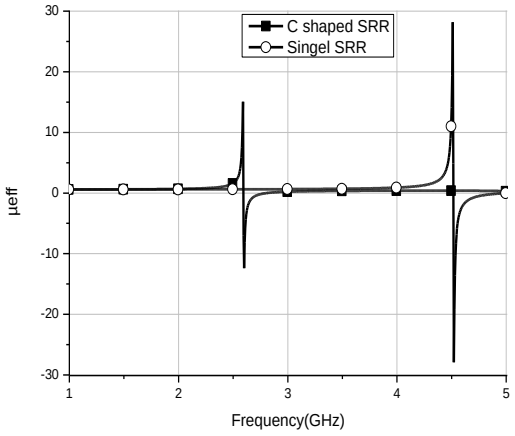
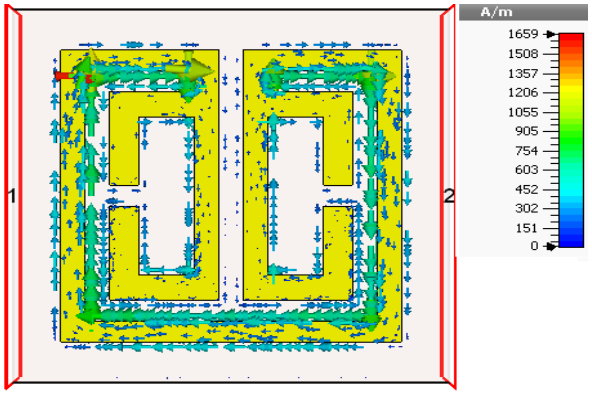
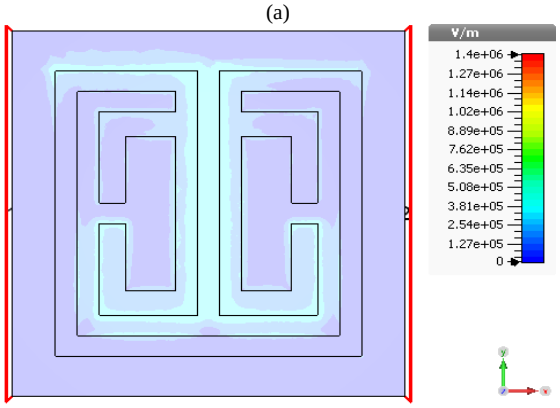


Fig. 3. The Effective permeability response of SSRR and C-shaped SRR.

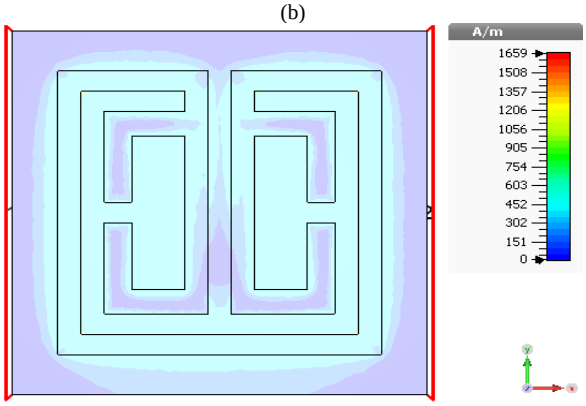
Fig. 3 illustrates the curve of the real part of the effective permeability of C-shaped SRR and SRR cell. We can find that where the CSRR unit cell resonates, the permeability gives negative values ranging from zero to -12.5.So for the frequency span 2.5-2.55 GHz. This negative region proves that the proposed metamaterial act as a Left-hand medium (LHM). Moreover, to have a deeper insight into the operation of the proposed CSRR, the distributions of the surface currents, magnetic field, and the electric field, are shown in Fig. 4(a)–4(c). All the distributions are presented out at the frequency which operates the proposed CSRR.



(a)



(b)



(c)

Fig. 4. Distributions of the (a) surface currents(b) electric field and (c) magnetic field in C-Shaped cell.

TABLE I. COMPARATIVE RESULTS BETWEEN THE SINGLE SRR AND C-SHAPED SRR STRUCTURE

Characteristics of metamaterials	**SRR**	**C-Shaped SRR**
Frequency (GHz)	4.63	2.6
S21 [dB]	-33.5	-24.4
S11 [dB]	-57	-51
μ_{eff}	From 0 to -28	From 0 to -12.5
FB [MHz]	85	40
Size % λ_0	λ_0/9.31	λ_0/16.48
E (Electric field)	7.4e^{05}V/m	1.4.e^{06}V/m
H(Magnetic field)	566 A/m	1658A/m
I(surface current)	564 A/m	1659A/m

The comparison of the characteristics of the C-Shaped SRR with the classic SRR falling into the same geometry form is presented in Table I. It is shown that the distribution of surface current of the proposed C-Shaped SRR at the lower frequencies is higher than the higher resonance frequencies of Classic SRR. As seen in the comparison table, the C-Shaped SRR not only occupies a smaller size in comparison to classic SRR but also offers a high distribution of, Electric field and magnetic field. This is due to the fact of two C-Shaped on SRR.

III. STUDY ON THE IMPROVEMENT OF THE RFID PATCH ANTENNA

A. Reference Patch Antenna

The design structure of the reference patch antenna is shown Fig. 5. It [17] is employed with *50 Ω* micro-trip line feeding technique on a Rogers RT 5880 dielectric substrate with a thickness of 1.6 mm, a permittivity ε = 2.2 and a tangential loss δ = 0.0009.The antenna is designed for RFID applications in the frequency band from 2.4 to 2.485 GHz. The overall size of the antenna is 45x38x1.6 mm2, it is calculated according to (5) and (6) [17].

Patch Width:

$$W = \frac{C}{2 f_r}\sqrt{\frac{2}{\varepsilon_r + 1}} \tag{5}$$

Patch Length:

$$L = \frac{C}{2 f_r \sqrt{\varepsilon_r}} - 2\Delta l \tag{6}$$

With:

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(\frac{1}{\sqrt{1 + \frac{12h}{W}}}\right) \tag{7}$$

$$\Delta l = 0.412h\left[\frac{\left(\varepsilon_{eff} + 0.3\right)\left(\frac{W}{h} + 0.264\right)}{\left(\varepsilon_{eff} + 0.258\right)\left(\frac{W}{h} + 0.8\right)}\right] \tag{8}$$

The physical parameters of the transmission line are determined using “LineCalc” in the Advanced Design System (ADS) software. The Geometrical parameters of the reference patch antenna are shown in Table II:

TABLE II. THE GEOMETRICAL PARAMETERS OF THE TRADITIONAL PATCH ANTENNA

Geometrical parameters	Index	Value (mm)
Patch Width	W	45
Patch Lenth	L	38
Feed line width	W_T	1
Feed line length	L_T	18
Feed line end length	L_F	5
Feed line end width	W_F	7.5

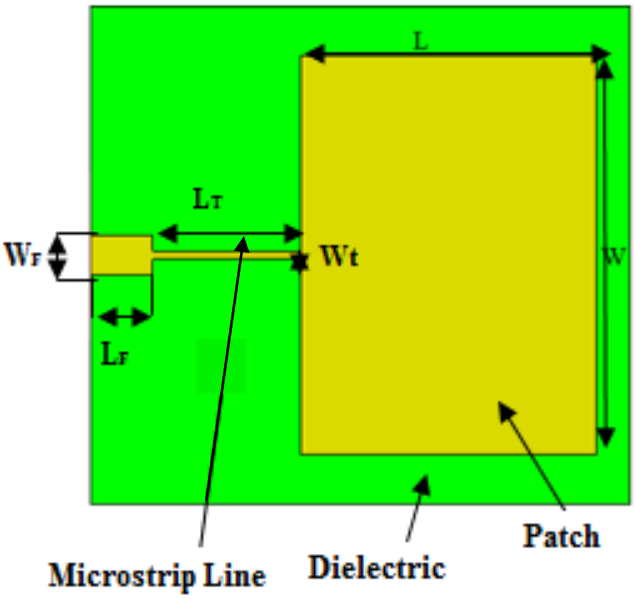


Fig. 5. Design of patch antenna reference.

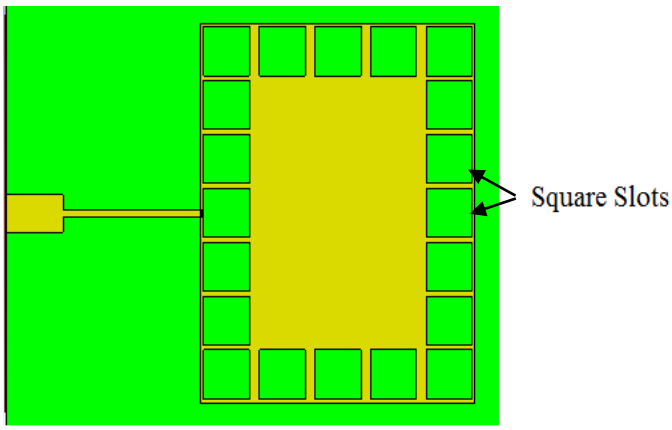
B. Inclusion of metamaterials in the environment of antenna

In RFID applications, numerous works have been integrated of metamaterials in the environment of the patch antenna in order to reduce of its size and improve of its performances. To bind an RFID antenna together with metamaterials, two techniques have been proposed. The first one is to integrate an array of complementary split ring resonators into the ground plane of patch antenna [16]. This technique is able to miniaturize the dimension of the antenna due to change in input impedance, but contrariwise the antenna gain is still a major problem. The achievable gain with this technique is low about 3 dBi at SHF Band, which has a negative influence on RFID performance such as read range. The other technique consists of placing the network of metamaterials around [15] or above [14] the patch antenna in order to improve the gain and bandwidth of the patch antenna, but this technique has some difficulties in antenna size. By the presence of metamaterials, the antenna becomes too big to attach to small devices. The main goal of the new technique is to increase the gain and miniaturize the dimension of patch antenna at the same time by using the proposed metamaterial.

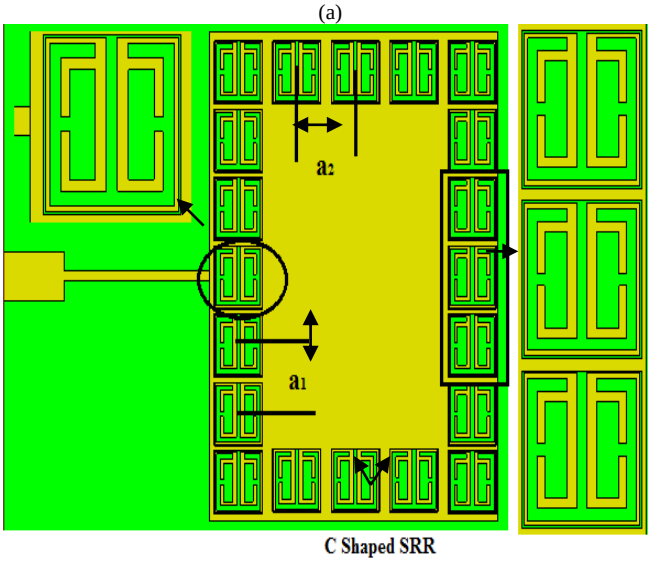
The concept of the new technique is based on the integration of an array of CSSR on the radiating surface of the antenna, the objective is to increase the distribution of surface-current, Electric field and Magnetic field on this surface. The antenna composed of two parts: the periodic square slot part and the CSRR load part. The SRR loaded part consists of twenty CSRRs which are engraved vertically and horizontally on square slots of the patch antenna where the size of each square slot is $7.2\times7.2mm^2$ (Fig 6(a)). The number of square slots is determined by the number of CSRR placed in the patch antenna. Moreover, to ensure a significant improvement in the performance of patch antenna, we must respect the constraint of periodicity during the integration of C-Shaped SRR in the antenna.

The network of C Shaped SRR must have a period $a < \frac{\lambda}{10}$

(a depends on the size of C-Shaped SRR) along the radiating surface of the antenna. For the reason that the dimension of patch width is not the same than that of patch length, the periodic C-shaped SRR has a vertically edge-to-edge spacing of a_1=7.5 mm (Fig. 6(b)) and horizontally edge-to-edge spacing equal to a_2=7.8mm (a_1 and a_2 are slightly lower than the wavelength) for maximizing the number of C-Shaped SRR Cells around the patch antenna and ensure better performance improvement.



(a)



(b)

Fig. 6. Design of Rectangular Patch antenna (a) with square slots (b) with C-shaped SRR Array.

C. Results and Discussions

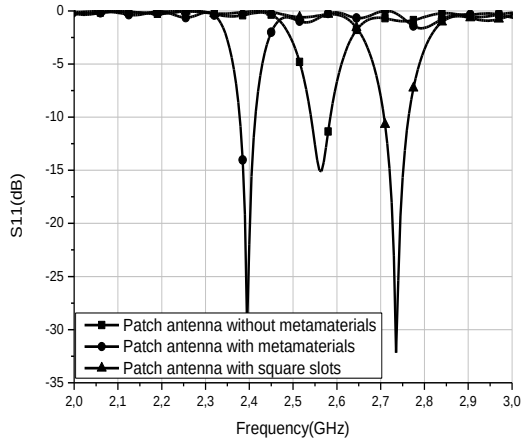


Fig. 7. Simulated S11 for the patch antenna without and with C-shaped SRR.

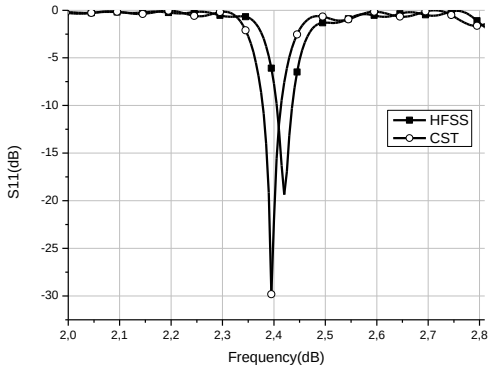


Fig. 8. Simulated S11 for the patch antenna with C-shaped SRR by HFSS and CST software.

Fig. 7 presents the simulated return loss in dB of the patch antenna in different cases: without metamaterials; with Square slots; with C-SRR. It can be observed that resonance frequency S11 is increased from 2.58 GHz to 2.74 GHz when we created a small square slots on the radiated part of patch antenna and downshifted from 2.74 GHz to 2.4 GHz with a better adaptation that reached -30dB when we integrated the proposed SRR on the slots by a degradation rate of 340 MHz (this result is validated by HFSS and CST software as can be seen in Fig. 8). This degradation is understood due to the strong coupling of metamaterials which produce an inductive and capacitive effect. So we can say that the resonant frequency S11 is inversely proportional to the number of C-shaped SRR cells.

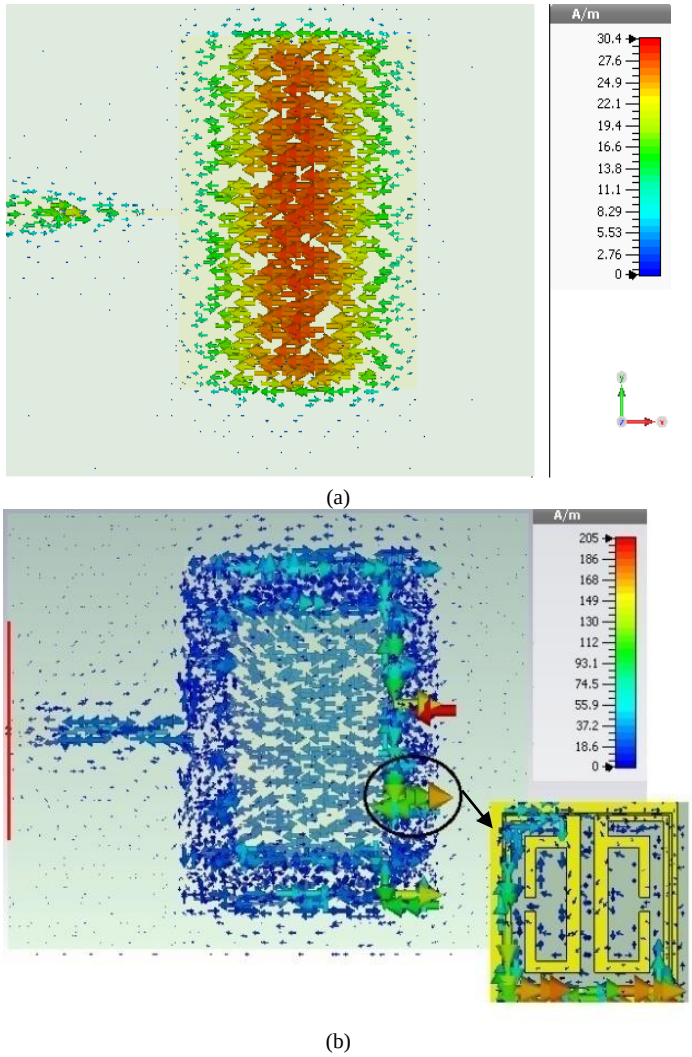
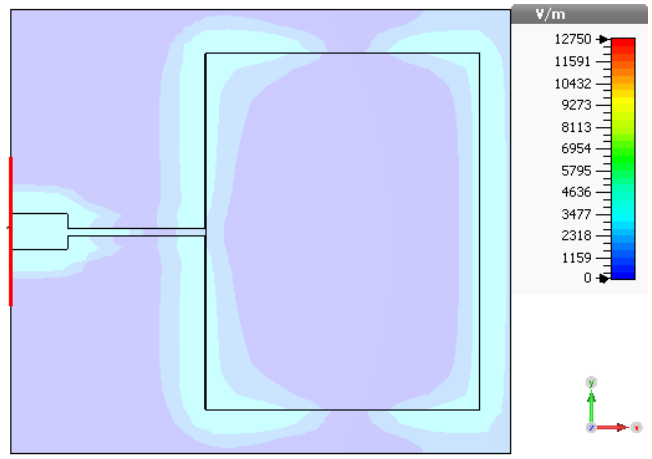
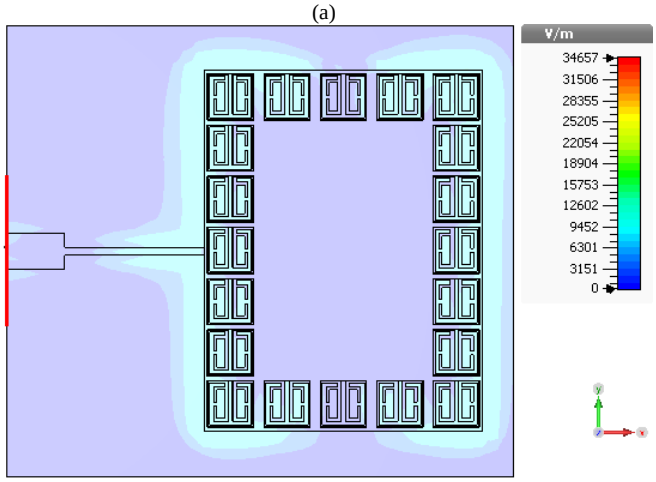


Fig. 9. Surface currents distribution in patch antenna (a) without C –Shaped SRR at 2.59 GHz (b) with C-shaped SRR Array at 2.4 GHz.

The simulated current distribution on the reference patch antenna is shown in Fig. 9(a). It is observed that the currents are asymmetrical and mainly distributed in the middle of the patch antenna. The surface current traveled through the reference patch antenna generates an Imax current of 30.4 A/m. By using the new technique on the patch antenna, the current is entering from the small square slots part and distributed in the CSRRs cell (Fig. 9(b)). The current density is changing rapidly at a radiating surface of the antenna at 2.4 GHz. It has led to an Imax current of 205 A/m, which is slightly higher to that current distribution of patch antenna without the use of C-SRR. So, we can conclude that the CSRRs part makes a vital contribution to the current movement.

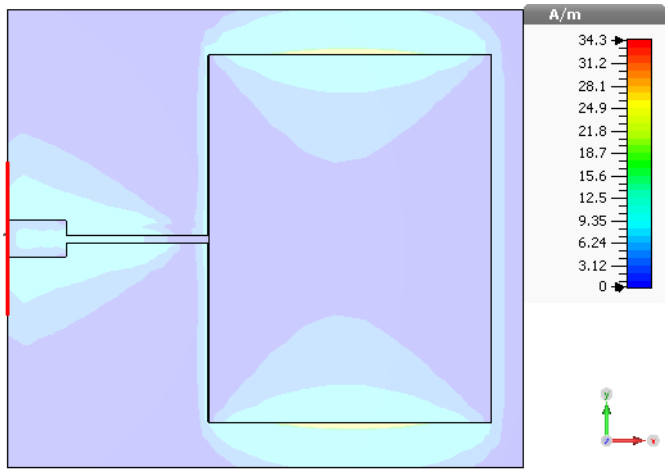


(a)

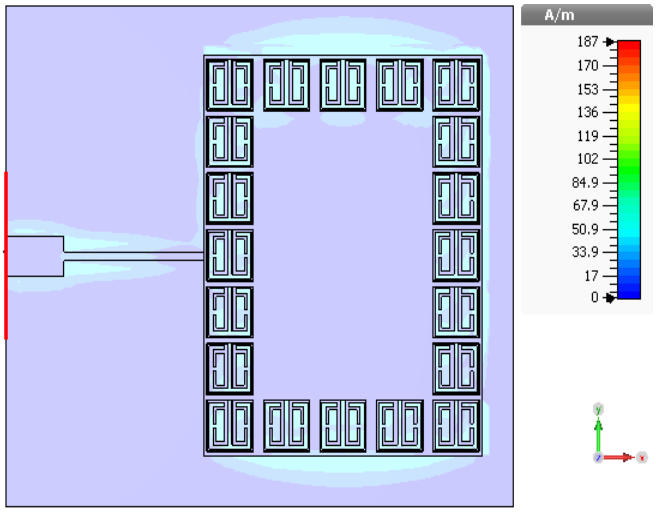


(b)

Fig. 10. Electric field distribution in patch antenna (a) without C –Shaped SRR at 2.59 GHz (b) with C-shaped SRR Array at 2.4 GHz.



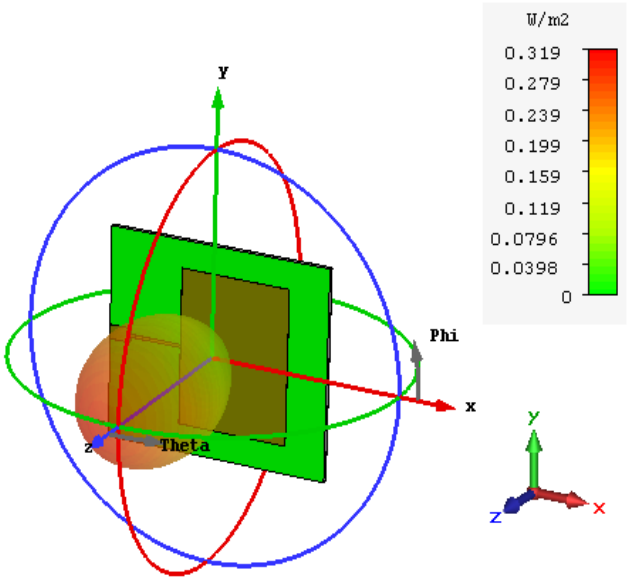
(a)



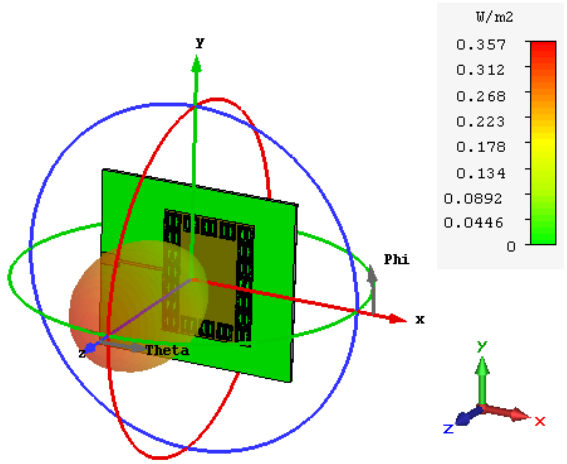
(b)

Fig. 11. Magnetic field distribution in rectangular patch antenna (a) without C –Shaped SRR at 2.59 GHz (b) with C-shaped SRR Array at 2.4 GHz.

Fig. 10 and 11 show the Peak values of the electric field and the magnetic field in the patch antenna with and without the use of C-shaped SRR at the frequency of 2.59 and 2.4 GHz. It is clear that at low frequencies (2.4 GHz) the Electric and magnetic fields are changed and the peak values are higher as compared with the peak values of reference patch antenna at 2.59 GHz. So, we can conclude that the proposed metamaterial is helping to increase the distribution of electric and magnetic fields of the patch antenna.



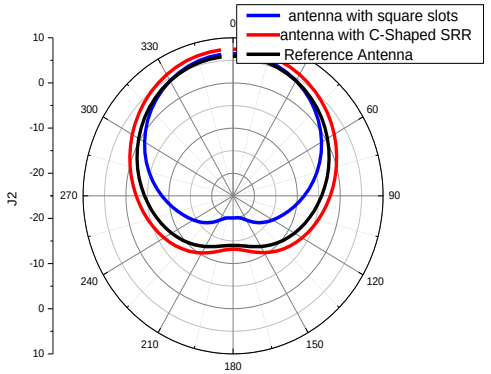
(a)



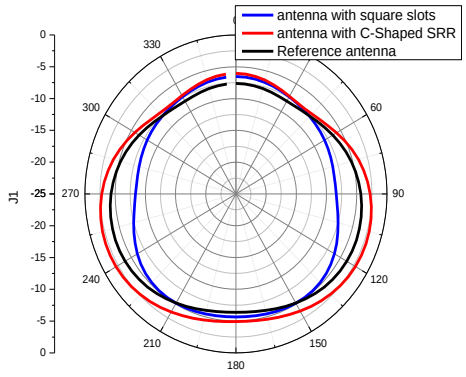
(b)

Fig. 12. 3D Simulated power pattern for patch antenna (a) without the use of C-SRR (b) with C-SRR.

Moreover, the increase in the electric field and the magnetic field acts directly on the total power of the antenna. Fig. 12 presents the radiated power field of the patch antenna with and without the use of C-Shaped split ring resonator. It can be seen that the metamaterials antenna produces a total emitted power field P equal to 0.357 W/m2 that is higher to that provided by the patch antenna alone.



(a)



(b)

Fig. 13. Simulation Radiation pattern of the patch antenna with and without C-Shaped SRR (a) E plane (b) H-plane.

Fig. 13 shows the simulated E-plane and H–plane radiation patterns of the antennas without metamaterials; with Square slots; with C-shaped SRR. It is observed that C Shaped SRR-loaded antenna presents gain of about 6.91 dB in the whole working band (2.39~2.43 GHz), which is about 1 dB more than the unloaded one. Consequently, the antenna efficiency in the use of C-Shaped SRR increases to 94% with a rate of improvement that exceeds 14% compared to the reference antenna (Fig. 14).

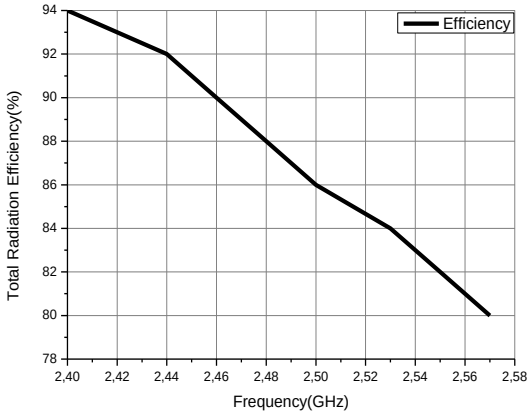


Fig. 14. Radiation efficiency curve of the proposed patch antenna.

The comparative table between the antenna patch without and with C-Shaped SRR is shown in Table III. From the results obtained, it shows that the C-Shaped SRR is successfully utilized to increase the gain and miniaturize the dimension of the RFID patch antenna that operates in the SHF band (2.4-2.485GHz).

TABLE III. COMPARATIVE RESULTS BETWEEN THE ANTENNA PATCH WITHOUT AND WITH C-SHAPED SRR

Characteristics of antennas	**Patch antenna without metamaterials**	**Patch antenna with C-shaped Array**
Resonance Frequency	2.57 GHz	2.4 GHz
S11	-15.5 dB	-29 dB
Electric field	12750V/m	34657V/m
Magnetic field	34.3 A/m	187A/m
Surface current	30.4 A/m	165A/m
Total Power	0.319W/m^2	0.357W/m^2
Gain	5.92 dB	6.91dB
Directivity	7.32dBi	7.34dBi
η	80%	94%
Size reduction	**	3%

D. Theoretical Demonstration

As we stated in the previous part (Results and discussions), the integration of a periodic network based on a new model of a split C-ring resonator on the square slots of radiating rectangular patch antenna increase the surface current distribution I_{max} and help to enhance the distribution of electric E_{max} and magnetic H_{max} fields which acts directly on

the total power of the antenna and consequently has a positive influence on the antenna frequency characteristics in term of gain; radiation efficiency; S11 adaptation.

Let us now discuss the theoretical study on the improvement of the RFID patch antenna. The currents I_0 and the charges Q_0 are the primary sources of the radiated electric and magnetic fields whose propagation of the wave is in function of $\frac{\exp(-i\beta r)}{r}$. Such as "r" presents the maximum distance of radiation and β indicates the phase constant that is defined by the following relation:

$$\beta = \frac{2\pi}{\lambda} \tag{9}$$

The imbalance that occurs in the distribution of the charges Q_0 negative / positive and the current I_0 traversing an antenna, allows the creation of the electric and magnetic fields delivered by the radiating element. As a result, we get electromagnetic waves propagated in space by forming a radiation pattern. The radiated electric E_{rad} and magnetic H_{rad} fields have a direct influence on the total radiated power density, which in turn results in this density being as high as possible because it correlates proportionally with the antenna performance.

The propagation of electric and magnetic fields in space can be determined by solving the following Maxwell equations [20]:

$$\begin{cases} div\vec{E} = \frac{\rho}{\varepsilon} & (10) \\ div\vec{B} = 0 & (11) \\ rot\vec{E} = -\mu\frac{d\vec{H}}{dt} & (12) \\ rot\vec{H} = \sigma\vec{E} + \varepsilon\frac{d\vec{E}}{dt} & (13) \end{cases}$$

So now, it is possible to determine any primary source characterized by a volume distribution of the current Iv and charge Qv at a point M by the vector potential $\vec{A}$ (14) and the scalar V (15) which are defined by the following relations [21], [22]:

$$\begin{cases} \vec{A}(M) = \frac{\mu}{4\pi}\int_V \vec{I}\ \frac{\exp(-i\gamma r)}{r}dv & (14) \\ V(M) = \frac{1}{4\pi\varepsilon}\int_V Q\ \frac{\exp(-i\gamma r)}{r}dv & (15) \end{cases}$$

According to these two parameters $\vec{A}$ and V, we can determine the radiated electric E_{rad} (15) and magnetic H_{rad} (16) fields and total emitted power density *p* (18) [18], [19].

$$\begin{cases} \vec{E}_{rad} = \frac{-d\vec{A}}{dt} - grad\vec{V} & (16) \\ \vec{H}_{rad} = \frac{1}{\mu}rot\vec{A} & (17) \end{cases}$$

$$\begin{cases} p(r,\theta,\varphi) = |E_{rad}| \times |H_{rad}| = \frac{|E_{rad}|^2}{Z_m} & (18) \\ Z_m = \frac{|E_{rad}|}{|H_{rad}|} \end{cases}$$

With Zm: Impedance of the medium propagating by the electromagnetic wave.

Finally, we can calculate the Radiation diagram in power *D* (Eq19) and the Gain in power *G* (Eq20) according to the expression of total radiated power density *p* using the following equations [18], [19]:

$$\begin{cases} D(\theta,\varphi) = \frac{P(\theta,\varphi)}{P_{\max}} = \frac{4\pi r^2 p(\theta,\varphi)}{P(\theta_0,\varphi_0)} & (19) \\ G(\theta,\varphi) = \frac{P(r,\theta,\varphi)}{P_{isotropie}} = \frac{4\pi r^2 p(r,\theta,\varphi)}{\int_0^{2\pi} d\varphi \int_0^{\pi} r^2 Sin(\theta_0) p(r,\theta_0,\varphi_0) d\theta} & (20) \end{cases}$$

IV. Comparative Study

In the final section, a comparison made between the proposed structure and others realized works in order to select the perfect technique. The reflection coefficients (S11) curves of the antennas corresponding to different types of loading structures are given in Fig. 15. It is possible to observe that by using the new technique the resonating frequency of the antenna shifts from 2.57 GHz to 2.4 GHz, thus obtaining a miniaturization of 3%, and an improvement with 1 dB compared to the reference antenna which means that the proposed technique does not only have the influence of the radiation pattern but also on the resonant frequency. From previous works [14]–[17], it is observed that better miniaturization has been achieved using the technique of implementing an array of complementary split ring resonators into the ground plane. This technique has a better compact size compared to the other existing structure by 30% size reduction, but the negative effect on the antenna performance. Similarly, in case of the antenna with CSRR, the reflection coefficients (S11) decrease and also shift towards the lower frequency with a reduction in antenna performance.

Fig. 16 shows a comparison of the radiation patterns of the antenna with and without metamaterials taken at 2.57 GHz, 2.58 GHz, 1.9 GHz, 2.32 GHZ and 2.4 GHz for both E plane and H plane. We can find that the strong gain in the whole working band is obtained by the technique of placing the substrate layer with Chiral Metamaterial (CM) above the antenna. The results demonstrate that 2 dB gain enhancement is achieved compared to the reference antenna, but

contrariwise the antenna size still a major problem. The frequency Characteristics comparison of the proposed antenna with the techniques used in [15]-[17] is presented in Table IV.

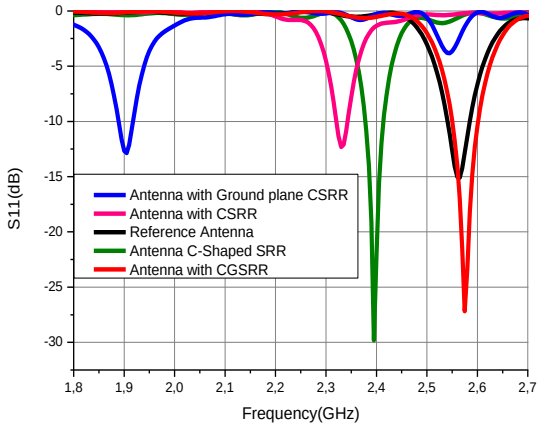


Fig. 15. Comparative analysis of S11.

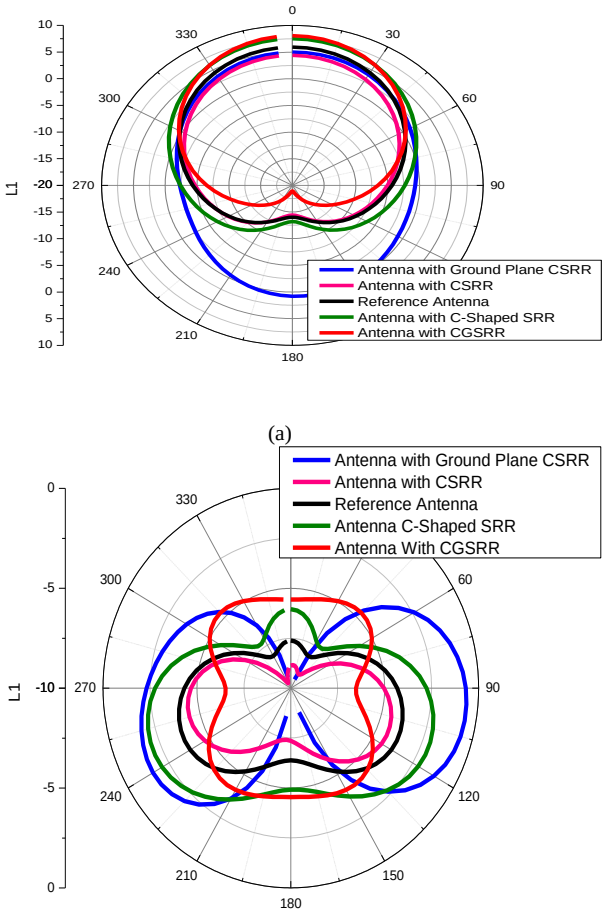


Fig. 16. Comparison of the radiation patterns of proposed antenna with the designs in [14-17] (a) E plane (b) H plane.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED ANTENNA WITH THE TECHNIQUES USED IN [15]-[17]

Antennas	**Fr GHz**	**S11 dB**	**Gain dB**	**Dir dBi**	**η %**	**Size reduction %**
Reference Antenna	2.57	-15.5	5.92	7.32	80	**
Ref [15]	2.58	-27	7.88	8.76	89	**
Ref [16]	1.90	-12.5	4.72	7.03	67	11.7
Ref [17]	2.32	-12.3	4.46	6.89	65	4.37
Proposed Antenna	2.4	-29	6.91	7.34	94	3

TABLE V. COMPARATIVE STUDY OF THE PROPOSED TECHNIQUE AND OTHER PREVIOUS WORKS

Antennas	**Miniaturization**	**Performance**
Antenna With GCSRR	- -	++
Antenna With Ground Plane CSRR	++	- -
Antenna with CSRR	+	- -
Antenna with C-Shaped SRR	+	+

Finally, it is noted from Table V that the proposed technique is the perfect approach because at the same time we can improve the performance and miniaturize the size of the patch antenna.

Based on this work several perspectives can be envisaged concerning the method of inclusion metamaterials in RFID antennas. It would be very interesting to realize multi-band antenna based on left-handed metamaterials more miniaturized and more efficient. Identically, another interesting perspective can be created by obtaining antenna with good frequency characteristics and with a circular polarization based on metamaterials.

V. CONCLUSION

This paper presents a new method to implement C Shaped SRR on the conductive surface of the rectangular patch antenna. This method is proposed to enhance the gain in the operating frequency and decrease the size of the patch antenna. In comparison with the rectangular patch antenna without metamaterials, the proposed method able to increase the antenna in terms of gain. Therefore the efficacy η and return loss of the patch antenna are improved.

ACKNOWLEDGMENT

The authors would like to thank the team Unit of Research Circuits and Electronic Systems High Frequency of the faculty of Science Tunisia for supporting this project.

REFERENCES

[1] D.Richards, H.Saunders and St.George, “Royal Air Force”, HM Stationery office, vol 2, pp. 1939-1945, November 1953.

[2] B.Violino and M.Roberti,”The history of RFID Technology”, RFID Journal, pp.1338, Jan 2005.

[3] A.Juels "RFID Security and Privacy: A Research Survey", IEEE Journal on Selected Areas in Communications, Vol. 24, Issue 2, pp. 381-394, February 2006.

[4] M.R.Rieback, B.Crispo and A.S. Tanenbaum “The Evolution of RFID Security,” IEEE Pervasive Computing, Vol. 5, No. 1, pp. 62-69, JANUARY–MARCH 2006.

[5] K.V.Seshagiri, Rao, P.V. Nikitin, and S.F. Lam, "Antenna Design for UHF RFID Tags: A Review and a Practical Application”, IEEE Transactions on Antennas and Propagation, Vol. 53, N°12, December 2005.

[6] L.Yang, A.Rida, R.Vyas, and M.M. Tentzeris, "RFID Tag and RF Structures on a Paper Substrate Using Inkjet-Printing Technology”, IEEE Transactions on Microwave Theory and Techniques, Vol. 55, N°12, pp. 2894-2901, December 2007.

[7] V.G Veselago, "Electrodynamics of substances with simultaneously negative electrical and magnetic properties", Sov, Phys, vol. 10, pp. 509-517, February 1968.

[8] J.B.Pendry, A.J.Holden, W.J.Stewart and I. Youngs, "Extremely low-frequency plasmons in metallic meso structures", Phys. Rev. Lett, vol. 76, pp. 4773-4776, June 1996.

[9] J. B. Pendry, A. J. Holden, D.J. 1.6 mm, and W.J. Stewart, "Magnetism from conductors and enhanced nonlinear phenomena", IEEE Trans. Microwave Theory Tech, vol. 47,pp.2075-2084, Nov 1999.

[10] S.Naoui, L.Latrach, and A.Gharsallah, “Electrical Modelling of Split Ring Resonators Operate in the UHF band”, International Conference on Computer Technologies Innovations & Applications, Tunisia, June 14-16, 2014.

[11] R.A Shelby, D. R. Smith, S. Shultz, “Experimental verification of a negative index of refraction”, Science, vol. 292, pp. 77-79, Apr. 2001.

[12] D. R. Smith, “The reality of negative refraction,” Phys.World, vol. 16, pp. 23–24, June 2003.

[13] D. R. Smith, J. B. Pendry, and M. C. K. Wiltshire, “Metamaterials and negative refractive index,” Science, vol. 305, pp. 788–792, Aug 2004.

[14] Y.Liu and all, "Investigation of Circularly Polarized Patch Antenna With Chiral Metamaterial", IEEE Antennas and Wireless Propagation Letters, Vol. 12, pp. 1359-1362, June 2013.

[15] M.M.Bakry, A.B. Abdel-Rahman and H.F. A. Hamed, “Metamaterials Gain and Bandwidth Improvement of Microstrip Patch Antenna using Complementary G-shape Split Ring Resonator”, 31st National Radio Science Conference, (NRSC2014), Egypt, pp.35-40, April 28-30, 2014.

[16] M.Elsdon and O.Yurduseven, “Direct-Fed Reduced Size Patch Antenna Using Array of CSRR in the Ground Plane”, Microwave and Optical Technology Letters, Vol. 57, No. 7, pp.1526-1529, July 2015.

[17] S.Naoui, L.Latrach, and A.Gharsallah, "Metamaterials Microstrip Patch Antenna For Wireless Communication RFID Technology", Microwave and Optical Technology Letters, Vol. 57, No. 5, pp 1060-1066, May 2015.

[18] Constantine A. Balanis, "Antenna Theory: Analysis and Design", 2nd edn, John Wiley & Sons, Ltd, Chichester, ISBN:0-471-59268-4,858 pages, 1996.

[19] Constantine A. Balanis, "Antenna Theory: Analysis and Design", 3rd edn, John Wiley & Sons, Ltd, Chichester, ISBN: 978-0-471-66782-7, 11-36, May 2005.

[20] Christophe Caloz and Tatsuo Itoh, “Electromagnetic Metamaterials: Transmission line theory and microwave applications”, John Wiley & Sons, Inc. ISBN-10: 0-471-66985-7, 2006

[21] Akhlesh Lakhtakia, Vasundara V. Varadan, and Vijay K. Varadan,” Field equations, Huygens’s principle, integral equations, and theorems for radiation and scattering of electromagnetic waves in isotropic chiral media”, JOSA A, Vol. 5, Issue 2, pp. 175-184,1988.

[22] David A. B. Millier,"Huygens's Wave Propagation Principle Corrected", Optics Letters, Vol.16, No.18, September 15, 1991.

Impact of Web 2.0 on Digital Divide in AJ&K Pakistan

Sana Shokat[1], Rabia Riaz[2],
Department of CS & IT, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan

Samaira Aziz[5]
Department of Mathematics, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan

Sanam Shahla Rizvi[3]
Department of Computer Sciences, Preston University, 15, Banglore Town, Shahrah-e-Faisal, Karachi, 75350, Pakistan

Raja Shoaib Hussain[6], Mohaib Zulfiqar Abbasi[7]
Department of Health Sciences, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan

Farina Riaz[4]
Independent Researcher

Saba Shabir[8]
Kashmir Institute of Economics, University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan

***Abstract*—Digital divide is normally measured in terms of gap between those who can efficiently use new technological tools, such as internet, and those who cannot. It was also hypothesized that web 2.0 tools motivate people to use technology i.e. social networking sites can play an important role in bridging digital gap. The study was conducted to determine the presence of digital divide in urban and rural areas of district Muzaffrabad, Azad Jammu & Kashmir. A cross-sectional community based survey was conducted involving 384 respondents from city Muzaffrabad and village Garhi Doppta. The existence of digital divide was assessed on the basis of the questionnaires given. Chi-square test was applied to find the association of different demographic and ICT related factors with internet usage. Despite the growing awareness there are possibilities of gender, age and area based digital divide. Outcomes of the survey affirmed that web 2.0 based web-sites are becoming popular and attracting people to use internet.**

***Keywords*—AJ&K; digital divide; ICT; web2.0; social networking; social inclusion and cohesion enabling approaches; net-living life styling personalization and optimization; subjective human and social factors; well-being through net living**

I. Introduction

Digital divide is defined as two interdependent gaps; gap between skills of people who use technology and gap caused by access to the technological tools [1]. It is very difficult to use the technology without its proper understanding. In [2] digital divide is defined as the disproportion among; the internet users, poor and the rich people, internet access in the developing countries and gender discrimination. Access, awareness, attitude and application is the 4A viewpoint given by [3], which emphasis on digital gaps at local as well as global level. Other factors that influence access to internet and communication technologies are income and gender. Thus, digital divide is not only limited to the hardware use of technology, but also to the use of technological software's and internet [4].

Pakistan is among the developing nations of the world. Cities like Karachi, Lahore and Islamabad are competing the world in technological advancement. Mubashir Akram, a Pakistani political communication specialist, indicated that Pakistan is having a total population of 180 million, among them only 36% population is living in cities and from the urban residents only 16% are using internet [5].

Azad Jammu and Kashmir (AJ&K) is a self-governing state under the constitution of Pakistan. Head of the state is President. It is located in the north of Pakistan and has a mountainous terrain. Pothwari is the most commonly spoken language. Azad Kashmir is famous for its beauty all over the world and it is full of natural resources. Kashmir is also called heaven on earth and tourism is one of its main industries. Azad Kashmir suffered from earthquake of 7.6 magnitudes in 2005 which totally destroyed its infrastructure.

Internet has become fundamental for performing everyday tasks from all aspects of human life. But unfortunately, it is found that senior citizens of Asia are using internet rarely, which can contribute to their social segregation. Pew's finding in Asia reported that only 25% older people in china, 17% in Malaysia, 14% in Thailand, 7% in India, 5% in Pakistan and 2% in Bangladesh own smart phones [6].

People in AJ&K have limited internet facilities and resources. In-spite of the literacy rate of 60% in AJ&K [7], little is known about prevalence of ICT (Information and Communication Technology) usage in its community. No study has been made so far to identify the presence of digital divide in AJ&K. The objective of the present study is to estimate the prevalence of digital divide in rural and urban population of AJ&K by taking "Weekly Internet Usage" as the key factor. Because "gap in internet usage" is a main contributor in digital divide and has been the subject of many scholarly debates.

Paper is divided into following sections. Section II introduces the different types of digital divides. Research questions and hypothesis are highlighted in Section III. Methodology used to evaluate the presence of digital divide is presented in Section IV where research finding and Internet usage along with Web 2.0 are discussed in detail in Section V and Section VI, respectively. We analyze association between hypotheses and findings in Section VII. Recommendations to reduce the digital divide in targeted community are proposed in Section VIII. Finally, the paper is concluded and future work is highlighted in Section IX.

II. Related Studies on Digital Divide

A. Age based Digital Divide

Internet has become an integral part of our lives. Internet prevalence have reached more than 80% in Germany, South Korea, USA and Switzerland and it has crossed 90% in Scandinavian countries. This ubiquity of internet access is due to socio-demographic dimensions such as income, education, gender and age [8].

Age based digital divide is categorized in first level of digital divide. According to [8]-[9] more than 95% youth in United States and Switzerland are taking technology benefits. From a survey conducted in Switzerland during the time period 1997-2016 it is quite clear that people of age 70+ are less frequent internet users and they need extra consideration.

B. Gender based Digital Divide

One of the major problems faced by women living in developing countries is gender discrimination [9]. This affects their access to internet and communication technologies. Most of females are also unemployed [3]. 25% women in Middle East countries [10] and 45% women's in Sub-Saharan Africa are lacking internet access [11] due to the reason that they are more indulged in accomplishing their household tasks [12]. UN data statistics showed that 60% women's are unemployed [13]. In [14] segregation from technology awareness, supporting men in comparison to women and economic issues are shown as the main obstacles in the way to technology usage.

C. Socio Economics Digital Divide

Open ended questions were asked to collect qualitative data in a study carried out in Oyo state and Yewa about exploring the impact of digital divide on computer usage and internet access.. Their results showed that affordability and lack of internet access are the main reasons behind digital divide [26]. Another research carried out in Australia focused on the existence of digital divide in young generation. Outcomes of the research indicated that young people are using more technological tools like mobiles, handheld devices, etc. it may affect the economy, academia and health [27]. In a report presented by OECD for exploring the existence of digital divide, government officials identified that electronic commerce can be improved by increasing the technological use [28]. OECD countries are motivating investors to increase the demand for communication technologies for economic growth. They intend to decrease the digital divide by individual capacity building and by improving the technological setups.

D. Area based Digital Divide

Age and gender are not the only glitches that are producing digital cavity but one of the most imperative elements is locality of an individual. A number of reasons were underlined in a study carried out in US and all over the world, problems identified are; less frequent use of technological tools in class rooms, unavailability of technology, inadequate bandwidth to attend online programs [15], [16]. If appropriate technological tools are provided to the teachers they can make a big revolution in bridging the gap [17].

E. Other Digital Divides

Trend of using social networking sites are growing rapidly. Social networking is not only a mean of communication between different people it also plays a significant role in establishing a geographically close romantic relationship [18], [19]. Another study was carried out to elaborate the problems that teachers face in order to cope with the technological changes in the society. They identified that teachers were unable to choose suitable technological tools for education [20]. Furthermore over population, use of English language, ecological factors and transportation are the main reasons of digital illiteracy in addition to problems like budget, resources, social norms and mentality of societies [21].

III. Research Questions and Hypothesis

On the basis of the background studies following research questions are formulated.

1) Are their multiple user groups based on age, area, computer literacy, gender, etc.?

2) What proportion of people are more frequent internet users?

3) How web 2.0 applications influence people towards internet use?

4) What is the internet access level of the people living in rural and urban areas?

5) Which social factors restrict people from using internet?

6) Which economical and other factors limit internet use?

Based on research questions we have tested the following hypothesis in urban and rural areas of district Muzaffrabad.

H1: Young generation is more inclined towards technology than older people.

H2: Gender difference between computer usages is increasing.

H3: Web 2.0 sites have attracted peoples towards using computer technology.

H4: Cities have better access to technology.

H5: Language is a factor that influences digital divide.

H6: There is association between demographic aspects and ICT factors.

IV. METHODOLOGY

The entire study was done in order to evaluate the presence of digital divide in Muzaffrabad city and village Garhi Doppta. Total population of district Muzaffrabad and garhi dopptta is approximately 0.8 Million. This study involved cross sectional data collection through a detailed open ended questionnaire based survey of 384 respondents in the area under consideration. Data was collected using systematic random sampling technique. However, first we used cluster sampling for residents of neighborhoods, blocks, and housing structures. Then within each selected area, we used systematic random sampling where each unit in the population was identified, and each unit had an equal chance of being in the sample. To select the sample of 600 houses, we picked every 18th house to cover large area of population. Instrument for data collection, in the form of a structured questionnaire, was designed to elicit information about the use of internet and web 2.0 sites. A printed questionnaire was distributed among the participants manually. Group of 10 volunteers helped in conducting the survey; they were well trained before. All the questions were explained in native language to participants who had problem understanding English. Interviews were also conducted where required. Initially the survey was given to 300 people from whom 174 responded back. Survey questionnaire were again distributed manually among 300 more people out of whom 210 returned the completely filled survey. Remaining 90 questionnaires were either incomplete or never received. So the total response rate was 64%. Every time, during survey distribution, questionnaire was explained completely. Problems faced by the participants like unawareness about the web 2.0 technologies were resolved by the volunteers. Survey consisted of two sections. First section collected demographic information and the second collected information regarding the usage of web 2.0 based websites. It took two months to complete the survey.

In order to check the reliability and consistency of questionnaire, Cronbach Alpha test was used [22], [23]. The reported value of Cronbach Alpha turned out to be 0.65, implying that questionnaire used for the study was reliable and consistent.

This survey facilitated in identifying major difficulties faced by the community while using ICT. Total 600 questionnaires were disseminated in Muzaffarabad city and village Garhi Duppatta while 384 people responded positively. Problems faced by the people were discussed and resolved during survey. The data was analyzed using SPSS version 14.0 [24].

Descriptive statistics was done using mean ± SD for continuous variables and frequencies with percentage for categorical variables. The Chi-square test [25] was applied to find the association of demographical and ICT factors with digital divide using (1). Chi-Square is the sum of the squared difference between observed (o_i) and the expected (e_i) data, divided by e_i. Furthermore 95% CI was calculated using (2).

$$\chi^2 = \sum_{i}^{n} (o_i - e_i)^2 / e_i \quad (1)$$

$$\hat{P} \pm 1.96\sqrt{\hat{P}(1-\hat{P})/n} \quad (2)$$

V. RESEARCH FINDINGS

The distribution of socio-demographic characteristics of the study population is shown in Table I. Survey was distributed random among respondents of various age ranges and gender evenly among city and rural population. Rural to urban ratio in AJ&K is 88:12. This fact is depicted in our survey results as 54.69% respondents belong to rural areas. 80% of the respondents are either employees or students thus we have 48.1% respondents who have mediocre computer knowledge, 37.7% novice and 14% as expert users. Literacy rate in AJ&K is very high as compare to rest of the country. 56.7% of the population under study was English literal, 27% were comfortable with English to some extent.

Tables II and III shows association of weekly internet usage with demographical and ICT factors respectively. These tables indicate various aspects of the digital divide in the region. In Table II, the association between demographical and weekly internet usage, the factor “gender” showed non-significance; the p-value is greater than level of significance (0.05). While all the remaining variables that are area, age group, computer knowledge, professional status and English literacy are highly associated with weekly internet usage. Thus our results, based on responses of various age groups, indicate that gender difference among computer users are decreasing. Another observation seen is that more females of age group 16-25 and 26-40 are involved in using computers than males. But in 40 onwards age group, number of male computer users are doubled than the females, as evident in Table IV. This shows that previously fewer females were involved in using the computer but now-a-days this gap is reducing. To find out the most significant ICT related factors that can cause digital divide in the area, Chi-square test was performed to check the association of different ICT factors with weekly internet usage. This confirms the validity of hypothesis of the study. Internet cost, computer access and mobile access all contribute in prevalence of digital divide.

Tables V and VI show the problems faced by participants while using internet and their purpose of internet usage respectively. It was observed that internet was most commonly used for emails, social networking sites as well as academics and research. Internet speed was the major problem faced by the participants.

TABLE I. DEMOGRAPHICAL CHARACTERISTICS OF THE STUDY COHORT

Variable	Number	(%age)	95% CI
Area			
Rural	210	54.69%	(49.70-59.66)
Urban	174	45.31%	(40.33-50.33)
Gender			
Female	197	51.30%	(46.30-56.30)
Male	187	48.69%	(43.69-53.69)
Age Range			
16-25	127	33.07%	(28.36-37.36)
26-40	129	33.59%	(28.86-38.86)
40 onwards	128	33.33%	(28.61-38.04)
Computer Knowledge			
Novice	145	37.76%	(32.91-42.60)
Mediocre	185	48.18%	(43.17-53.17)
Expert	54	14.06%	(10.58-17.53)
Professional Status			
Employee	167	43.48%	(38.53-48.44)
Student	140	36.46%	(31.64-41.27)
House wife	47	12.24%	(8.96-15.51)
none of these	30	7.81%	(5.12-10.49)
English Literacy			
Yes	218	56.77%	(51.81-61.72)
NO	62	16.14%	(12.46-19.82)
To Some extent	104	27.08%	(22.63-31.52)

TABLE II. ASSOCIATION OF WEEKLY INTERNET USAGE WITH DEMOGRAPHICS PARAMETERS OF THE STUDY COHORT

Variables		Weekly Internet Usage						
		0 hrs	*1-4 hrs*	*5-10 hrs*	*More*	*Chi Square*	*df*	*Sig*
Area	***Rural***	80	69	40	21	23.48	3	0.000
	%	20.83	17.97	10.42	5.47			
	Urban	33	54	51	36			
	%	8.59	14.06	13.28	9.38			
Gender	***Female***	64	65	48	20	7.479	3	0.580
	% of total	16.67	16.93	12.50	5.21			
	Male	49	58	43	37			
	% of total	12.76	15.10	11.20	9.64			
Age Group	***16-25***	28	39	32	28	20.873	6	0.002
	% of total	7.29	10.16	8.33	7.29			
	26-40	31	46	35	17			
	% of total	8.07	11.98	9.11	4.43			

	40 onwards	54	38	24	12			
	% of total	14.06	9.90	6.25	3.13			
Computer Education	*Novice*	84	39	12	10	156.139	6	0.000
	% of total	21.88	10.16	3.13	2.60			
	Mediocre	28	77	60	20			
	% of total	7.29	20.05	15.63	5.21			
	Expert	1	7	19	27			
	% of total	0.26	1.82	4.95	7.03			
Professional Status	*Employee*	31	60	51	25	62.184	9	0.000
	% of total	8.07	15.63	13.28	6.51			
	Student	33	48	30	29			
	% of total	8.59	12.50	7.81	7.55			
	House wife	30	7	8	2			
	% of total	7.81	1.82	2.08	0.52			
	None of these	19	8	2	1			
	% of total	4.95	2.08	0.52	0.26			
English Literacy	*Yes*	16	80	69	53	138.83	6	0.000
	% of total	4.17	20.83	17.97	13.80			
	NO	44	11	7	0			
	% of total	11.46	2.86	1.82	0.00			
	To Some extent	53	32	15	4			
	% of total	13.80	8.33	3.91	1.04			

TABLE III. ASSOCIATION OF WEEKLY INTERNET USAGE WITH AVAILABILITY/ACCESS TO ICT FACILITIES

Variables		**Weekly Internet Usage**						
		0 hrs	***1-4 hrs***	***5-10 hrs***	***More***	***Chi square***	***df***	***Sig***
Internet Cost	*1000*	5	80	39	15	277.063	9	0.000
	%	1.30	20.83	10.16	3.91			
	2000	8	23	43	24			
	%	2.08	5.99	11.20	6.25			
	More	12	7	6	17			
	%	3.13	1.82	1.56	4.43			
	< than 1000	88	13	3	1			
	%	22.92	3.39	0.78	0.26			
Computer Available at home	*Yes*	48	100	90	56	114.939	3	0.000
	%	12.5	26.0	23.4	14.6			
	No	65	23	1	1			
	%	16.93	5.99	0.26	0.26			
Availability of Mobile Phone	*Yes*	89	112	87	56	22.588	3	0.000
	%	23.18	29.17	22.66	14.58			

	No	24	11	4	1			
	%	6.3	2.9	1.0	0.3			
If internet is provided at low cost number of users will increase	***Strongly Agree***	18	44	58	44	88.695	9	0.000
	%	4.69	11.46	15.10	11.46			
	Agree	76	70	32	12			
	%	19.79	18.23	8.33	3.13			
	Disagree	14	8	1	0			
	%	3.65	2.08	0.26	0.00			
	Strongly Disagree	5	1	0	1			
	%	1.30	0.26	0.00	0.26			
Internet is available at schools or offices	***Yes***	18	44	58	44	137.131	6	0.000
	%	4.69	11.46	15.10	11.46			
	No	76	70	32	12			
	%	19.79	18.23	8.33	3.13			
	Some times	14	8	1	0			
	%	3.65	2.08	0.26	0.00			

TABLE IV. AGE AND GENDER BASED COMPUTER EXPERTISE

Computer Expertise	**Age Group**		
	16-25	***25-40***	***40 onwards***
Novice (F)	8.63%	9.14%	15.23%
Mediocre (F)	21.83%	20.81%	17.26%
Expert (F)	2.54%	2.54%	2.03%
Novice (M)	12.83%	13.37%	16.58%
Mediocre (M)	14.97%	11.23%	9.63%
Expert (M)	5.35%	10.16%	5.88%

TABLE V. PROBLEMS FACED USING INTERNET

Speed	**Availability**	**Computer Expertise**	**Busy Schedule**
61.61%	34.82%	19.64%	22.32%

TABLE VI. PURPOSE OF USING INTERNET

Job Search	**Email**	**Chatting**	**Surfing**	**Banking**	**Submitting Bills**	**Academics and Research**	**Social Networking**	**Other**
29.46%	60.71%	39.29%	26.79%	3.57%	3%	55%	48%	14%

VI. INTERNET USAGE AND WEB 2.0

Most frequent use of internet was found for emails and social networking sites. As these help people to stay connected with their work as well as families (Table V). In our survey, we also inquired about various features of Web 2.0 that are used most by the participants. Also they were asked to mention if these features have helped them in becoming computer fluent. Their responses depicts that Facebook is the famous social networking site among both the females and males with total 32% of respondents using only this site. YouTube was second on this ranking with 7% users. 44% of participants used more than one web 2.0 based website as shown in Fig. 1.

Fig. 2 shows that 70% females and 71% males agree on the theory that social networking sites have helped them in becoming a fluent computer and internet user. Results clearly indicate that web 2.0 based sites play a major role in bringing people towards the use of ICT that not only reduces the digital gap but also results in productivity and improves overall economic growth and development.

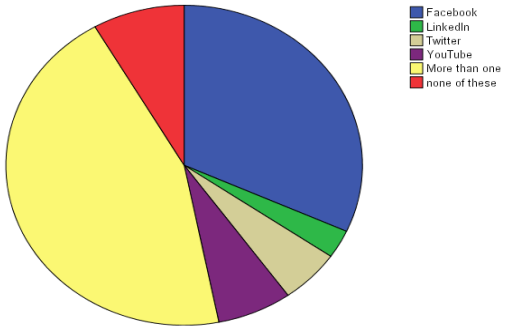


Fig. 1. Usage pattern of web 2.0 applications.

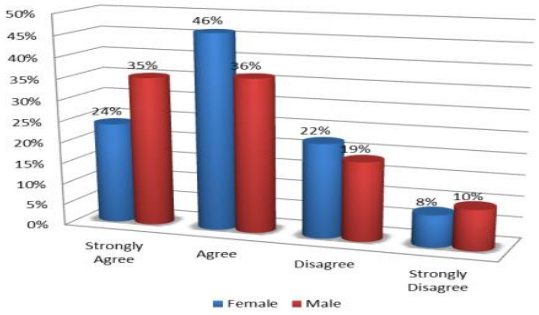


Fig. 2. Social networking sites vs. fluent internet user.

VII. Analysis

The first hypothesis "young generation is more inclined towards technology than older people" has been supported by the survey results. Results also indicate that gender difference among computer users are decreasing day by day, so second hypothesis is not supported by the survey results. Third hypothesis;" web 2.0 sites have attracted peoples towards using computer technology" is proven by the research conducted as most of the young participants are using internet to use Facebook and social networking sites. Fourth hypothesis, "cities have better access to technology" is also backed by survey results as people living in cities have better access to technology as compare to the rural areas. The fifth hypothesis; "language is a factor that influences digital divide" is not supported in the targeted community. Other results revealed that usability problems in website like Facebook and LinkedIn also influences the digital divide. If the websites are user friendly than more people will feel comfortable in using these websites which in turn will help in reducing the digital divide. The overall results provide a prominent association between demographic aspects and ICT factors.

VIII. Recommendations

The study commenced is a preliminary work done to investigate the attitude of the people towards internet usage. Subsequent steps were taken to involve more people into the study to enhance the quality of the collected data. On the basis of the collected data, to reduce the digital divide in the targeted community following recommendations are made; proper infrastructure should be provided at low cost in all rural areas, to introduce the benefits of using technology proper awareness sessions should be given to the community, sufficient number of computers should be provided in schools and offices, proper training of using different technologies should be provided, speed of the internet should be increased, focus the Digital Literacy from basic education level, internet cost reduction, promotion of online banking and online shopping can bring people closer to ICT. The positive aspects of social networking sites should be highlighted to avoid any kind of miss use of these services

IX. Conclusion and Future Work

Digital divide not only has a great impact on economic development but it also emphasis on development of a community. This research is conducted in the major districts of AJ&K. AJ&K is considered as underdeveloped and hard area. The main reasons of digital disparity in the targeted community are age and lack of infrastructure in rural areas.

Results showed that more females are getting technology oriented, gender based digital divide has decreased, youth is more inclined towards using technology as compare to their old age counter parts. Language is not a big barrier as English is being taught from early schools, computers are now becoming the essential parts of each house hold, and internet is available at schools and offices. However, internet cost for home usage is still high and if the cost of the internet is reduced internet users will increase.

The main purpose for internet usage is social networking, email and chatting. Main problems faced by internet users include speed and availability especially in rural areas. Web 2.0 sites now encouraged people to use internet. Proper training to the people will increase the use of technology.

This work can be enhanced by targeting visually impaired and handicapped audience.

References

[1] Kularski, C.; Moller, S. The digital divide as a continuation of traditional systems of inequality. *Sociology*, **2012**, *5151*, 1-23.

[2] Norris, P. Digital divide: Civic engagement, information poverty, and the internet worldwide. Cambridge University Press, Chicago, 2001.

[3] Unites Nations (UN) e-Government Survey 2016. Available online: http://unpan3.un.org/egovkb/Portals/egovkb/Documents/un/2016-Survey/unpan048065.pdf (accessed on 04 May 2017).

[4] Fuchs, C.; Horak, E. Africa and the digital divide. *Telematics and Informatics*, 25(2), **2008**, 99-116.

[5] Mundi, I. Index Mundi. Available online: http://www.indexmundi.com/pakistan/demographics_profile.html (accessed on 04 May 2017).

[6] Digital Divide: The unconnected Senior Citizens of Asia. Available online: http://propakistani.pk/2015/05/25/digital-divide-the-unconnected-senior-citizens-of-asia/ (accessed on 04 March 2017).

[7] G.O. AJ&K, IT Policy Strategy. Available online: http://itb.ajk.gov.pk/index.php?option=com_content&view=article&id=25&Itemid=6 (accessed on 04 March 2017).

[8] ITU Individuals using the internet 2013–2016. Available online: http://www.itu.int/ITU-D/ict/statistics/ (accessed on 14 March 2017).

[9] Thomas, N.F. The digital divide has grown old: Determinants of a digital divide among seniors. *New Media & Society*, **2014**, *18(2)*.

[10] Madden, M.; Lenhart, A.; Duggan, M.; Cortesi, S.; Gasser, U. *Teens and technology 2013*. Washington, DC: Pew Internet & American Life Project, 2013.

[11] Antonio, A.; Tuffley, D. The gender digital divide in developing countries. *Future Internet*, **2014**, *6*(4), 673-687.

[12] Hilbert, M. Digital Gender Divide or Technologically Empowered Women in Developing Countries? A Typical Case of Lies, Damned

Lies, and Statistics. Proc. of the Women's Studies International Forum, 2011, 34(6), 479-489. Pergamon.

[13] Unites Nations (UN) e-Government Survey 2012. Available online: http://unpan3.un.org/egovkb/Portals/egovkb/Documents/un/2012-Survey/unpan048065.pdf (accessed on 04 March 2017).

[14] Telecentre women: Digital literacy campaign 2016. Available online: http://www.telecentre.org/programs/women/ (accessed on 04 March 2017).

[15] Dugdale, A. One fifth of women in the developing world think Internet use is inappropriate for them 2013. Available online: http://www.fastcompany.com/3004797/one-fifth-women-developingworld- countries-think-internet-use-inappropriate-them (accessed on 04 March 2017).

[16] Gill, K.; Brooks, K.; McDougall, J.; Patel, P.; Kes, A. Bridging the gender divide. How technology can advance women economically. *International Centre for Research on Women*, 2010.

[17] Broadley, T. Digital revolution or digital divide: Will rural teachers get a piece of the professional development pie?. *Education in Rural Australia*, **2010**, 20(2), 63–76.

[18] Muruli, B.K.S. Dimensions of digital divide and digital opportunities: An indian scenairo. *Asia Pacific Journal of Library and Information Science*, **2012,** 55-68.

[19] Joy, B. The use of social networking sites for relationship maintenance in long-distance and geographically close romantic relationships. *Cyberpsychology, Behavior, and Social Networking*, **2015**.

[20] Connolly, T. M.; Hainey, T.; Baxter, G. J.; Stansfield, M. H.; Gould, C.; Can, C.; ... & Dimitrova, N. Teachers' views on Web 2.0 in education: An evaluation of a large-scale European pilot. Proc. of Next Generation Web Services Practices (NWeSP), 2011, 517-522.

[21] Kumar, B. Dimensions of digital divide and digital opportunities: An Indian scenario. *Asia Pacific Journal of Library and Information Science*, **2012**, *2*(1), 55-68.

[22] Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika*. **1951**, 16(3), 297–334.

[23] What Does Cronbach's Alpha Mean?. UCLA: Statistical Consulting Group. Available online: http://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/ (accessed January 22, 2017).

[24] SPSS 14.0 Brief Guide, Marcus Hearne, the University of Virginia. Available online: http://www.tau.ac.il/cc/docs/spss14manuals/SPSS%20Brief%20Guide%2014.0.pdf (accessed January 22, 2017).

[25] Weisstein, E.W. Chi-Squared Test. From MathWorld--A Wolfram Web Resource. Available online: http://mathworld.wolfram.com/Chi-SquaredTest.html (accessed January 22, 2017).

[26] Tayo, O.; Thompson, R.; Thompson, E. Impact of the digital divide on computer use and internet access on the poor in Nigeria. *Journal of Education and Learning*, **2016**, 5(1), 1-6.

[27] Harris, C.; Straker, L.; Pollock, C. A socioeconomic related digital divide exists in how, not if, young people use computers. PloS one, **2017**, 12(3).

[28] Understanding the digital divide. OECD publishers, France, 2001. Available online: http://www.oecd-ilibrary.org/science-and-technology/understanding-the-digital-divide_236405667766 (accessed September 09, 2017).

Studying the Impact of Water Supply on Wheat Yield by using Principle Lasso Radial Machine Learning Model

Muhammad Adnan
Institute of Manufacturing Information and Systems, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

Naheed Akhter
Department of Computer Science, GC University Faisalabad, Pakistan

M. Abid, M. Ahsan Latif, Abaid-ur-Rehman
Department of Computer Science, University of Agriculture Faisalabad, Pakistan

Muhammad Kashif
Department of Computer Science, Allama Iqbal Open University, Islamabad, Pakistan

***Abstract*—Wheat plays a vital role in the food production as it fulfills 60% requirements of calories and proteins to the 35% of the world population. Owing to wheat importance in food, wheat demand is increasing continuously. Wheat yield is committed to the availability of water supply. Due to climatic and environmental variations of different countries, water supply is not available in constant and desire quantity that is necessary for better wheat yield. So, there is a strong relationship and dependency that exists between water supply and wheat yield. Therefore, water supply is becoming an issue because it directly effects wheat yield. In this research, a Principle Lasso Radial (PLR) model is proposed using Machine Learning technique to measure the effect of water supply on wheat yield. In this Principle Lasso Radial (PLR) model, various experiments are conducted with respect to the performance metrics, i.e. relative water contents, waxiness, grain per spike and plant height. Principle Lasso Radial (PLR) model's produced reduced dimensional data with respect to performance metrics. That data is provided to Radial Basis Neural Network (RBNN), and it showed regression values R under different water supply conditions. Principle Lasso Radial (PLR) model achieved an accuracy of 89% among variance Machine Learning techniques.**

Keywords*— Radial basis function (RBF); Radial Basis Neural Network (RBNN); ANN; lasso; principle component analysis (PCA)

I. INTRODUCTION

It will predict wheat yield under stress and normal conditions by using ML. This model will enable us to predict the factor measurements which are necessary to have the required yield in future. The dependency of attributes of the wheat will be measured by using ML. The proposed model will predict the growth under stress and normal condition. Genotype with 1600 sets under different conditions will be tested by ML model. These sets of 1600 different values will be presented varyingly under normal and stress conditions. The ML model will predict and tell the yield which can be obtained by using different features under these conditions. The model will map and find regression and classify these attributes with respect to wheat yield.

The Machine learning technique enables us to find both the anticipatory as well as relational problems. This technique is applied to analyze the water effect on wheat crop and traits for better agriculture production. The machine learning algorithm is applied for the classification of yield component in order to get high yield. Wheat is the key and basic part of the food. It fulfills and provides a large number of calories and vitamins to human beings. It is a large element of food. Because of its significance, its demand has been enlarged for the last many decades. It is very fundamental to fabricate high yield of wheat to meet the supplies. Wheat also contains critical amino acids, minerals, and vitamins, beneficial photochemical and dietary fiber apparatus essential for the human diet, and these are predominantly enriched in whole-grain foodstuffs. It is not only a food factor but it also contains many proteins, fats and power in it which are vital and positive for human body [1].

The Agriculture sector in Pakistan is a major source of livelihood; 21% of Growth Domestic Production depends on it. Country's 44% workforce is related to this sector. Wheat is the essential part of food in Pakistan. Therefore, it grows almost in all parts of the country. As the population is increasing rapidly, there is need of growing wheat at an extensive level. The population is increasing with the rate of 3.1% yearly actions [2]. The area specified for wheat production is more than any other crop. This is because its consumption is more than any other crop. It requires a specific amount of water and minerals for growth [3].

Many environmental factors like the presence of Salt in the soil, water stress caused by lesser water in the plant and waterlogging poorly affect plant development. The climatic

change also contacts the plant growth and has an important impact. It is not possible for a plant to set itself fully free from ecological circumstances because of environmental variation habit [4]. Plant's cells, genetics, biophysics, nutrition, photosynthesis, fertility rate and hormones are affected by it. Typically, reaction and environmental acceptance rate of different plants are different from one another. Not all the plants act the same way under different growth situations. Some plants can face unfavorable environment whereas the others cannot do this [5].

II. Related Work

The phenomenon of stress altercation can be seen in two ways that are stress tolerance and stress avoidance. Both have different perspective and strategy about stress. In avoidance mechanism, plants grow speedily and become more talented to pass up stress collision. as the plant is at the peak level of its life, so it can easily accomplish at its final growth stage and can avoid stress environment. On the other hand, the patience method is little risky and dominant. At tolerance level, plants are taught by Nature to live under diverse environmental and stress full conditions. Even though in this plan, plants have to negotiate on their growth and enlargement procedure [6].

Dahikar and Rode [7] analyzed the crop prediction on the basis of different environmental conditions. Demographical conditions, temperature, and many other factors affected the prediction system. It was important to understand the environmental effect because crop production was directly affected by this phenomenon. They used feed-forward back-propagation ANN model to discover the crop yield under different stress conditions including water, ionic, weather and many other parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. This network can take a choice on the source of the training set of data which has been provided to it.

Emamgholizadeh et al. [8] premeditated that in the agricultural research the most vital purpose of breeding was the production of seed yield. In this research, two techniques were used; ANN and Multiple Regression Models (MLR). Both techniques were used to forecast same seed yield on the basis of premeasured features of the plant like maximum flowering days, the height of the plant in centimeters, numbers of capsules of each plant, etc. The results were experienced by using both MLR and ANN techniques and findings showed that ANN was more correct with respect to the root mean square error and founded coefficient.

Adnan et al. [12] use the machine learning methods to measure the evapotranspiration rate. Evapotranspiration diverges with the climate change and as the climate has a lot of deviation geographically, the pre-developed systems have not used all available meteorological data hence not robust models. In their study, a model is developed to evaluation evapotranspiration with extra authentic and accurate reduced meteorological parameters using different machine learning methods. The dataset with reduced dimension is modeled through time series neural network. Time series neural network delivered better results as compared to other existing approaches.

III. Contributions

In ML, there are many approaches that can be used for the best analysis of the water stress. But the current study used Stepwise regression and Principle component analysis (PCA). Both the techniques are suitable for reducing the dimension. By reducing the data set, the well dependent elements can be obtained which will make the analysis easy. Then the reduced dependent data will be analyzed by using the Radial Basis Neural Networks (RBNN). Some work related to this use of machine learning approach has been performed. Water factor was studied deliberately and very critically.

IV. Experimental Setup

Experimental material: The data related to wheat crop was collected in two years. In this procedure, the plant was grown in four different seasons. This was done at the department of Plant Breeding and Genetics, University of Agriculture, Faisalabad. The area where the selected population was grown consisted of a 5-meter wide long plot and with 15 cm and 30 cm plant to plant and row to row distances, respectively. The plant production procedure was examined under stressed and normal conditions. Under normal condition, the water quantity provided to the land was sufficient as according to the area allocated. When the growth of wheat plant was analyzed under the stressed condition, the quantity of water was declined and not enough water necessary for growth was supplied. By reducing the water supply or creating an artificial shortage of water, plant growth was analyzed.

The data for heading, relative water content (RWC) and proline were recorded before anthesis while peduncle length (PL), extrusion length (EL), awn length (AL), plant height (PH), seed size (SZ), grains per spike (gr/s), and yield/plant (Y) were recorded at maturity. The machine learning techniques were applied to the collected data. It was meant to find out the relation between the yield and the water-stressed conditions. Following three approaches were used for this purpose.

A. Principle Component Analysis (PCA)

Principle Component Analysis can simplify the problem by replacing a group of variables with a single new variable. The principal component analysis is a quantitatively rigorous method for achieving this simplification. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The first principal component is a single axis in space. The first plotted line on the graph represents the greater variation among all the data sets. When projected each observation on that axis, the resultant values form a new variable. And the variance of this variable is the maximum one among all possible choices of the first axis.

$$P_e = \sum_{k=1}^{n} dkE_{ke} \quad (1)$$

This is how our data will process using this equation, where "e" is the principal component. All our traits will be treated as separate principal components. Here in (1) "Pe" will be output generated for each component. "k" is the input data

set that is related to the specific component. In this case, each trait will be treated as a separate input band. "n" is the total number of input values. Here the PCA, during its process, calculates the eigenvector value that is represented by "E".

The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis. The full set of principal components is as large as the original set of variables. But it is commonplace for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data [9].

Step 1 - Standardize: The data that is collected from the field is not able to process our function directly. It needs 'new' in a ready form so that we can obtain a precise and efficient result. The data units are not same for every data component during the collection process. So before processing, the data must be standardized into an acceptable form, in our case that is performed in PCA Matlab function.

Step 2 - Calculate covariance: Covariance is a measure of how much random variables change together. In our data, the variables are large in quantity. The value of covariance among all those will tell about the actual relation of data with the outcome.

$$\sum n_i = 1(X_i - \bar{x})\,(Y_i - \bar{y})\ n\text{-}1 \qquad (2)$$

With $\bar{x}$ and $\bar{y}$ denoting the means of X and Y, respectively. X denotes the input variable and Y denotes the output variable. All these above following variable are treated as Y. The covariance among the variables is shown below.

Step 3 - Selecting Principal Components: PCA reduced the data dimension on the basis of dependency or variation of variables. In our data, PCA will reduce the eleven traits into a specified number of trait variables on the basis of their dependency on yield that is our outcome.

$$\Sigma v = \lambda v \qquad (3)$$

Where in (3),

Σ= Covariance matrix , v=Eigenvector, λ=Eigen value.

Here each trait will be principle component and will be decided by the eigenvectors directions on axis since they have all the same unit length. We can check that the eigenvectors, the Eigen-value calculation is correct through the equation.

Step 4 - Transforming the samples onto the new subspace: In the last step, we use-dimensional matrix W that we just computed to transform our samples onto the new subspace via the equation. These new traits were used for future estimation of the wheat yield.

$$Y = Wt \times X \qquad (4)$$

B. Lasso and Elastic Net

Lasso is a regularization technique for performing linear regression. Regularization is a technique used excessively to solve the compound problems. Lasso has the capability and functionality to reduce the dimensions of the data. Therefore, it resembles ridge regression. Lasso in its property can compress the variables into a suitable number of values and this property makes it a shrinkage estimator. It produces small coefficient which acts as an estimator in the data processing. In lasso method, the estimator produces lesser and minimum value of mean square error. While on another hand, when a generalized least squares estimator is applied, it produces a large value as compared to lasso estimator.

As a number of penalty terms increases, lasso coefficient values become closer to zero. This technique is different from ridge regression. This means that the lasso estimator is a smaller model, with fewer predictors. As such, the lasso is an alternative to stepwise regression and other model selections and dimensionality reduction techniques [10].

Elastic net is a lasso related technique. It is a hybrid of ridge regression and lasso regularization. It is a kind of mixture of both. Like lasso, the elastic net can generate reduced models by generating zero-valued coefficients. Empirical studies have suggested that the elastic net technique can outperform lasso on data with highly correlated predictors.

The lasso technique solves this regularization problem. For a given value of λ, a non-negative parameter, lasso solves the problem

$$\min_{\beta o,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_{i-\beta_0-x_i^T\beta}\right)2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \qquad (5)$$

Where in (5) "N" is the number of observations. Observation is the quantity of our data which has been collected during the field experiment. "yi" is the response at observation. It means that against every observation, we got some related response when it was made. "xi" is data. Xi is data values related to our traits. It will be represented as a vector of p values at observation i. "λ" is a positive regularization parameter corresponding to one value of Lambda. The parameters β0 and β are scalar and p-vector, respectively.

The elastic net technique solves this regularization problem. For a α strictly between 0 and 1, and a nonnegative λ, elastic net solves the problem

$$\min_{\beta o,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_{i-\beta_0-x_i^T\beta}\right)2 + \lambda P_\alpha(\beta)\right) \qquad (6)$$

Here, the parameter is same as the lasso when the value of α remains 1. "α" is the processed outcome of the entire data that is given during the formalization.

C. Data Modeling

Another approach used in our work was that of Data Modeling. It was applied to a data set that was collected after obtaining the result of the PCA and the stepwise regression. Radial basis neural network is the next type of the machine learning method that is very useful in finding the regression among different variables. Radial basis neural network is based on the number of layers and activation, training function

with a specific number of neurons in training of a data. It is used because of its rapid training and learning process. It provides good interpolation of the data set as compared to other neural networks [11].

a= radbas (ll W-p ll b) (7)

Here in (7) the network output is the output layer consisted of a single neuron. The activation function used here is "radial bias" and it is made of weight W and input vector product with b bias.

V. Results and Discussion

We provided data set in the principal component analysis method showing in Fig. 1. It categorizes all variables as according to its variation in the dataset. In this method, we got different plotted lines. Each line in the plotted area represents principle component importance of the mentioned variable according to its variance. The dataset consisted of eleven variables. The PC1 showed 24% of total variance in the first component. That shows its importance and significant impact on the total data set. That component showed that it distinguished behavior because of its greater variation among the data set, and for this reason, we are taking this PC into account. On the graph, the eight principle component showed more than 85% of the variation in the total data set. Thus, it reduced the dimension of the data by eliminating three other variables. Those three variables showed a minor variation in the dataset. Therefore, they have been discarded.

Now, the component that was derived on the graph has Coefficient values. Coefficient value has a significant importance in the formation of the effect of principle component Fig. 1. The PCA generalizes the plotted graph from higher to smaller as according to coefficient value. This value indicates the impact of the specific trait on the outcome variable. The study was to examine the factor that had close relation and dependent behavior to the yield.

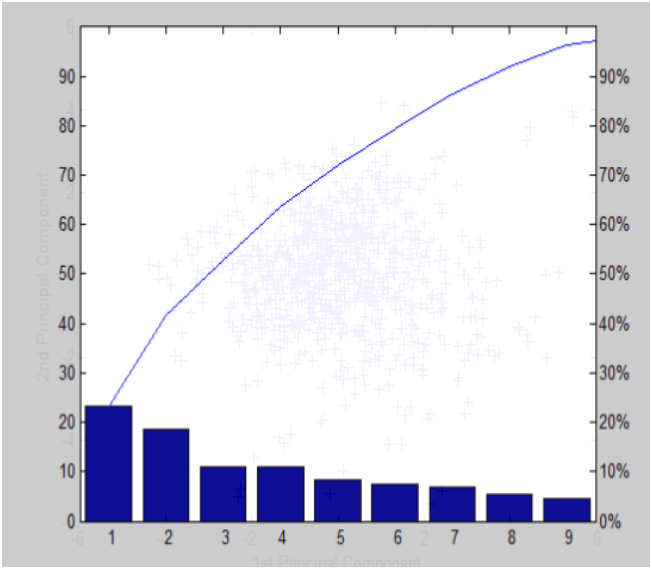


Fig. 1. PCA under normal water conditions.

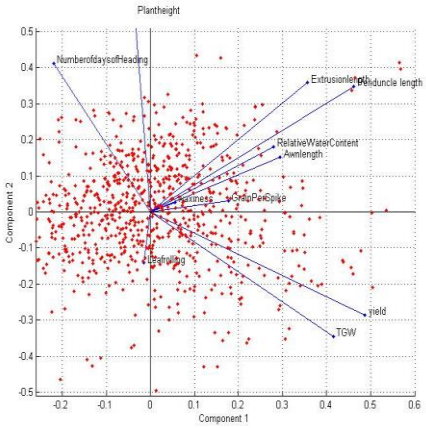


Fig. 2. PCA under normal water conditions using 2Dimensional graph.

The graph in Fig. 2 shows PC1 along x-axis and PC2 along the y-axis. From the graph, it can be visualized that trait "yield" has higher coefficient value among all other variables. So in 1st principle component "yield" has higher influence that shows that under the normal condition yield has higher relation and can be obtained at its maximum level. Whereas in the 2nd component along the y-axis, the coefficient value of "Plant height" is greater as compared to all other variables. Its participation in the variation of the 2nd component is higher. In the comparison of the PCA 3 and PCA 4, it was noticed that the variation in PC3 was because of the higher value of "waxiness" that was noticed 0.67 and the variation in PC4 were because of "RWC" that was 0.56 (Fig. 1). Other principal components also showed their coefficient values on the graph. So in the normal condition of the water, yield plant height and other variables have stronger and dependent relation.

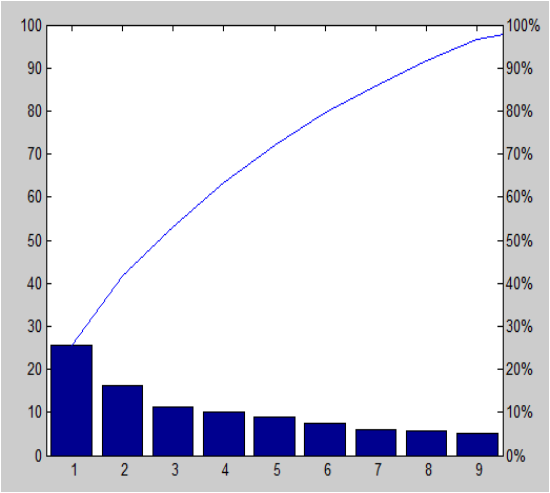


Fig. 3. Variation of data under stress water conditions.

In the second experimental phase, we provided eleven variables to the principal component analysis method as shown in Fig. 3. These were collected under water stress condition. It categorizes all variables according to its variation in the dataset. It has the same number of data set but was noted under stress condition of the water. Here the first

principle component that has the higher variation among the all data set, showed 25% of the total variation among all. That shows its overall impact on data. If we observe the graph in Fig. 3, it can be seen that the "pendulence length" has higher variation among all other traits.

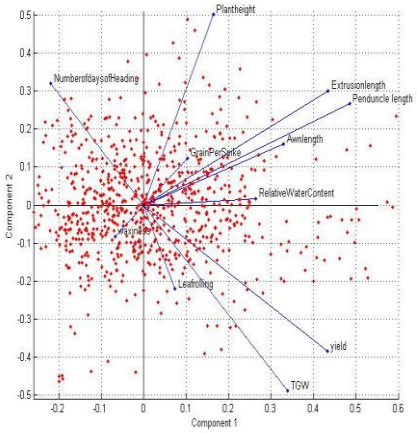


Fig. 4. PCA under stress water conditions using 2Dimensional.

So in the first principle component, the involvement of the "pendulence length" is more effective as compared to other traits or variables. The second trait that shows its dependency on the graph was "yield". In 2nd principal component that is along the y-axis, the variable "plant height" has greater coefficient value. In stress data of the principal component 3 and PC4, the "grain yield" and "plant height" values were observed at the highest level among all other variables. We got 9 principle components out of 11. Each has different variations and impacts. These components showed 90% of the total variation. It means that they are effective in the specific condition where the data has been collected in Fig. 4.

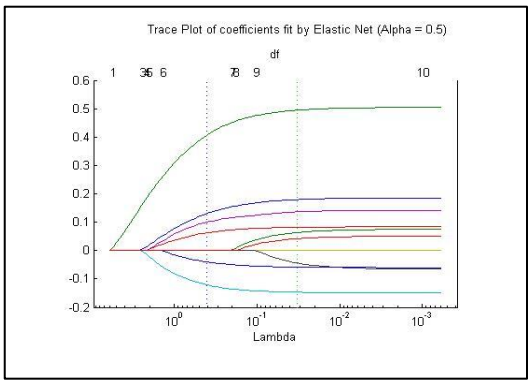


Fig. 5. Elastic net shows wheat traits variation in normal water conditions.

In lasso and elastic net, the expression of lambda is used. We used cross-validation to find the optimal value of regularization parameter λ with both minimum MSE shown in Fig. 5. Here the total values are plotted on the graph. Each line on the graph has separated color and represents the separate value. The dashed vertical lines represent the Lambda value with minimal mean squared error (on the right). Here when we applied data to the lasso technique. The first line of the graph was of TGW. That trait is one from our large data set. As the lambda value against "tgw" is decreasing the MSE value is increasing. The smaller the predicator lambda, the greater is the MSE value; hence the dependency among variables will be less shown in Fig. 6.

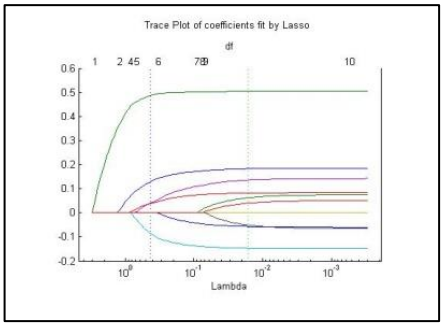


Fig. 6. Lasso shows wheat traits variation in normal water conditions.

The next variable is "grain per spike" its MSE is increased towards the right side of the graph as lambda is decreasing. So the point where it has larger none zero MSE values can be the best value of lambda to reduce the dimension of data and for finding the dependency of variables with yield. Where their lambda is between 10^0 and 10^{-1} at that value, there exist a minimum of the non zero values of MSE. So the "tgw", "g/spk", "rwc", "heading", "endurance lenght" are in a strong dependent relation with the yield under normal water condition shown in Fig. 6.

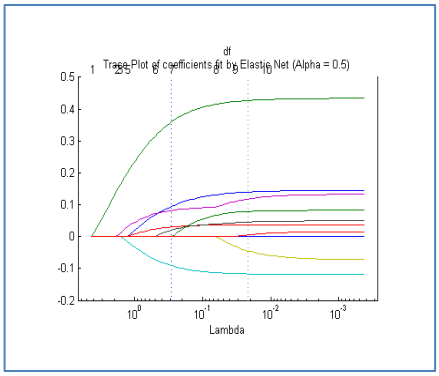


Fig. 7. Elastic net shows wheat traits variation in stress water conditions.

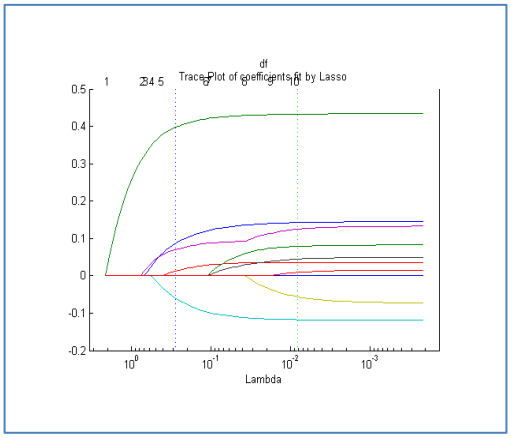


Fig. 8. Lasso shows wheat traits variation in stress water conditions.

Here, when we applied our data on the lasso and elastic net under stress condition, there produced a few results shown in Fig. 7 and 8. This result was based on the MSE values that showed the dependency of the variables to the yield. The values under stressed condition showed the different result in the lasso elastic graph. It was observed that the "TGW", "pendulous length", "grain per spike", "heading" and "relative water content" had the higher values of the mean square when the lambda was declined. So a strong relationship between these variables and yield under stressed condition was observed. It means these variables were affected by the stressed condition as shown in Fig. 7 and 8.

A. Data Modeling in Normal water condition

In RBNN, we trained the neural network under the data set of normal water condition. Here, a single-layered architecture was used. This model consists of 100 numbers of neurons in hidden layer and has one output layer which is conventionally contained a single neuron.

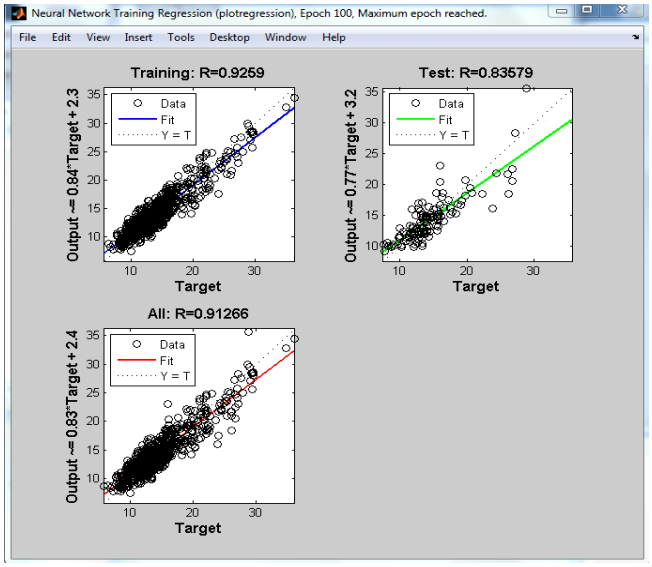


Fig. 9. Regression graph under normal water conditions.

Here radial basis function (RBF) was used as the activation function. A number of epochs in that model was 100. In radial basis neural network, when we use "trainlm" or "trainbr" activation function, the result obtained from this experiment is shown as in Fig. 9. Here the network is trained using the data set of those values which were observed under normal water conditions. Radial basis network works according to a total number of epochs unless the error is 0. Here is the plot of regression graph, the value of regression R was 0.91266. It was observed according to the radial basis function. The regression value indicates that traits and yield dependency was more than 90% and the yield relied completely on water.

The water factor showed significant impact on the wheat crop production i.e. the more water we provide, the more production of wheat we achieve. Just as yield was affected, different traits were also affected the same way in proportion to the supply of water as shown in Fig. 9.

TABLE I. RADIAL BASIS NEURAL NETWORK FOR WATER NORMAL CONDITION OBSERVATIONS

	Pr#	No of neuron	Activation Function	Training Function	R
Normal Condition	1	40	RBF	Trainlm	0.894
	2	40	RBF	Trainbr	0.904
	3	100	RBF	Trainlm	0.823
	4	100	RBF	Trainbr	0.912

We also applied some different techniques and methods in this neural network model to have different Regression values as shown in Table I. It helps to find the best relation between the traits and yield because of water. It also shows neural network architecture model is the best model to determine such observations. Different training functions were used for training dataset. In the first experiment namely activation function and "trainlm" training function, 40 neurons were used along with "RBF" model. By this experiment, we got 0.894 regression value. Then we trained the network model with the same architecture, and the output as regression value was 0.904. In next experiment, we used 100 neurons with activation function"RBF" and training function "trainlm". The regression result we got was 0.823 Table I showing above.

B. Data Modeling in Stress water Condition

TABLE II. RADIAL BASIS NEURAL NETWORK FOR WATER STRESS CONDITION OBSERVATIONS

	Pr#	No of neuron	Activation Function	Training Function	R
Stress Condition	5	40	RBF	Trainlm	0.876
	6	40	RBF	Trainbr	0.899
	7	100	RBF	Trainlm	0.856
	8	100	RBF	Trainbr	0.896

In radial basis neural network model for the measurement of yield under stress water condition, we also implemented some more techniques. Those techniques consisted of the use of a different number of neurons and different training and activation functions [11].

First, we used 40 neurons and RBF as activation function, "trainlm" used as training data. In this, the result we achieved, was 0.876 regression value. Then with the same architecture, we did little change and used "trainbr" as training function. This time the regression value we achieved was 0.899. In the next experiment, the number of neurons was increased up to 100 and "trainlm" was used for training showing in Table II. The R regression value was 0.856. Same as on the similar architecture we applied "trainbr" training function and got 0.896 regression R value. All techniques and models presented different and identical values. It shows the dependency of yield and traits on the water Table II showing above.

VI. CONCLUSION

After applying different machine learning techniques on the field collected data that was observed, the growth ratio of wheat is found affected by water stress. Moreover, the wheat production has been reduced under the water stress condition. Many traits showed higher variation under stress condition. The variation shows that if there is lack of the availability of the water, then fewer yields will produce. Plant height and leaf rolling were affected under stress condition. Yield and TGW also showed significantly different ratio as compared to normalvalues. “Awnlength”, “pendulacnelength”, “extractionlength”, “noofdaysheading” remain same under both conditions showing in Fig. 3. It means these four specific traits are survived under both water conditions. This factor showed significant impact on the wheat crop yield. It was also observed that the traits of the wheat were influenced by the water conditions showing in Fig. 4. Water condition showed significant effect on the growth and production of the wheat. The growth of traits, which influenced the yield, was reduced during that process. There was a significant relationship observed between different water condition and wheat and traits.

REFERENCES

[1] P. R. Shewry, "Wheat," Journal of experimental botany, vol. 60,pp. 1537-1553, 2009

[2] L. Akhtar, et al., "A review of hundred years of wheat research and development in Punjab (1911-2010)," Pakistan Journal of Science, vol. 62, pp. 128-134, 2010

[3] F.Joint, et al., Evaluation of certain food additives and contaminants: seventy-seventh report of the Joint FA: World Health Organization, 2013

[4] T. Umezawa, et al., "Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future," Current opinion in biotechnology, vol. 17, pp. 113-122, 2006.

[5] R. Munns and M. Tester, "Mechanisms of salinity tolerance," Annu. Rev. Plant Biol., vol. 59, pp. 651-681, 2008.

[6] E. A. Bray, "Classification of genes differentially expressed during water deficit stress in Arabidopsis thaliana: an analysis using microarray and differential expression data," Annals of botany, vol. 89, pp. 803-811, 2002.

[7] M. S. S. Dahikar and S. V. Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach."

[8] S. Emamgholizadeh, et al., "Seed yield prediction of sesame using artificial neural network," European Journal of Agronomy, vol. 68, pp. 89-96, 2015.

[9] L. J. van der Maaten, et al., "Matlab toolbox for dimensionality reduction," MICC, Maastricht University, 2007.

[10] O. Gonzalez-Recio, et al., "Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits," Livestock Science, vol. 166, pp. 217-231, 2014

[11] A. Galieni, et al., "Effects of nutrient deficiency and abiotic environmental stresses on yield, phenolic compounds and antiradical activity in lettuce (Lactuca sativa L.)," Scientia Horticulturae, vol. 187, pp. 93-101, 2015.

[12] Adnan, Muhammad, M. Ahsan Latif, Abaid-ur-Rehman and Maria Nazir. "Estimating Evapotranspiration using Machine Learning Techniques. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol 8, no 9, pp. 108-113, 2017.

An Improvement of FA Terms Dictionary using Power Link and Co-Word Analysis

El-Sayed Atlam
Information Sceinec and Intelligent System, Faculty of Enginerring, Tokushiam, Japan
Computer Science Division, Mathematical department, Tanta, Egypt

Fayed Ghaleb
Department of Mathematics, Faculty of Science, Ain Shames University, Cairo, Egypt

Dawlat A. El A.Mohamed
Department of Mathematics, Faculty of Science, Ain Shames University Cairo, Egypt

Doaa Abo-Shady*
Computer Science Division, Mathematical department, Tanta, Egypt

***Abstract*—Information retrieval involves obtaining some wanted information in a database. In this paper, we used the power link to improve the extracted field association terms from corpus by the proposed algorithm to support the machine to take the right decision and attach the candidate words in their convenient position in dictionary of the field association terms. Power Link is used as a quantitative tool to compute the co-citation relation among two words depending on the co-frequency and distances among instances of the words. In this paper, concept of the Power Link as well as modifications of the rules is used to classify the scientific papers into its proper field. In this paper, instead of whole document, a given document will be divided into three parts, namely, title, abstract and body. A given term will be given a weight that depends on the location of the term inside a specific document. The greatest weight will be given to the title then the abstract then the body, respectively. Results show an improvement in precision, recall and F measure.**

***Keywords*—Information retrieval; FA terms; co-word analysis; power link; precision; recall**

I. INTRODUCTION

Information retrieval (IR) defined as the activity of finding information resources related to an information need from a group of information resources. Searches can depend on whole document text or other content-based indexing. To provide automatic information retrieval systems, we can use several different retrieval techniques based on Field Association (FA) Terms and this paper concentrate on the concept of FA terms with co-word analysis [3].

Humans can understand the field of the scientific papers through detecting the particular terms, these terms called FA terms. Field of a document can be classified as: a super field, a sub field and terminal field, and the representation scheme of the document field called field tree [12]. For example, the path <Science& Technology/ COMPUTER/ Programming> expresses super field < Science& Technology > having sup field < COMPUTER > and terminal field < Programming > and the field code of this path can be defined by K.12.5.

FA terms are collected according to how well they refer to particular field. For example, "Communication network" and "compiler" are FA Terms of sup-field < COMPUTER >. As an FA Term may relate to more than one field, there are five levels used to rank FA terms as in [12]:

Level 1: The terms that specified to only one subfield and called Perfect FA terms.

Level 2: The terms that specified to more than one subfield and in only one super-field and called Imperfect FA terms.

Level 3: The terms that specified to one super-field and called Super FA Terms.

Level 4: The terms that specified to more than one subfield of more than one super field and called Cross FA Terms.

Level 5: The terms that do not assign any subfield or super-field and called Non FA Terms.

To choose the helpful FA terms need to consider the relations among simple and compound FA terms and field ranking. So, we need to use the co-word analysis and the Power Link concepts [18].

The co-word analysis is a quantitative study of relations between elements (i.e., terms or noun phrases or topics or fields). The inclusion and proximity indexes are used to compute the strength of relations among elements, these indexes depended on the co-occurrence frequency of elements. Co-word analysis focus on the dynamics of science as an outcome of actor methods. Changes in the content of a topic area are the common impact of a great number of individual strategies. This method must let us in principle to identity the actors and describe the global dynamic as in [11].

In [6], author presented an approach using the passage retrieval to improving constructing FA terms dictionary. They suggested a new method for locating FA terms using passage (parts of a document text) method instead of locating them from the full documents.

In [10], author provided the algorithm based on Power Link concept which explained and computed the relation among two words depended on the co-frequency and the relative locations of various successive instances. If words have nearer relative locations then the Power Link become bigger for those words.

In [13], author presented a method based on the Power Link concept to improve the classification of search engines results. This method depends on ranking the terms in a given field.

Depending on the absolute frequencies reflects the documents length rather than the weight of words, so recent works depend on normalized frequencies instead of absolute frequencies [10], [13], [19] and [20]. Also, recent works used the co-occurrence frequencies to reflect the relation between terms [4]. Power Link method uses the normalized frequencies, co-occurrence frequencies and considered the relative distances between terms.

While Power Link algorithm considers the whole documents, and gives the same weight for all parts of scientific paper, we will give different weight for different parts of a given scientific paper. In this work, the Power Link algorithm will be implemented, in addition to the another algorithm detect the pre-defined errors in Pre-text processing step presented by [7] to improve the quality of results and purge files from the resulting errors.

After collecting the corpus, in the pre-processing phase, every scientific paper will be divided into three parts, title, abstract and body. Each part will be given a different weight based on its importance. The title contains the most related terms to the topic and reflects the field of the document more than other parts. The abstract contains related terms to each other and reflects the field of the document more than the remainder body. So, we propose to give the terms that occur in the title the highest weight, then the abstract and give the body the least weight in the processing phase, the Power Link will be used to improve the FA terms dictionary. As a result, the proposed idea improved the Perfect FA terms (Level 1) and not improved in results of Imperfect and super FA terms (Level 2 and 3) so, level 1 is enough in our data. This idea can be used in many applications in information retrieval field.

The precision, recall and F measure values referred that the presented algorithm produced in average 0.90%, 0.85% and 0.87% respectively which means that the algorithm effective performance. The F value refers the strength of the algorithm.

The rest of article proceeds as the following: In Section 2, we presents a summary discuss of some definitions and modified algorithm. Sections 3 provide the modified algorithm for determining the Perfect FA terms (Level 1). Section 4 includes the results and discussion then in Section 5.

II. DEFINITIONS

A. Power Link Analysis

Power Link is a quantitative tool to determine the co-citation relationship among two terms depending on the frequency and the distances among instances of the terms [21]. In this paper, we used the Power Link as a tool to improve the extracted field association terms from corpus by the proposed algorithm.

The Power Link algorithm presented in [10] was provided calculations for how tow terms tend to occur altogether in a specific corpus. The Power Link value among two terms was high, if these terms are related together strongly.

The link between any two terms t1 and t2 in document D can calculated by the function of power link $LT(t_1, t_2)$ defined in Section 3.

B. Continuity and Transition Theme

Continuity and transition theme is a method to detect or determine the field of each part of a given document. The features of a subject are given based on continuity and transition. The theme field is defined as the field that a sentence presents, which is denoted by F_{theme} [14]. F_{theme} is preserved by *continuity* or changed by *transition* through sentences [9].

Let F_{theme} is field of sentence S that includes FA terms, then the power link among S and F_{theme} is computed by the field that gives $\max_j P(S, F_j)$ where, $P(S, F_j)$ is the Power Link among S and whole fields which expressed by the formula $P(S, F_j) = \sum_{i=1}^{n} P(FA_i, F_j)$ for each FA Term in F. So, the existing sentence is attached to the same passage If it has the equal F_{theme} as the previous sentence, or has no F_{theme}, or has no field. And S is delimited and a new passage starts if the existing sentence S has a different F_{theme} from the previous sentence, for more details see [5], [8] and [10].

Here, we can detect the three parts (title, abstract and body) by determining the head word of every part (i.e., abstract and introduction). If the head words are not present or repeated then we need to apply continuity and transition theme in this case. Always the first sentence on any document is the title that contains the most related words together and indicates to the field of the paper, the second paragraph usually is the abstract that contains a summary of all important information about the paper. So it contains the most important FA terms that indicate to the field of the paper and the power link between these terms should be high. So according to the previous rules we can detect and extract the abstract part from the document.

C. Real Word Spell Checker

Many words with multiple meanings exist in the English language. Technically, almost every word has a multiple meaning. How often do you go into the dictionary to look up a word, and find that only one meaning is listed next to it? Practically never! Many words have slightly varying meanings, or they can be used as different parts of speech.

For example (right: You were right./Make a right turn at the light, type: He can type over 100 words per minute./That dress is really not her type), (ate/eight, blew/blue, fair/fare, no/know).

To solve these problems, some algorithms were proposed to automatically detect such errors in syntax or meaning. In this work, to avoid these problems, we use the Real Word Spell Checker algorithm in Pre-text processing step. This

method depends on automatic building of errors that called confusion sets for a specific terms dictionary and corresponding corps. For more details see [7].

1) Algorithm for Calculating the Perfect FA Terms (PFAT)

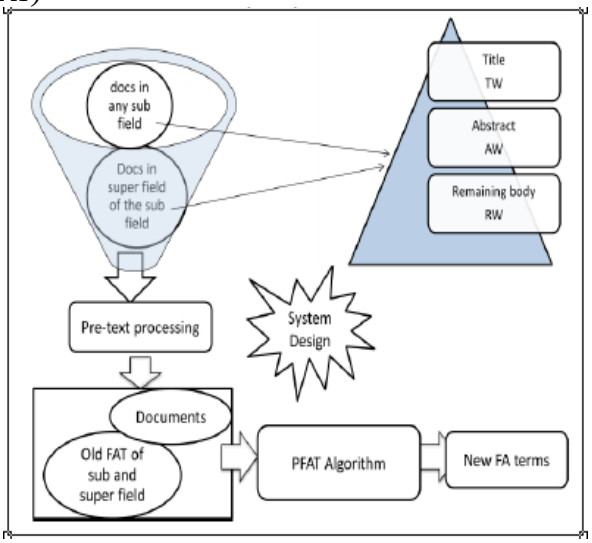


Fig. 1. System design.

Inputs:

a) Documents in any specified sub field and its super field after indexing to new terms candidate (ranked depending on their occurrence in document after stemming, removing stop words) to extract the new field association terms from them, as in Fig. 2 and Table I.

b) FA terms dictionary (by traditional algorithm by [12]) of sub and super field to be used in the Power Link calculations among them and the candidate terms in each document. Also, data of super field will be used to calculate the concentration ratio.

Programming Wireless Sensor Networks:
Fundamental Concepts and State of the Art
Abstract
Wireless sensor networks (WSNs) are attracting great interest in a number of application domains concerned with monitoring and control of physical phenomena, as they enable dense and untethered deployments at low cost and with unprecedented exibility.
1. INTRODUCTION
Wireless sensor networks (WSNs) are distributed systems typically composed of embedded devices, each equipped with a processing unit, a wireless communication interface, as well as sensors and/or actuators. Many applications have been proposed to date that show the versatility of this technology, and some are already ending their way into the mainstream.
2 L. Mottola and G.P. Picco
available to application developers [OnWorld ; CONET]. However, of the several experiences reported in the literature where WSN applications have been deployed in the real-world, only a few exceptions rely on some high-level programming support [Ceriotti et al. 2009; Buonadonna et al. 2005; Whitehouse et al. 2004].

Fig. 2. Sample of Docs Input.

Output:

A new set of improved FA terms:

We can demonstrate the system design and proposed algorithm in Fig.1 and 3 by the four main steps: 1) Power Link calculation, 2) Compute the Candidate Terms Frequency, 3) compute the concentration ratio, and 4) Compute the Precision and Recall Values, more details about those steps will be discussed in the following sub sections.

TABLE I. CANDIDATE TERMS AFTER INDEXING

program	**cost**	**technolog**
wireless	**unpreced**	**alreadi**
sensor	**exibl**	**nding**
network	**introduct**	**way**
fundament	**wireless**	**Mainstream**
concept	**sensor**	**mottola**
state	**network**	**picco**
art	**wsn**	**avail**
abstract	**distribut**	**applic**
wireless	**system**	**develop**
sensor	**typic**	**onworld**
network	**compos**	**conet**
wsn	**embed**	**howev**
attract	**devic**	**sever**
great	**equip**	**experi**
interest	**process**	**report**
number	**unit**	**literatur**
applic	**wireless**	**wsn**
main	**commun**	**applic**
concern	**interfac**	**deploy**
monitor	**well**	**real**
control	**sensor**	**world**
physic	**actuat**	**except**
phenomena	**mani**	**reli**
enabl	**applic**	**high**
dens	**propos**	**level**
Unteth	**date**	**program**
deploy	**show**	**Support**
low	**versatil**	

2) Power Link Calculations

For each candidate term t in each document D compute the following Power Link calculations:

a) Compute the Power Link between the term t and the sub field <S>:

$$\mathrm{LTS}(t, <S>) = [\sum_i \mathrm{LTD}(t, D_i, <S>) * \mathrm{crs}(t, <S>)] \; / \; nd \quad (1)$$

where: D_i is includes at least one FA term belong to $< S >$.

$\mathrm{LTD}(t, D_i, <S>)$ is the Power Link between t and D that will be compute it in b.

$\mathrm{crs}(t, <S>)$ is co-occurrence of term t and $< S >$ S.T: $\mathrm{crs}(t, <S>) = |\{f_i : f_i \in <S>, min\ cr\ (f_i, t) > 0\}| + 1$ is the number of FA terms identify $< S >$ and appear in D that the t appears.

nd is the number of documents that includes FA terms that identify $< S >$ and t.

b) Compute the Power Link between t and D respect to a given terminal field $< S >$:

$$\text{LTD}(t, D, < S >) = \sum_{f_i \in <S>} \frac{LT(t, f_i)}{n} \quad (2)$$

S.T.: n is number of $f_i = f_1, f_2, f_3, \ldots$ that are FA terms that identify $< S >$ and in D.

$\text{LT}(t, f_i)$ is link between t and f_i that will be compute it in (c).

c) Compute the Power Link between two terms t_j and f_i based on dividing the document:

Firstly: we have two constant terms (stems) in every doc are "abstract" and "introduct" according to the corpus are scientific papers.

let $\text{word1}_k =$ "abstract" and $\text{word2}_m =$ "introduct".

S.T: k is the index of word1 in D.

m is the index of word2 in D.

j is the index of t in D.

There are three cases to compute LT according to term t position:

Suppose ($\text{TW} > \text{AW} > \text{RW}$ are the title, abstract and reminder body weights respectively) so:

case 1: if $j > k$ (S.T: t position is in the title) then

$$\text{LT}(t_j, f_i) = \frac{|D| \times \text{cr}(t_j, f_i)}{\text{average } L(t_{j_r}, f_{i_s})} * \text{TW} \quad (3)$$

else,

case 2: if $k < j < m$ (S.T: t position is in the abstract) then

$$\text{LT}(t_j, f_i) = \frac{|D| \times \text{cr}(t_j, f_i)}{\text{average } L(t_{j_r}, f_{i_s})} * \text{AW} \quad (4)$$

else,

case 3: if $j > m$ (S.T: t position is in the body of paper) then

$$\text{LT}(t_j, f_i) = \frac{|D| \times \text{cr}(t_j, f_i)}{\text{average } L(t_{j_r}, f_{i_s})} * \text{RW} \quad (5)$$

where: $|D|$ is the number of different terms in document D, co-occurrence frequency $\text{cr}(t_j, f_i)$ of t_j and f_i in D and $L(t_{j_r}, f_{i_s})$ is the distance between any two successive instants t_{j_r} and f_{i_s} of t_j and f_i, such that there are no other instants of the term t_j or f_i between the instants t_{j_r} and f_{i_s} in D, note that, the extremes values are neglected. TW, AW and RW are reflects how much the relation between terms in each part of a document (i.e. Title, abstract and body, respectively). such that TW bigger than AW and AW bigger than RW because as usual the terms are more related together in the title more than the abstract also more than the body of the scientific researches and its values are determined by experiments.

Also, we used the continuity and transition to determine the abstract in case if the doc has problems to detect this part.

3) Compute the Candidate Terms Frequency

The frequency of a term t in a sub field $< S >$ is denoted by $F(t, < S >)$ then

$$F(t, < S >) = \sum_{D_i} f(t, D_i) \quad (6)$$

S.T: D_i is a document that includes FA terms that identify $< S >$ and $f(t, D_i)$ is defined as this formula:

$$f(t, D_i) = \log(\text{dtf}(t) + 1) / \left(\sum_{y \in D_i} \log(\text{dtf}(y) + 1) * \frac{U}{(1 + 0.0115 * U)}\right)$$

S.T.: $\text{dtf}(t)$ is number of times that term t occur in D_i .

$\left(\sum_{y \in D_i} \log(\text{dtf}(y) + 1)\right) =$ sum of $\log(\text{dtf} + 1)$ for whole terms in the D_i .

The local information and the normalization factor are given as these parts $\log \frac{\text{dtf}(t)+1}{\sum_{x \in d} \log(\text{dtf}(x)+1)}$ and $\frac{U}{(1+0.0115*U)}$ respectively [2].

U is the number of unique terms in D_i.

This formula is derived from the classic known formula 'TF $*$ IDF' (Term Frequency-Inverse Document Frequency) of Salton and used it in this algorithm instead of the traditional methods [12], [15], [16] and [6] that used the absolute frequency that only depend on the number of a term repetition in the document and not effective enough [1].

4) Compute the Concentration Ratio

The concentration ratio $\text{PL}(t, < S >)$ that based on the frequency and Power Link calculations can be used to judge whether or not the term t is a *Perfect FA term* and defined as:

$$\text{PL}(t, < S >) = \frac{F(t, <S>) * LTS(t, <S>)}{F(t, <S`>) * LTS(t, <S`>)} \quad (7)$$

Where $F(t, < S >)$ and $LTS(t, < S >)$ are frequency and Power Link calculations that will be computed in previous steps , since $< S >$ is the sub field, $< S` >$ is the super field of this sub field and by using threshold α to judge the levels of FA terms. Such that, If PL is less than value of α then t is not perfect term else t is perfect term.

5) Compute the Precision and Recall Values

To test the efficiency of the system we used the measurement of precision and recall to reach the best result of FA terms and its measure are

$$P_i = \frac{Number\ of\ Relevant\ FATs\ extracted\ by\ system}{total\ no.of\ FATs\ extracted\ by\ system} \quad (8)$$

$$R_i = \frac{\text{Number of Relevant FA terms extracted by system}}{\text{Total Number of FA words extracted Manually}} \quad (9)$$

where the termination condition of algorithm is $P_i - P_{i-1} < \mathcal{E}1$, $R_i - R_{i-1} < \mathcal{E}2$ where $\mathcal{E}1$ and $\mathcal{E}2$ are most low value as we wish, such that:

if $P_i - P_{i-1} < \mathcal{E}1$ and $R_i - R_{i-1} < \mathcal{E}2$ then algorithm will be the terminated, else repeat the processes.

(***End of algorithm)***

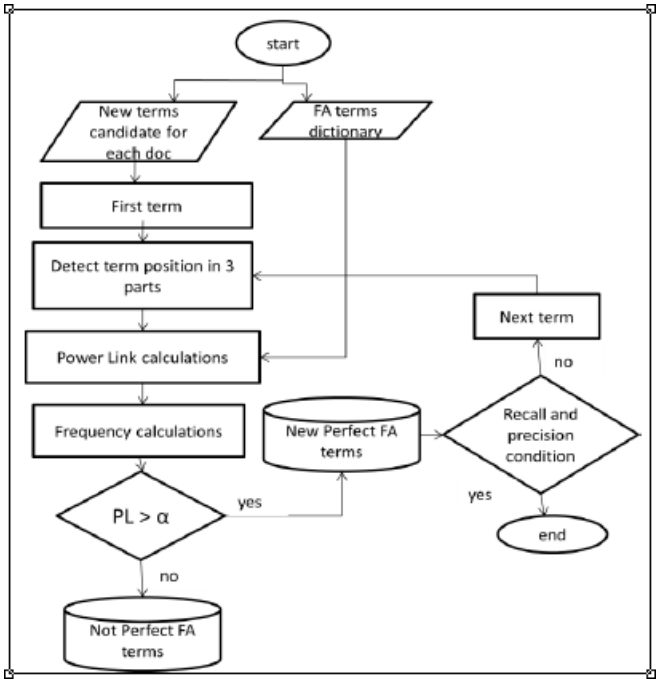


Fig. 3. PFAT Algorithm.

III. Experiments and Results

The experiments used to validate the advantage of the newly approach and that was the main purpose of it. Furthermore, we choose a most efficient weights along group of trials to provide good algorithm performance. Also, we write the code of our system by Python language that can be easily satisfied for any process on the text but there was a lot of challenges in pre-processing the text files to be formulated, like to convert from PDF file to txt file where some data can lost and there are not function in python can read from PDF file.

In this paper, we focus on a super field science and technology(K) and its sub-field Computer (K. 12) with corps size 12.2 MB about 4741 candidate terms were extracted.

Used the Real Word Spell Checker algorithm in pre-processing step led to discovery and correction 5% errors of the terms. Also we detect the three parts in 100 documents by use the continuity and transition theme. After the comparative analysis of the power link algorithm presented by [10], the proposed algorithm and some research information systems on scientific researches, it was recognized that giving different weights for each part could be improved selection of Perfect FA Terms (Level 1) but not improved of level 2 and 3 in our data. Table II show samples of perfect and not perfect FA terms that resulting from proposed algorithm (PFAT) and traditional algorithm [10], note that terms "Data, keyword and system" are detected as perfect by old method but they are not perfect FA terms in <Science& Technology\ Computer> field. We use $\mathcal{E} = 0.001$ and the threshold value = 0.9 that showed the best one for the concentration values in [17]. So, this threshold used as a fixed threshold for the concentration values in all loops and the average values of precision and recall are 0.90% and 0.85% respectively, as in Fig. 4. The results showed that the power links by weights do better than the random that produced the values of precision and recall in average 0.80% and 0.70%, respectively, as in Fig. 5. This means that, in this random data, the algorithm has efficiency 100% and to ensure the strong of the results, F is also calculated using the formula.

$$F = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

The average of new value of F is 0.87% while it was 0.74% using traditional method which refers a high performance of the system.

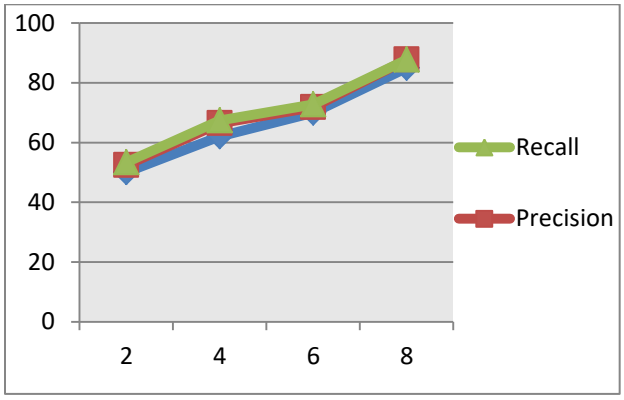


Fig. 4. Precision, recall and F measure by new approach.

TABLE II. Comparison of New and Traditional Approaches.

Term Samples	Old system	New system
Network	Perfect term in K.12	Perfect term in K.12
Data	Perfect term in K.12	Not Perfect term in K.12
Softwar	Perfect term in K.12	Perfect term in K.12
Keyword	Perfect term in K.12	Not Perfect term on K.12
Hardwar	Perfect term in K.12	Perfect term in K.12
System	Perfect term in K.12	Not Perfect term in K.12
Algorithm	Perfect term in K.12	Perfect term in K.12
Memori	Perfect term in K.12	Not Perfect term in K.12
Control	Perfect term in K.12	Not Perfect term in K.12
Structur	Perfect term in K.12	Not Perfect term in K.12
Code	Not Perfect term in K.12	Perfect term in K.12
Goal	Not Perfect term in K.12	Not Perfect term in K.12
Select	Not Perfect term in K.12	Not Perfect term in K.12
Imag	Not Perfect term in K.12	Not Perfect term in K.12

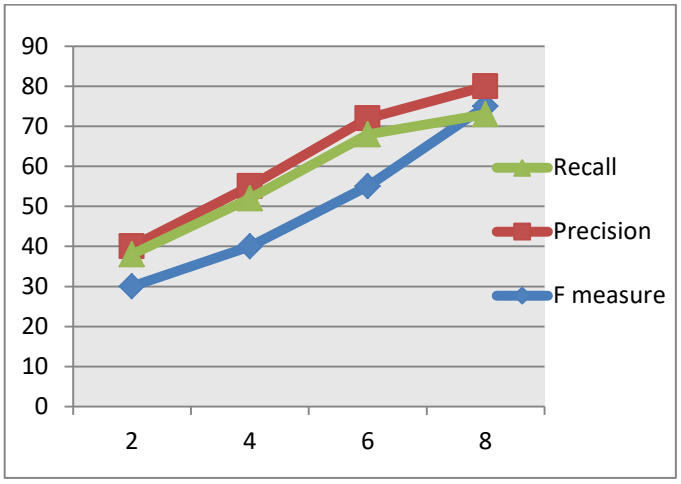


Fig. 5. Precision, recall and F measure by traditional approach.

IV. Conclusion

In this work we proposed an approach to produce an improvement FA terms dictionary by used Power Link concept and give different weights to terms according to their position in the document. The precision achieved using the new method 0.90. Hence, the algorithm succeeded to improve the values of precision by 10% than traditional approach.

Future work could focus on the importance to consider the difference between languages and cultures between English and Arab countries in the Middle East. Different languages can be implemented by doing some natural language processing & speech recognition researches using English, Japanese and Arabic languages. Also, this method can used in Building a comprehensive FA terms dictionary and can apply it in many of the applications especially in text summarization, text classifications, Extraction, filtering and machine translation.

Furthermore, we can apply the Power Link analysis using different weights not only on the scientific research but also any type of unstructured documents.

References

[1] Song S-K, Myaeng SH,"A novel term weighting scheme based on discrimination power obtained from past retrieval results, " Inf Process Manag, 2012 . http://dx.doi.org/10.1016/j.ipm.2012.03.004

[2] Cummins R, O'Riordan C, "Evolving local and global weighting schemes in information retrieval, " J Inf Retr vol.9, pp. 311–330, 2006.

[3] Chun H-W, Jeong C-H, Song S-K, Choi Y-S, Jeong D-H, Choi S-P, Sung W-K "Smart searching system for virtual science brain," LNCS 6890, pp. 324–332, 2011.

[4] Mahmoud Rokaya *, Elsayed Atlam, Masao Fuketa,Tshering C. Dorji, Jun-ichi Aoe, "Ranking of field association terms using Co-word analysis," Information Processing and Management, 2007.

[5] Mahmoud B. Rokaya, "Automatic text extraction based on field association terms and power links," *International Journal of Computer and Information Technology (ISSN: 2279 – 0764) vol. 02– no 06, November 2013.*

[6] Uddin, S., Elmarhomy, G., Atlam, E., Fuketa, M., Morita K. and Aoe, J., "Improvement of automatic building field association term dictionary using passage retrieval," *Information Processing & Management Journal*, vol. 43, 2007.

[7] Mahmoud Rokaya, AbdAllah Nahla, Sultan Aljahdali, "Context-sensitive spell checking based on field association terms," IJCSNS International Journal Of Computer Science And Network Security, vol. 12 no. 3 pp. 64-68, 2012.

[8] Oi Mean Foong1, Alan Oxley1 And Suziah Sulaiman1, "challenges and trends of automatic text summarization," International Journal Of Information And Telecommunication Technology, vol. 1, no 1, pp 34-39, 2010.

[9] Mahmoud Rokaya, "Automatic summarization based on field coherent passages," International Journal of Computer Applications, Published by Foundation of Computer Science, New York,USA, October 2013.

[10] Mahmoud Rokaya and El-Sayed Atlam, "Building of field association terms based on links," *Int. J. Computer Applications in Technology, vol. 38, no. 4, 2010.*

[11] Callon, M., Courtid, J. and Ladle, F, "Co-word analysis as a tool for describing the network of interactions between basicand technological research: the case of polymer chemistry," *Science Metrics*, vol. 22, no. 1, pp.155–205, 1991.

[12] Atlam, E., Morita, K., Fuketa, M. and Aoe, J, "Documents similarity measurement using field association terms," *Information Processing and Management*, Vol. 39, pp.809–824, 2003.

[13] Mahmoud Rokaya, "Improving Ranking of Search Engines Results Based on Power Links," *IPASJ International Journal of Information Technology (IIJIT)*, vol 2, no 9, September 2014.

[14] Lee, S., Shishibori, M., Sumitomo, S., Aoe, J., " Extraction of field-coherent passages," Information Processing And Management, vol. 38, pp. 173-207, 2002.

[15] Fuketa, M., Lee, S., Tsuji, T., Okada, M. and Aoe, J, "A document classification method by using field associationwords," *Information Science*, vol. 126, pp.57–70, 2000.

[16] Elmarhomy, G., Atlam, E., Fuketa, M., Morita K., Sumitomo, T. and Aoe, J, "Automatic deletion of unnecessary field association word using morphological analysis," *Journal of Computer and Mathematics*, vol. 83, no. 3, pp.247–262, 2006.

[17] Atlam, E., Elmarhomy, G., Morita, K., Fuketa, M. and Aoe, J, "Automatic building of new field association word candidates using search engine," *Information Processing &Management Journal*, vol. 42, no. 4, pp.951–962, 2006.

[18] Atlam E-S., Ghaleb F., Taha A., Ismail A., "A new retrieval method based on time seriesvariation using field association terms," Mathematical Methods in the Applied Sciences, in press.

[19] Mahmoud Rokaya, Dalia I. Hemdan, "Bibliometric cartography of nutrition science researches based on Power Links analysis," International Information Institute, vol.19, no.9(B), 2016.

[20] Mahmoud Rokaya, "Arabic Semantic Spell Checking Based on Power Links," International Information Institute, vol.18, no.11,2015.

[21] Mahmoud Rokaya, "Spam Reduction Based on Power Link Analysis," International Information Institute, vol.19, no.6(A),2016.

An Improved Social Media Analysis on Three Layers: A Real Time Enhanced Recommendation System

Mohamed Amine TALHAOUI, Hicham AIT EL BOUR, Reda MOULOUKI, Saida NKIRI, Mohamed AZOUAZI
Laboratoire Mathématiques Informatique et Traitement de l'Information MITI
Hassan II University, Faculty of Sciences Ben m'Sik Casablanca, Morocco

***Abstract*—The Internet can be considered as an open field for expression regarding products, politics, ideas, and people. Those expressive interactions generate a large amount of data pinned per users and groups. In that scope, Big data along with various technologies, such as social media, cloud computing, and machine learning can be used as a toolbox to make sense of the data and give the opportunity to generate efficient analysis and studies of the individuals and crowds regarding market orientation, politics, and industry. The recommendation system for this acts as the pillars of technology, in the field of sentiment analysis and predictive analysis to make sense of user's data. However, this complex operation comes at the price of this. To each analysis, there is a personalized architecture and tool. In this paper, a novel design of a recommender system is provided powered by sentiment analysis and predictive models applied onto an example of data flow from the social media Twitter.**

Keywords—Twitter; machine learning; sentiment; Lambda; recommendation; Big data; opinion mining

I. INTRODUCTION

A significant amount of available data on Internet represent a rich source of knowledge extraction, in particular, the social media like Twitter or Facebook. How this data processing is performed enables data scientists to deliver relevant and exploitable results for economic, social, industrial, government policies or business purposes.

In this paper, the work is focused on providing for a company valuable information about its products. A proposition on how to extract data from Twitter based on keywords and store it in Hadoop Ecosystem [1] to carry out a sentiment analysis. This processing allows us to categorize tweets as positive, negative or neutral using scoring methods. The proposed system provides as well as a relevant way to detect the qualities and defects of the specific product using Word Clouds. Machine Learning is also a crucial aspect of this work; it enables us to predict Satisfied customer, Customer not satisfied, prospect and suspect for a distinct product.

For that several techniques are used:

A. Machine Learning

The purpose of machine learning [2] is to induce a system into learning from the past or present and apply that knowledge to make predictions or decisions regarding unknown future events. In the most general terms, a supervised machine learning task is composed of three stages: build the model, assess and tune the model, and next put the model into production.

After an analysis of some ML supervised algorithms such as:

- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Neural Networks (NN)

A choice to use NB is because it has been shown to perform surprisingly well with minimal amounts of training data that most other classifiers, would find significantly insufficient.

Naïve Bayes is a classification [3] algorithm that relies on strong assumptions of covariance independence in the Bayes theorem. The Naïve Bayes classifier assumes independence between dependent predictive variables and a Gaussian distribution of the numerical predictors with mean and standard deviation calculated from the training dataset. Naïve Bayes models are commonly used as an alternative to decision trees for classification problems. When constructing a Naïve Bayes classifier, each row of the training dataset containing at least one NA will be ignored completely. If the test data set has missing values, these predictors are omitted in the probability calculation during prediction.

B. Opinion Mining

Sentiment analysis [4] or Opinion mining can be described as the computational study of people's opinions, judgments, attitudes, and emotions toward entities and their aspects. The entities usually relate to products, services, organizations, individuals, events, etc. and the aspects are attributes or components of the entities. With the growth of social media (i.e., reviews, forum discussions, and blogs) on the Web, individuals and organizations are increasingly using the opinions in these media for decision-making.

Sentiment detection [5], as usual, is a classic problem of text classification. Unlike other text classification tasks, the goal is not to identify topics, entities, or authors of text but to rate the expressed sentiment typically as positive, negative, or neutral. Most approaches used for sentiment detection have also been used for other text classification tasks and usually involve techniques from machine learning, computational linguistics, and statistics. Typically, different approaches from these fields are used for sentiment detection. Linguistic considerations [6] range from tokenizing the to-be- classified texts to other syntactic analyses. Statistical considerations typically involve frequencies of tokens or phrases, e.g., the occurrence of many "positive" words in a text, or similar

statistics. The respective features then usually are coupled with machine learning algorithms to classify the sentiment of arbitrary texts.

Due to the diverse applications [7] in mining and retrieval, and since Twitter is one of the most abundant sources of opinion, a lot of different approaches to sentiment detection in tweets have been proposed. Various methods use different feature sets ranging from standard word polarity expressions or unigram features also applied in general sentiment detection, to the usage of emoticons and uppercases, word lengthening, phonetic features, multi-lingual machine translation, or word embedding. The task usually is to detect the sentiment expressed in a tweet as a whole. But it can also be to identify the sentiment in a tweet concerning a given target concept shown in a query. The difference is that a negative tweet might not say anything about the target concept and must thus be considered neutral concerning the target concept.

In this section, a detailed study of the different technologies used to elaborate the architecture followed by a section to go through the components and specify the role of each one then the third section to present the proposed architecture as well as the algorithms modeled to achieve it. Section four will establish the fundamental view of the results obtained and the last section to conclude and discuss the next perspectives.

II. Technological Spectrum

A. Ease of Use Big Data and Components

1) Hadoop distributed file system (HDFS)

A file system designed [8] to store massive volumes of data across various commodity hardware configurations called nodes.

2) MapReduce

An engine [9] for data processing. A MapReduce job consists of two components, a map phase, which takes raw data and arranges it into a paradigm of key/value pairs, and a reduce stage which processes data in parallel.

3) Apache Pig

A tool/platform [10] that can analyze large datasets and represent them as data streams.

4) Flume

Flume, developed initially by Cloudera, is today a project [11] of the Apache software foundation.

Flume is a product that can inject large volumes of data into Hadoop in real time. By design, close to that of an HDFS cluster, Flume is:

- Reliable: in case of failure of one of its components, Flume can continue to feed HDFS.
- Evolutive: Flume performance can be increased by adding scaling outs.
- Extensible: by default, Flume can ingest data from various sources (local files, HDFS files, system logs, stdout ...) and, if necessary, additional connectors can be developed.

Flume is composed of agents. Each agent has a source (source), a destination (sink) and a channel:

A source can be a data source (firewall, mail server, web server ...) or another agent.

A destination may be another agent or an HDFS file.

A channel is a path followed by data between a source and a destination: a channel can write its data to RAM or disk, depending on the user's needs regarding performance and reliability, the process is described in Fig. 1.

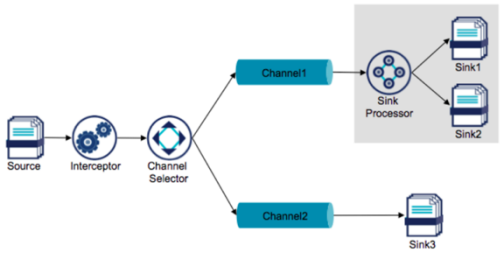


Fig. 1. Flume data process.

Flume offers different levels of reliability:

- Support: In the event of a node failure, all data being received or sent by this node is lost; it is the mode that offers the least guarantees and generates the least overhead.
- Store on failure: In this mode, a source node expects to receive an acknowledgment from the destination node; if not, the data to be transferred is saved on the hard disk until the destination node returns "life sign" or is replaced by another node. This mode does not protect against silent or cascade failures.
- End-to-end: This method ensures that data supported by a source node will reach its final destination, provided that the source node does not experience a crash immediately after data support. This is the mode that offers the most guarantees and generates the most overhead.

5) Mahout

One of the significant well-known tools [12] for ML. It is known for having a wide selection of robust algorithms, math environment named Samsara, which carries linear algebra, statistical operations, and data structures. The goal of the Mahout-Samsara project is to help users build their own distributed algorithms, rather than just a library of already-written implementations. They still propose a comprehensive suite of algorithms for MapReduce.

The algorithms included in Mahout, focus primarily on classification, clustering, and collaborative filtering, and have been shown to scale as much as the size of the data increases. Additional tools include topic modeling, dimensionality reduction, text vectorization, similarity measures, a math library, and more. One of Mahout's most usually cited assets is its extensibility as shown in Fig. 2 as a comparative study of the different ML frameworks, and many have achieved positive results by building o of the baseline algorithms.

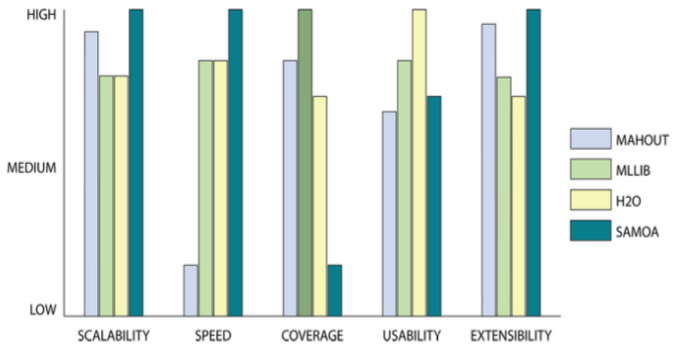


Fig. 2. Mahout vs the others (comparative).

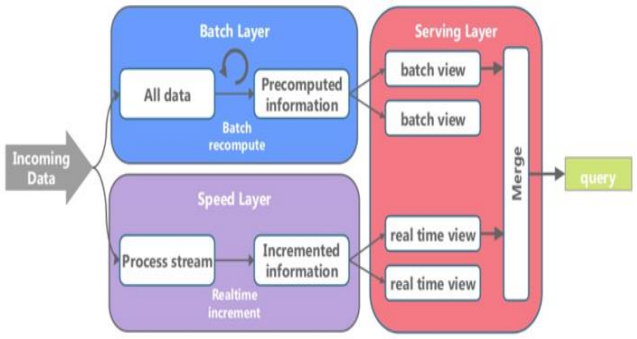


Fig. 3. Lambda architecture layers.

III. Approach

Data analysis is essential for planning and creating recommendation systems or decision-making systems to optimize the underlying infrastructure. This involves not only the processing of data online, looking for specific events but also the historical data sources that may be needed to find data profiles influencing decisions.

Researchers have often merged techniques with other tools to develop field-related solutions. In particular [13], the Twitter API had been extended to third researchers to deploy their analysis of flow data from Twitter to improve business practices. However, unique solutions that allow multiple users in different environments to write and use optimized data processing applications are still needed. Nevertheless, there is a need for tailored solutions for online and batch data processing that maintain non-functional attributes such as network costs and complexities.

Presented as a software design model, the Lambda architecture [14] unifies real-time and batch processing in a single framework. It was founded to be a hybrid system. The model is suitable for applications where there are delays in the collection and availability of data in dashboards, which requires the validity of data for on-line processing as it happens. The model also allows batch processing for older data sets to find patterns of behavior according to user needs.

Some of the critical requirements in the construction of this architecture include:

- Fault tolerance against hardware breakdowns and human mistakes.
- Support for a multiple diversity of use cases that involve low latency queries and updates.
- Linear scalability, which means that the launch of additional machines will solve the problem without degrading the performance of resources.
- Scalability so that the system is manageable and can quickly adapt to new features.

From a high-level viewpoint, the figure below shows the underlying architecture of how the lambda architecture works.

Three layers are treated as detailed in Fig. 3:

1) Batch layer

Batch deduct for large quantities of data sets. It provides management of the Master Dataset; an immutable and exclusive raw data set, but also includes pre-computation of arbitrary query functions, called batch views.

2) Serving layer

Real-Time calculation (Speed Time) to minimize latency by performing calculations in real time as the data arrives. This layer indexes the batch views so that they can be queried in Ad-Hoc with low latency.

3) Speed layer

Responses to requests, interfacing, query and providing the results of calculations. This layer accepts all requests that are subject to low latency requirements. Using fast and incremental algorithms, Speed Layer processes only recent data.

Each of these layers can be obtained using various large data technologies.

To technically reflect the aspect of the Lambda Architecture in the solution provided in Fig. 4 this equivalent Lamda design with the respective big data component to each layer.

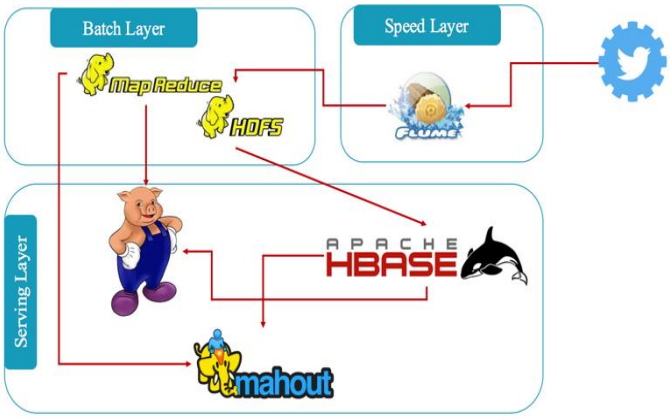


Fig. 4. Equivalent Lambda architecture.

IV. Solution

As a solution, this architecture describing the process of the several steps from capture to data analysis in Fig. 5:

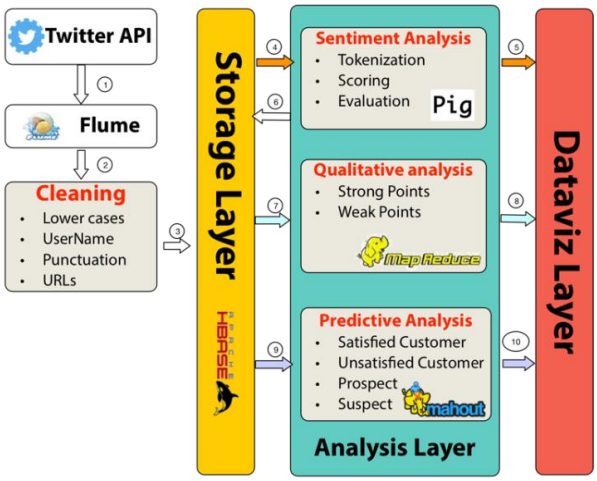


Fig. 5. Solution architecture process.

A. Data Capture

The development of data mining applications requires a covering of issues related to data access, mainly when application policies result in a lack of data to analyze. In other words, the data is not always open.

The exploitation of social networks is a little less worrying compared to other corporate environments as most social media platforms offer well-designed language agnostic APIs that provide access to the data needed.

The availability of this data obviously depends on how users share their data and how they allow us to access it. For example, Facebook users can decide the level of privacy of information about their public profile and the details that can be displayed only to their friends. Profile information such as birthday, current location and work history (and many others) can all be reported individually as private or public. Similarly, when trying to access this data through the Facebook API, users who sign up to the application can grant us access only to a limited subset of the requested data.

On the other hand, the Twitter API is open and allows access to all user data. Using Flume, by retrieving data from various services and transport them to centralized stores (HDFS and HBase) [15]. More precisely, extracting data from the Twitter service and store it in HDFS using Apache Flume as illustrated in the following Fig. 6.

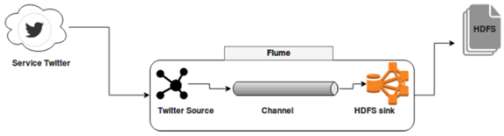


Fig. 6. Twitter data capture.

To get Twitter tweets, it is necessary to create a Twitter application, then start the services of Hadoop and finally configure Apache Flume.

B. Data Pre-processing

Data pre-processing involves eliminating unwanted fields and special characters so that the feeling analysis is as reliable as possible. This is called noise elimination or cleaning. To do this, the following adjustments will be satisfied:

- LowerCase: To convert tweets to lowercase.
- URLs: To remove all URLs via a match or a regular replacement expression with a generic word URL.
- @username: To remove "@username" via regex matching or replace it with a generic word like AT-USER.
- Additional Punctuation and White Spaces: Punctuation at the beginning and end of the tweets should be removed. We also replace several white spaces with unique white space.
- Eliminate unnecessary fields: As previously stated, the tweet contains information and fields that will not provide any benefit to

C. Storage

As previously explained, storage must be at HBase level. Now all tweets are cleaned and ready to be stored. To get started, firstly by preparing the database and configure it, create region servers, master backups, and tables. The files are stored in the tables of HBase, and the tables can be managed using the 'hbase shell' commands.

D. Data Analysis

1) Opinion Mining

The opinion mining will be done with the Apache Pig tool. This one will provide us the classification of the tweets according to the feelings; positive if the polarity of the total words constituting the tweet is positive, neutral if it is zero and negative elsewhere. The data (tweets) are stored in Hbase. To perform sentiment analysis, using the Hbase table where the tweets are stored. The first thing to do with a tweet is what is called Tokenization or splitting.

This operation consists of dividing a string (which contains a group of words) into a row and returns a result holding the output of the split operation. Here is an illustration of the output of this treatment.

- First, starting with loading tweets in Pig, and the next step is the extraction of the id and the text of the tweet from the complete tweet.
- Working with the function tokenize of Pig which refers to the words needed to work with.
- Now since having the tweets or rather the words, the treatment sentiment analysis can start.
- This processing is done as follows, and we classify the words treated from -5 to 5 using the AFFIN dictionary [16] (is a dictionary that contains 2500 words ranged from -5 to 5: from 0 to -5 it is negative, 0 neutral, from 0 to 5 this is positive).
- Loading the AFFIN dictionary.
- The next step is to link each token of the previous step by its ranking word which is in the AFFIN dictionary.
- Extracting that id, text, and score.
- The next step is the calculation of the score of the whole tweet, starting with the grouping of the words of each tweet by the id of the latter:

- Calculating the average of all the tweet, And since having the average, the filter can be applied according to user's needs.

Performing the rating on all the words of each tweet as follow:

Algorithm Algorithm for input and sentiment analysis

Input: KeyWords

Output: (id,sentiment)

Extraction :

1: Cleaning

Tokenization :

2: Token(tweet)

Loading :

3: Load(AFFIN Dictionary)

Rating :

4: Rate(word ,dictionary)

Group Rating :

5: GroupRate(id ,word)

LOOP Process

6: **for each** GroupRate **do**
7: AVGrate tweetRating
8: **if** (tweetRating > 0) **then**
9 print(id ,positive)
10: **else**
11: **if** (tweetRating < 0) **then**
12 print(id ,negative)
13 **else**
14 print(id ,neutral)
15: **end if**
16: **end if**
17: **end for**

2) Qualifying Analysis

The qualification service is a service that allows visualizing the different keywords with varying sizes according to their frequencies in the set of tweets. Indeed, it is also a MapReduce program which looks for the occurrence of a keyword in the tweets and calculates its frequency of occurrence, then gives each word a value.

This value represents the proportion of size with which it will be written or displayed. The processing is done on all tweets, resulting in a Wordcloud. As sentiment analysis processing, it is an R program that will be able to perform a Word Cloud based on the output of the MapReduce program that is stored in a file in HDFS. Indeed, R contains a library called 'wordcloud', the latter as its name indicates it is responsible for the creation of WordCloud.

3) Predictive Analysis

This service is dedicated to the prediction of user profiles such as Satisfied Customer, Unsatisfied Customer, Prospect, Suspect.

Mahout is the right tool for this system, since it implements the Naive Bayes algorithm, which was chosen as the prediction algorithm during the comparative study, to define the classification mechanism, because the algorithm is of the supervised type, we will start with a set of observations of many variables, until we can assign a new observation to a particular category.

Using this set, the classifier determines the probability that a user belongs to one of the declared categories for each word. To calculate the probability that a user belongs to one of the categories, it multiplies the individual probability of each of its elements in that category. The category with the highest probability is the category that the user is most likely to belong to.

V. Experimental Evaluation

All the experiments were performed using a 2,5 GHz Intel Core i7 processor and 16GB of RAM running on a Cloudera virtual machine.

A. Opinion Mining

From this analysis, we will get all tweets, both positive and negative. To classify them, the positive tweets are those with a rating between 1 and 5. Obviously, the negative tweets will have a score between -5 and -1. The rest is considered neutral.

In Fig. 7, we can see the tweet text field and its score. We managed to classify the tweets according to the emotion in the text field.

```
((argan and grapeseed what a lovely combination I bet it smell wonderful), 3.
5)
((wonderful natural hair oil for damaged hair damaged hair growth argan gomed
iderm), -0.5)
((give away give away give away giveaway argan arganoil  I hate oil),-4.0)
((fx11 argan based mascara to help your eyelashes flourish! pop in today and
try),1.5)
((try the organic natural argan face cream from aromabguk order from us moroc
co ),0.5 )
((love argan oil? remember thait only made in morocco if icomes from another
country its probably nothe real thing), 2.0)
```

Fig. 7. Sentiment analysis

Now all that remains is to present them on a graph. It is a MapReduce program that will do this; it counts all tweets with a positive score and those with a negative rating, then passes the result to a program written in R and finally has an understandable diagram. In this work, here is the result obtained in Fig. 8.

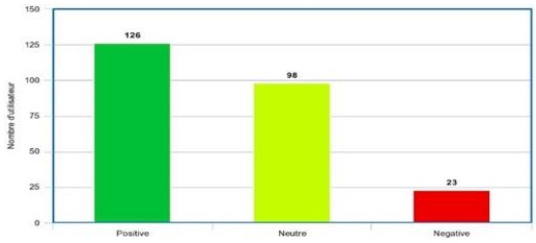


Fig. 8. Scoring Algorithm results

B. Qualifying Analysis

To go further, we were able to isolate positive tweets from negative tweets and then perform the same treatment on each of them. This will allow us to detect problems with the product just by looking at the negative WordCloud. On the other hand, it will help us to know the strong points of the product through WordCloud positive. Here is the result in Fig. 9.

Fig. 9. WordCloud result.

C. Predictive Analysis

Practically speaking, we started by preparing a Training Set. The most tedious task was to manually sort all tweets by category. We have chosen to classify a hundred tweets retrieved via the keyword (number of elements recommended for the correct functioning of the algorithm), the result of this sorting is stored in HDFS, in fact, it is a file that contains the following fields: User id that published the tweet, tweet feeling (positive, negative) obtained from the analysis of the previously processed feeling, another field that indicates if the user is already a customer of the company or not, and finally, to transform his data into a Training Set, we add the last field that represents the profile of the user (satisfied customer, unsatisfied customer, prospect, and suspect). The classifier determines for each keyword, the probability that a user belongs to one of the categories considered (satisfied customer, unsatisfied customer, suspect, and prospect) which allows us to deduce the profile of the person who published this tweet.

Table I below shows an extract of the data from the Training set:

TABLE I. TRAINING SET DATA

Ids	Sentiment	Client	Profil
id1	Positive	Yes	Satisfied
Id2	Negative	Yes	Non Satisfied
Id3	Negative	No	Suspect
Id4	Positive	No	Prospect
…	…	…	…

Finally, we use Mahout to transform this Training Set to a vector, the tweets-vectors directory of the HDFS contain sequences on which the Naive Bayes algorithm will be based for its calculations as described in Fig. 10.

```
$ mahout testnb -i train-vectors -m model -l labelindex -ow -o tweets-testing -c
[...]
Summary
-------------------------------------------------------
Correctly Classified Instances          :        135       97.8260%
Incorrectly Classified Instances        :          3        2.1739%
Total Classified Instances              :        138

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       <--Classified as
33      0       0       0       |  33          a     = clientSatisfait
0       18      0       1       |  19          b     = clientNonSatisfait
0       0       48      0       |  48          c     = prospect
1       1       0       36      |  38          d     = suspect
```

Fig. 10. Final classified results.

VI. CONCLUSION

In this paper, we provided a new architecture for a real-time recommendation system based on a social media case study: Twitter.

This architecture provides a new vision of massive data processing and opens the door to new research horizons. Analyzing the full scenario for social data processing.

The recommendation system applications have been designed and implemented in such a way that they complement each other, to become involved in a functional structure for Big Data analysis. The system has demonstrated its ability to perform a quantitative (and qualitative) analysis of the user's behavior to better understand their needs.

As a new perspective, we will concentrate the research on the user behavior field and loyalty by implementing the aspects of gamification for an improved recommendation.

REFERENCES

[1] R. Mukherjee, "A Comparative Review Of Data Warehousing ETL Tools With New Trends And Industry Insight," 2017 IEEE 7th Int. Adv. Comput. Conf., pp. 944–949, 2017.

[2] T. P. Pavlic and S. C. Pratt, Handbook of Human Computation. 2013.

[3] J. Chin, V. Callaghan, and I. Lam, "Understanding and personalising smart city services using machine learning, The Internet-of-Things and Big Data," 2017 IEEE 26th Int. Symp. Ind. Electron., pp. 2050–2055, 2017.

[4] H. Blockeel and K. U. Leuven, Encyclopedia of Machine Learning and Data Mining. 2017.

[5] N. Schrape, Gamification and Governmentality. 2014.

[6] D. Hutchison, LNCS 8462 - Persuasive Technology 2014. 2014.

[7] E. Fredericks and D. R. Schneider, Supporting Real Time Decision-Making, vol. 13. Boston, MA: Springer US, 2011.

[8] Y. Mao and W. Min, "Storage and Accessing Small Files Based on HDFS," in Proceedings of International Conference on Computer Science and Information Technology, 2014, pp. 565–573.

[9] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. 6th Symp. Oper. Syst. Des. Implement., pp. 137–149, 2004.

[10] S. Wadkar and M. Siddalingaiah, "Data Processing Using Pig," in Pro Apache Hadoop, Berkeley, CA: Apress, 2014, pp. 241–269.

[11] D. Vohra, "Apache Flume," in Practical Hadoop Ecosystem, Berkeley, CA: Apress, 2016, pp. 287–300.

[12] S. Rallapalli, R. R. Gondkar, and G. V. Madhava Rao, "Cloud Based K-Means Clustering Running as a MapReduce Job for Big Data Healthcare Analytics Using Apache Mahout," 2016, pp. 127–135.

[13] J. Littman, D. Chudnov, D. Kerchner, C. Peterson, Y. Tan, R. Trent, R. Vij, and L. Wrubel, "API-based social media collecting as a form of web archiving," Int. J. Digit. Libr., Dec. 2016.

[14] J. Kroß, A. Brunnert, C. Prehofer, T. A. Runkler, and H. Krcmar, "Stream Processing on Demand for Lambda Architectures," 2015, pp. 243–257.

[15] N. Maheswari and M. Sivagami, "Large-Scale Data Analytics Tools: Apache Hive, Pig, and HBase," in Data Science and Big Data Computing, Cham: Springer International Publishing, 2016, pp. 191–220.

[16] S. Paul and S. Deshmukh, "Domain Independent Joint Sentiment And Topic Detection Using AFINN And Other Lexicons," IOSR J. Comput. Eng. Ver. I, vol. 17, no. 4, pp. 2278–661, 2015.

Machine Learning Method to Screen Inhibitors of Virulent Transcription Regulator of *Salmonella* Typhi

Syed Asif Hassan
Department of Computer Science, Faculty of Computing and Information Technology Rabigh (FCITR)
King Abdulaziz University, Jeddah, Saudi Arabia

Atif Hassan
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur,
Kharagpur, West Bengal, India

Tabrej Khan
Department of Information Sciences,
Faculty of Computing and Information Technology Rabigh (FCITR)
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—The PhoP regulon, a two-component regulatory system is a well-studied system of *Salmonella enterica* serotype typhi and has proved to play a crucial role in the pathophysiology of typhoid as well as the intercellular survival of the bacterium within host macrophages. The absence of PhoP regulon in the human system makes regulatory proteins of PhoP regulon for target specific for future drug discovery program against multi-drug resistant strains of *Salmonella enterica* serotype typhi. In recent years, high-throughput screening method has proven to be a reliable source of hit finding against various diseases including typhoid. However, the cost and time involved in HTS are of significant concern. Therefore, there is still a need for an expedient method which is also reliable in screening active hits molecules as well as less time consuming and inexpensive. In this regards, the application of machine learning (ML) based chemoinformatics model to perform HTS of drug-like hit molecules against MDR strain of *Salmonella enterica* serotype typhi is the most applicable. In this study, bagging and gradient boosting based ML algorithm was used to build a predictive classification model to perform virtual HTS of active inhibitors of the PhoP regulon of *Salmonella enterica* serotype typhi. The eXtreme Gradient Boosting (XGBoost) based classification model was comparatively accurate and sensitive in classifying active drug-like inhibitors of PhoP regulon of *Salmonella enterica* serotype typhi.**

Keywords*—Typhoid; PhoP regulon, classification model; machine learning (ML) algorithm; eXtreme Gradient Boosting; random forest; sensitivity; accuracy

I. INTRODUCTION

Typhoid is an endemic disease of developing nations caused by *Salmonella enterica* serotype typhi. According to recent WHO data, around 21 million people are infected with *Salmonella* Typhi, and approximately 2, 22, 000 people die annually across the globe [1]. The multidrug-resistant (MDR) strain of *S. typhi* has spread rapidly and has become a major endemic problem in South East Asia and Indian subcontinent [1]. Therefore, the target based screening of novel anti-typhoidal compound with a higher potential to destroy MDR strains of *S. typhi* causing MDR typhoid fever is of prime importance.

The two-component (PhoQ-PhoP) regulon is a crucial virulence regulatory system of *S. typhi* regulating the expression of more than 120 different genes involved pathogenicity of *S. typhi* within the host cells [2]. The PhoP regulon consists of an environmental sensor histidine kinase (PhoQ) that in response host defensins (abundant in macrophages), low level of periplasmic Mg^{2+} ions, acidic pH is activated upon autophosphorylation at the conserved histidine residue present in the cytoplasmic domain of PhoQ protein [3]. Consequently, the PhoP a response regulator of the PhoP regulon is phosphorylated at the aspartate residue present at the conserved N-terminal domain of PhoP protein by accepting a phosphate group from PhoQ protein [4]. The phosphorylated PhoP regulates the transcription of corresponding genes involved in the intracellular survival [5]-[6] and virulence of the *S. typhi* within host cells [7]-[9]. The PhoP/PhoQ operon based virulence regulatory system is not present is present only in a bacterial system; therefore the PhoP regulon has gained significance as a potential target for antibacterial drug discovery program.

ML algorithms are robust and fast in dealing with high dimensional data. Since the chemical dataset used for screening of drug-like lead molecules during the earlier stages of drug discovery involves high dimensional data, i.e., comprising a large number of two dimensional and three-dimensional chemical attributes. Therefore, ML-based methods are most appropriate in categorizing inactive and active compounds from a given library of chemical compounds. Chemoinformatics models based on ML algorithms has been suitably applied in the past to screen as well as rank active hit molecules during the lead molecule identification stages of drug discovery and development program. In this regard,

Garcia-Sosa et al. 2012 [10] used multivariate logistic regression methods to classify active and inactive drug-like molecules. On the other hand, Korkmaz et al. 2014 [11] used a combination of various feature selection method with Support Vector Machine (SVM) to discriminate active and inactive drug-like molecule from on similar dataset. Further, Korkmaz et al. 2015 [12], proposed a web tool (MLViS) using the best ML-based classification algorithms to screen active drug-like molecule during the early stages of drug discovery protocols. A comparative study to evaluate the performance of SVM and Neural network (NN) based classification model to discriminate between a drug-like and nondrug-like was conducted by Zernov et al. [13] and Byvatov et al. [14]. They both showed that SVM based model performed better in classifying drug-like molecule from the non-drug-like molecule. Similarly, SVM algorithm based classification model was also used to classify inhibitors of cytochrome P450 [15], lymphocyte-specific protein tyrosine kinase [16] and butyrylcholinesterase [17]. On the other hand, other ML algorithms such as k-Nearest Neighbor (KNN) [18], NN [19-20] and Naïve Bayes (NB) [2], were used to classify active inhibitors from non-inhibitor molecules. Likewise, SVM algorithm based predictive model was used by Rathke *et al.* [22], Wassermann *et al*. [23], Jorissen and Gilson [24], and Agarwal et al. 2010 [25] to evaluate chemical compound based on their activity. Similarly, Abdo et al. 2010 [26] and Plewczynski et al. 2009 [27] have applied Random Forest (RF), and Bayesian neural network (BNN) based predictive model for predicting the activity of chemical molecules. Additionally, Harlen et al. 2012 [28] used Random Forest (RF) algorithm to build a predictive model to classify active and inactive chemical molecule against the PhoP regulon of *S. typhi* from the HTS bioassay dataset. The accuracy, sensitivity, and specificity obtained using the RF classifier based model was 81.5%, 87.7%, and 81.5%, respectively. In this context, an improved classification model built using the supervised ML-based algorithms (XGBoost and RF) have been proposed to classify active inhibitors of PhoP transcriptional regulatory system protein (PhoP) with higher accuracy, sensitivity and specificity than proposed and built by Harlen et al. 2012 [28]. Since the number of the active molecule with a potential to inhibit PhoP regulon was less as compared to their inactive counterpart in the AID-1850 dataset, therefore, the dataset was balanced using Synthetic Minority Over-sampling Technique (SMOTE) algorithm prior applying supervised ML algorithm for model building. The basic idea of the proposed model is to build a less expensive and robust predictive classification model which will be potent in screening active inhibitors of PhoP regulon and thus will save time and money for identifying lead molecule during the early stages of anti-tuberculosis drug discovery program. The present original research article is sectioned into four sections: In order to overcome the problem associated with the high-cost experimental screening protocols, the current research work in Section II. Firstly, defines the dataset used for building the chemoinformatic classification model; secondly apply SMOTE algorithm to balance the dataset since the AID-1850 dataset is highly imbalanced, and finally discuss various Classification algorithms namely RF and XGBoost used to construct the current supervised classification model. Section III explains the results of the statistical model performance evaluators for the classification model build using balanced bioassay dataset, and Section IV provide the concluding statements about the proposed classification model as well as the future scope of the proposed model. A pictorial representation of the workflow diagram involved in constructing the supervised classification model for classifying active inhibitors of PhoP protein is summarized and is shown in Fig. 1.

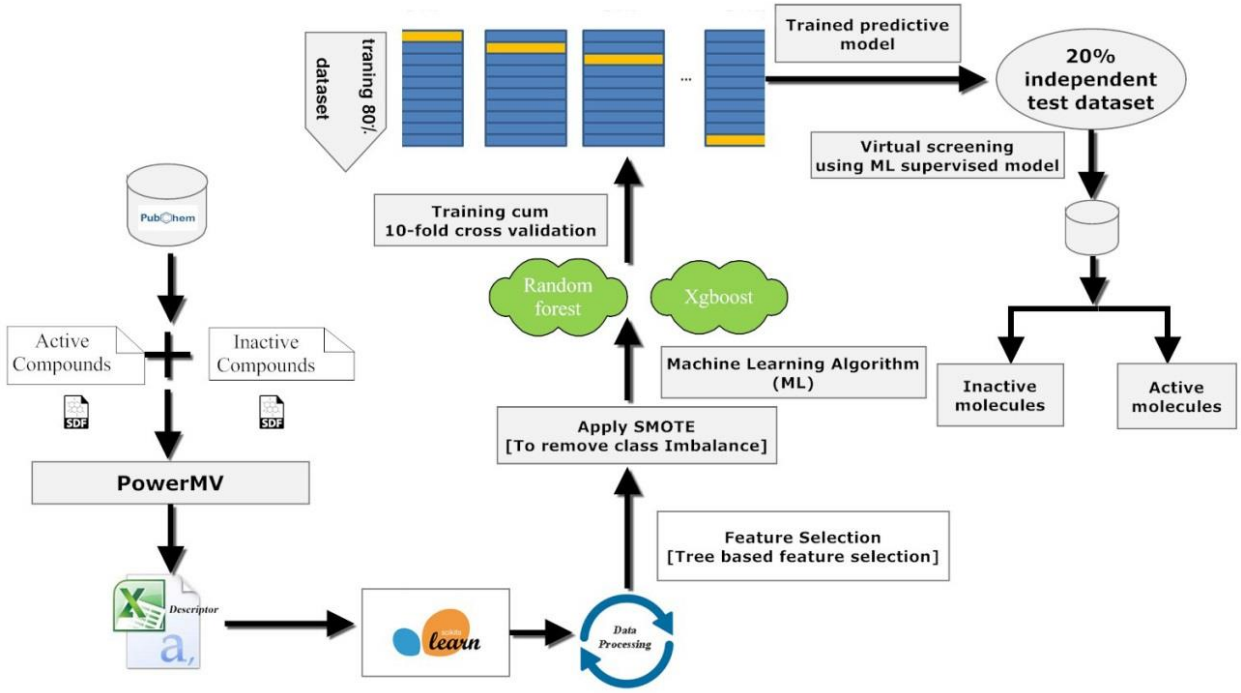


Fig. 1. A pictorial representation of workflow to illustrate the methods required to build a supervised classification model to screen active inhibitors of PhoP operon proteins.

II. Materials and Methods

This part describes the dataset and defines the process to pre-process and balance the dataset. This part also explains the ML algorithm used to build the supervised classification model as well as describes the statistical evaluators used to access the performance of the ML-based classification models.

A. Data Source

The dataset AID=1850 was obtained from the PubChem BioAssay Database of the National Center for Biotechnology Information (https://pubchem.ncbi.nlm.nih.gov/bioassay/1850). A total of 306568 compounds were screened for a compound that inhibits the PhoP operon in *S. typhi.* The compounds based on a percentage of inhibition were classified into the active and inactive molecule. The molecules which showed > 30% inhibition in the confirmatory PhoP dose response assay were considered as active while the molecules which showed less than 30% of inhibition in the dose-response test were deemed to be inactive. Therefore, the trial generated 1021 active inhibitors of PhoP regulon and 305404 inactive compounds in the confirmatory bioassay.

B. Attribute Generation

The Structural-data files (SDF) of both active and inactive compounds from AID-1850 dataset were downloaded from PubChem bioassay dataset [29]. The molecular descriptor file for both active and inactive compounds of the AID-1850 dataset was generated using PowerMV, a Graphical User interface (GUI) based software for molecular descriptor generation and visualization [30]. A Perl script based Mayachemtool[1] was used to split the sizeable structural-data file of the inactive compounds into smaller structural-data files. Using PowerMV, 179 molecular descriptors (attributes) were generated from the chemical-data record for each active and inactive compounds of the AID-1850 dataset. Bit string and continuous calculation method were used create the molecular descriptor file for each active and inactive compounds present in the AID-1850 dataset. A total of 147 molecular descriptors generated based on pharmacophore fingerprinting were represented as a bit string, i.e., 0 and 1. Where the bit string "1" signifies the occurrence of a specific feature/ fragment and "0" represent the absence of that specific fragment/feature. Twenty four weight burden number and eight chemical properties based continuous molecular descriptors were generated using PowerMV. A list of eight property based descriptors namely the number of rotatable bonds, Polar surface area (PSA), XLogP, molecular weight, a molecule containing the toxic group (bad group indicator) and blood-brain indicator represented by 1 and 0. Here the discrete value "1" displays the ability of the molecule to cross the Blood-Brain Barrier (BBB) and while "0" indicates the inability of the compound to pass the BBB. The molecular descriptor file of each active and inactive compounds consisting of 179 descriptors was combined and saved as Comma-separated Value (CSV) file for further processing. An extra column depicting the outcome (bioactivity) of each instance (compound) was appended. The inhibitors of PhoP compounds were given a nominal value "active," and the non-inhibitors of PhoP were labeled as "inactive."

C. Processing of Clinical Dataset

1) Data Preparation

The molecule id column was removed from each matrix, as it does not contribute to the feature list. The combined file CSV consisting of active and inactive molecules was preprocessed to remove the duplicate instances. As a result, 352 duplicate samples including 11 active compound samples were removed from the dataset. A quick count shows that a total of 300992 samples were present with the majority class being the inactive compounds and occupying 98.34% of the sample space whereas the 1010 active compounds being the minority class hold only 1.66% of the total sample space. Further, the final dataset after removal of duplicate instances was subjected to filtration of non-informative attributes to improve the efficiency of the model generated using ML tools [31], [32]. The removal of non-informative attribute reduces the feature space of AID-1850 dataset to 154 attributes. The final list of 154 attributes are enlisted and shown in Table I (Supplementary File).

2) Dimensionality Reduction

Using all features of a given dataset is not an efficient model building process as higher dimensions add to the complexity of the final classifier which leads to longer computation time while unimportant features reduce the model performance/accuracy. Since the feature space of the dataset was 154, therefore a tree-based feature selection module[2] to reduce the dimensionality of dataset to only 43 features listed in Table II (Supplementary File). When an attribute is used as a decision node in a tree, its relative rank/depth can determine how important it concerns the prediction of the target variable. Since the features used at the top of a decision tree affect the final prediction of a large number of input data. Thus, the fraction of samples that they influence can be used to estimate the importance of each feature against one another. By creating some randomized trees and averaging the importance value of each feature, a more robust feature selection model with lower variance can be constructed. A pictorial representation of the scoring based selection of attributes using Tree-based feature selection method is shown in Fig. 2.

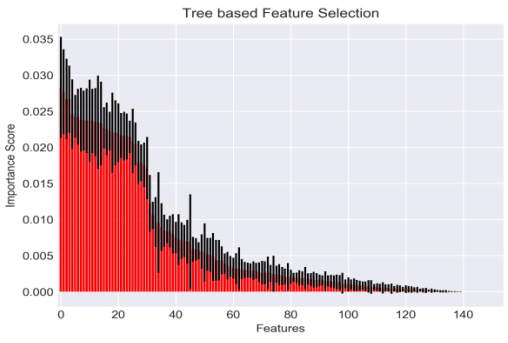


Fig. 2. Illustrates the score obtained by the 154 features of the AID-1850 dataset using Tree-based feature selection module of Scikit-learn package.

[1] http://www.mayachemtools.org/

[2] http://scikitlearn.org/stable/modules/feature_selection.html#tree-based-feature-selection

3) Class Balancing

A dataset is called imbalanced if the numbers of target classes are not nearly equally represented. As the present dataset was highly imbalanced with the majority class taking up 98.34% of the total sample space, SMOTE algorithm was used to balance the AID-1850 dataset by creating synthetic instances from the minority class from the AID-1850 dataset. SMOTE is an oversampling method which creates synthetic instances of the minority class rather than oversampling the class through replacement. Oversampling is done by randomly choosing a minority sample from the given data and finding its k nearest neighbors [33], [34]. In the present study, k equal to 5 was used. A sample is generated on the line segment joining any or all k neighbors by multiplying the difference between the selected feature vector (instance) and its nearest neighbor with a random value in the range of 0-1. In the present study, a random value 0.5 was selected and multiplied by the feature vector under consideration leading to the generation of a new sample. Similar action is performed for all/any neighbors which effectively forces the minority class decision region to become more general. The final dataset, after completing class balancing consisted of 50% active and 50% inactive instances. The pseudo code for generating a synthetic sample is as follows:

Let a feature vector $\vec{a}$ represents the instance under consideration. Find its k nearest neighbors and select one of them. Let this instance be represented as a feature vector $\vec{b}$. Then the new sample $\vec{c}$ will be equal to

$$\vec{c} = \vec{a} + (\vec{b} - \vec{a}) * rand(0, 1) \quad (1)$$

Where, "rand (0, 1)" represents a random value between 0 and 1.

D. Data Partitioning and Cross-validation Procedures

The balanced dataset was segmented into train and test sets as 80% and 20% respectively. The train set would be used to train the model while the test set, data never before seen by the model, would be used for testing its accuracy/performance. The train set was used for k-fold cross-validation; in the present case, k is 10. Thus, in 10-fold cross validation, one fold is used for testing purpose while the rest 9 (k-1) folds are used for training. This process is repeated until all folds have been tested. The average accuracy taken over all folds gives a more reliable measure than a single training and testing phase. This action can also be used to verify if the final, trained model overfits the test set or not.

E. Model Building Algorithm

Classification is the process of segregating a sample, based on its attributes into the given target classes. In this study, two algorithms namely Random forest and XGBoost were used to perform classification where both of which were based on the concepts of decision trees. A decision tree is a flowchart-like tree structure whose internal nodes are attributes which are used to split the tree further based on a threshold and whose leaf nodes are the target classes. The feature selected at each level for further splitting is determined by calculating information gain for each attribute and the one with maximum information gain is selected. Calculation of the information gain for each feature is done by calculating the entropy of all possible values of an attribute and then finding its information gain. Entropy is the amount of homogeneity of a sample while information gain is the difference of entropies before and after a split. The attribute which yields the maximum information gain is chosen as the root node.

$$E(X) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (2)$$

Entropy using the frequency table of one attribute

$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad (3)$$

Entropy using the frequency table of two attributes

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (4)$$

Information Gain

1) Random Forest (RF)

A random forest is a forest/collection of decision trees and works on the concept of bagging. Some un-correlated classifiers/learners can be used together to form a better classifier with less variance and reduce overfitting [35]. In a random forest classifier, the training samples are divided into some random subsets based on which decision trees are created. The target class for a particular instance is then selected based on the maximum voting scheme wherein each tree outputs a class, and the one with the highest votes is chosen as the target class for the given sample. This reduces variance in the present model while also decreases overfitting, a problem which usually occurs in decision trees.

2) XGBoost

XGBoost is an ensemble algorithm which is majorly used in kaggle competitions as it provides excellent performance out of the box and has some parameters for tuning. XGBoost is based on the concept of gradient boosting [36]. Gradient Boosting is a technique which uses an ensemble of weak classifiers, models which perform slightly better than random guesses, to create a strong classifier. Here CART (Classification and Regression Trees) was used for preparing an ensemble algorithm for classification. In boosting, a model is built to optimize a differentiable loss function, at runtime. In the next stage/iteration, a new model is developed to optimize the loss function from the previous step further. This process continues until a threshold is reached. In this way, the errors committed by the earlier models can be corrected by the models in the next stage. XGBoost works on the idea of gradient boosting but differs from the fact that it uses regularized models to control overfitting which gives better performance. At the same time, XGBoost or eXtreme Gradient Boosting is called so because it utilizes other computational techniques such as cache access patterns, data compression, etc. to push computational boundaries and achieve state of the art performance regarding speed and accuracy.

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (5)$$

Where, g_i and h_i are inputs defined by

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (6)$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (7)$$

$f_t(x)$ is given by

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \rightarrow \{1, 2, \ldots \ldots, T\} \quad (8)$$

Where w is the vector of scores on the nodes, T is the number of leaves and q is a function assigning each data point to the corresponding leaf, While $\Omega(f_t)$ is the regularizer term, given by

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (9)$$

The ML classification algorithm, the data preprocessing and post-processing data analysis are performed in Scikit-learn tool for data mining and analysis. Scikit-learn tool is a state-of-the-art implementation of many supervised and unsupervised ML algorithms written in and for Python. It has an easy-to-use interface and is a go-to choice for exploratory data analysis. It is increasingly used by academicians as it allows for more time on designing of algorithms than their implementation [37]. Due to the exceedingly imbalance characteristics of the AID-1850 data, the Imbalanced-learn tool was used tackle the problem of class imbalance in the AID-1850 dataset. The imbalanced-learn toolkit is an implementation of some popular re-sampling techniques which is useful for datasets with highly imbalanced classes [38]. It is a Python package and is compatible with Scikit-learn tool and can be downloaded from http://contrib.scikit-learn.org/imbalanced-learn/stable/.

F. Evaluation of Model Performance

The ML-based predictive model trained using XGBoost, and RF classifier was evaluated using statistical model performance evaluators present in Scikit-learn data mining tool. For two-class problem the 2x2 confusion matrix consists of the following sections: (1) True Positive (TP) in this study, is the active inhibitors of PhoP appropriately classified by classification model as active class; (2) False Positive (FP) actually non-inhibitors of PhoP operon but incorrectly classified as active by the predictive model; (3) True Negative (TN) actual non-inhibitor of PhoP (inactive molecule) correctly classified by the model as non-inhibitor (inactive) and lastly, (4) False Negative (FN) actually inhibitor molecule (active molecule) but incorrectly classified by the classification model as non-inhibitor (inactive). In this context, the True Positive Rate (TPR) is defined as the proportion of TP (i.e., active inhibitors of PhoP regulon predicted correctly by the classification model) from the total population of inhibitors of PhoP regulon and is estimated as TP/TP + FN. Similarly, False Positive Rate (FPR) is defined as the fraction of FP (i.e., erroneously categorized as active inhibitor molecule) when compared to the total number of predicted inactive chemical molecule and the FPR is estimated as FP/FP+TN.

Sensitivity, another model statistical evaluator, represents the capacity of the ML-based predictive model to correctly classify the inhibitors (True Positives) of PhoP regulon from the instances given in AID-1850 dataset and is estimated as (TP/TP+FN)*100. On the other hand, specificity refers to the capability of the ML-based predictive model to classify inactive (non-inhibitor) molecules from AID-1850 dataset correctly and is estimated as (TN/TN+FP)* 100.

Accuracy is another statistical evaluator to measure the ability of the model to correctly classify the TN and TP instances from the total number of predicted instances (TP+TN+FP+FN) and is calculated as

$$([TP+TN])/ [TP+TN+FP+FN])*100 \quad (10)$$

The ideal accuracy value for any classification model is one. The Receiver Operating Characteristics (ROC) graph defines the consistency of the model to efficiently discriminate between two classes by Area under the Curve (AUC) value. The AUC value is generated by making a graph between True Positive Rate (TPR) on the Y-axis and False Positive Rate (FPR) on the x-axis. The estimated AUC value is the likelihood that the model will assign a higher score to an arbitrarily selected inhibitor compound than to an arbitrarily picked non-inhibitor compound. The score of AUC value ranges from 0 to 1, therefore the model with a score close or equal to 1 will be considered reliable in predicting active compound from the AID-1850 chemical dataset and vice versa for a model with an AUC value close to zero.

G. Determination of Statistically Significant Difference between Models

Estimation of a significant statistical difference between the models generated using XGBoost and RF in predicting active molecule from AID 1850 dataset was determined using two-sample unpaired *t*-test [39]. The accuracy value of each test-fold obtained during the ten-fold training cum-cross validation of XGBoost and RF model were grouped and tested for significant difference using two-sample t-tests at a confidence interval of 95%.

III. Results and Discussion

The HTS dataset AID-1850 was obtained from the publically available bioassay database of PubChem-NCBI. The HTS dataset consisted of 1021 active molecules (inhibitors of PhoP regulon) and 305404 inactive molecules (non-inhibitors of PhoP regulon). The SDF files of the inactive and active molecules were taken from the bioassay database of PubChem. Due to the larger size, the SDF file of the compounds was fragmented into smaller files using MayaChemTool. A CSV file consisting of 179 molecular descriptors for each active and inactive molecule were generated using PowerMV. Finally, each descriptor files of both active and inactive molecule were randomly merged into one CSV file. The Final merged molecular descriptor CSV file of both active and inactive molecule was preprocessed in Scikit-learn platform to remove duplicate instances and noninformative attributes. Since only specific attributes among the total attributes contribute significantly in accurately predicting the desired class. Therefore, in this context Tree-based feature selection module of Scikit-learn package was used to remove noninformative features or attributes from the final merged molecular descriptor file of active and inactive molecules of AID-1850 HTS chemical dataset. The last molecular descriptor file of both active and inactive molecule

consisted of 43 attributes (molecular descriptors) contributing most towards model building were screened using tree-based feature selection method. While an outcome column was labeled as "class" having two discrete values "0" and "1". Here the value "1" denotes an inactive molecule and "0" signifies an active molecule.

The modified molecular descriptor file of both active and inactive molecule with 43 attributes was split into 80% training data and 20% independent test data. Due to the imbalanced nature of the dataset, SMOTE algorithm was applied to generate synthetic instances from the minority class (active molecule) to create a balance between the two classes (i.e., active and inactive molecule) of the dataset. Therefore, the final dataset post SMOTE application consisted of 50% active and 50% inactive instances. The balanced dataset was fragmented into 20% independent test dataset and 80% training dataset. The classification models built using XGBoost and RF algorithms were trained using the 80% training dataset. Furthermore, the classification models trained using 80% training data were tested for their ability to classify active and inactive molecule from the 20% independent test dataset. A comparative performance evaluation of the predictive models with and without SMOTE is tabulated in Table III. All the results tabulated in Table III for each classifier is determined using 20 % independent test data. Due to the imbalanced nature of the dataset, the models tend to be predisposed to the majority class (inactive molecule).The biasness of the models for the majority class can be observed from the results of both FNR and sensitivity. Here, sensitivity reflects the capability of the model to appropriately categorize the True Positive (active molecule) instances from AID-1850 HTS dataset while FNR reflects the prejudiced nature of models for the majority class inactive molecule). In this regard, the models tested on an imbalanced dataset shows a higher rate of FN's (79.6% for RF and 85.7% for XGBoost) and a lower percentage of sensitivity (20.4% for RF and 14.1% for XGBoost). The above results show that the predictive classification models built and tested on imbalance dataset are biased towards the majority class (inactive molecule). However, contrasting results are obtained post-SMOTE application. The sensitivity and the FNR for the classification models constructed using RF and XGBoost algorithm and tested on balanced dataset show a lower value for FNR (0.7 for RF and 2.4 for XGBoost) and a higher percentage for sensitivity (99.2% for RF and 97.8% for XGBoost) as shown in Table III. The higher percentage of sensitivity for RF-based model show its higher ability to accurately predict active molecule (TP) instances from total positive instance (TP+FN) present in AID-1850 HTS dataset when compared to XGBoost based predictive model. The ability of the predictive system to accurately predict TP post SMOTE application can also be determined by the TPR. The TPR value before the application of SMOTE was 25.4% for RF and 14.1% for XGBoost based predictive models. Subsequently, the TPR value post SMOTE application is 99.2% for RF and 97.6% for XGBoost, based classification models. Further, the accuracy and specificity of the models tested on the imbalanced test data have similar value for each model (i.e., 98.5% accuracy for each model and 99.8% specificity for each model). Furthermore, post SMOTE application the RF-based classification model had similar accuracy and specificity value (99.2%), which was comparatively better than accuracy and specificity, obtained using the XGBoost classifier based predictive model. Since, the RF classifier based classification model achieves higher percentages of specificity and sensitivity in detecting TN (inactive molecules) and TP (active molecules) samples, respectively from a balanced AID-1850 chemical dataset. Therefore, the classification model built using the RF algorithm is considered as in ideal model to screen active PhoP regulon inhibitors from a given balanced chemical HTS dataset. Evaluation of the ability of the classifiers to selectively classify TP from the FN instances present in the balanced dataset is another significant statistical model performance evaluator. In this regard, the AUC value is the probability that classifier based model will give a higher score to an arbitrarily selected positive sample (active molecule) when compared to an arbitrarily selected negative sample (inactive sample). The AUC value for a classification model is calculated by plotting a ROC curve between the TPR in the y-axis and FPR in the x-axis. The AUC value for a classification model ranges from 0 to 1. Thus a model with an AUC value close or equal to 1 is considered as an ideal model to selectively choose a positive instance from mixed instances of positive (active molecule) and negative (inactive molecule) instances in a given dataset.

TABLE III. THE PERFORMANCE EVALUATION OF RF AND XGBOOST CLASSIFIER BASED SUPERVISED CLASSIFICATION MODELS

SMOTE	Classifier	Area Under the Curve (AUC)	Accuracy	True Positive Rate (%)	False Positive Rate (%)	True Negative Rate (%)	False Negative Rate (%)	Specificity (%)	Sensitivity (%)
Not Using SMOTE	Random Forest	0.86	98.533	25.398	0.151	99.849	79.602	99.849	20.398
	XGBoost	0.93	98.525	14.136	0.126	99.874	85.864	99.874	14.136
Using SMOTE	**Random Forest**	**0.99**	**99.237**	**99.243**	**0.768**	**99.231**	**0.757**	**99.232**	**99.243**
	XGBoost	0.97	97.75	97.636	2.138	97.862	2.364	97.862	97.636

Fig. 3 shows the comparative ROC plot of RF and XGBoost classification model built and tested on balanced 20 % independent tests dataset. It can be apparently comprehended from Fig. 3 that the AUC value of RF algorithm-based model which is 0.99 is comparatively higher than that of XGBoost classifier based predictive model. Thus, the RF classifier based predictive model is considered as an efficient model for selectively classifying positive instances (active PhoP inhibitor molecules) from a given balanced chemical dataset. The statistically significant difference between the RF and XGBoost based classification models in accurately predicting the active and inactive molecule was determined using two sample Unpaired t-test, and the result is tabulated in Table IV. An exceptionally statistically significant two-tailed P value is less than 0.0001 was obtained when the mean of ten-fold accuracy values of the RF-based trained model was compared with XGBoost classification trained model at 95 % confidence interval.

The present proposed classification model based on RF classifier is more sensitive and accurate in classifying active PhoP inhibitors molecules from the AID-1850 dataset as compared to classification model proposed by Kaur et al. 2016 as observed from the present findings tabulated in Table V. The accuracy AUC, specificity, and sensitivity of the current RF based model is higher as compared to accuracy AUC, specificity, and sensitivity obtained by the base classifiers used by Kaur et al. 2016 to build a predictive model to classify inhibitors of PhoP operon from the AIS-1850 dataset.

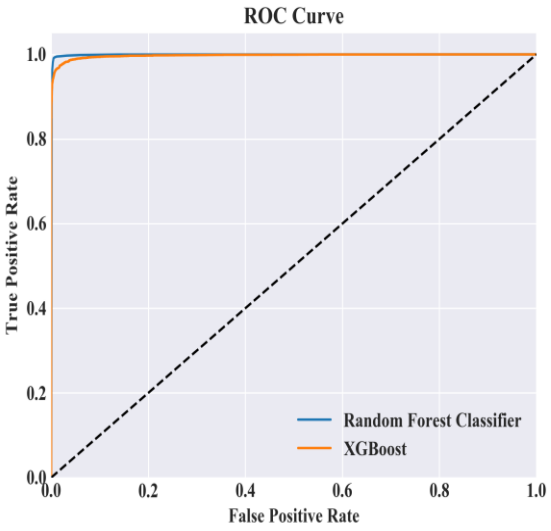


Fig. 3. Comparative ROC plot of RF and XGBoost algorithm based supervised classification tool on balanced dataset.

Moreover, the FNR of the present study is lower as compared to the FNR obtained by predictive models proposed by Kaur et al. 2016. Therefore, the current model based on RF-based classifier built and tested on the balanced dataset is far more superior in screening a real positive (active PhoP inhibitor molecule) from a given AID-1850 dataset.

TABLE IV. TWO-SAMPLE UNPAIRED T-TEST TO DETERMINE SIGNIFICANT DIFFERENCE BETWEEN RF AND XGBOOST CLASSIFIER BASED PREDICTIVE MODEL

Algorithms	**Paired Differences**							
	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**		**t**	**Df**	**Sig. (2-tailed P-value)**
				Lower	***Upper***			
RF and XGBoost	0.0166900490	.00196	0.00062	0.0156290243	0.0177510737	33.0478	18	<0.0001

TABLE V. COMPARATIVE PERFORMANCE EVALUATION OF THE RF CLASSIFIER BASED SUPERVISED CLASSIFICATION MODEL WITH ANOTHER PREDICTIVE MODEL FOR SCREENING ACTIVE INHIBITORS OF PHOP REGULON PROTEIN

ML classifier based classification model	**AUC**	**Accuracy**	**FNR**	**sensitivity**	**specificity**
RF (Kaur et al. 2016)	91.5	81.5	12.3	87.7	81.5
RF (Hassan et al. 2018)	**0.99**	**99.2**	**0.8**	**99.2**	**99.2**

Here, AUC=Area Under the Curve; FNR=False Negative Rate

IV. CONCLUSION AND FUTURE SCOPE

In the current study, ML algorithm is used to build a supervised classification model to classify active PhoP inhibitor molecules from the balanced AID-1850 HTS dataset. The capability of the predictive model to distinguish between the active and inactive classes of the AID-1850 dataset was determined by specific attributes selected using the tree-based feature selection module of Scikit learn package. The final 43 descriptors based dataset was processed using SMOTE algorithm to remove the class imbalance present in the AID-1850 dataset. Several statistical assessors were used to assess the performance of RF and XGBoost classifier based classification model in screening true inhibitors of PhoP regulon protein from a given dataset. The comparative performance evaluation of both XGBoost and RF classifier based predictive model revealed that RF classifier based model showed better ability to predict active PhoP inhibitors from the preprocessed balanced AID-1850 dataset. Moreover, the present RF classifier based model is far more superior in predicting active inhibitors of PhoP regulon protein from AID-1850 when compared to the model proposed by Kaur et al. 2016. Therefore, the present predictive model will be a step forward in screening novel drug-like inhibitors of PhoP a virulent two-component regulatory system of *S. typhi*. Moreover, in future, a web-based real-time predictive system will be built based on the results of the present model to efficiently classify active inhibitors of PhoP operon protein

from the sizeable molecular library of molecules from various chemical databases.

ACKNOWLEDGMENT

We are thankful to the Dean Prof. Omar Barukab of the Faculty of Computing and Information Technology Rabigh (FCITR) of King Abdulaziz University, Jeddah for providing the computational facility to perform current experiments.

REFERENCES

[1] S. Now and V. P. Editions, "Release of the 2017 Global Vaccine Action Plan Reports Past Meetings / Workshops," no. November, 2017.

[2] Kato and E. A. Groisman, "The PhoQ/PhoP regulatory network of Salmonella enterica.," Adv. Exp. Med. Biol., vol. 631, pp. 7–21, 2008.

[3] D. Shin and E. A. Groisman, "Signal-dependent Binding of the Response Regulators PhoP and PmrA to Their Target Promoters in Vivo," vol. 280, no. 6, pp. 4089–4094, 2005.

[4] M. E. Castelli, A. Cauerhff, M. Amongero, F. C. Soncini, and E. G. Vescovi, "The H box-harboring domain is key to the function of the Salmonella enterica PhoQ Mg2+-sensor in the recognition of its partner PhoP.," J. Biol. Chem., vol. 278, no. 26, pp. 23579–85, Jun. 2003.

[5] L. R. Prost and S. I. Miller, "The Salmonellae PhoQ sensor: mechanisms of detection of phagosome signals.," Cell. Microbiol., vol. 10, no. 3, pp. 576–582, Mar. 2008.

[6] B. Blanc-Potard and E. A. Groisman, "The Salmonella selC locus contains a pathogenicity island mediating intramacrophage survival.," EMBO J., vol. 16, no. 17, pp. 5376–5385, Sep. 1997.

[7] W. S. Pulkkinen and S. I. Miller, "A Salmonella typhimurium virulence protein is similar to a Yersinia enterocolitica invasion protein and a bacteriophage lambda outer membrane protein.," J. Bacteriol., vol. 173, no. 1, pp. 86–93, Jan. 1991.

[8] E. Garcia Vescovi, F. C. Soncini, and E. A. Groisman, "Mg2+ as an extracellular signal: environmental regulation of Salmonella virulence.," Cell, vol. 84, no. 1, pp. 165–174, Jan. 1996.

[9] L. Guo, K. B. Lim, C. M. Poduje, M. Daniel, J. S. Gunn, M. Hackett, and S. I. Miller, "Lipid A acylation and bacterial resistance against vertebrate antimicrobial peptides.," Cell, vol. 95, no. 2, pp. 189–198, Oct. 1998.

[10] T. Garcia-Sosa, M. Oja, C. Hetenyi, and U. Maran, "DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties.," J. Chem. Inf. Model., vol. 52, no. 8, pp. 2165–2180, Aug. 2012.

[11] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "Drug/nondrug classification using Support Vector Machines with various feature selection strategies.," Comput. Methods Programs Biomed., vol. 117, no. 2, pp. 51–60, Nov. 2014.

[12] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development.," PLoS One, vol. 10, no. 4, p. e0124600, 2015.

[13] V. V Zernov, K. V Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V Pletnev, "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions.," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 2048–2056, 2003.

[14] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 1882–1889, 2003.

[15] F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P. W. Lee, and Y. Tang, "Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers.," J. Chem. Inf. Model., vol. 51, no. 5, pp. 996–1011, May 2011.

[16] C. Y. Liew, X. H. Ma, X. Liu, and C. W. Yap, "SVM model for virtual screening of Lck inhibitors.," J. Chem. Inf. Model., vol. 49, no. 4, pp. 877–885, Apr. 2009.

[17] J. Fang, R. Yang, L. Gao, D. Zhou, S. Yang, A.-L. Liu, and G. Du, "Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery.," J. Chem. Inf. Model., vol. 53, no. 11, pp. 3009–3020, Nov. 2013.

[18] D. W. Miller, "Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning," J. Chem. Inf. Comput. Sci., vol. 41, no. 1, pp. 168–175, Jan. 2001.

[19] Ajay, W. P. Walters, and M. A. Murcko, "Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules?," J. Med. Chem., vol. 41, no. 18, pp. 3314–3324, Aug. 1998.

[20] J. Sadowski and H. Kubinyi, "A scoring scheme for discriminating between drugs and nondrugs.," J. Med. Chem., vol. 41, no. 18, pp. 3325–3329, Aug. 1998.

[21] H. Sun, "A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing.," J. Med. Chem., vol. 48, no. 12, pp. 4031–4039, Jun. 2005.

[22] F. Rathke, K. Hansen, U. Brefeld, and K.-R. Müller, "StructRank: A New Approach for Ligand-Based Virtual Screening," J. Chem. Inf. Model., vol. 51, no. 1, pp. 83–92, Jan. 2011.

[23] M. Wassermann, H. Geppert, and J. Bajorath, "Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors.," J. Chem. Inf. Model., vol. 49, no. 3, pp. 582–592, Mar. 2009.

[24] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine.," J. Chem. Inf. Model., vol. 45, no. 3, pp. 549–561, 2005.

[25] S. Agarwal, D. Dugar, and S. Sengupta, "Ranking chemical structures for drug discovery: a new machine learning approach.," J. Chem. Inf. Model., vol. 50, no. 5, pp. 716–731, May 2010.

[26] Abdo, B. Chen, C. Mueller, N. Salim, and P. Willett, "Ligand-based virtual screening using Bayesian networks.," J. Chem. Inf. Model., vol. 50, no. 6, pp. 1012–1020, Jun. 2010.

[27] Plewczynski D, Grotthuss MV, Rychlewski L, Ginalski K. Virtual high throughput screening using combined random forest and flexible docking. Comb Chem High T Scr. 2009;12: 484–489.

[28] H. Kaur, M. Ahmad, and V. Scaria, "Computational Analysis and In silico Predictive Modeling for Inhibitors of PhoP Regulon in S. typhi on High-Throughput Screening Bioassay Dataset," Interdiscip. Sci. Comput. Life Sci., vol. 8, no. 1, pp. 95–101, 2016.

[29] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," Nucleic Acids Res., vol. 37, no. Web Server issue, pp. W623–W633, Jul. 2009.

[30] K. Liu, J. Feng, and S. S. Young, "PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation," J. Chem. Inf. Model., vol. 45, no. 2, pp. 515–522, Mar. 2005.

[31] H. Kaur, R. Chauhan, and S. K. Wasan, "A Bayesian Network Model for Probability Estimation," in Encyclopedia of Information Science and Technology, Third Edition, IGI Global, pp. 1551–1558.

[32] H. Kaur, R. Chauhan, M. A. Alam, S. Aljunid, and M. Salleh, "SpaGRID: A Spatial Grid Framework for High Dimensional Medical Databases BT - Hybrid Artificial Intelligent Systems," 2012, pp. 690–704.

[33] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[34] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning BT - Advances in Intelligent Computing," 2005, pp. 878–887.

[35] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[36] T. Chen and C. Guestrin, "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, pp. 785–794.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J.

Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2012.

[38] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," vol. 18, pp. 1–5, 2016.

[39] Fadem, Barbara (2008). High-Yield Behavioral Science (High-Yield Series). Hagerstwon, MD: Lippincott Williams & Wilkins. ISBN 0-7817-8258-9; "The Probable Error of a Mean" (PDF). Biometrika. 6 (1): 1–25. 1908.

SUPPLEMENTARY FILE

TABLE I. ALL FEATURES SORTED ACCORDING TO IMPORTANCE

WBN_GC_L_1.00	NEG_04_NEG
WBN_GC_L_0.25	HBA_04_HYP
WBN_GC_L_0.50	ARC_05_HYP
WBN_LP_H_0.25	POS_03_POS
WBN_LP_L_0.75	HBA_03_HBA
WBN_GC_L_0.75	ARC_06_HYP
WBN_GC_H_0.75	HBD_05_HBD
WBN_GC_H_0.25	HBD_05_HBA
WBN_LP_L_0.25	POS_04_HBA
WBN_EN_H_0.25	ARC_07_HYP
WBN_EN_H_0.50	NEG_04_ARC
WBN_LP_H_0.75	NEG_05_HYP
WBN_LP_H_0.50	ARC_02_HYP
WBN_EN_L_0.75	HBA_06_HYP
WBN_LP_L_1.00	POS_02_HYP
WBN_EN_L_0.25	POS_05_POS
BadGroup	POS_05_HBA
WBN_GC_H_0.50	HBD_07_HBA
WBN_EN_L_0.50	NEG_02_ARC
NumRot	ARC_04_HYP
WBN_LP_L_0.50	POS_07_HBA
WBN_LP_H_1.00	HBD_04_HBA
WBN_EN_H_1.00	POS_04_HBD
WBN_EN_L_1.00	HBA_07_HBA
WBN_EN_H_0.75	POS_07_HBD
PSA	HBA_02_HYP
WBN_GC_H_1.00	HBA_05_HBA
XLogP	HBA_07_HYP
MW	HBD_04_HBD
NumHBA	HBD_06_HBA
NumHBD	POS_06_HBA
BBB	POS_05_HYP
ARC_05_ARC	HBA_03_HYP
HBA_03_ARC	HBD_06_HBD
POS_04_ARC	HBA_04_HBA
ARC_03_ARC	HBD_07_HBD
ARC_06_ARC	POS_05_HBD
ARC_04_ARC	HBD_03_HYP
HBA_05_ARC	NEG_05_ARC
HBA_04_ARC	HYP_01_HYP
HBD_02_ARC	NEG_05_HBA
ARC_01_ARC	POS_03_HYP
ARC_02_ARC	NEG_07_ARC
POS_07_ARC	NEG_06_ARC
POS_05_ARC	HBA_05_HYP
POS_06_HYP	HBD_07_HYP
HBA_06_ARC	HYP_03_HYP
HBD_03_ARC	NEG_04_HBD
HBA_07_ARC	HBD_06_HYP
ARC_07_ARC	NEG_03_HBD
POS_02_ARC	POS_07_HYP
POS_06_ARC	NEG_07_HYP
HBD_05_ARC	NEG_03_ARC
POS_03_ARC	NEG_07_HBD
HBD_04_ARC	POS_03_HBA
ARC_03_HYP	POS_04_HYP
NEG_02_NEG	HBD_02_HYP
HBD_06_ARC	HBD_05_HYP
POS_04_POS	NEG_06_HBA
HBA_06_HBA	POS_06_HBD
HBD_07_ARC	HYP_05_HYP

POS_06_POS	NEG_06_HYP
HYP_02_HYP	NEG_06_POS
NEG_07_HBA	HYP_06_HYP
HBD_04_HYP	NEG_01_HBD
HYP_04_HYP	NEG_05_POS
POS_07_POS	HBD_03_HBA
NEG_02_HYP	NEG_02_HBD
HBD_03_HBD	POS_03_HBD
HYP_07_HYP	NEG_03_POS
NEG_04_HBA	NEG_07_NEG
NEG_01_NEG	NEG_06_NEG
NEG_03_HBA	NEG_05_HBD
NEG_06_HBD	POS_02_HBD
NEG_04_POS	NEG_03_NEG
NEG_05_NEG	NEG_04_HYP
NEG_03_HYP	NEG_07_POS

TABLE II. TOP 43 FEATURES SORTED ACCORDING TO IMPORTANCE

WBN_GC_L_1.00	WBN_EN_H_1.00
WBN_GC_L_0.25	WBN_EN_L_1.00
WBN_GC_L_0.50	WBN_EN_H_0.75
WBN_LP_H_0.25	PSA
WBN_LP_L_0.75	WBN_GC_H_1.00
WBN_GC_L_0.75	XLogP
WBN_GC_H_0.75	MW
WBN_GC_H_0.25	NumHBA
WBN_LP_L_0.25	NumHBD
WBN_EN_H_0.25	BBB
WBN_EN_H_0.50	ARC_05_ARC
WBN_LP_H_0.75	HBA_03_ARC
WBN_LP_H_0.50	POS_04_ARC
WBN_EN_L_0.75	ARC_03_ARC
WBN_LP_L_1.00	ARC_06_ARC
WBN_EN_L_0.25	ARC_04_ARC
BadGroup	HBA_05_ARC
WBN_GC_H_0.50	HBA_04_ARC
WBN_EN_L_0.50	HBD_02_ARC
NumRot	ARC_01_ARC
WBN_LP_L_0.50	ARC_02_ARC
WBN_LP_H_1.00	

Comparative Performance of Deep Learning and Machine Learning Algorithms on Imbalanced Handwritten Data

A'inur A'fifah Amri, Amelia Ritahani Ismail,* Abdullah Ahmad Zarir
Department of Computer Science, Kulliyyah of ICT
International Islamic University Malaysia
P.O. Box 10, 50728 Kuala Lumpur, Malaysia

***Abstract*—Imbalanced data is one of the challenges in a classification task in machine learning. Data disparity produces a biased output of a model regardless how recent the technology is. However, deep learning algorithms, such as deep belief networks showed promising results in many domains, especially in image processing. Therefore, in this paper, we will review the effect of imbalanced data disparity in classes using deep belief networks as the benchmark model and compare it with conventional machine learning algorithms, such as backpropagation neural networks, decision trees, naïve Bayes and support vector machine with MNIST handwritten dataset. The experiment shows that although the algorithm is stable and suitable for multiple domains, the imbalanced data distribution still manages to affect the outcome of the conventional machine learning algorithms.**

Keywords*—Deep belief networks; support vector machine; back propagation neural networks; imbalanced handwritten data; classification

I. Introduction

Imbalanced class in a dataset occurs when the dataset is not in the same amount of values among the parameters or classes. The majority class of the dataset is when the class has the most instances. The minority class of the dataset is when the class has the least instances. A few disadvantages prompted by imbalanced class data in a classification are over fitting, deficient class model and wrongly classified. Over fitting is a result of accuracy bias due to overwhelming data values in one class compared to missing values of another class. The model might give a high accurate result, but it is biased to the majority class.

The approach that will be focused on this paper is a review on the effects of imbalanced class in a handwritten dataset towards deep learning and machine learning algorithms. Deep learning is a part of machine learning algorithms that are recently introduced to solve complex, high-level abstract and heterogeneous datasets, especially image and audio data. There are several types of deep learning architectures, which are deep neural network (DNN), convolutional Neural Network (CNN), deep belief networks (DBN) and convolutional deep belief networks (CDBN). In this paper, we will focus on two deep learning algorithms, which are CNN and DBN. CNN is composed of one or more convolutional layers with fully connected layers at the end of it. CNNs are used in computer vision and acoustic modeling for automatic speech recognition (ASR). A deep belief network (DBN) is a probabilistic, generative model made up of multiple layers of hidden units. It can be seen as a composition of simple learning modules of Restricted Boltzmann Machine (RBM) that make up each layer.

Conventional machine learning algorithms such as back propagation neural network (BPNN), support vector machine (SVM), Naïve Bayes and decision trees are also included in the experiment to enhance performance comparison value between deep learning and traditional machine learning algorithms when an imbalanced class handwritten data is used as the training set.

This paper is organized as follows. Section 2 clarifies the definitions of imbalanced data, the effects of imbalanced data have for classification tasks and the application of any deep learning algorithms used to counter this problem. Basic concepts and the applications of DBN, CNN, BPNN, SVM, Naïve Bayes and decision tree algorithms are described in the same section. Section 3 explains the experimental setup of imbalanced class data classification using deep learning and machine learning algorithms. Section 4 interprets the result analysis of the experiments and conclusions are presented in Section 5.

II. Related Work

Encouraging results have been received upon the application of deep learning algorithms in text recognition [1], audio classification [2] and even abstract high-level domains such as emotional recognition [3]. However, these are applied to data that are distributed evenly. Not many imbalanced data problems have been solved using a deep learning method.

According to some papers [4]-[7], imbalanced class in a dataset refers to the disparity of data dispensation between the classes. The class that has more training values is called the majority class and the class that has the least or most missing data values are called the minority class [5]. Minority data class is a realistic problem that the real-world situation faced because most of the time, data are scarce, despite its importance. The examples of minority classes in real world problem are credit fraud detection [8] and cancer abnormalities diagnosis [6], [8]. It can be expensive if the new data needs labeling [9]. Unfortunately, most algorithms devised shown stable and promising performance when using

This research is supported by the International Islamic University Malaysia under the Research Initiative Grants Scheme (RIGS): RIGS16-346-0510

balanced data in classification tasks but showed otherwise when imbalanced data is used[4]. Prediction of minority class is presumed to have a higher error rate compared to the majority class and its test examples are often wrongly classified as well [10].

The imbalanced class could cause deficient classification models [6], [7]. The algorithm that performs on balanced dataset will not perform as good when using an imbalanced dataset [4], regardless how good the model is. In a work, an imbalanced multimedia dataset was used on CNN [5], and it shows that the error rate "fluctuate" compared to when using a balanced dataset, where the error rate continues to decrease. In a paper [6], the author used SVM as the main algorithm and showed that the effect of data disparity results in a "high false negative rate". Another paper [11] modified kNN algorithms to counter the effect of imbalanced data to the algorithm. Bootstrapping is often used to improve the algorithm performance when imbalanced data is used [6], [9].

A. Deep Belief Networks

To comprehend DBN, the concept of Restricted Boltzmann Machine (RBM) must first be explained. The architecture of RBM is it consists of a bidirectional connection between hidden layers and visible layers. This feature allows the weight to be connected exclusively and allows deeper extraction between the neurons. RBM is a probabilistic model[2] and a two-layer, bipartite, undirected graphical model with a set of binary hidden random variables (units) h of dimension K, a set of (binary or real-valued) visible random variables (units) v of dimension D, and symmetric connections between these two layers represented by a weight matrix ($W \in R^{D \times k}$) [12]. Two main RBM often used are Bernoulli, where visible and hidden layers are binary, and Gaussian is where the visible units are allowed to use real number values [3].

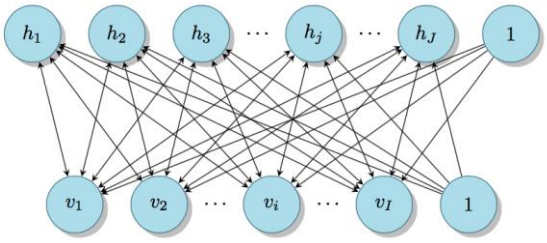


Fig. 1. Example of RBM architecture schematic design [12].

Fig. 1 above presents the schematic design of RBM architecture. RBM is made up of stochastic visible units and stochastic hidden units that are connected to each other [13].

A deep belief network (DBN) is a probabilistic, generative model made up of multiple layers of hidden units. It can be seen as a composition of simple learning modules that make up each layer. DBN is made up of stacked Restricted Boltzmann Machine (RBM) used greedily as depicted in Fig. 2. However, such feature results DBN to be computationally expensive and time-consuming because the number of layers DBN needs to go through is a lot.

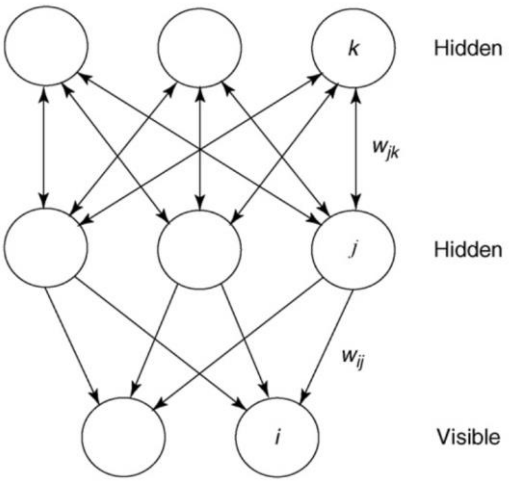


Fig. 2. A stacked RBM or known as DBN [14].

According to Le & Provost [3], training a DBN is expensive in terms of computation because pre-training took 11 minutes per epoch and fine-tuning takes up 10 minutes per epoch. DBN is used in emotions recognition [3] by learning high-level features. Face verification is also using DBN, despite the usage of CNN, the hybrid algorithm aims to achieve robustness in verifying similarities of different faces [15]. DBN is also used to model natural images [16] by learning multiple layers of unlabeled data.

B. Convolutional Neural Networks

Convolutional neural networks (CNN) consists of one or more convolutional layers [4], [5], [17], alternating with subsampling layers and by the end of the network, optionally, a fully connected MLP [4]. Basically, CNN architecture must consist of one or more convolutional, pooling and a fully connected layers on top [5]. The convolutional layers are responsible for feature extraction and is called feature map [4], [5], [17] and sometimes feature detection [18]. After convolutional layer, it is often paired up with a pooling layer that will perform a pooling function based on the inputs it received from the previous convolutional layer [4]-[7]. The pooling layer is also known as a subsampling layer, and it will alternate with a convolutional layer because it computes the statistics of the convolutional layer. The pooling layer will perform pooling functions and is called min-pooling, max-pooling layers or etc. according to its context of problem-solving. The pooling function will *"downsample"* the input it received from its convolutional layer [5]. Such will carry on until the end of the network. At the end of the series of alternation, a fully connected MLP will be added. It works as a classification module for the network [4]. This layer will receive all neurons from its previous layers whether they are convolutional or pooling and connect them with its own neurons [5]. The architecture is depicted in Fig. 3.

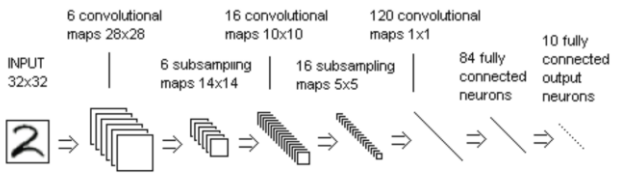


Fig. 3. Example of a convolutional neural network reading an image input [49].

However, the implementation of convolutional and subsampling layers in a CNN plus the method of the network training differs in every CNN [19]. It depends on the context of problems that are attempted to solve.

C. Back Propagation Neural Networks

Artificial neural networks (ANN) are modeled after the human brain networks [20], [21]. It is widely known used for supervised learning and recognizing patterns from input dataset by weight adjustments [20]. There are many examples of ANN such as feed forward and radial basis function (RBF). ANN's ability to scrutinize nonlinear data and to design complex models has allows it to be applied in studies of different fields [21], [22]. Fig. 4 shows an example of ANN architecture.

The most common neural network algorithm used is the back propagation neural network (BPNN). BPNN has three layers, which consists of an input layer, a hidden layer and an output layer [20]. The layers are made up of interconnected nodes by a weighted association and the number of nodes in the layers depends on the problem domain [21]. The input layer will accept the data for training or testing and pass the weights to the connected hidden layer. Hidden layer can be one or more and it will continue calculating the weights it received and send it to the output layer where the result is produced. BPNN compares its real output and target output and adjusts its weight according to the error and propagates back to its network.

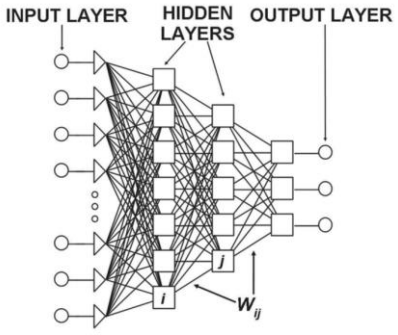


Fig. 4. An example of a neural network with two hidden layers [21].

Arora [23] implemented back propagation neural network classifier to categorize Devnagari handwritten classes and compared its performance with SVM using the same handwritten dataset. The experiment result for BPNN performance is 90.44% for testing dataset accuracy. Another work [24] proposed a diagonal based feature extraction and used a "feed forward back propagation" neural network to classify the data based on the new feature extraction and achieved 96.52% with 54 features and 97.84% with 69 features accuracy rate.

In tackling imbalanced data, Cao and et al. [25] presented a cost sensitive back propagation neural network for a multiclass imbalanced data, as opposed to the "limited" binary class imbalanced data [20].

D. Support Vector Machine

According to Arora [23], SVM can be defined as a 'binary classifier', where the outcome will be divided into two groups based on the optimum hyperplane. Fig. 5 depicts the definition of SVM in pictorial form.

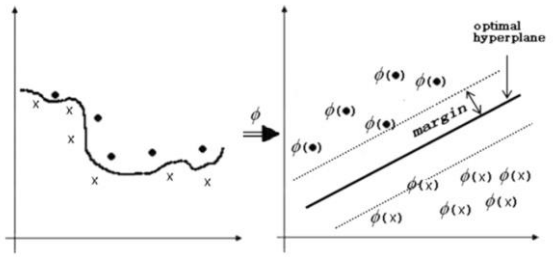


Fig. 5. SVM mapping nonlinear problem to linear using optimum hyperplane [46].

Niu & Suen [26] implemented a hybrid of SVM and CNN for classifying MNIST handwritten digits dataset. Feature extraction is done using CNN and SVM acts as a 'recognizer'. Arora [23] compared the performance of SVM and ANN using the Devnagari handwritten recognition problem. SVM performance in the experiment achieved 92.38% for testing accuracy.

In countering the imbalanced data classification problem using SVM, its weight and activation function are manipulated in order to increase the classification accuracy [27]. Tang and et al. [28] stated that SVM outperforms other conventional classifiers when a moderate imbalanced data is used. Even so, when a high imbalanced data is used instead, SVM classifier can still produce a biased result. Most works using SVM to counter imbalanced data only focused on the performance and not efficiency, hence, SVM can be a slow classifier [28]. However, Zou and et al. [29] stated that SVM could not perform imbalanced data classification successfully based on the works of Cristianini & Shawe-Taylor [30].

In Big Data domain, Koturwar and et al. [31] stated that SVM has the ability to balance massive data correctly. Feature extraction using SVM is good as it can be done promptly using SVM kernel instead of a feature extraction process that results to data lost [31].

One of the disadvantages of SVM classifier is its training and execution is very complex caused it to be implemented in mostly small category set problem[23]. According to Koturwar and et al. [31], large training data makes SVM inefficient and costly, as SVM is not scalable to huge size data. When the training data is noisy and imbalanced, it can affect the outcome of SVM due to its high training execution and low generalization error [31].

E. Naïve Bayesian

Naive Bayes (NB) is a supervised probabilistic classifier that is based on the Bayes' theorem with the assumption the attributes of the data are discrete [32]-[34]. NB calculates the conditional probability of the features and choose the class with the highest value [34].

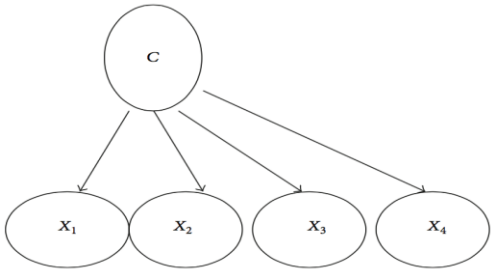


Fig. 6. An example of Naive Bayes structure [35].

Bal and et al. [35] suggests that the NB is made up of one classification node that acts as the parent nodes for all the rest of the nodes as shown in Fig. 6. According to Kumar and et al. [33], Bayes theorem suggests that a problem case to be classified is represented by vector x = (x_1,..., x_n) with n independent features. It brings to the instance probability, p(C_k|x_1,..., x_n) for each K possible outcomes. The equation is summarised as below:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (1)$$

where,

p(C_k) = probability of class *k*,

$p(x|C_k)$ = probability of query *x* given class *k*,

$p(x)$ = probability of query *x*.

This allows the supervised learning to be implemented solely on logical and statistical calculation [36]. Therefore, NB is suitable as a solution for predictive and diagnostic problem [36]. Due to its ability to determine hypothesis by calculating probabilities, NB is robust to input data noises [36]. NB provides stable performance for a bank dataset [32] with an accuracy rate of 89%. Dey et al. (2016) stated that the performance comparison in sentiment analysis of movie and hotel reviews datasets, NB algorithm outperform k-NN with over 80% accuracy rate. Ahmed and et al. [37] proposed a hybrid of NB and Apriori algorithm to detect SMS spam and achieved the accuracy of 98.7% as compared to 97.4% accuracy using traditional NB. In another classification task, Sapkale and Nair [36] used NB as a method to improve domain classification of Google search results. The experiment resulted shorter performance time with the same domain classification rate.

In imbalanced class dataset problem, Imran and et al. [38] applied NB on an imbalanced educational data by using Weka tool and achieved accuracy rate of 68.2432%. Sharma and et al. [39] reviewed the recognition performance of NB algorithm on handwritten Gujarati character data and acquired 96.43% of classification accuracy and F-measure. In another work, a comparison of performance using NB for writers' identification through their handwriting in English language was done [40]. The accuracy result based on aggregated feature attained by NB is 85%. Sarangi and et al. [41] recorded the experiment involving handwritten Odia numerals by performing LU factorization as feature extraction and then classify the dataset using NB. Although the experiment focuses on feature extraction instead of the classifier, overall accuracy result from number 0 until 9 are between 74.39% and 85%.

F. Decision Trees

Decision trees (DT) produce an output based on the series of binary decision in the model called in the form of dendritic graph [42], [43]. It presents all possible output with the path leading to the output [35] as shown in Fig. 7. Tree pruning is a method of downsizing tree size by eliminating nodes that does not give accuracy in result [42].

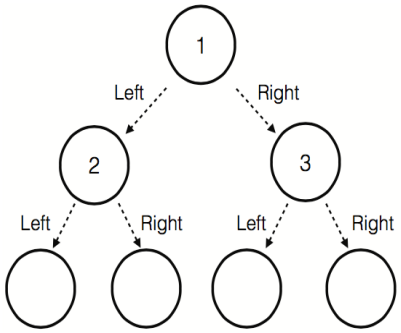


Fig. 7. An example of a decision tree structure deciding output path [44].

According to Y. Zhang and et al. [42], DT is suitable for decision analysis as it can show the strategy to achieve a solution. Analysis using DT is simple because the connection between the input and output is clear [43], [45]. Besides that, DT is able to operate on both numerical and categorical data [43].

One of the few disadvantages of DT in Big Data domain is that the large sets of data will cause more time take to construct a traditional DT [31]. Menickelly et al. [43] mentions that DT is not robust to different training data, which can result low accuracy performance.

One application of DT is in the medicine field where the algorithm is implemented to classify Parkinsonian Syndromes using FDG-PET brain dataset. The algorithm correctly classifies with ranges from 47.4% to 80.0% of accuracy [45].

III. Experimental Setup

For this experiment, an imbalanced dataset that is suitable for classification task is selected. Then, the source code of CNN and DBN is modified to suit the dataset, which is extended from [48]. Then, the preliminary results CNN and DBN are recorded and further evaluated. The experimental dataset used in this experiment is MNIST handwritten digit dataset as many experiments have used the dataset as a benchmark [1], [12], [47]. The dataset is preprocessed and consists of 4 files, 2 training files, and 2 testing files.

The training set contains 60000 examples, and the test set 10000 examples. However, since the main aim of this paper is to review the data disparity and the algorithm's performance, the data has been modified to a smaller size but imbalanced. The labels values are 0 to 9. Pixels are organized row wise and the values are between 0 and 255. 0 means background (white), 255 means foreground (black). The images were centered in a 28x28 image. Data distribution is described in Table I below together with their percentages.

TABLE I. DATA DISTRIBUTION OF MNIST DATASET

Labels	Number of data	Imbalance Percentage (%)
0	500	100
1	45	9
2	150	30
3	250	50
4	150	30
5	35	7
6	25	5
7	200	40
8	350	70
9	15	3

IV. ANALYSIS OF RESULTS

Accuracy rate calculates the number of correct predictions out of the number of all predictions. Classification error shows is the number of wrongly predicted number out of all predictions. Kappa statistics takes into account the correct predictions made by chance and is between 0 and 1. Weighted mean recall calculates class recall or sensitivity for each class. Weighted mean precision calculates through class precisions for individual classes. Absolute error presents the average absolute deviation of the prediction from the actual value. Relative error is the average of the absolute deviation of the prediction value.

Table II presents the results of DBN, CNN and DNN. The accuracy rate of DBN is 92.5% and is the highest accuracy among the three deep learning algorithms. The classification error is 7.5%, which is a promising result. CNN has 10% accuracy rate and 90% classification error rate. DNN achieved accuracy rate of 27.91% and classification error of 71.57%.

For DBN, the kappa statistics result is 0.893, which is very high. CNN has kappa statistics of 0.0. Kappa statistics for DNN is -0.001, which is below than 0. Hence, it means that the two observers are agreeing even less.

The weighted mean recall for DBN is 90.4% or 0.9 and is agreeable since it is more than 0.5. For CNN, its weighted mean recall is 0.1% or 0.1, which is low. The weighted mean recall for DNN is 9.91% or 0.0991, which is not good as it is less than 0.5.

For DBN, the weighted mean precision is 91.5% or 0.915, which is good since it values more than 0.5. CNN achieved 1.0% or 0.01 for its weighted mean precision. The result is not good as it is less than 0.5 compared to DBN. The weighted mean precision for DNN is 5.79% or 0.0579, which is not good as it is also less than 0.5.

TABLE II. RESULTS FOR DEEP LEARNING ALGORITHMS

	DBN	CNN	DNN
Accuracy rate	92.5%	10%	27.91%
Classification error	7.5%	90%	71.57%
Kappa	0.893	0.0	-0.001
Weighted mean recall	90.4%	0.1%	9.91%
Weighted mean precision	91.5%	1.0%	5.79%
Absolute error	0.301	2.9	0.721
Relative error	30.1%	290%	72.13%
Root mean squared error	1.14	3.52	0.84
Squared error	1.3	12.39	0.706
Processing time	3 hours 47 minutes	37 minutes 26 seconds	2 hours 2 minutes 35 seconds

The absolute error for DBN is 0.301. The result is quite low. CNN has the highest absolute error, which are 2.9. DNN achieved 0.721 in absolute error, which are high although not as high as CNN.DBN achieved relative error of 30.1% and is the lowest among the three deep learning algorithms. CNN has 290% relative error and is the highest. Relative error for DNN is 72.13% or 0.7213 and it is quite high as well.

Root mean squared for DBN is 1.14 and its squared error achieved 1.3. CNN has root mean squared value of 3.52 and squared error of 12.39. Root mean squared error of DNN is 0.84 and its squared error achieved 0.706. All of the values are high as they are more than 0.5. However, DNN has the least root mean squared and squared error as compared to DBN and CNN.

The processing time for DBN is 3 hours and 47 minutes. It is the longest processing time compared to CNN and DNN. CNN has the shortest processing time at 37 minutes 26 seconds. The processing time for Deep Learning to compute the dataset is 2 hours 2 minutes and 35 seconds.

Table III presents the results of BPNN, SVM, Decision tree and Naïve Bayes. The accuracy rate of BPNN is 23.9% with 77.03% classification error. SVM has 23.43% accuracy rate and 77.21% classification error. Decision tree has 29.07% accuracy rate which is the highest among four algorithms with the lowest classification error, which is 70.93%. Naïve Bayes has the lowest accuracy rate at 12.32% and the highest classification error at 87.69%.

TABLE III. RESULTS FOR MACHINE LEARNING ALGORITHMS

	BPNN	SVM	Decision tree	Naïve Bayes
Accuracy rate	23.9%	23.43%	29.07%	12.31%
Classification error	77.03%	77.21%	70.93%	87.69%
Kappa	0.032	0.010	0.000	0.017
Weighted mean recall	10.97%	10.18%	10.00%	15.43%
Weighted mean precision	11.17%	10.07%	2.91%	9.42%
Absolute error	0.826	0.772	0.823	0.877
Relative error	82.59%	77.21%	82.28%	87.67%
Root mean squared error	0.847	0.879	0.827	0.936
Squared error	0.717	0.772	0.685	0.876
Processing time	8 hours 16 minutes 2 seconds	3 minutes 14 seconds	28 seconds	5 seconds

Kappa statistics for BPNN is 0.032. SVM has kappa value at 0.010 while decision tree has 0 for its kappa statistics value. Lastly, Naïve Bayes achieved 0.017 for its kappa value. All four algorithms have very low kappa value, but the highest kappa statistics value is attained by BPNN.

Weighted mean recall for sensitivity of BPNN achieved 10.97% or 0.1097 and 10.18% or 0.1018 for SVM. Both algorithms have low sensitivity classifying the imbalanced class handwritten dataset. Decision tree has 10.00% or 0.1 value for its weighted mean recall. Naïve Bayes attained the highest weighted mean recall at 15.43% or 0.1543.

The weighted mean precision for all the algorithms are very weak as they are less than 0.5. BPNN achieved 11.17% or 0.1117 for its weighted mean precision, which is the highest among the four algorithms. SVM attained 10.07% or 0.1007 for its weighted mean precision. Decision tree has 2.91% or 0.0291, which is the lowest weighted mean precision obtained by the rest of the algorithms. The weighted mean precision for Naïve Bayes is at 9.42% or 0.0942.

The results of absolute error for all four algorithms are high because they are more than 50% rate. BPNN has absolute error of 0.826. SVM has the lowest absolute error out of four algorithms, which is 0.772. Decision tree has 0.823 absolute error values and Naïve Bayes achieved 0.877 for absolute error and is the highest.

Relative errors for all the algorithms are high as well since they achieved more than 50% rate. The relative errors for all the algorithms are similar to their respective absolute error. BPNN relative error rate is at 82.59% or 0.8259. SVM has the lowest relative error at 77.21% or 0.7721. Decision tree has relative error rate at 82.28% or 0.828 and Naïve Bayes has relative error rate of 87.67% or 0.8767.

Root mean squared error for all four algorithms is inflated as they almost achieve 100% or 1.0 rate. BPNN attained 0.847 root mean squared error rate while SVM has 0.879 for its root mean squared error rate. Decision tree has the lowest root mean squared error among the four algorithms, which is 0.827. Meanwhile, Naïve Bayes has the highest root squared mean error among the four algorithms at 0.936, which is near 1.0.

The squared error for BPNN is 0.717 and SVM is at 0.772. Decision tree has the lowest squared error out of the four algorithms at 0.685. Naïve Bayes has the highest squared error at 0.876.

The processing time varies for all the algorithms. BPNN has the most expensive processing time at 8 hours 16 minutes and 2 seconds. SVM took 3 minutes and 14 seconds to classify the data accordingly while decision tree took 28 seconds. Naïve Bayes is the least expensive out of the four algorithms as it took 5 seconds to compute.

V. CONCLUSIONS

Imbalanced class dataset affects the outcome despite the stability of the algorithm. The complexity of handwritten form of data also influenced the results of the algorithms. Therefore, all the algorithms have really low accuracy rate, which is below 50% and high classification error with poor performance. However, DBN managed to achieve high accuracy rate and low error rate according to the performance metrics as compared to the other algorithms. As a conclusion, DBN algorithm is stable and robust when an imbalanced handwritten dataset is utilized as an input.

ACKNOWLEDGMENT

This research is supported by the International Islamic University Malaysia under the Research Initiative Grants Scheme (RIGS): RIGS16-346-0510.

REFERENCES

[1] Wang, T., W, D. J., Coates, A., & Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. ICPR.

[2] Dieleman, S., mon Brakel, P., & Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. ISMIR.

[3] Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. IEEE.

[4] Hensman, P., & Masko, D. (2015). The impact of imbalanced training data for convolutional neural networks (Unpublished doctoral dissertation). KTH Royal Institute of Technology.

[5] Yan, Y., Chen, M., Shyu, M.-L., & Chen, S.-C. (2015). Deep learning for imbalanced multimedia data classification. IEEE International Symposium on Multimedia.

[6] Liu, Y., Yu, X., Huang, J. X., & An, A. (2010). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Elsevier Ltd.

[7] Fernandez, A., Garcıa, S., & Herrera, F. (2011). Addressing the classification with imbalanced data: Open problems and new challenges on class distribution. Springer-Verlag Berlin Heidelberg.

[8] Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter

[9] Berry, J., Fasel, I., Fadiga, L., & Archangeli, D. (2012). Training deep nets with imbalanced and unlabeled data. Interspeech.

[10] Weiss, G. M., & Provost, F. (2001). The effect of class distribution on classifier learning: An empirical study. Technical Report ML-TR-44.

[11] Liu, W., & Chaw, S. (2011). Class confidence weighted knn algorithms for imbalanced datasets. Springer.

[12] Lopes, N., Ribeiro, B., & Goncalves, J. (2012). Restricted boltzmann machines and deep belief networks on multi-core processors. IEEE World Congress on Computational Intelligence.

[13] Mohamed, A., Yu, D., & Deng, L. (2010). Investigation of full-sequence training of deep belief networks for speech recognition. Interspeech.

[14] Hinton, G. E. (2007). Learning multiple layers of representation. Elsevier Ltd.

[15] Sun, Y., Wang, X., & Tang, X. (2013). Hybrid deep learning for face verification. IEEE International Conference on Computer Vision.

[16] Aurelio, M., Krizhevsky, R. A., & Hinton, G. E. (2010). Factored 3-way restricted boltzmann machines for modeling natural images. Proceedings of the 13th In- ternational Conference on Artificial Intelligence and Statistics.

[17] Abdel-Hamid, O., Deng, L., & Yu, D. (2013). Exploring convolutional neural network structures and optimization techniques for speech recognition. Interspeech.

[18] Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. Elsevier Science Ltd.

[19] Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible high performance convolutional neural networks for image classification.

[20] Zhang, D., & Xu, W. (2014). A data-distribution based imbalanced data classification method for credit scoring using neural networks. IEEE.

[21] Amato, F., López, A., Peña-Méndez, E. M., Vanˇhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine.

[22] Drew, P. J., & Monson, J. R. T. (2000, January). Artificial neural networks. Surgery. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National Science Review.

[23] Arora, S., Bhattacharjee, D., Nasipuri, M., Malik, L., Kundu, M., & Basu, D. K. (2010, May). Performance comparison of SVM and ANN for handwritten devnagari character recognition. IJCSI International Journal of Computer Science Issues, 7.

[24] J.Pradeep, E.Srinivasan, & S.Himavathi. (2011). Diagonal based feature extraction for handwritten character recognition system using neural network. IEEE.

[25] Cao, P., Li, B., Zhao, D., & Zaiane, O. (2013). A novel cost sensitive neural network ensemble for multiclass imbalance data learning. IEEE.

[26] Niu, X. X., & Suen, C. Y. (2011, October). A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Elsevier Ltd, 1318–1325.

[27] Hwang, J. P., Park, S., & Kim, E. (2011). A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Elsevier Ltd, 8580–8585.

[28] Tang, Y., Zhang, Y.-Q., Chawla, N. V., & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems.

[29] Zou, S., Huang, Y., Wang, Y., Wang, J., & Zhou, C. (2008). Svm learning from imbalanced data by GA sampling for protein domain prediction. IEEE Computer Society.

[30] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

[31] Koturwar, P., Girase, S., & Mukhopadhyay, D. (2015). A survey of classification techniques in the area of big data. arXiv preprint arXiv:1503.07477.

[32] Patil, T. R., & Sherekar, M. S. S. (2013). Performance analysis of naive bayes and j48 classification algorithm for data classification. International Journal Of Com- puter Science And Applications.

[33] Kumar, J., Prasad, S. S., & Pal, S. (2016). Irism @ ntcir-12 temporalia task: Experiments with maxent, naive bayes and decision tree classifiers. Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies.

[34] T, T., MS, V., & S, K. (2013). Supervised learning approach for tamil writer identity prediction using global and local features. International Journal of Research in Engineering and Technology, 2(5).

[35] Bal, M., Amasyali, M. F., Sever, H., Kose, G., & Demirhan, A. (2014). Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. The Scientific World Journal.

[36] Sapkale, D., & Nair, D. P. S. (2016). An improved domain classification of Google search results using abstract: naive Bayes classifier. International Journal of Engineering Science and Computing.

[37] Ahmed, I., Guan, D., & Chung, T. C. (2014). Sms classification based on naïve bayes classifier and apriori algorithm frequent itemset. International Journal of Machine Learning and Computing.

[38] Imran, M., Afroze, M., Sanampudi, D. S. K., & Qyser, D. A. A. M. (2016, June). Data mining of imbalanced dataset in educational data using weka tool. International Journal of Engineering Science and Computing.

[39] Sharma, A., Thakkar, P., Adhyaru, D. M., & Zaveri, T. H. (2016). Features fusion based approach for handwritten gujarati character recognition. Nirma Univeristy Journal Of Engineering And Technology.

[40] V, R. S., & S, V. M. (2015). Predicting the identity of a person using aggregated features of handwriting. Advances in Image and Video processing.

[41] Sarangi, P. K., Ahmed, P., & Ravulakollu, K. K. (2014). Naïve bayes classifier with lu factorization for recognition of handwritten Odia numerals. Indian Journal of Science and Technology.

[42] Zhang, Y., Wang, S., Phillips, P., & Ji, G. (2014). Binary pso with mutation operator for feature selection using decision tree applied to spam detection. Elsevier Ltd..

[43] Menickelly, M., Gunluk, O., Kalagnanam, J., & Scheinberg, K. (2016). Optimal gener- alized decision trees via integer programming. Industrial and Systems Engineering.

[44] Baumann, F., Vogt, K., Ehlers, A., & Rosenhahn, B. (2015). Probabilistic nodes for modelling classification uncertainty for random forest. 14th IAPR International Conference on Machine Vision Applications (MVA).

[45] Mudali, D., Teune, L. K., Renken, R. J., Leenders, K. L., & Roerdink, J. B. T. M. (2015). Classification of parkinsonian syndromes from fdg-pet brain data using decision trees with ssm/pca features. Computational and Mathematical Methods in Medicine.

[46] Ren, J. (2012). Ann vs. svm: Which one performs better in classification of mccs in mammogram imaging. Elsevier Ltd.

[47] Ciresan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. IEEE.

[48] Amri, A.A, Ismail, A. R, Zarir, A. A. (2017). Convolutional Neural Networks and Deep Belief Networks for Analysing Imbalanced Class Issue in Handwritten Dataset. International Journal on Advanced Science, Engineering and Information Technology (IJASEIT), 7(6), pp 2302-2307.

[49] Mrazova, I, & Kukacka, M. (2012). Can Deep Neural Networks Discover Meaningful Pattern Features? Procedia Computer Science 12, pp 194-199.

Insulator Detection and Defect Classification using Rotation Invariant Local Directional Pattern

Taskeed Jabid
Computer Science and Engineering Department
East West University
Dhaka, Bangladesh

Tanveer Ahsan
Computer Science and Engineering Department
International Islamic University Chittagong
Chittagong, Bangladesh

***Abstract*—Detecting power line insulator automatically and analyzing their defects are vital processes in maintaining power distribution systems. In this work, a rotation invariant texture pattern named rotation invariant local directional pattern (RI-LDP) is proposed for representing insulator image. For this at first, local directional pattern (LDP) is applied on image which can encode local texture pattern into an eight bit binary code by analyzing magnitude of edge response in eight different directions. Finally this LDP code is made robust to rotation by meticulously rearranging the generated another binary code which named as rotation invariant local directional pattern (RI-LDP). Insulator detection is carried out where this RI-LDP based histogram act as a feature vector and support vector machine (SVM) plays the role of the classifier. The detected insulator image region is further analyzed for possible defect identification. For this, an automatic extraction method of the individual insulator caps is proposed. The defect in segmented insulators is analyzed using LDP texture feature on individual cap region. We evaluated the proposed method using two sets of 493 real-world insulator images captured from a ground vehicle. The proposed insulator detector shows comparable performance to state-of-the-arts and our defect analysis method outperforms existing methods.**

Keywords*—Insulator detection; insulator defect analysis; local direction pattern (LDP); rotation invariant local directional pattern (RI-LDP); support vector machine (SVM)

I. INTRODUCTION

A moment without electricity cannot be imagined by our modern-day life. We, the citizen of modern-day society cannot imagine a moment without electricity. It is so much endemic to our every day exertion that the uninterrupted distribution of electricity is must. The possibility of uninterrupted distribution of electricity is associated with number of factors including problem free power line channel. That's why regular inspection and maintenance of power line distribution system is required. Among the number of components in power line distribution systems, insulator is one of the key components for stable power supply. According to the statistical data of the national power company, the breakdown of insulator is the most frequent cause of the failure of power system [1]. Therefore, rigorous insulator inspection for defect identification is necessary to ensure an uninterrupted power distribution. Until recently, the insulator inspection relies mostly on manual labor by climbing the pole which is dangerous and time-consuming [2]. However, advancement of vision based object detection technology promotes automatic insulator monitoring system which can detect the position of insulator and identify possible defect in the detected insulator [3] from captured images and/or videos.

A large number of insulator localization and defect identification works are done on aerial images and/or videos. Some of them utilize local image descriptor for recognizing insulators from clutter background. One of them is proposed by Oberweger et al. [4] where a circular descriptor generated by local gradient information. They utilize a unique voting scheme for accurate localization of insulator caps. On the other hand Liao and An [5] describe a local interest point by both gradient and gray level feature. They also utilize multiple circular image patches to describe a local interest point. This multiscale and multifeature (MSMF) descriptor is also able to handle some degree of rotation variation. Some researchers used characteristic shape features of insulators for the detection. In [6] insulator structure is modeled using Haar-like feature to enable rapid feature extraction. They generate a 3D insulator model which minimizes the scarcity of positive samples. Vertical projection curve is derived from video sequence by Bingfeng Li et al. [7] for SVM based insulator detection. Wang et al. [3] propose a Gabor based feature extraction method for insulator recognition. False positive which appears due to clutter background is discarded by morphological operation. Tiantian et al. [8] utilized local textured feature local binary pattern (LBP) and fuse it with histogram of oriented gradient (HOG) feature for insulator detection from aerial videos. Zhao et al [9] detect insulator using infrared images by taking advantage of high-level discriminative Convolution Neural Networks.

Even though there are a number of proposed systems for automatic insulator detection from aerial images, the high price and low stability make them less practical. Moreover, aerial vehicles are more susceptible to ill weather conditions (e.g., strong wind). Therefore, most of the systems utilize ground vehicles for insulator detection and subsequent defect analysis. Edge histogram descriptor (EHD) feature along with Kalman filter is used by Li et al. [1] to recognize the insulator. Jabid et al. proposed local textured based insulator detection [10]. A SVM based insulator condition analysis using wavelet transformation feature is proposed by Murthy et al. [11]. They also illustrate suitability of wavelet transformation based feature for identifying good insulator from bad ones using hidden Markov model in [12].

Developing a successful insulator monitoring system is a challenging problem due to the large variations of the appearance of insulator caused by scale, viewpoint, color, and occlusion (example insulator images are shown in Fig. 1). Cluttered backgrounds also increase the complexity of the problem and often increase the computational load and decrease the success rate in detection. Regarding the viewpoint, arbitrary in-plane and out-of-plane rotational angles make the detection problem highly challenging. Most of the existing insulator detection methods address only a subset of the variations without having the capability to handle all of them.

Fig. 1. Real life insulator appearance in different orientation.

In this work, an insulator monitoring system is proposed which can detect the insulators from images captured through the ground vehicle and subsequently analyze for potential defect. As the proposed insulator detection system utilizing a novel rotation invariant texture feature namely rotation invariant local directional pattern (RI-LDP), it can handle highly cluttered images with insulators appeared in arbitrary orientation. Image pyramid based multiscale detection approach is used to overcome the scale variation. The proposed method utilizes a novel rotation invariant texture encoding method to describe image which helps better detection of insulator region even if insulator appeared in arbitrary orientation.

The main strength of proposed encoding is two folds: firstly it describes local image texture by comparing relative strength of eight directional gradients. As gradient provides relatively robust information in adverse imaging situation, hence the proposed code inherently becomes more stable than other intensity based feature. Secondly, to make the feature robust to rotation, the code is rearranged according to direction of highest gradient direction. The effectiveness of the proposed RI-LDP is substantiated by the higher detection accuracy when classification is carried out utilizing support machine (SVM) classifier. Furthermore, we an automatic insulator defect analysis system is proposed which can automatically partition each cap of an insulator from its core and subsequently analyze each cap for the defect. Our defect analysis system can identify defected insulators and can categorize them into five common defects, i.e., Cracks, Contamination, Whitening, Bullet Damage, and Alligatoring effects.

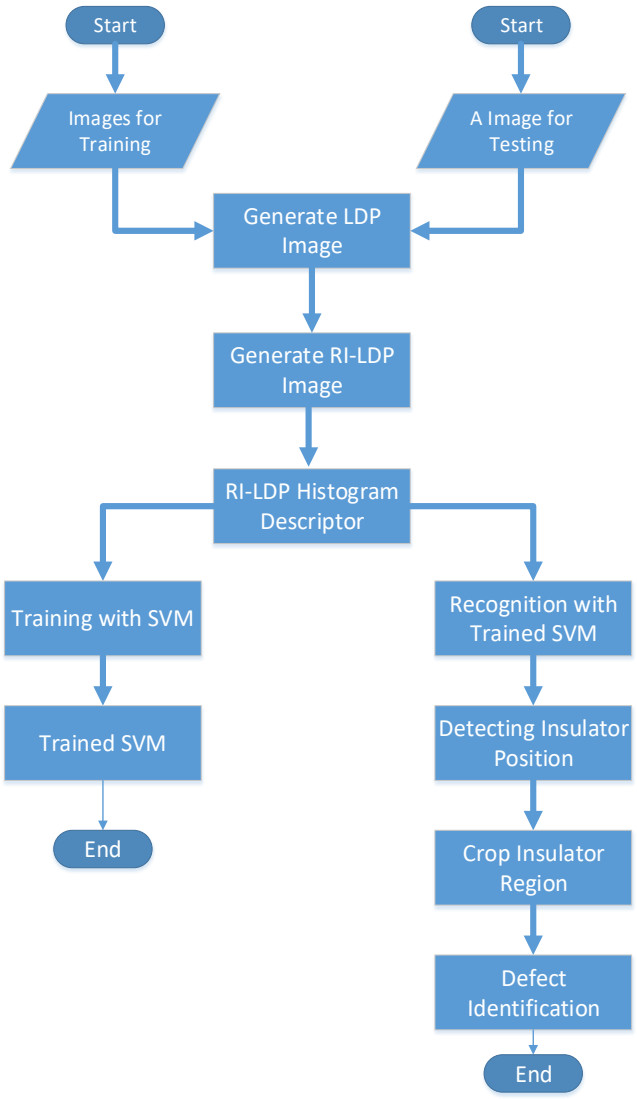


Fig. 2. Flowcharts of the proposed rotation invariant local directional based insulator detection and defect identification.

II. Insulator Detection using Rotation Invariant Local Directional Pattern (RI-LDP)

We can locate power line insulators in different size, shape, and/or texture. However there are also some trivialities. Every insulator contains a repeating patterned called cap. Number of caps may vary significantly which made the insulator length variable. Insulators are made of ceramic and/or glass which make sure low textured appearance. However, these low textured insulators mostly appear in clutter background which made it difficult the recognition. The aspect ratios of insulators vary in a certain range as long as the images are taken within tolerable viewing angle. All these properties may entice researcher to see the insulator detection problem as a generalize object detection using local texture feature [13]. However, in reality the insulators appear in varying orientation which makes the detection process complex. The rotation variation can be handled by explicitly normalizing the rotated object or using a rotation invariant feature. In this work, a

rotation invariant local featured named rotation invariant local directional pattern (RI-LDP) is proposed which enables the detector to detect insulator caps even if those are not in same orientation. However, this rotation invariance is not sufficient to detect the whole string of insulator caps as a single insulator. Therefore, we propose a novel post processing method which finally combines series of insulator caps as a single insulator. The overall step of the proposed method is shown in Fig. 2. In the following sub-sections, the basic LDP code generation process is briefly described, and then the proposed rotation invariant (RI-LDP) is explained in details.

$$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \quad \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \quad \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$$

East M_0 North East M_1 North M_2 North West M_3

$$\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \quad \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \quad \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \quad \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

West M_4 South West M_5 South M_6 South East M_7

Fig. 3. Eight different Kirsch edge masks.

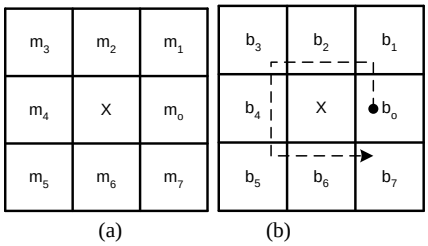


Fig. 4. (a) Eight directional edge response positions; (b) LDP binary bit positions.

A. Local Directional Pattern (LDP)

Local Directional Patten (LDP) is a local texture pattern which efficiently describes spatial structure of a tiny image patch. Typically, a local texture is calculated by analyzing image pixel value and/or gradient direction of a tiny image region. However, LDP utilizes the gradient magnitude of different direction and by analyzing all directional responses generates an eight bit binary number to describe the image patch [14], [15]. One of the main strength of LDP is utilization of edge responses instead of image pixel as edge responses are typically more robust in adverse imaging situation. Therefore, local image primitive like dark pot, corner, junction, edge, bright spot etc. can be encoded by LDP with lesser influence of external impurity. LDP can be generated for any pixel by analyzing a 3x3 images region centering at that code pixel. For this, eight directional edge responses values $\{m_i\}$, $i = 0,1,...,7$ are computed by Kirsch masks M_i in eight different orientations centered on its position. The eight different masks are shown in Fig. 3.

The relative magnitudes of response values carry noteworthy information. The higher magnitude in a particular direction ensures prominent texture in that direction. However, the relatively lower magnitudes also carry some important information. Therefore, we are interested to know the *k* most prominent directions in order to generate the LDP. Here, the top *k* directional bit responses b_i are set to 1. The remaining (*8-k*) bits of 8-bit LDP pattern is set to 0. Finally, the LDP code is derived by (1). Fig. 4 shows the mask response and LDP bit positions, and Fig. 5 shows an exemplary LDP code with *k=3*.

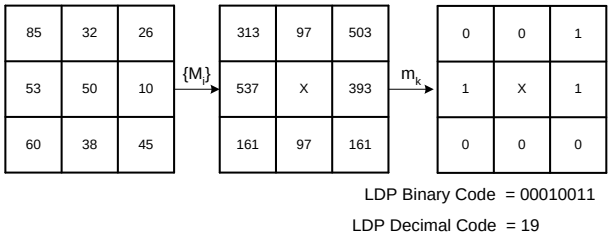


Fig. 5. Example of LDP code generation with k=3.

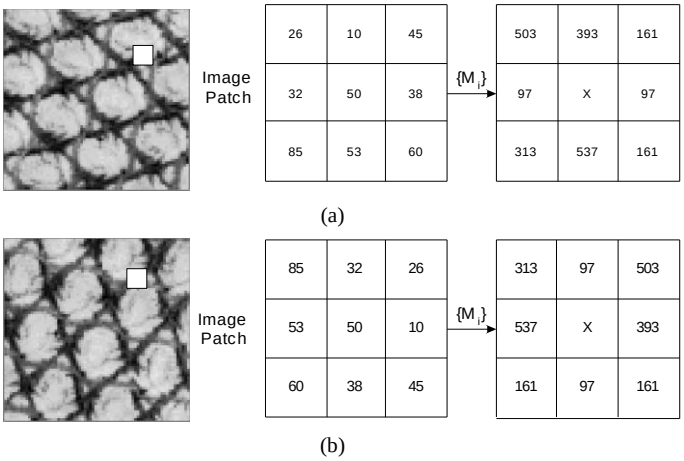


Fig. 6. Modified edge responses value after rotation of the image. (a) Original image along with eight edge response values, (b) Rotated image along with changed edge responses values.

$$LDP_k = \sum_{i=0}^{7} b_i(m_i - m_k) \times 2^i \,. \tag{1}$$

$$b_i(a) = \begin{cases} 1 & a \geq 0 \\ 0 & a < 0 \end{cases} \tag{2}$$

where, m_k is the *k*-th highest value among all eight directional response values. LDP texture code provides similar pattern in adverse imaging situation like change in illumination and/or noise presence [16] because inherently directional response magnitudes are more stable than intensity values.

LDP based descriptor is calculated after computing LDP code for each pixel (r,c) of the input image I of size $M \times N$. LDP descriptor generation process is motivated by other texture based feature descriptor where histogram of individual feature plays the role of image descriptor [17], [18]. In accordance, we also generate a LDP histogram H using for describing that image. The equation of this histogram generation is shown with (3) and (4).

$$H(\tau) = \sum_{r=1}^{M} \sum_{c=1}^{N} f\left(LDP_k(r,c), \tau\right). \tag{3}$$

$$f(a,\tau)=\begin{cases}1 & a=\tau\\ 0 & \text{otherwise}\end{cases} \quad (4)$$

where, τ is the all possible LDP code. The number of possible LDP code varies as our choice of *k* value and can be expressed by $C_k^8=\frac{8!}{k!(8-k)!}$.

B. Rotation Invariant Local Directional Pattern (RI-LDP)

Change in rotation appearance of an image affects the pixel value of the image. Though the spatial distribution of image pixels is modified, the relative appearance of pixels is not affected by mere rotation (see Fig. 6). The image clearly exhibits that the edge direction of the object is altered due to alteration of the image orientation. Consequently, direction of highest gradient magnitude will be modified and that will lead to a completely different LDP code. As the content of the image is same but appeared in different orientation, we require a steady LDP code to achieve rotation invariant detection. For achieving this, we analyzed the gradient magnitude of all directions and found that though the direction of highest magnitude changes, but the relative position of other lower rank magnitude with highest one is unchanged. This trend is explained using a small image patch shown in Fig. 6. In Fig. 6(a), it shows an image along with eight directional gradient magnitudes. This image is rotated by 90^0 in a counterclockwise direction and modified gradient magnitudes in Fig. 6(b). By observing these two set of gradient magnitude, we found that the relative position of gradient magnitude is preserved, it just shifted by the corresponding image rotation value. In this example it shifted two places in rightward direction which agreed 90^0 clockwise rotations. In addition to that, it is well known fact that image rotation in spatial domain is equivalent to circular shift in feature vector [19]. Based on these observations we proposed a simple method for achieving the rotation invariant LDP code. The proposed method performs the circular shift operation to the original LDP code until the bit representing the most prominent edge response is aligned with the least significant bit b_0 as shown in Fig. 7.

For example, if the 8 directional responses of a pixel is given by the set {3, 5, 2, 2, 8, 4, 3, 2}, which stars with m_0, then the original LDP code is 00110010. The rotation invariant code for the pixel can be obtained by shifting the directional response as {8, 4, 3, 2, 3, 5, 2, 2}. The resulting code is 00100011. This rotation normalization method tries to evaluate the objects with different rotational appearance by aligning the inherent texture pattern along with the most prominent texture property and then compare. This rotation invariant LDP code which is denoted by RI-LDP can be generated with (5).

$$\text{RI-LDP} = \text{ROR}(\text{LDP}, d-1) \quad (5)$$

where ***d*** is the bit position of the strongest edge response and ROR defined circular shift of the bit pattern.

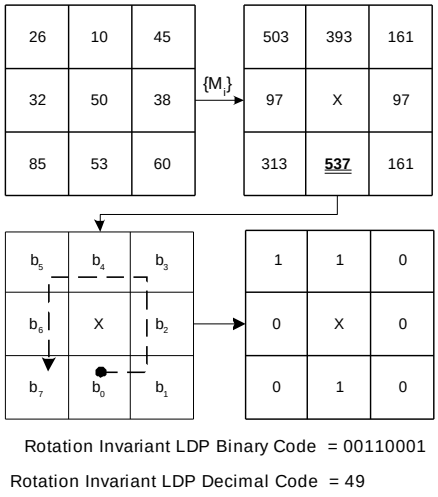


Fig. 7. Steps of rotation invariant LDP code generation.

The LDP operator can produces at most C_k^8 different code, because out of eight bit data *k* bit must be set 1. However, when circular shift on the LDP code is applied, the left most bit certainly become 1. Aftermath, out of remaining seven bit data (*k-1*) bit need be set 1. That's why number of possible rotation invariant LDP code reduces to C_{k-1}^7. In consequence rotation invariant LDP descriptor will be a histogram with C_{k-1}^7 bins in comparison of C_k^8 bin in original LDP. This LDP histogram based descriptor is then used to classify between insulator region and non-insulator region of an image within a sliding window framework. The two class classification is carried out using support vector machine (SVM).

C. Support Vector Machine (SVM)

Support vector machine (SVM) is one of the most popular machine learning techniques. SVM which is proposed by Vapnik et al. [20] is shown effectiveness as supervised pattern recognition. During classification, SVM try to separate a given set of labeled data with the best hyperplane by transforming it into higher dimension. SVM choose the hyperplane when the distance from hyperplane to the closest data point of each class is maximized. Feature vector for any image processing problem is non-linear; hence SVM apply complex but easy to implementable kernel without using potentially infinite dimensional feature vector. The hyperplane found by the SVM in the high dimensional feature space corresponds to a nonlinear boundary in the input space.

The choice of appropriate kernel and different parameter associate to the selected kernel is very critical during classification performance using SVM. In literature, we found that linear, polynomial and radial basis function (RBF) kernels are the most frequently used during image based object detection. The choice of inappropriate kernel parameter can

leads to very poor classification accuracy. Hence grid-search approach is applied for selecting the parameters [21]. The parameter of specific kernel that produces best classification accuracy is picked.

Typical SVM based object detection requires number of training samples grouped into two sets – one is positive sample and other one is negative sample. These two sets are used to train a binary classifier which can subsequently predict where an unknown sample belongs. However, to make the training computationally viable, only a subset of possible sample data is used. D. King [22] proposed Maximum Margin Object Detection (MMOD) method which tries to optimize the classifier using all the candidate windows available in the image. In this work, dlib's [23] structural SVM based algorithm is used which enables us to train on all the sub-windows in every image.

III. Insulator Defect Analysis

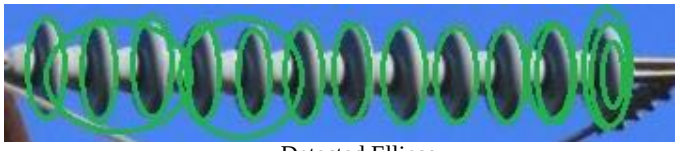
We proposed an insulator defect analysis system which can evaluate detected insulator region for identifying the potential defect. There are different kinds of defects which deteriorate the effectiveness of the insulators. Some of those defects modify the electrical property of the insulator rather than appearance. These types of defect are out of our vision-based identification system's scope. Rather we only focused on the defects which alter the appearance of the insulators like change the color, shape and/or texture. Some of those defects affect core region while some others affect in the cap region. In addition to that sometimes defects may appear in some caps while others remain in good condition. Hence, each cap should be individually segmented from the core region. After that, we can analyze each caps or core region for identifying the defects. Hence, the defect analysis work can be separated into two parts: 1) insulator partitioning, 2) defect identification.

A. Insulator Partitioning

Individual caps are detected by detecting elliptical regions on the detected insulator region followed by clustering based on orientation and size of the ellipses. At first, elliptical shaped regions in the insulator are detected by a method proposed by Fornaciari [24]. All the detected elliptical shaped arcs are labeled into four groups and estimate the ellipse parameters using the decomposed parameter space. At first, edge pixels are classified in two main directions (i.e., positive and negative) according to their gradient phases. Edge pixels with the same gradient phases are grouped together and classified as an arc according to their convexity.

This method is capable to detect those ellipses whose arcs are visible and can be detected in at least three quadrants. Consequently, this method looks for combinations of three arcs, called triplets, each belonging to a different quadrant. A selected triplet forms a *candidate* ellipse. As from the triplet information, we may already know its center; we estimate the remaining three parameters in a decomposed Hough space requiring three 1D accumulators. Candidate ellipses are then validated according to the fitness of the estimation with the actual edge pixels.

Original Image



Detected Ellipse

Fig. 8. Detected ellipse from the insulator region.

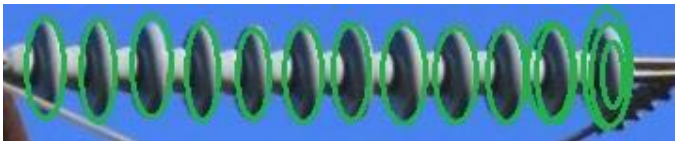


Fig. 9. Detected ellipse after selecting ellipses of largest group.

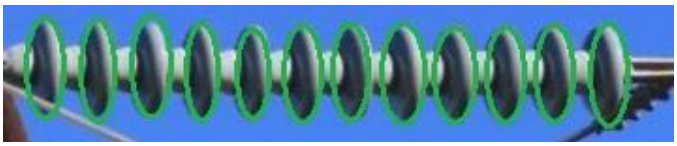


Fig. 10. Separately detected each insulator.

However, due to possible clutter in the image some non-cap region can be detected as an ellipse (see Fig. 8). We have to eliminate those non-caps region and keep only the true cap regions. To achieve this, we cluster the ellipses based on their orientation and size and ellipses belongs to the largest cluster are retained. Sometimes, multiple ellipses can be detected around a single cap as shown in Fig. 9. We keep only one ellipse if there are multiple overlap ellipses depending on the ellipse parameters of the other ellipses of the same group. The result after this step is shown in Fig. 10.

B. Defect Identification

After partitioning insulator, each cap is analyzed for defect identification. A number of different defects may appear in the insulators. In this work, we attempt to identify defects which alter appearance i.e. color, shape and/or texture of the insulator. Defects which change only physical property of the insulator but not appearance is out of scope this research work. Therefore, in this work we consider the following defects: 1) whitening; 2) aligatoring effect; 3) bullet damage; 4) contamination; and 5) crack.

All of the above mentioned defects change the cap's surface area which is normally smooth. Therefore, we can identify whether any cap is altered by the analyzing local texture of the cap's region. In this respect, we calculate the LDP histogram from each of the cap region for identifying whether cap is affected by any of these defects. This calculated LDP histogram is then feed into a multi-class classifier for identify specific defect. However, due to lack of availability of defected sample, sophisticated classifier like SVM or CNN is not

utilized rather simple nearest-neighbor classifier is used in classify the defect type. As our proposed LDP descriptor is histogram based feature, we need to calculate similarity or dissimilarity between two histograms. For this we can choose any one from Histogram intersection, Log-likelihood statistics and/or Chi square statistics (χ^2). In this work Chi square statistics is used for comparing two histograms.

$$\chi^2(S,M) = \sum_i \frac{\left(S_i - M_i\right)^2}{\left(S_i + M_i\right)} \quad (6)$$

IV. Experimental Result

In this section, insulator detection and defect identification results are presented and discussed. As there is no publicly available dataset for insulator detection, we develop our own evaluation dataset with two different image resolutions. Low-resolution dataset contains 298 images where in total 1470 insulators appeared in those images. However, the high-resolution dataset contains 395 images (2592×1944 pixels) with 722 labeled insulators. The segmented ground-truth information is generated by manually tagging the insulators.

A. Result of Insulator Detection

The performance of the insulator detector is evaluated by finding detection accuracy of the marked insulators. As insulator appears in arbitrary orientation in the dataset, each insulator is marked by a rotated rectangle. To decide whether detected window is correct detection or not, well-known Pascal scores [25] has been used. Pascal score is calculated from the overlap of our generated bounding box $\mathbf{B_{cl}}$ to the ground-truth $\mathbf{B_{gt}}$ by $P(B_{cl}, B_{gt}) = \frac{area(B_{cl} \bigcap B_{gt})}{area(B_{cl} \bigcup B_{gt})}$. An object is considered detected if $P(B_{cl}, B_{gt}) > 0.5$

a) Detection in Low-Resolution Images: Our objective is to maximize the number of correct detections and minimize the number of false detections. The two used evaluation metrics are precision, the fraction of correct detections to the total number of detections made by our detector, and recall, the fraction of correctly detected objects to the number of annotated objects. Our detector provides a score for each detection, and with average detection score, we achieved recall rate 51.29% with 81.82% precision in low-resolution image dataset. But in low-resolution images, our objective is detecting correct insulators more even if it provides some false detection. Because those false alarms can be eliminated by later steps in high resolution. By lowering detection threshold, we can relax the detection criteria which in turn increase recall rate in the expense of precision. The result is shown in Table III by varying detection threshold. With detection threshold -0.3, our system can detect more than 80% of insulators from low-resolution images. Some of the insulators which are not detected appeared close to already detected insulators. Therefore, when the second camera zoomed in towards the detected insulators the missed insulators will also appear in high-resolution images. Consequently, those missed insulators can be detected in high-resolution.

b) Detection in High Resolution Images: In high-resolution image, the objective is same as low resolution i.e., to maximize the number of correct detections and minimize the number of false detections. However in high-resolution, we cannot tolerate high false alarm as we do not have any further steps which will eliminate those. With average detection score, the proposed system achieved recall rate 95.74% with 89.94% precision. Table I shows recall and precision rate of the proposed method and other comparable method. It clearly demonstrate the suitability of the proposed method over other state of the art techniques. Table I shows that recall rate of the proposed method is just behind the method of Oberweger et al. However, the precision rate is much higher than that of Oberweger et al.

TABLE I. Precision and Recall Curve of Proposed Method and Other Methods

Method	Recall	Precision
Oberweger et al [4]	98%	33%
Liao and An [5]	91%	87%
Wu and An [2]	86.47%	85.59%
Proposed Method	95.74%	89.94%

For further analyze the performance of proposed method, we compare precision recall (PR) curve. As our detector provides a score for each detection, we can vary this score to elaborate the trade-off between recall and precision metrics. Fig. 11 exhibits the precision recall curve of our method and Oberweger's method which exhibits the superiority of the proposed system.

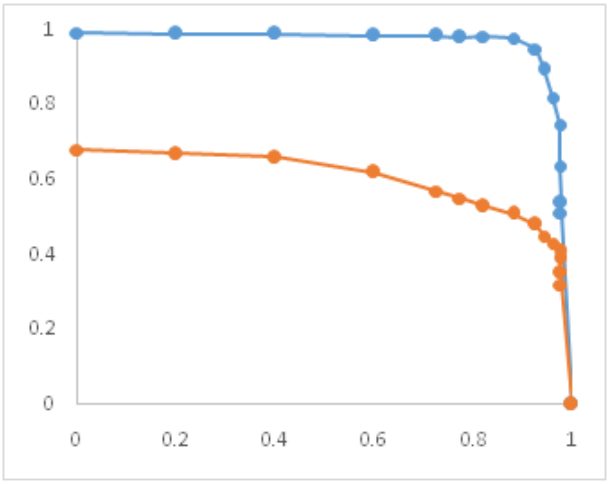


Fig. 11. Precision-recall curve for insulator detection.

By further scrutinizing the detected insulator images, we found that some of the insulators are considered as not detected due to low overlap value. In our evaluation, the required Pascal criterion was 0.5 which is very strict due to the fact that we use rotated bounding boxes. However the originally Pascal criterion is intended for axis aligned bounding box. Subjectively speaking, an overlap score of 0.5 fits the insulator very well, whereas a lower threshold might be good enough for visually consider as a true detection. Therefore, lowering

Pascal criterion may improve our detection performance in expense of some loss in precision and the result with varying overlap threshold is shown in Table II.

TABLE II. INSULATOR DETECTION RESULT WITH VARYING PASCAL OVERLAP THRESHOLD

Overlap Threshold	Recall	Precision
0.50	95.74%	89.94%
0.40	94.79%	89.39%
0.30	94.94%	89.22%

B. Result of Insulator Defect Analysis

A proper insulator partitioning is essential for effective defect analysis. By insulator portioning, we mean separation of each cap from one another so that we can analyze each cap independently. Therefore we first evaluate the insulator partitioning and then the defect identification itself.

1) Result of Insulator Partitioning: For the evaluation of insulator portioning, we compare number of caps correctly segmented ($\mathbf{N_{sg}}$) by proposed method with ground truth number of visible caps ($\mathbf{N_{gt}}$)in the insulator image. As evaluation criterion of correct cap segmentation, we compare ground truth cap area with segmented cap area with Pascal score. If Pascal score is above 0.8 then we consider the cap as a correct detection. Finally, an insulator is considered as correctly partitioned if the following condition is true.

$$abs(N_{sg} - N_{gt}) <= 1$$

The appearance of insulators varies a lot due to image viewing angle. The appearance of caps can be near frontal to almost side view. We group the insulator images in three categories depending on the viewing angle of the image.

- Group1: Viewing angle of the insulators are almost frontal.
- Group 2: Viewing angle of the insulators varies from $\sim 10^0$ to $\sim 50^0$.
- Group 3: Viewing angle of the insulators are more than 50^0.

It is almost impossible to segment the caps from insulator of group1 as caps are not individually visible. However, the core part is visible in this orientation. So we detect core region when images are captured in this orientation. When the viewing angle increased from near frontal, the cap region becomes separable. Therefore, we partition each insulator of these two groups and result of partitioning is shown in Table III.

TABLE III. PERFORMANCE OF INSULATOR PARTITIONING

Group No	Number of Insulator Images	Number of Insulator Correctly Partitioned	Partition Accuracy
Group 2	76	73	96.05%
Group 3	17	8	47.05%

2) Result of Defect Identification: To the best of our knowledge, there is currently no publicly available dataset for insulator defect identification. Hence, we use our own insulator database to analyze the defect identification performance. There are five defected insulator images for each type of defect. Furthermore, we add another 20 non-defected images to the database. Accordingly, there are 45 images in our defect identification database with six caps in each insulator making total 6x45 = 270 insulator cap. Out of these 270 caps, 20x6 = 120 caps are in good condition and other 250 (25x6) contains one of those five defects. The overall defect identification result is shown in Table IV. So far, we have discussed the average defect identification accuracy of five different defects. To get a better picture of the identification accuracy of individual defect types, the confusion matrix is shown in Table V.

TABLE IV. PERFORMANCE OF INSULATOR DEFECT IDENTIFICATION

Number of Caps	Number of Caps correctly classified	Defect identification accuracy
270	244	90.37%

TABLE V. CONFUSION MATRIX OF 6-CLASS DEFECT CLASSIFICATION

	White.	Alliga.	Bullet.	Contam.	Crack	No Def.
Whitening	28	0	0	0	1	1
Alligatoring	2	26	0	0	2	0
Bullet	0	1	25	0	3	1
Contamination	0	1	0	28	0	1
Crack	0	2	1	0	27	0
No Defect	1	0	0	4	5	110

V. CONCLUSION

In this work, we have presented a rotation invariant local directional feature showing encoding scheme for representing texture of insulator image. The proposed rotation invariant local directional pattern (RI-LDP) feature shows efficacy in detecting rotated insulator. We also show the effectiveness of the proposed method in the low-resolution image which allows us coarse to fine detection approach. Experimental results illustrate that suitability of the proposed method over other state-of-the-art methods. After successful detection and segmentation of the insulators from images with the cluttered background, their condition is analyzed by a gradient based feature named local directional pattern (LDP). The insulators are partitioned into the individual cap for better defect identification. The results of defect identification show that almost all defective insulators can be separated from the good insulator. Thus the proposed insulator analysis system would improve the reliability of power supply and the requirement of technical manpower can also be substantially reduced as non-technical persons can be employed for acquiring images and feeding to the system for analysis.

REFERENCES

[1] W. G. Li, G. Ye, F. Huang, S. Wang, and W. Z. Chang, “Recognition of insulator based on developed MPEG-7 texture feature,” in Proc. IEEE ICISP, pp. 265–268, 2010.

[2] Q. Wu and J. An, "An active contour model based on texture distribution for extracting inhomogeneous insulators from aerial images", IEEE Trans. Geosci. Remote Sens., vol. 52, no. 6, pp. 3613-3626, 2014.

[3] X. Wang and Y. Zhang, “Insulator Identification from Aerial Images Using Support Vector Machine with Background Suppression” in Proc.

International Conference on Unmanned Aircraft Systems (ICUAS) June 7-10, pp. 892 – 897, 2016.

[4] M. Oberweger, A. Wendel, and H. Bischof, "Visual recognition and fault detection for power line insulators," in Proc. 19th Computer Vision Winter Workshop, pp. 1–8, 2014.

[5] S. Liao, J. An, "A robust insulator detection algorithm based on local features and spatial orders for aerial images," IEEE Geosci. Remote Sens. Lett. vol. 12, no. 5, pp. 963–967, 2015.

[6] Y. Zhai, Y. Wu, H. Chen, and X. Zhao, "A method of insulator detection from aerial images," Sensors & Transducers, vol. 177, no. 8, 2014.

[7] B. Li, D. Wu, Y. Cong, Y. Xia, and Y. Tang, "A method of insulator detection from video sequence," in 2012 International Symposium on Information Science and Engineering (ISISE), pp. 386–389, 2012.

[8] Y. Tiantian, Y. Guodong, Y. Junzhi, "Feature Fusion Based Insulator Detection for Aerial Inspection", Proceedings of the 36th Chinese Control Conference, July 26-28, 2017, Dalian, China

[9] Z. Zhao, X. Fan, G. Xu, L. Zhang, Y. Qi, and K. Zhang, "Aggregating Deep Convolutional Feature Maps for Insulator Detection in Infrared Images" IEEE Access, vol. 5, 2017.

[10] T. Jabid and M.Z. Uddin, "Rotation invariant power line insulator detection using local directional pattern and support vector machine", IEEE Conference on Innovations in Science Engineering and Technology (ICISET), 2016.

[11] V.S. Murthy, K. Tarakanath, D.K. Mohanta, S. Gupta, "Insulator condition analysis for overhead distribution lines using combined wavelet and support vector machine (SVM)," IEEE Transactions on Dielectrics and Electrical Insulation, vol. 17, no. 1, pp. 89–99, 2010

[12] V. S. Murthy, S. Gupta and D. K. Mohanta, "Digital image processing approach using combined wavelet hidden Markov model for well-being analysis of insulators", IET Image Process., vol. 5, no. 2, pp. 171-183, 2011.

[13] D.A. Lisin, M.A. Mattar, M.B. Blaschko, E.G. Learned-Miller, M.C. Benfield, "Combining local and global image features for object class recognition," in Workshop on Learning in Computer Vision and Pattern Recognition at IEEE CVPR, 2005.

[14] T. Jabid, M. H. Kabir, and O. S. Chae, "Robust Facial Expression Recognition based on Local Directional Pattern," ETRI Journal, vol. 32, no. 5, October, 2010. Pp. 784-794.

[15] T. Jabid, M. H. Kabir, and O. S. Chae, "Local Directional Pattern (LDP) for Face Recognition," IEEE International Conference on Consumer Electronics, January 2010.

[16] T Jabid, MH Kabir, O Chae, "Local Directional Pattern (LDP) A Robust Image Descriptor for Object Recognition", Proceedings of the IEEE Advanced Video and Signal Based Surveillance (AVSS), pp. 482-487, August 29-September 1.

[17] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", Int'l J. Computer Vision, vol. 2, no. 60, pp. 91-110, 2004.

[18] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005.

[19] D. Zhang, A. Wong, M. Indrawan, and G. Lu, "Content-based image retrievalusing gabor texture features," in IEEE Transactions PAMI, 2000, pp. 13‑15.

[20] C. Cortes and V. Vapnik, "Support vector networks", Machine Learning," vol. 20, no. 3, pp. 273-297, 1995.

[21] C. -W. Hsu and C. -J. Lin, "A comparison on methods for multiclass support vector machines," IEEE Trans. Neural Networks, vol. 13, no. 2, 2002, pp. 415-425.

[22] D. E. King, "Max-margin object detection", CoRR, 2015

[23] D. E. King, "Dlib-ml: A machine learning toolkit", Journal of Machine Learning Research, pp. 1755-1758, 2009

[24] M. Fornaciari, A. Prati, and R. Cucchiara, "A fast and effective ellipse detector for embedded vision applications," Pattern Recognition, vol. 47, no. 11, pp. 3693–3708, 2014.

[25] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge", IJCV, vol. 88, no. 2, pp. 303–338, 2010.

Machine-Learning Techniques for Customer Retention: A Comparative Study

Sahar F. Sabbeh
Faculty of computing and information sciences, King AbdulAziz University, KSA
Faculty of computing and information sciences, Banha University, Egypt

***Abstract*—Nowadays, customers have become more interested in the quality of service (QoS) that organizations can provide them. Services provided by different vendors are not highly distinguished which increases competition between organizations to maintain and increase their QoS. Customer Relationship Management systems are used to enable organizations to acquire new customers, establish a continuous relationship with them and increase customer retention for more profitability. CRM systems use machine-learning models to analyze customers' personal and behavioral data to give organization a competitive advantage by increasing customer retention rate. Those models can predict customers who are expected to churn and reasons of churn. Predictions are used to design targeted marketing plans and service offers. This paper tries to compare and analyze the performance of different machine-learning techniques that are used for churn prediction problem. Ten analytical techniques that belong to different categories of learning are chosen for this study. The chosen techniques include Discriminant Analysis, Decision Trees (CART), instance-based learning (k-nearest neighbors), Support Vector Machines, Logistic Regression, ensemble–based learning techniques (Random Forest, Ada Boosting trees and Stochastic Gradient Boosting), Naïve Bayesian, and Multi-layer perceptron. Models were applied on a dataset of telecommunication that contains 3333 records. Results show that both random forest and ADA boost outperform all other techniques with almost the same accuracy 96%. Both Multi-layer perceptron and Support vector machine can be recommended as well with 94% accuracy. Decision tree achieved 90%, naïve Bayesian 88% and finally logistic regression and Linear Discriminant Analysis (LDA) with accuracy 86.7%.**

Keywords*—Customer relationship management (CRM); customer retention; analytical CRM; business intelligence; machine-learning; predictive analytics; data mining; customer churn

I. Introduction

For any business, customers are the basis for its success and revenue and that is why companies become more aware of the importance of gaining customers' satisfaction. Customer relationship management (CRM) supports marketing by selecting target consumers and creating cost-effective relationships with them. CRM is the process of understanding customer behavior in order to support organization to improve customer acquisition, retention, and profitability. Thus, CRM systems utilize business intelligence and analytical models to identify the most profitable group of consumers and target them achieve higher customer retention rates. Those models can predict customers with high probability to churn based on analyzing customers' personal, demographic and behavioral data to provide personalized and customer-oriented marketing campaigns to gain customer satisfaction. The lifecycle of business – customer relationship includes four main stages: 1) *identification*; 2) *attraction*; 3) *retention*; and 4) *development*.

1) Customer identification/acquisition: This aims to identify profitable customers and the ones that are highly probable to join organization. Segmentation and clustering techniques can explore customers' personal and historical data to create segments/sub-groups of similar customers [1], [2].

2) Customer attraction: The identified customer segments / sub-groups are analyzed to identify the common features that distinguish customers within a segment. Different marketing techniques can be used to target different customer segments such targeted advertising and/or direct marketing [3].

3) Customer retention: This is the main objective of CRM as retaining existing customers is at least 5 to 20 times more cost effective than acquiring new ones depending on business domains [4], [5]. Customer retention includes all actions taken by organization to guarantee customer loyalty and reduce customer churn. Customer churn refers to customers moving to a competitive organization or service provider. Churn can be for better quality of service, offers and/or benefits. Churn rate is an important indicator that all organizations aim to minimize. For this sake, churn prediction is an integral part of proactive customer retention plan [6]. Churn prediction includes using data mining and predictive analytical models in predicting the customers with high likelihood to churn/defect. These models analyze personal and behavioral customer data for tailored and customer-centric retention marketing campaigns [7].

4) Customer development: The main objective of this phase is to increase the amount of customer transactions for more profitability. For this sake, market basket analysis, customer lifetime value, up, and cross selling techniques are used. Market basket analysis tries to analyze customers' behavior patterns to maximize the intensity of transactions [8], [9]. Analyzing customer lifetime value (CLTV) can help identifying the total net income expected from customer [10]-[12]. Up and/or Cross selling include activities that increase the transactions of the associated services/products [13], [14].

Customer retention and churn prediction have been increasingly investigated in many business domains, including, but not limited to, telecommunication [15]-[18],

banking [19]-[21], retail [22] and cloud services subscriptions [23], [24]. Different statistical and machine-learning techniques are used to address this problem. Many attempts have been made to compare and benchmark the used techniques for churn prediction. In [28], [66] a comparison between (Decision trees, Logistic regression and Neural Network) models was performed. The study found that neural network perform slightly higher than the other two techniques. Another comparison between a set of models against their boosted versions is discussed in [67]. This study included two-layer Back-Propagation neural network (BPN), Decision Trees, SVM and Logistic Regression. The study showed that both decision trees and BPN achieved accuracy 94%, SVM comes next with 93% while Logistic Regression failed with accuracy 86%. Additionally, study showed 1-4% performance improvement in the boosted versions. In [68] the study investigated the accuracy of different models (Multi-layer perceptron (MLP) and Decision Tree (C5)). The study showed that MLP achieves accuracy of 95.51%, which outperforms C5 decision tree 89.63%.

Most of comparisons in the literature did not consider a study that covers the various categories of learning techniques. The bulk of the models applied for churn prediction fall into one of the following categories:

1) Regression analysis, 2) Decision tree–based, 3) Support Vector Machine, 4) Bayesian algorithm, 5) Instance – based learning, 6) Ensemble learning, 7) Artificial neural network, and 8) Linear Discriminant Analysis.

This study presents a comparative study of the most used algorithms for predicting customer churn. The comparison is held between algorithms from different categories. The main goal is to analyze and benchmark the performance of the models in the literature. The selected models are:

1) Regression analysis: logistic regression.
2) Decision tree–CART.
3) Bayes algorithm: Naïve Bayesian.
4) Support Vector Machine
5) Instance – based learning: k-nearest Neighbor.
6) Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest.
7) Artificial neural network: Multi-layer Perceptron.
8) Linear Discriminant Analysis.

A. Contribution

The key contribution of this paper is the analysis of most common learning techniques in the state of the arts and the evaluation of their accuracy.

This paper is organized as follows: Section 2 presents a state of the arts of data mining techniques for churn prediction and briefly discusses the evaluated techniques. In Section 3, methodology of the study is discussed, Results and discussion are given in Section 4 and finally Section 5 concludes this work.

II. Machine-Learning for Churn Prediction

Machine-learning techniques have been widely used for evaluating the probability of customer to churn [25]. Based on a survey of the literature in churn prediction, the techniques used in the bulk of literatures fall into one of the following categories 1) Regression analysis; 2) Tree – based; 3) Support Vector Machine; 4) Bayesian algorithm; 5) Ensemble learning; 6) Sample – based learning; 7) Artificial neural network; and 8) Linear Discriminant Analysis. A brief introduction of the chosen algorithms is presented in this section.

1) Regression analysis: Regression analysis techniques aim mainly to investigate and estimate the relationships among a set of features. Regression includes many models for analyzing the relation between one target/response variable and a set of independent variables. Logistic Regression (LR) is the appropriate *regression analysis* model to use when the dependent variable is binary. LR is a predictive analysis used to explain the relationship between a dependent binary variable and a set of independent variables. For customer churn, LR has been widely used to evaluate the churn probability as a function of a set of variables or customers' features [26]-[33].

2) Decision Tree: Decision Tree (DT) is a model that generates a tree-like structure that represents set of decisions. DT returns the probability scores of class membership. DT is composed of: a) **internal Nodes:** each node refers to a single variable/feature and represents a test point at feature level; b) **branches,** which represent the outcome of the test and are represented by lines that finally lead to c) **leaf Nodes** which represent the class labels. That is how decision rules are established and used to classify new instances. DT is a flexible model that supports both categorical and continuous data. Due to their flexibility they gained popularity and became one of the most commonly used models for churn prediction [27]-[29], [33]-[36].

3) Support Vector Machine: Support Vector Machine (SVM) is a supervised learning technique that performs data analysis in order to identify patterns. Given a set of labeled training data, SVM represents observations as points in a high-dimensional space and tries to identify the best separating hyperplanes between instances of different classes. New instances are represented in the same space and are classified to a specific class based on their proximity to the separating gap. For churn prediction, SVM techniques have been widely investigated and evaluated to be of high predictive performance [37]-[41].

4) Bayes Algorithm: Bayes algorithm estimates the probability that an event will happen based on previous knowledge of variables associated with it. Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem. It adopts the idea of complete variables independence, as the presence/absence of one feature is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB is a supervised learning technique that bases its predictions for new instances based on the analysis of their ancestors. NB

model usually outputs a probability score and class membership. For churn problem, NB predicts the probability that a customer will stay with his service provider or switch to another one [42]-[46].

5) Instance – based learning: Also known as ***memory-based learning***, new instances are labeled based on previous instances stored in memory. The most widely used instance based learning techniques for classification is K-nearest neighbor (KNN). KNN does not try to construct an internal model and computations are not performed until the classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. Instance is labeled with the class most common among its neighbors. KNN determine neighbors based on distance using Euclidian, Manhattan or Murkowski distance measures for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training instances (k) that are the closest to the new point, and assign label from these. Despite its simplicity, KNN have been applied to various types of applications. For churn, KNN is used to analyze if a customer churns or not based on the proximity of his features to the customers in each classes [17], [51].

6) Ensemble – based Learning: Ensemble based learning techniques produce their predictions based on a combination of the outputs of multiple classifiers. Ensemble learners include bagging methods (i.e. Random Forest) and boosting methods (i.e. Ada Boost, stochastic gradient boosting).

a) Random Forest

Random forests (RF) are an ensemble learning technique that can support classification and regression. It extends the basic idea of single classification tree by growing many classification trees in the training phase. To classify an instance, each tree in the forest generates its response (vote for a class), the model choses the class that has receive the most votes over all the trees in the forest. One major advantage of RF over traditional decision trees is the protection against overfitting which makes the model able to deliver a high performance [47]-[50].

b) Boosting – based techniques (Ada Boost and Stochastic Gradient Boosting)

Both AdaBoost (Adaptive Boost) and Stochastic Gradient Boosting algorithms are ensemble based algorithms that are based on the idea of boosting. They try to convert a set of weak learners into a stronger learner. The idea is that having a weak algorithm will perform better than random guessing. Thus, Weak learner is any algorithm that can perform at least a little better than random solutions. The two algorithms differ in the iterative process during which weak learners are created. Adaboost filters observations, by giving more *weight* to problematic ones or the ones that the weak learner couldn't handle and decrease the correctly predicted ones. The main focus is to develop new weak learns to handle those misclassified observations. After training, weak learners are added to the stronger learner based on their alpha weight (accuracy**)**, the higher alpha weight, the more it contributes to the final learner. The weak learners in AdaBoost are decision trees with a single split and the label assigned to an instance is based on the combination of the output of all weak learners weighted by their accuracy [56].

On the other hand, gradient bosting gives importance to misclassified/difficult instances using the remaining errors (pseudo-residuals) of the strong learner. At each iteration, errors are computed and a weak learner is adjusted to them. Then, the contribution of the weak learner to the strong one is the minimization of the overall error of the strong learner [57]. For churn prediction Adaboost [58]-[60] and Sochastic gradient [61], [62] have been used for churn prediction.

7) Artificial neural network: Artificial Neural Networks (ANNs) are machine-learning techniques that are inspired by the biological neural network in human brain. ANNs are adaptive, can learn by example, and are fault tolerant. An ANN is composed of a set of connected nodes (neurons) organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs. The weights on connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANN can address complex problems, such as the churn prediction problem. Multilayer perceptron (MLP) is an ANN that consists of at least three layers. Neurons in each layer use supervised learning techniques [52], [53]. In the case of customer churn problem, MLP has proven better performance over LR [21], [27], [28], [54], [55].

8) Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is a mathematical classification technique that searches for a combination of predictors that can differentiate two targets. LDA is related to regression analysis. They both attempt to express the relationship between one dependent variable and a set of independent variables. However, unlike regression analysis, LDA use continuous independent variables and a categorical dependent variable (target). The output label for an instance is estimated by the probability that inputs belong to each class and the instance is assigned the class with the highest probability. Probability in this model is calculated based on Bayes Theorem. LDA can be used for dimensionality reduction by determining the set of features that are the most informative. LDA has been used in for different classification tasks including customer churn [63]-[65].

III. METHODOLOGY

The first step before applying the selected analytical models on the dataset, explanatory data analysis for more insights into dataset was performed. Based on the observations data was preprocessed to be more suitable for analysis.

1) Data: The used dataset for the experiments of this study is a database of customer data of a telecommunication company. The dataset contains customers' statistical data including 17 explanatory features related to customers' service usage during day, international calls, customer service calls. 14% of the observations have the target variable "yes" and 86% observations have the value "No". The dataset variables of customer transactions and their descriptions are presented in Table I and Fig. 1 shows the distribution of each feature.

TABLE I. CUSTOMER FEATURES IN DATASET

Variable	DataType	Description
AccountLength	Integer	how long account has been active
Int'l Plan	categorical	International plan activated (yes, no)
VMail Plan	categorical	Voice Mail plan activated (yes , no)
VMailMessage	Integer	No. of voice mail messages
DayMins	Integer	Total day minutes used
DayCalls	Integer	Total day calls made
DayCharge	Integer	Total day charge
EveMins	Integer	Total evening minutes
EveCalls	Integer	Total evening calls
EveCharge	Integer	Total evening charge
NightMins	Integer	Total night minutes
NightCalls	Integer	Total night calls
NightCharge	Integer	Total night charge
IntlMins	Integer	Total International minutes used
IntlCalls	Integer	Total International calls made
IntlCharge	Integer	Total International charge
CustServCalls	Integer	Number of customer service calls made
Churn	categorical	Customerchurn(yes=churn,No=nochurn)

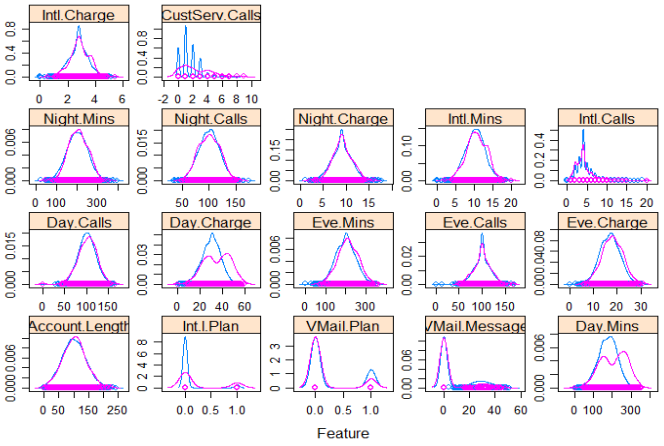


Fig. 1. Features' distribution.

2) Data preprocessing: Preprocessing includes three steps: a) data transformation, b) data cleaning and c) feature selection.

a) Data Transformation

Two of the explanatory variables (Int.l.plans and VMail Plan) were transformed from binominal form (yes/no) into binary form (1/0) to be more suitable for the selected models.

b) Data cleaning

This stage includes missing data handling/imputation: Some of the selected algorithms cannot handle missing data such as SVM. That's why missing value can be replaced by mean, median or zero. However, missing data replacement by statistically computed value (imputation) is a better option. The used dataset included missing values in some the numerical variables (Day Charge, Eve Mins, Intl Calls, Intl Charge and Night Charge) and two categorical variables (VMail Plan, Int'l Plan). Numerical data were replaced using random forest imputation technique [69]. And binary values were imputed using the techniques in [70]

c) Feature selection

Before model training, feature selection is one of the most important factors that can affect the performance of models. In this study, the importance of the used variables was measured to identify and rank explanatory variables influence on the target/response. This allows dimensionality reduction by removing variables/predictors with low influence on the target. Random forest technique can be used for feature selection using mean decrease accuracy. Mean decrease measures the impact of each feature on model accuracy. The model permutes values of each feature and evaluates model accuracy change. Only features having higher impact on accuracy are considered important [71]. Another well-known feature selection technique Boruta [72] was used. It is an improvement on RF. It considers all features that are relevant to the target variable whereas, most of techniques follow a minimal optimal method. Additionally, it can handle interactions between features [72]. Both techniques were applied to rank predictors based on the mean importance from Boruta and the mean decrease error calculated by random forest. Results shown in Table II shows that both models agree on the top three variables with the same rank (custServ.Calls, Int.l.Plan,Day.Mins). Both models agree on the next six features with different ranks (Day.Charge, VMail.Message, Intl.Calls, Eve.Charge, Intl.Mins and Eve.Mins). Both models give very low rank to the same four variables (Day.Calls,Night.Calls, Eve.Calls and Account.Length). Results are shown in Table II and Fig. 2.

3) Simulation Setup: For this study, the selected models are used to generate predictions using the dataset containing 3333 samples with 13 predictors and one response variable. 10-fold cross validations were used for models training and testing. Training and testing datasets are randomly chosen with cross validation 60% for training and 40% for testing. Each module requires initial parameters that are set as follows:

TABLE II. FEATURES MEAN IMPORTANCE

Feature	mean Importance	Mean Decrease Error	decision
CustServ.Calls	65.402	120.650	Confirmed
Int.l.Plan	47.719	80.223	Confirmed
Day.Mins	42.494	48.016	Confirmed
Day.Charge	34.429	37.424	Confirmed
VMail.Message	22.767	34.782	Confirmed
Intl.Calls	22.038	43.199	Confirmed
Eve.Charge	21.630	27.489	Confirmed
Intl.Mins	20.679	29.462	Confirmed
Eve.Mins	18.646	23.221	Confirmed
VMail.Plan	16.999	19.903	Confirmed
Intl.Charge	16.725	22.014	Confirmed
Night.Mins	9.787	15.141	Confirmed
Night.Charge	8.741	13.944	Confirmed
Day.Calls	0.301	0.227	Rejected
Night.Calls	-0.292	1.155	Rejected
Eve.Calls	-0.804	-0.443	Rejected
Account.Length	-1.067	-1.407	Rejected

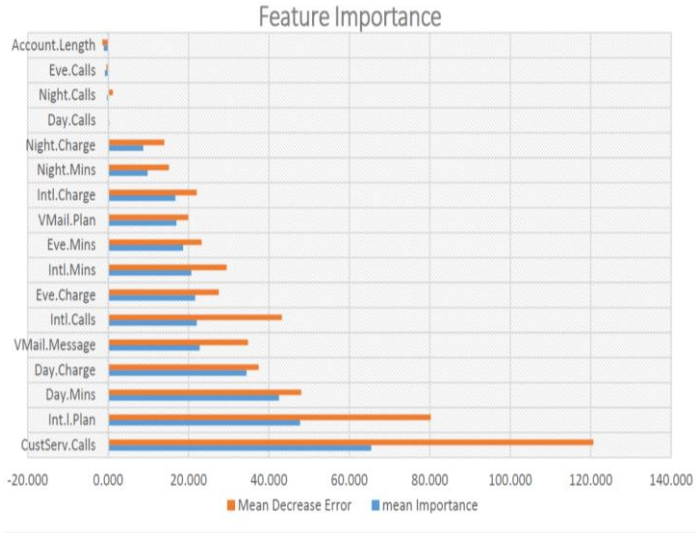


Fig. 2. Feature mean importance.

a) Decision Tree (CART)

One parameter is used for decision tree, CP which is a complexity parameter used to control the optimal tree size. Accuracy is used to choose the optimal model. The final (cp) value used for the model was: 0.07867495 as shown in Table III.

TABLE III. CART COMPLEXITY VARIABLE AND ACCURACY

Cp	Accuracy
0.07867495	0.8793827
0.08488613	0.8607829
0.08902692	0.8565868

b) Support Vector Machine

In order to train SVM, two main parameters are required: C and Sigma. The C parameter affects the prediction. It indicates the cost of penalty. Large value For C means high accuracy in training and low accuracy in testing. While small value for C indicates unsatisfactory accuracy. While sigma parameters has a more influence than C on classifications, as it affects hyperplane partitioning. A too large value of sigma leads to over-fitting, while small values lead to under-fitting [73]. Cross-validation was performed to select and tune performance parameters. The values that gave the highest accuracy were sigma = 0.06295758 and C = 1 as shown in Table IV.

c) K-nearest Neighbor

In KNN, one parameter needs to be tuned. K is the number of instances/neighbors that are considered for labeling an instance to a certain class. Cross validations were performed using different k values. Results shown in Table V shows that the highest accuracy is obtained using k=7.

d) AdaBoost

For Ada boost mode, *nIter* - represents the number of weak learners to be used. Grid search was used to determine the best accuracy. Results show that highest accuracy is at nIter=100 as shown in Table VI.

e) Random Forest

A forest of 500 decision trees has been built using the Random Forest algorithm. Error rate results indicate that after 100 trees, there is no significant error reduction. Another parameter is *mtry that* indicates number of predictors sampled for spliting at each node. Results in Table VII show that the optimal performance is at mtry = 7.

TABLE IV. ACCURACY USING DIFFERENT C VALUES

C	Accuracy
0.25	0.8997878
0.50	0.9171926
1.00	0.9261972

TABLE V. 10- FOLD ACCURACY OF SVM

k	Accuracy
5	0.8922857
7	0.8949884
9	0.8937899

TABLE VI. NITER FOR ADABOOST MODEL

nIter	Accuracy
50	0.9507984
100	0.9517002
150	0.9504990

TABLE VII. MTRY FOR RANDOM FOREST MODEL

Mtry	Accuracy
2	0.9409020
7	0.9502023
13	0.9474996

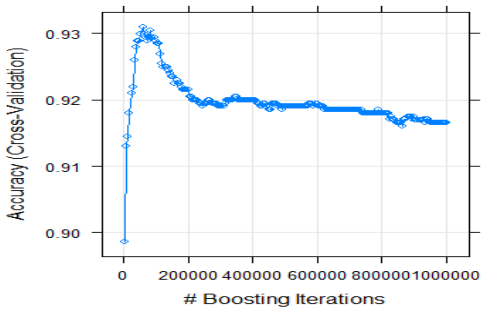


Fig. 3. Boosting iterations of SGB model.

f) Stochastic gradient boost

The model was tuned to calculate the number of trees that achieves the best accuracy. The parameter was initially 5000 to 1000000. Results show that after 60000 ntrees, there's no significant change in accuracy as shown in Fig. 3.

g) MLP ANN

Multi-layer perceptron neural network was built using: 13 inputs, 2 outputs and one hidden layer with 5 neurons. The initial weight matrix was randomly generated. The learning function is "Std_Backpropagation" and the learning rate = 0.1. The resulted weight matrix after epochs' network training is shown in Table VIII.

IV. Results and Discussion

Accuracy is used to evaluate the model performance. Accuracy indicates the ability to differentiate the credible and non-credible cases correctly. It's the proportion of true positive (TP) and true negative (TN) in all evaluated news:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where,

TP: is the total number of customers correctly identified as churn.

FP: is the total number of customers incorrectly identified as churn.

TN: is the total number of customers correctly identified as no-churn.

FN: is the total number of customers incorrectly identified as no-churn.

Results of applying the cross validation in all models are shown in Table IX and Fig. 4.

TABLE VIII. Weight Matrix of MLP

	Hidden2.1	**Hidden2.2**	**Hidden2.3**	**Hidden_2_4**	**Hidde 2_5**	**churn**	**No_churn**
Input_1	-3.657	2.147	-4.093	1.868	-7.117	0.000	0.000
Input_2	-1.407	0.719	-0.587	0.022	1.720	0.000	0.000
Input_3	-0.569	-0.261	0.400	0.251	1.524	0.000	0.000
Input_4	-2.200	-0.067	0.022	0.131	-3.674	0.000	0.000
Input_5	-1.604	-0.733	1.327	-0.493	-1.328	0.000	0.000
Input_6	-0.504	0.364	1.028	0.240	-0.250	0.000	0.000
Input_7	-1.149	-0.757	-0.723	-0.664	-2.153	0.000	0.000
Input_8	-0.675	-0.416	0.122	-0.116	-0.961	0.000	0.000
Input_9	-0.468	0.121	0.054	-0.175	-0.328	0.000	0.000
Input_10	-0.311	-1.649	0.176	-0.838	-0.053	0.000	0.000
Input_11	-0.056	-0.672	-1.017	3.285	-0.263	0.000	0.000
Input_12	-0.052	-0.412	0.369	0.649	-0.063	0.000	0.000
Input_13	0.857	-1.078	-1.965	-0.134	0.287	0.000	0.000
Hidden2_1	0.000	0.000	0.000	0.000	0.000	-0.426	0.411
Hidden2_2	0.000	0.000	0.000	0.000	0.000	-4.107	4.104
Hidden2_3	0.000	0.000	0.000	0.000	0.000	-3.332	3.326
Hidden2_4	0.000	0.000	0.000	0.000	0.000	-3.322	3.320
Hidden2_5	0.000	0.000	0.000	0.000	0.000	-5.024	5.036

TABLE IX. Accuracy of Models during 10 Folds

LR	**CART**	**SVM**	**Naïve**	**KNN**	**Adabost**	**SGB**	**RF**	**LDA**
0.856	0.901	0.940	0.865	0.904	0.961	0.940	0.949	0.841
0.874	0.886	0.922	0.871	0.895	0.961	0.931	0.946	0.862
0.844	0.880	0.904	0.865	0.898	0.943	0.890	0.946	0.832
0.862	0.889	0.919	0.865	0.901	0.946	0.945	0.964	0.868
0.874	0.904	0.937	0.871	0.916	0.964	0.946	0.964	0.868
0.856	0.865	0.922	0.883	0.874	0.949	0.930	0.946	0.859
0.859	0.874	0.919	0.865	0.886	0.934	0.931	0.931	0.850
0.859	0.862	0.940	0.874	0.895	0.955	0.925	0.952	0.850
0.865	0.853	0.934	0.865	0.883	0.955	0.955	0.949	0.856
0.868	0.880	0.925	0.880	0.898	0.958	0.916	0.955	0.853

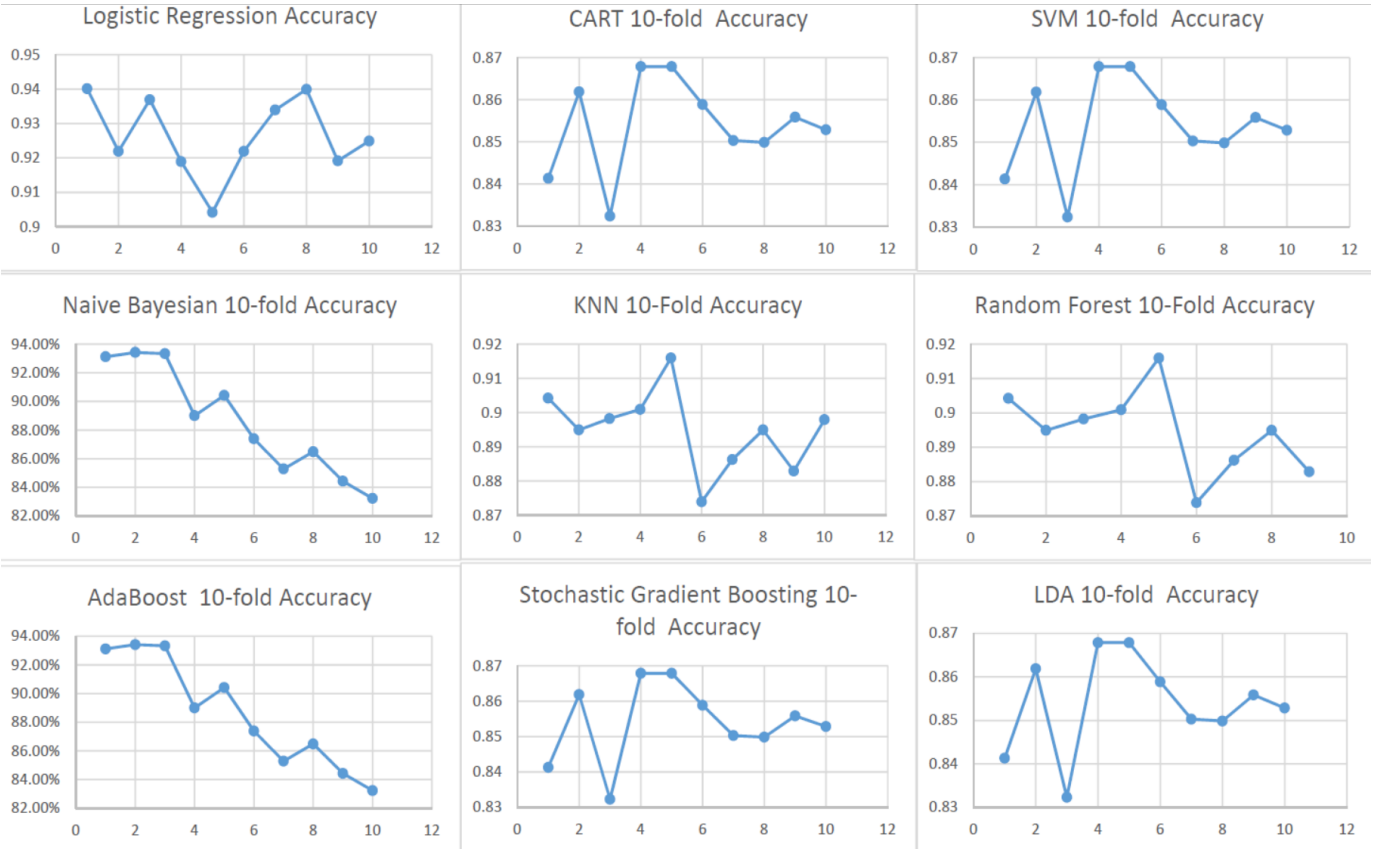


Fig. 4. 10-Fold accuracy of the selected models.

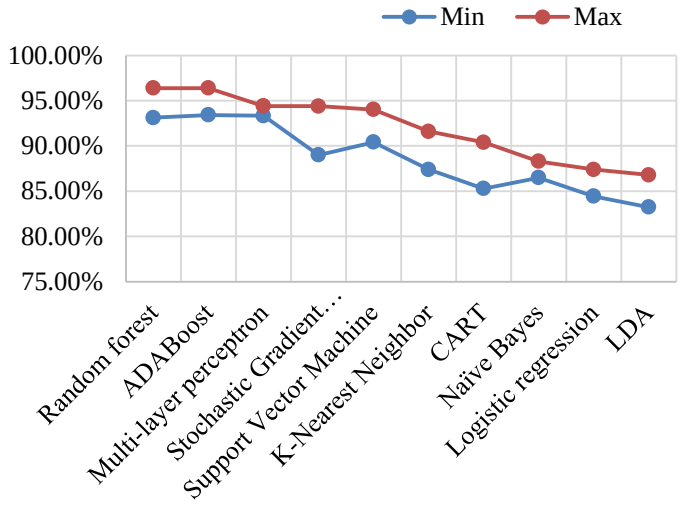


Fig. 5. Accuracy of the selected models.

Minimum and maximum accuracies for all of the selected models are summarized in Table X and Fig. 5. Results of the study show that ensemble based learning techniques (RF and AdaBoost) achieved the highest performance with approximately 96%. Both MLP and SVM can be recommended as well with 94% accuracy. DT achieved 90%, NB 88% and finally LR and LDA with accuracy 867% as shown in Fig. 5.

TABLE X. ACCURACY OF THE SELECTED MODELS

Model	Min	Max
Random forest	0.931138	0.963964
ADABoost	0.934132	0.963964
Multi-layer perceptron	0.93329	0.944
Stochastic Gradient Boosting	0. 8900000	0.9439552
Support Vector Machine	0.904192	0.94012
K-Nearest Neighbor	0.873874	0.915916
CART	0.852853	0.903904
Naïve Bayes	0.864865	0.882883
Logistic regression	0.844311	0.873874
LDA	0.832335	0.867868

V. CONCLUSION AND FUTURE WORK

This study tries to present a benchmark for the most widely used state of the arts for churn classification. The accuracy of the selected models was evaluated on a public dataset of customers in Telecom Company. Based on the findings of this study, ensemble – based learning techniques are recommended as both Random forest and Ad boost models gave the best accuracy. However, the study can be extended by including hybrid models and deep learning models. Other performance metrics can be used for performance evaluation. Timing measures of the models can also be a major indicator for performance. Models can also evaluate against different datasets from different domains.

REFERENCES

[1] Guha, Sudipto, and Nina Mishra. "Clustering data streams:- In Data Stream Management". Springer Berlin Heidelberg, 2016.

[2] Brito, Pedro Quelhas, Carlos Soares, Sérgio Almeida, Ana Monte, and Michel Byvoet. "Customer segmentation in a large database of an online customized fashion business." Robotics and Computer-Integrated Manufacturing , Elsevier ,2015.

[3] Abolfazl Kazemi, Mohammad Esmaeil Babaei, "Modelling Customer Attraction Prediction in Customer Relation Management using Decision Tree: A Data Mining Approach", journal of Optimization in Industrial Engineering , 2011.

[4] The Chartered Institute of Marketing, Cost of customer acquisition versus customer retention (2010).

[5] Colin Shaw, CEO, Beyond Philosophy, 15 Statistics That Should Change The Business World – But Haven't, Featured in: Customer Experience, June 4, 2013.

[6] Ramakrishna Vadakattu ; Bibek Panda ; Swarnim Narayan ; Harshal Godhia " Enterprise subscription churn prediction" ,IEEE International Conference on Big Data (Big Data), 2015.

[7] Miguel A.P.M. Lejeune "Measuring the impact of data mining on churn management", Internet Research , Vol. 11 Issue: 5, pp.375-387,

[8] [8] [8] Jain S., Sharma N.K., Gupta S., Doohan N. (2018) Business .Strategy Prediction System for Market Basket Analysis. In: Kapur P., Kumar U., Verma A. (eds) Quality, IT and Business Operations. Springer Proceedings in Business and Economics. Springer, Singapore. 2017. DOI https://doi.org/10.1007/978-981-10-5577-5_8

[9] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," Procedia Computer Science, vol. 85, pp. 78–85, 2016

[10] Berger, P.D. and Nasr, N.I. (1998), "Customer lifetime value: marketing models and applications", Journal of Interactive Marketing, Vol. 12 No. 1, pp. 17-29.

[11] Cheng, C.-H. and Chen, Y.-S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications, Vol. 36 No. 3, pp. 4176-4184.

[12] Huang, S.C., Chang, E.C. and Wu, H.H. (2009), "A case study of applying data mining techniques in an outfitter's customer value analysis", Expert System Application, Vol. 36 No. 3, pp. 5909-5915

[13] Nishant Saxena, ESCORT (Enterprise Services Cross-sell Optimization Using Rigorous Tests of Association), Advances in Economics and Business 5(5): 239-245, 2017

[14] Anita Prinzie , DirkVan den Poel, "Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models". European Journal of Operational Research, 170, 710–734. 2006.

[15] Rahul J. Jadhav, Usharani T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2 , February 2011

[16] Amin A., Khan C., Ali I., Anwar S. "Customer Churn Prediction in Telecommunication Industry: With and without Counter-Example". In: Gelbukh A., Espinoza F.C., Galicia-Haro S.N. (eds) Nature-Inspired Computation and Machine-learning. MICAI 2014. Lecture Notes in Computer Science, vol 8857. Springer, Cham,2014.

[17] A. Keramatia, R.Jafari-Marandi, M.Aliannejadi, I.Ahmadian, M.Mozaffari, U.Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques",Applied Soft Computing Volume 24,Pages 994-1012. 2014.

[18] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Khalid Alawfi, Amir Hussain ,Kaizhu Huang, "Customer churn prediction in the telecommunication sector using a rough set approach", Neurocomputing Volume 237, Pages 242-254,2017.

[19] Ben lanHe , Yong Shi, Qian Wan, XiZhao, "Prediction of Customer Attrition of Commercial Banks based on SVM Model",2nd International Conference on Information Technology and Quantitative Management, ITQM, Procedia Computer Science Volume 31, Pages 423-430, 2014.

[20] Keramati, A., Ghaneei, H. & Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining", S.M. Financ Innov (2016) 2: 10. https://doi.org/10.1186/s40854-016-0029-6

[21] Alisa Bilal Zorić, "PREDICTING CUSTOMER CHURN IN BANKING INDUSTRY USING NEURAL NETWORKS", nterdisciplinary Description of Complex Systems 14(2), page:116-124, 2016

[22] M. Clemente, V. Giner-Bosch, and S. San Matías, "Assessing classification methods for churn prediction by composite indicators", Dept. of Applied Statistics, OR & Quality,Universitat Politècnica de València, Camino de Vera s/n, 46022 Spain, 2010

[23] Anthony E. R. Sukow, Rebecca Grant, "Forecasting and the Role of Churn in Software-as-a-Service Business Models", iBusiness, Vol. 5 No. 1A, 2013, pp. 49-57.

[24] Yizhe Ge ; Shan He ; Jingyue Xiong ; Donald E. Brown, "Customer churn analysis for a software-as-a-service company", In the proceedings of Systems and Information Engineering Design Symposium (SIEDS), 2017

[25] Rahul J. Jadhav and Usharani T. Pawar. "Churn prediction in telecommunication using data mining technology" , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, 2011.

[26] Andrew h. Karp, using logistic regression to predict customer retention, 1998.

[27] M.C. Mozer ; R. Wolniewicz ; D.B. Grimes ; E. Johnson ; H. Kaushansky. "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunication Industry". IEEE Transactions on Neural Networks, Volume: 11, Issue: 3, 2000.

[28] Afaq Alam Khan, Sanjay Jamwal, and M.M.Sepehri, Applying data mining to customer churn prediction in an Internet Service Provider". International Journal of Computer Applications Volume 9–No.7,2010

[29] Guoxun Wang , Liang Liu, Yi Peng , Guangli Nie, Gang Kou, G., and Yong Shi. "Predicting credit card holder churn in banks of China using data mining and MCDM". IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2010.

[30] Abbas Keramati, Seyed M.S. Ardabili, "Churn analysis for an Iranian mobile operator". Telecommunications Policy Volume 35, Issue 4, Pages 344-356,2011.

[31] B.E.A. Oghojafor,G.C. Mesike, C.I. Omoera and R.D. Bakare, "Modelling telecom customer attrition using logistic regression", African Journal of Marketing Management Vol. 4(3), pp. 110-117, 2012.

[32] Sebastian H. T, Wagh R. Churn Analysis in Telecommunication Using Logistic Regression. Orient.J. Comp. Sci. and Technol; Vol. 10, No. (1): Pgs. 207-212, 2017.

[33] Guangli Nieae, Wei Rowec Ling ling, Zhangab Yingjie, Tiana YongShi, "Credit card churn forecasting by logistic regression and decision tree", Expert Systems with Applications Volume 38, Issue 12, November–December 2011

[34] Luo Bin ; Shao Peiji ; Liu Juan, "Customer churn prediction based on the decision tree in personal handyphone system service". International Conference on Service Systems and Service Management, 2007.

[35] Coussement, K., Benoit, D. F., & Van den Poel, D. Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Systems with Applications, volume:37. No.3, pp:2132-2143. 2010.

[36] Michel Ballings, Dirk Van den Poel, "Customer event history for churn prediction : how long is long enough?", Journal Expert Systems with Applications: An International Journal archive Volume 39 Issue 18, December, 2012 Pages 13517-13522

[37] Kristof Coussement, Dirk Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter selection techniques". Expert Systems with Applications, Volume 34, Issue 1, Pages 313-327, 2008.

[38] Hur Y., Lim S. Customer Churning Prediction Using Support Vector Machines in Online Auto Insurance Service. In: Wang J., Liao XF., Yi Z. (eds) Advances in Neural Networks . Lecture Notes in Computer Science, vol 3497. 2005.

[39] Jing Zhao. and Xing-hua Dang. Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer

churn as the example" , the 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008.

[40] Stefan Lessmann, Stefan VoB, "A reference model for customer-centric data mining with support vector machines". European Journal of Operational Research, volume 199, issue 2, pages:520-530,2009.

[41] Ali Dehghan, Theodore B. Trafalis, "Examining Churn and Loyalty Using Support Vector Machine", Business and Management Research,Vol. 1, No. 4; 2012.

[42] Benlan He, Yong Shi, Qian Wan, Xi Zhao, "Prediction of customer attrition of commercial banks based on SVM model", the 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.

[43] Lee, H., Lee, Y., Cho, H., Im, K., and Kim, Y. S. (2011). Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model. Decision Support Systems, 52(1):207{216.

[44] Huang, B., Kechadi, M. T., and Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1):1414-1425.

[45] Clement Kirui, Li Hong, Wilson Cheruiyot, and Hillary Kirui. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining". JCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.

[46] Catalin CIMPOERU, Anca ANDREESCU , "Predicting Customers Churn in a Relational Database", Informatica Economica vol. 18, no. 3/2014

[47] Bart Larivière, Dirk Van den Poel. Predicting customer retention and proftability by using random forests and regression forests techniques. Expert Systems with Applications, volume 29 issue 2, Pages:472-484, 2005.

[48] Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying. "Customer churn prediction using improved balanced random forests". Expert Systems with Applications, volume 36 issue 1, pages:5445–5449, 2009.

[49] Adnan Idris, Muhammad Rizwan, Asifullah Khan, "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies" Computers & Electrical Engineering, volume 38 issue 6, 2012.

[50] Sadaf Nabavi, Shahram Jafar, Providing a Customer Churn Prediction Model Using Random Forest and Boosted Trees Techniques (Case Study: Solico Food Industries Group, Journal of Basic and Applied Scientific Research, volume 3, issue 6, pages:1018-1026, 2013

[51] Ionut B. Brandusoiu Gavril Toderean, Predicting Churn In Mobile Telecommunications Industry", Acta Technica Napocensis Electronics and Telecommunications, Volume 54, Number 3, 2013.

[52] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain] Mechanisms. Spartan Books, Washington DC, 1961

[53] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.

[54] Hwang H., Jung T., Suh E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26 (2004) 181-188.

[55] Ferreira J., Vellasco M., Pachecco M., Barbosa C.: Data mining techniques on the evaluation of wireless churn. ESANN2004 proceedings – European Symposium on Artificial Neural Networks Bruges (2004). ISBN 2-930307-04-8, p 483-488.

[56] Schapire R.E. Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Springer, Berlin, Heidelberg, 2013.

[57] Jerome Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, 29(5), 2001, 1189-1232.

[58] Shao Jinbol, Li Xiu, Liu Wenhuang, "The application of AdaBoost in customer churn prediction. In the proceedings of the international Conference on Service Systems and Service Management 2007.

[59] Adnan Idris, Asifullah Khan, yeon soo Lee, "Genetic programming and adaboosting based churn prediction for telecom. In the proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1328–1332, 2012.

[60] Wu ; Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost", In the proceedings of the 13th international conference on Service Systems and Service Management (ICSSSM), 2016.

[61] Lemmens, Aurelie and Gupta, Sunil, "Managing Churn to Maximize Profits" (May 8, 2017). http://dx.doi.org/10.2139/ssrn.2964906

[62] Gerald Fahner , Stochastic Gradient Boosting Approach to Daily Attrition Scoring Based on High-dimensional RFM Features"2015 Fair Isaac Corporation. Confidential.

[63] Yaya Xie ; Xiu Li, "Churn prediction with Linear Discriminant Boosting algorithm ", International Conference on Machine-learning and Cybernetics, , 2008

[64] A.Keramatia, R.Jafari-Marandia, M.Aliannejadib, I.Ahmadianc, M.Mozaffaria, U.Abbasiad, "Improved churn prediction in telecommunication industry using data mining techniques", Applied Soft Computing Volume 24, Pages 994-1012,2014 .

[65] Naveen Kumar Rai,Vikas Srivastava, Rahul Kumar, "Churn Prediction Model Using Linear Discriminant Analysis (LDA)", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 18, Issue 5,PP 86-93, 2016.

[66] Tom Au, Shaomin Li, Guangqin Ma. Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques". Journal of Comparative International Management, Vol. 6, No. 1, 10-22, 2003.

[67] T. Vafeiadisa, K. I. Diamantarasb, G. Sarigiannidisa, K. Ch. Chatzisavvasa, A comparison of machine-learning techniques for customer churn prediction" Simulation Modelling Practice and Theory, 2015.

[68] Samira Khodabandehlou, Mahmoud Zivari Rahman, "Comparison of supervised machine-learning techniques for customer churn prediction based on analysis of customer behavior", Journal of Systems and Information Technology, Vol. 19 Issue: 1/2, pp.65-93, https://doi.org/10.1108/, 2017.

[69] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study". American Journal of Epidemiology , 179(6):764–774, 2014.

[70] Munevver Mine Subasia, Ersoy Subasib, Martin Anthony, Peter L.Hammer, "A new imputation method for incomplete binary data", Discrete Applied Mathematics Volume 159, Issue 10, 2011.

[71] Breiman,L. "Random Forests", Machine-learning Volume 45, Issue 1, pp 5–32, 2001.

[72] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki, "Boruta – A System for Feature Selection", Fundamental Informaticae volume101, pages:271–285, 2010.

[73] M. Pardo, G. Sberveglieri, "Classification of electronic nose data with support vector machines", Sens. Actuators B: Chem. 107 (2005) 730–737.

Crowd Counting Mapping to Make a Decision

Enas Faisal, Azzam Sleit
Computer Science Department
The University of Jordan
Amman-Jordan

Rizik Alsayyed
Business Information Technology
The University of Jordan
Amman-Jordan

***Abstract*—Congestion typically occurs when the number of crowds exceeds the capacity of facilities. In some cases, when buildings have to be evacuated, people might be trapped in congestion and cannot escape from the building early enough which might even lead to stampedes. Crowd Congestion Mapping (CCM) is a system that enables organizations to find information about the crowd congestion in target places. This project provides the ability to make the right decision to determine the reasons that led to that and to do the appropriate procedures to avoid this from happening again by optimizing locations and dimensions of the emergency exits less congested path on the target places. The system collects crowd congestion data from the locations and makes it available to corporations via target map. The congestion is plotted on target place map, for example, the red line for highly congested location, the pink line for mildly congested location and green line for free flow of humans in the location.**

***Keywords*—Crowd; map; image processing; human detection; threshold; recognition**

I. Introduction

There is great interest in the surveillance system, especially in cases of high congestion as the ability to identify objects and follow them or to detect the times that become congestion. These things have become important but difficult to measure, so there is concentrate on computer vision algorithms to try to solve these problems.

Work on the crowd has gained much attention in recent years for a variety of applications such as video surveillance, public safety design and traffic control. Many researchers have handled more than one aspect of the analysis of crowded scenes such as counting [9], [10], detecting anomalies [12], Segmentation [11], and many other aspects [13]-[15], [19], [20].

To help institutions and centers make appropriate decisions in the event of congestion in advance, a system for that is needed. In this paper, a system was proposed for the identification of cases of crowded gatherings that may occur at certain times, which may lead to security problems, interruptions or other problems.this system is based on image feature extraction and Feed-Forward Neural Networks (FFNN). This system plotted gathered information on the map to explore the hazardous locations to take a decision in this situation.

The rest of the paper is organized as follows. Section II discusses the related work, The proposed Algorithm is explained in Section III, Section IV discusses and analyzes the results. Finally, the conclusions are drawn in Section V.

II. Related Work

In the field of computer vision, many motions detection algorithms are introduced in literature to detect abnormal movement in dynamic congestion, but building an automatic detection system is still a challenge.

Brostow and Cipolla in [4] present a system to detect specific people in congestion, but this system has encountered problems when have a noise or objects exist in congestion such as kiosk or antiques.

Also, Pathan et al. [5] worked out a system in which they identified all erroneous movements during the congestion stages. But this system does not produce good results when there is high congestion where it is not possible to recognize the number of people because they use subtractions process in the detection people.

Moreover, Krausz and Bauckhage in [2] proposed a system works automatically for identifying the critical situation during an increased number of the public by using system alarms, but this system showed error in the detection even at normal state. Dee and Caplier in [6] suggest a system depend on representing the movement pattern of the crowd.

Shang et al. [10] proposed an end-to-end model using CNN to handle the process of counting in crowd images by simultaneously learning local and global count on the full sized input images. Onoro-Rubio and Lopez-Sastre in [16] proposed Hydra CNN to addressed the scale issue by proposing a scale aware counting model by uses a pyramid of image patches extracted at multiple scales to perform the final density prediction. Zhang et al. [17] proposed Multi-column Convolutional Neural Network (MCNN) architecture to extract features generated by filters of different scales to generate the final prediction for its crowd density map.

Boominathan et al. in [18] proposed a novel deep learning framework by a combination of deep and shallow to tackle the issue of scale variation for estimating crowd density to predict the density map for a given crowd image.

Rohit et al. in [21] they proposed epsilon Support Value Regression (SVR) fusion-based approach to help detect and sort out people in images of the highly blocked crowd by influence information on the global construction of the crowd scenes and identifying people in these scenes.

III. Proposed Method

First of all, we acquired the video feed from surveillance cameras of four locations in City Mall in Amman city. In order

to detect from the video, we applied various images processing procedures. The extracted image frames were divided into several blocks. Then analyze each block by finding the circular or the semicircular shape of edge detection.

The result was analyzed the collected objects based on bright and dark color. Then the objects were counted and people density can be plotted on map. The architecture of the proposed system is presented via flow diagram as shown in Fig. 1.

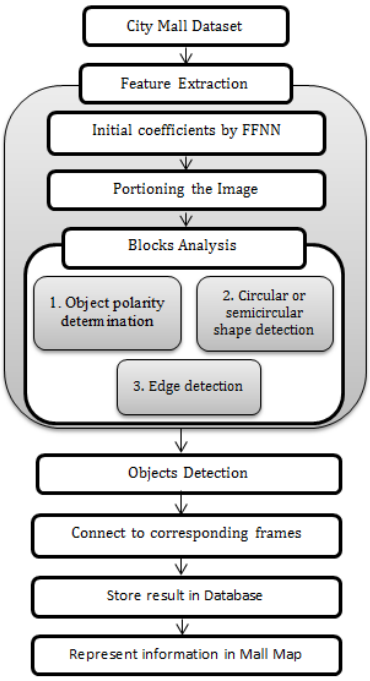


Fig. 1. Proposed algorithm architecture.

A. Feature Extraction

1) Initial coefficients by FFNN: We use Feed Forward Neural Network (FFNN), which contains a set of layers, the first layer is defined the variables of the network, and the last layer is the output layer. Between these two layers are called hidden layers. We use FFNN as a way, instead of randomly determining the initial coefficients (see Fig. 2), where:
Rmin = Minimum radius of the interested object (human head).
Rmax = Maximum radius of the interested object (human head).
EdgTh = Edge detection threshold.
Sens = Sensitivity level for interested object detection.

Therefore, instead of having the initial coefficients randomly and changing their values by trial and error. We use FFNN to accurately measure them, based on our knowledge about the accurate number of persons in the image. So any new test image can be analyzed more accurately.

2) Partitioning the image: A given a training video frame of City Mall for Ground floor i, where i = 1, 2...N and N denotes the total number of training frames, we first partition the frame into K blocks regions (see Fig. 3). The frame is dividing into 8 blocks and the coordinates of each block by (1), (2):

$$BL = FL/8 \tag{1}$$

$$BW = FW/8 \tag{2}$$

Where BL is the block length, BW is the block width, FL is frame length, and FW is a frame width.

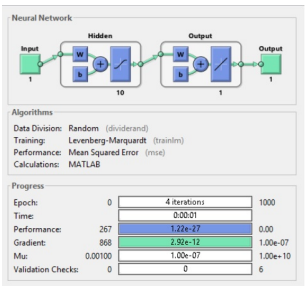


Fig. 2. ANN configuration.

B. Blocks Analysis

For every blocks region we apply the object polarity determination then we find the circular and the semicircular shapes after that we apply edge detection to reduce the false positive.

Fig. 3. Image before and after blocking.

1) Object polarity determination: In this process we detect the color by the polarity of the image horizontally, and determine the real bright and real dark objects, by (3) and (4) below:

$$\frac{\sum_{21}^{c-21}(\sum_{-20}^{20}\sum_{1}^{c}\sum_{1}^{r}(\left|p(r-1_{-p(r)})\right|) > 20)}{41} \tag{3}$$

$$if(Eq(3)) < \frac{(\sum_{1}^{c}\sum_{1}^{r}(|p(r-1)-p(r))|) > 20)}{c} \tag{4}$$

Where c is the column index, r is the row index and p is the current pixel, this equation investigates the pixels for each

41 column and compares them with the next 42 columns. This way we can determine the polarity of the image horizontally, and determine the real bright and real dark objects.

2) Circular or semicircular object detection: After portioning the frames for blocks and applying the object polarity, then we will find the circular and semicircular shape by using Hough Circle transform. Hough transform is used to determine which circuits are present in an image given, whether the circuit is a wheel, head or other[1]. In this paper, we use the circular Hough transform (CHT) [7] and randomized Hough transforms [8]. How the CHT works is based on (5):

$$(x-a)^2 + (y-b)^2 = r^2 \tag{5}$$

Where a and b are the coordinates of the circle center, and r is its radius. In Fig. 4 the solid line represents the circle and the dashed lines represent the hypothetical circles from the points on the edge of the actual circle.

The hypothetical circles in the image space can be presented in a parameter space as a circular cone. If r is increased and equal to r, the hypothetical circles will intersect at point c, and a circle can thereby be detected. The randomized Hough transform is also used to detect ellipses. The ellipse detection method requires five parameters: a center coordinates (x_0, y_0), a rotation angles α, and the half-length of the major and minor axes, a and b, respectively. These parameters are defined by the equations below:

$$x_0 = \frac{x_2 + x_1}{2}, y_0 = \frac{y_2 + y_1}{2} \tag{6}$$

$$a = \frac{\sqrt{(x_2 + x_1)^2 + (y_2 + y_1)^2}}{2}, b^2 = \frac{a^2 d^2 sin^2 T}{a^2 - d^2 cos^2 T} \tag{7}$$

$$\alpha = arctan \frac{y_2 - y_1}{x_2 - x_1} \tag{8}$$

Calculate the parameters in (6) across ellipse geometry across (7). Here the coordination (x_1, y_1) and (x_2, y_2) are endpoints of maximizing axis, and these equations are repeated on all pixels in the image to determine the ellipse.

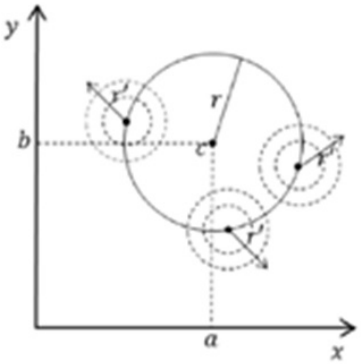


Fig. 4. The Hough transforms [1].

3) Edge detection: After finding circular and semicircular shapes we apply the Canny Edge Detection to reduce false positive. The Canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images [3]. We apply the process of canny edge detection algorithm at the output images from the previous phase, as shown in Fig. 5. With this phase, we finish the feature extraction process and final result appears in Fig. 6.

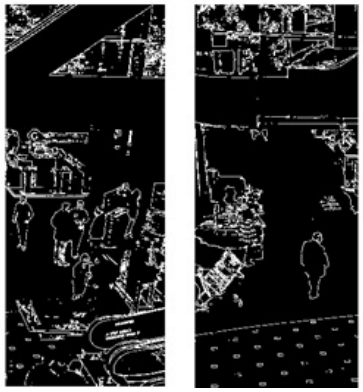


Fig. 5. Some block after edge detection.

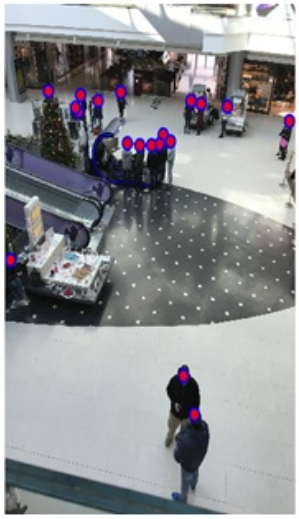


Fig. 6. The final result of crowd detection.

C. Objects Detection

As soon as we determine the location of the objects of interest, the algorithm itself returns the coordinates of all the detected objects and stores them in the vector. These coordinates we use them in "plot" command to put circles over them. Moreover, the number of the elements inside this vector equals the number of interesting objects.

D. Connect to Corresponding Frames

After counting the number of objects in each frame, a link is made between each frame with the appropriate location in the mall. This algorithm applies to the four locations on the ground floor of the mall. After that the number of objects in each frame is collected for each location to determine the

place where it is most crowded to apply a set of procedures to mitigate them in the future.

E. Thresholding

The process of determining the location that has the highest percentage of crowded depends on the threshold, which is based on the capacity of the area, as defined in (9):

$$Areacapacity = Imagearea/personoccupationarea \quad (9)$$

At the threshold algorithm as shown in Algorithm 1, we have assumed arbitrary values to calculate the threshold, since we have excluded half of the location area as an arbitrary value, assuming that humans are not present there, such as kiosks, decoration, stairs, etc. Thus, we assume that considered the threshold for the other half as the place of the capability of human existence.

As shown in Algorithm 1: line 1 calculate the area of the location as C and in line 2 divide the area by two to calculate the Ca as the area where humans are able to exist. In line 3-4, threshold Th computes as the human area Ca divided by average human area as ε.

In line 5-10: threshold Th used to decide the crowded locations by comparing the value of Th with φ, if φ value more or equal to Th value then this location has highly crowded and this location needs procedures to mitigate the hazardous companies with congestion, but if φ value between the half of Th value and Th-1 then this location has medium crowded, the location considered not crowded if φ value less than Th-1.

φ= the AVG detected # of objects $\forall$ frames in one Location.
Algorithm 1: Threshold Algorithm
C = Full image area
Compute Ca */where Ca = C / 2
ε = Avg human area
Th = Ca / ε
If φ > = Th Then
Highly Crowd
Elseif φ >= Th /2 and φ ¡= Th-1 Then
Medium Crowd
Elseif φ < = Th-1 Then
Not Crowd

F. Plotting Information on Map

The information of object density of particular location in City Mall was plotted on City Mall map, where the Green color indicated to free flow of people, red line indicated high crowd and pink line indicated mild crowd as shown in Fig. 7.

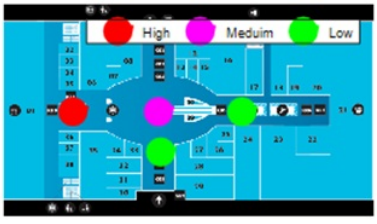


Fig. 7. Map after colored.

IV. Results and Discussion

The experiments conducted over MATLAB environment. The images it has been taken from various cameras of four locations placed at different angles on the ground floor of City Mall (Amman city).

The evaluation of the effectiveness of the proposed method for the four locations has been tested as part of the study. As the given data in Table I, the number of humans detected by the system was compared with the actual number of humans in image frames. The result was 96% accurate for the camera at the location 4 and 99%-98% accurate for the camera on the other three locations.

TABLE I. Detection Rates for Each Location

City Mall Locations	Average Performance
Location 1	98%
Location 2	98%
Location 3	99%
Location 4	96%

The dataset for each location is visualizing the real number of crowds in each image frame against the measured number of proposed algorithm as shown in Fig. 8.

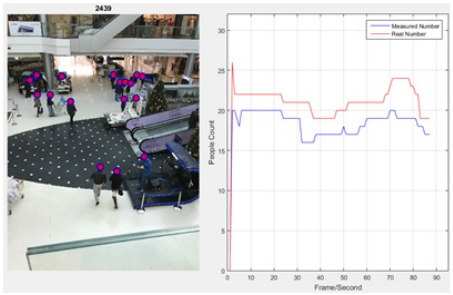


Fig. 8. Rate of Measured number vs. Real number.

The cent percent accuracy was not obtained because of various reasons. When the brightness is low this approach has a problem to differentiate between the human head and objects. And when their dark background with a person wraps his back to the camera (his face reversing the camera) made a difficulty to distinguish between the color of the hair and the background during the process of polarity. We expect that illumination might help in solving this problem at future work.

The other problem that we encountered is the people who stand behind the object (such as kiosks, etc.) wherein the visible eye there is difficulty in distinguishing the presence of that person or not.

The location of the camera affects the size of the head radius for people are close or far from the camera, which creates a problem in the training process of Neural network, because we give a range of radius between the largest and smallest value, and therefore may affect the detection of humans with object have radius matched with given range (such as the wheel of the kiosk).

We recommend using strong lighting and do not use a dark background to be bright as much as they can. Make coverage radius of the camera for human heads within a small range, therefore no big gap within the same range. The process of

calculating the crowd ratio depends on the environment since the environment is determined the level of the crowd as it varies between games at the stadium or in a mall. In this work we assumed a congestion ratio based on the mall environment.

V. Conclusion

Overcrowd occur in institutions and malls in excess of their ability to withstand these numbers, especially in times of holidays and other occasions. In these times the crowd is getting so critical, leading to problems in the stampede among the crowd. In this paper, crowd counting mapping approach was proposed, which based on image feature extraction and FFNN. In this method, high congestion was detected early to make the appropriate decision to deal with this situation so that no problems occur in the future. After testing, high detection rates were achieved, their accuracy could be as high as 98%. The proposed algorithm uses a powerful set of features which proved to be effective in the crowd detection and we believe that our approach would be integrated with any organization surveillance system. In the future we want to improve neural network training using one of the meta-heuristic algorithms to give more accuracy in detecting people.

References

[1] Young-Jin Cha , Kisung You , Wooram Choi ,Vision-based detection of loosened bolts using the Hough transform and support vector machines, Elsevier, 2016.

[2] B. Krausz and C. Bauckhage, Automatic detection of dangerous motion behavior in human crowds, in Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on, pp. 224229,2011.

[3] Paul Bao, Lei Zhang, and Xiaolin Wu, Canny Edge Detection Enhancement by Scale Multiplication, IEEE transactions on pattern analysis and machine intelligence, VOL. 27, NO. 9, SEPTEMBER 2005.

[4] G. Brostow and R. Cipolla, Unsupervised bayesian detection of independent motion in crowds, in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1, pp. 594601, 2006.

[5] S. Pathan, A. Al-Hamadi, and B. Michaelis, Crowd behavior detection by statistical modeling of motion patterns, in Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of, pp. 8186, 2010.

[6] H. Dee and A. Caplier, Crowd behaviour analysis using histograms of motion direction, in Image Processing (ICIP), 2010 17th IEEE International Conference on, pp. 15451548, 2010.

[7] H.K. Yuen, et al., Comparative study of Hough transform methods for circle finding, Image Vis. Comput. 8 (1) (1990) 7177.

[8] C. Basa, M. Talo, R. Brad, Randomized Hough transform for ellipse detection with result clustering. Computer as a Tool, 2005, EUROCON 2005. The International Conference on. Vol. 2, IEEE, 2005.

[9] A. B. Chan and N. Vasconcelos. Counting people with lowlevel features and bayesian regression. IEEE Transactions on Image Processing, 21(4):21602177, 2012.

[10] C. Shang, H. Ai, and B. Bai. End-to-end crowd counting via joint learning local and global count. In IEEE ICIP, pages 12151219. IEEE, 2016.

[11] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. arXiv preprint arXiv:1411.4464, 2014.

[12] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino. Novel dataset for fine-grained abnormal behavior understanding in crowd. In IEEE International Conference on AVSS, pages 95101. IEEE, 2016.

[13] J. Shao, C. Change Loy, and X. Wang. Scene-independent group profiling in crowd. In IEEE CVPR, pages 22192226, 2014.

[14] J. Shao, K. Kang, C. C. Loy, and X. Wang. Deeply learned attributes for crowded scene understanding. In Proceedings of the IEEE CVPR, pages 46574666. IEEE, 2015.

[15] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In IEEE ICCV, pages 24232430. IEEE, 2011.

[16] D. Onoro-Rubio and R. J. Lopez-Sastre. Towards perspective-free object counting with deep learning. In ECCV, pages 615629. Springer, 2016.

[17] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Singleimage crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE CVPR, pages 589 597, 2016.

[18] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In Proceedings of the 2016 ACM on Multimedia Conference, pages 640644. ACM, 2016.

[19] V. Sindagi and V. Patel. CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting. (AVSS) , IEEE, 2017.

[20] V. Sindagia, V. Patelb. A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation. Elsevier, 2017.

[21] Rohit, V. Chauhan, S. Kumar, S. K. Singh .Human Count Estimation in High Density Crowd Images and Videos. Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2016.

Quality of Service Impact on Deficit Round Robin and Stochastic Fair Queuing Mechanism in Wired-cum-Wireless Network

Fahim Khan Khalil, Samiullah Khan
Farooq Faisal, Mahmood Nawaz, Farkhanda Javed
Institute of Business Management Sciences
The University of Agriculture Peshawar
Peshawar-Pakistan

Matiullah, Zia ullah, Muhammad Shoaib, Faqir Usman Masood
Department of Basic Sciences and Islamiat
Department of Computer System Engineering
University of Engineering and Technology
Computer science Department, Qurtuba University
Peshawar-Pakistan

Fawad Ali Khan, Rafidah MD Noor
Department of Computer System and Technology
University of Malaya, Kuala Lumpur
Malaysia

***Abstract*—The deficient round robin (DRR) and stochastic fair queue (SFQ) are the active queue mechanism (AQM) techniques. These AQM techniques play important role in buffer management in order to control the congestion in the wired-cum-wireless network by dropping packets during the buffer overflow or near to overflow. This research study focus on the performance evaluation of the DRR and SFQ using different scenarios such as increasing number of node scenario, pause time scenario and mobility scenario. We evaluate the performance of DRR and SFQ based on two parameters such as average packet delay and average packet dropped. In case of increasing number of nodes, the SFQ has outperformed than DRR by having comparatively low per packet delay. DRR has higher packet dropped ratio as compare to SFQ. In mobility and pause time scenario, SFQ has less per packet delay while DRR has less packet dropped ratio These results revealed that DRR performance was affected by an increase in the number of nodes in a network. The DRR send the packet in a round-robin fashion without caring about the bandwidth of a path due to which the packet dropped ratio was high. On another hand, the SFQ has comparatively outperformed in all scenarios by having less per packet delay. SFQ become aggressive by dropping more data packets during buffer overflow. In short, SFQ will be preferred for a network where the congestion occurred more frequently.**

***Keywords*—Active queue management; deficit round robin; stochastic fair queuing**

I. Introduction

To provide a wide range of connectivity to its mobile nodes wireless networks are connected with infrastructure networks. Such a scenario is known as wired-cum-wireless networks. Access Point is used as fixed base stations between wired and wireless networks [1]. In mobile ad hoc network (Manet), a number of efficient routing protocols are used, i.e. ad hoc-on demand distance vector (AODV), dynamic source routing (DSR), DRR and SFQ whose performance degraded in wired-cum-wireless scenario. In wired-cum-wireless networks, mobile host performances are affected during handoff time causes packet loss and degrade throughput [2].

Active Queue Management main function is to remove congestion, predictable queuing delay, and high link utilization, but an AQM scheme should promote high network stability, robustness, responsiveness, and scalability. To define robustness, it is important that the AQM algorithm executes constantly well under intense and unfavorable network conditions (like when changes in network parameters occur it does not perform any effect) [25]. It showed more improvement when AQM parameters tuned when there is a change in traffic load. Responsiveness defines as the speed of convergence to an equilibrium. Stability means that AQM algorithm performs static whenever there is a change in network condition. Scalability is important in AQM as it performs its functions firmly and steadily when speed and number of routers increases and number of links also increase [23].

This research focus on performance evolution of DRR and SFQ routing protocols in wired-cum-wireless scenario. DRR used different sizes of packets without caring their mean size. The packet left from one round due to its large size will be prioritized for transmission in the second round. SFQ used hashing technique in packet selection for transmission. This technique was able to map packets to the corresponding queue. SFQ reserved one queue for each flow with condition that queues number should be less than flows count [26].

The rest of the paper is organized as follows: Section II summarizes some existing AQM techniques. Section III introduced DRR and SFQ mechanism. Simulation configuration, performance analysis parameters and results analysis are mentioned in Section IV. Finally, the conclusion with future recommendation are mentioned in Section V.

II. Related Work

In wired-cum-wireless network, the intermediate nodes receive and forward the packets to the destination. These nodes store extra packets in the internal memory of intermediate nodes called buffer. When incoming traffic rate is larger than outgoing traffic the buffer space becomes full which results in buffer overflow by dropping data packets. This degrades the quality of service (QoS) in wired-cum-wireless scenario by increasing the delay and packet drop rate. The research problem is to find out the effect of buffer overflow on AQM techniques in wired-cum-wireless network in order to cope with abnormal delay and packet drop rate [27].

AQM techniques are used to overcome the congestion in the network by managing the packets in the buffer. In order to reduce the congestion in the network, the AQM uses intelligent packet drop system when the buffer is full or near to full. The decisions of packet drop are taken by various algorithms such as DRR and SFQ. The first AQM scheme i.e. Random Early Detection (RED) was introduced in 1993.These schemes were known as active due to its vigorously signal congestion to sources, explicitly by making packets or implicitly by dropping packets, while Drop-Tail queue is known as passive due to its dropping nature of packets when the queue is full. In 1998, the Internet Engineering Task Force (IETF) commended the deployment of AQM in internet routers. Their main task was to improve the performance and the prevention of congestion collapse which may arise from the growth of non-responsive traffic on the internet [3].

Shreedhar and Varghese (1996) found that in fair queuing process each flow was passing through a device which shared network resources [24]. Adopting such kind of network is not possible due to high expenses. DRR was a new approach in fair queuing. This scheme achieved nearly perfect fairness in terms of throughput, requires only *O(l)* work to process a packet and was simple enough to implement in hardware. DRR could also be implemented in those scheduling problems where servicing cannot be broken up into smaller units and to distributed queues.

Kortebi and Roberts (2005) studied the performance of DRR and priority deficit round robin PDRR [15]. PDRR is an extension to DRR and used in highly dynamic networks. PDRR is more scalable than DRR. The number of flows to be looped is less than hundred at any kind of link speed. Its latency period is also very small for streaming packets. PDRR required few additional instructions to be implemented as compare to DRR [28].

Rind et al. (2006) studied that IEEE 802.11 network had various encounters like connectivity and performance problems [2]. Different routing protocols were used i.e. Destination-Sequenced Distance Vector (DSDV). TCP and UDP performances were checked in wired cum wireless LAN using DSDV protocol. TCP give accurate results when numbers of moving nodes are less. File transfer rate, buffering of video and audio produced encouraging results while it is down in case of voice over IP. While UDP show better results in the case of voice over IP [19].

McKenney (2009) introduce fairness queuing method to separate dense network users from overloading prompted users [20]. According to fairness queuing method each conversation is mapped to its particular queue. There are other methods which implemented other mapping techniques but they are slow and lot of memory required. To mitigate these issues SFQ introduced. In this algorithm no exact mapping is required and also suitable for firmware implementation [29].

Cooper and Meghanathan (2010) had tried to investigate the effects of different mobility models on non-disjoint and link-disjoint multipath routing algorithms for MANET [9]. Gauss-Markov model produced least number of multi-paths and maximize lifetime per multi path. Random Direction mobility model proved the smallest lifetime per multi-path routing.

Liu et al. (2010) studied that a fair scheduling mechanism had an excellent ability and having low complexity [17]. In a communication network, it was observed that self-similar traffic was persistently presented. An analytical model was found to best for judgment of packet size effect on performance. The developed model was considered to best for analysis individual traffic flows.

Lin and Hamdi (2010) studied Fair-queuing algorithms which rely greatly on multiple queuing structures or suffered from the high time complexity which is difficult to implement in large scale due to the access delay of DRAM [16]. FQ algorithm has suffered from at least one of the difficulties in 802.16 networks, i.e. high time complexity, flow aggregation and lack of scalability. To face these challenges, they proposed a two-stage FQ algorithm, namely BRR. Their scheme worked in two steps. Furthermore they discussed the process of enqueuing and dequeuing separately [30].

Maan and Mazhar (2011) try to prove the difference between MANET and other wireless and wired networks [18]. The main difference which distinguishes between them was mobility. Here, the author compare performance of three mobility models i.e. RWP, reference point group mobility and column mobility model in MANET.Noon et al., (2011) considered the round robin (RR) algorithm to be more widely used adopted algorithm and discussed its flaws [21]. Choosing the optimal time quantum is a bottleneck in a RR algorithm. The processing time of CPU is too high for the time quantum. So selecting the proper quantum time is a major issue to solve the processing time. To overcome this problem a new approach called AN algorithm is designed which was based on dynamic time quantum instead of a fixed time quantum. Instead of the user, the operating system itself chooses the time quantum for itself. It solves the time quantum problem and improves operating system performance and increases the run time of RR.

Jonit and Baba (2011) had mentioned more analysis on schedule in order to have better output on scheduler performance [13]. Two different types of scheduler i.e.FIFO and DDR were compared to check their performance in 802.16. Both schedulers give same throughput for all parameters except for variable packet size. FIFO scheduler has less delay as compare to DDR. Both schedulers can be chosen based on the need of the Internet provider.

Patel et al. (2012), investigated different congestion control scheduling algorithm i.e. RED, SFQ and random exponential marking (REM) [22]. These algorithm are tested

for the delay, throughput, and queue length parameters. Red is used to increase traffic in the network. SFQ is used for the prevention of busty flow and provide fair access to the network and REM was used for congestion measurement. Red shows better results in terms of delay. REM was best in throughput and loss ratio. S FQ shows an average in loss ratio. Among all the three algorithms REM is consider as best algorithm for a congested network.

Garg et al. (2013) analyzed the importance of management schemes for Internet and MANET [10]. Two hybrid routing protocols i.e. Zone Routing Protocol (ZRP) and LANMAR. The comparison of the system proved that the LANMAR system is more feasible and beneficial for the use. Veni and Latha (2013) studied that each device and equipment in MANET model move freely and independently in any direction. In such model, the main challenge is continuously maintaining the required and proper information. MANET has the capability to change the location and adapt itself according to requirement. In such case, the network is decentralized where all network activities are incorporated in mobile path.

According to Gupta et al. (2013) MANET system contained a variety of mobile node which can change topology easily. It did not require centralized infrastructure [**?**]. Three different routing protocols i.e. AODV, DSR, and DSDV were used in various mobility models. If mobility rate is high DSDV performance will be low. The better delivery ratio of DSR and AODV depends on discovery of a route through which data can be transmitted to the ultimate destination. These routing protocols adopt different mechanism in case of frequent link failure due to mobility.

Goyal and Kakar (2013) studied that various mobile nodes collectively form an ad hoc network and communication with a wireless link [11]. MANET was connected with wireless transmitter and receivers. The author compared the performance of four reactive routing protocols i.e. Dynamic MANET on Demand (DYMO), location aided routing (LAR) AODV, DSR. LAR protocol outperformed by having less jitter and end to end delay. Alsahag et al. (2014) studied the rapid advancement of technologies in computer networking and required QOS to manage its overall performance [6]. A Worldwide Interoperability for Microwave Access (WIMAX) network has been studied and a special bandwidth allocation technique has been applied. One of the major issue in WiMAX is its scheduling algorithm. In real-time and non-real-time application WiMAX did not fulfill the requirements of QOS which lead to insufficient allocation of bandwidth, latency, and throughput. To overcome these issues the author proposed a new scheme called FADRR for mobile WiMAX. For real and non-real time application FADRR uses fuzzy logic to allocate bandwidth that guarantees the optimal bandwidth for each flow.

Alsahag et al. (2014) studied the rapid advancement of technologies in computer networking and requiring quality of service QOS to manage its overall performance [6]. A WIMAX network has been studied and a special bandwidth allocation technique has been applied. One of the major issues in the WiMAX was its scheduling algorithm is real-time and non-real-time application that did not fulfill the requirement of the quality of services which lead to insufficient allocation of bandwidth, latency, and throughput. This paper introduces a new scheme called FADRR is propped for mobile WIMAX. FADRR used fuzzy logic approach and different service flows in BS. FADRR used a deadline based approach to allocating bandwidth for real and non-real time application and this bandwidth allocation is done by mean of a fuzzy logic system that guarantees the optimal bandwidth for each flow by taking latency and throughput parameters under its consideration. FADRR has also been evaluated for a number of different algorithms like MDRR and CDRR by taking jitter, delay, throughput and fairness under its consideration for different classes like arts, reps, nrtps, ugs and be but FADRR proved to be the best among all. The simulation results analyzed that FADRR was efficient in real time applications in respect to QOS while it's a lot fair allocation to non-real time application and improves the overall system performance.

Chitkara and Ahmad (2014), studied that laptops, wireless telephones, and wireless sensors were mostly used nowadays [7]. A wireless node having no infrastructure and no central administration were used in MANET. Topology was also changed very frequently between the nodes. Different routing techniques and different strategies were implemented in MANET. In MANET, the authors have studied different characteristics, advantages, application and challenges in Manets.

In short, there were many AQM techniques proposed in the literature that was specially designed for wired network or wireless networks such as RED, Droptail [5] Blue [22] and ECN [8]. These AQM techniques had their own pros and cons in various scenarios as discussed above. In literature survey, there were no specialized techniques that were well designed for wired cum wireless environment. On another hand, the wired cum wireless networks popularity increased day by day due to technological advancements such as laptops, smartphones, and tablets. This literature study revealed the importance of DRR and SFQ utilization in wired cum wireless network which can help us in formulating wired cum wireless base specialized AQM technique.

III. Active Queue Management

The main function of AQM is to remove congestion, predictable queuing delay, and high link utilization. There are three queue management schemes used in AQM components. namely congestion indicator, congestion control function, and feedback mechanism [4]. When there is congestion in the network the queue management uses congestion indication to decide when there is congestion. Whereas how to remove the congestion or what must be done when there is congestion, it's the duty of congestion control function. The function of the feedback mechanism is the congestion signal used to aware the source to adjust its transmission rates [23]. These components are shown in Fig. 1.

DRR and SFQ are AQM techniques play important role in buffer management in order to control the congestion in the wired-cum-wireless network by dropping packets during the buffer overflow or near to overflow [14].

A. Deficit Round Robin

Simple round robin service uses the constant time for packet scheduling along different paths. The major flaw in simple round robin scheduler is to avoid large packet size in the

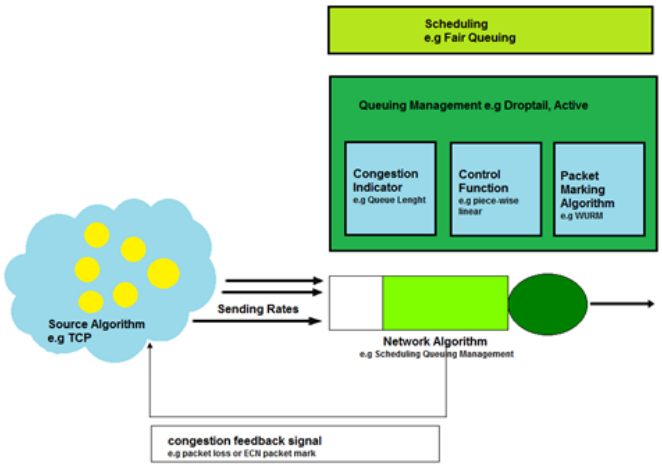


Fig. 1. General active queue management components.

first round. Variation in packet size creates unfairness which can be removed when the time is constant. This modification of round robin service is called DRR [24].

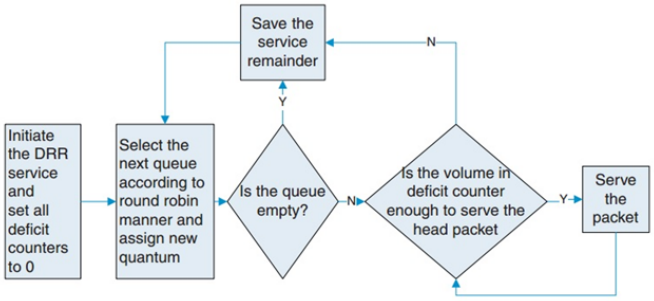


Fig. 2. Flow diagram of DRR mechanism.

DRR scheduling transmits traffic flows in a round robin manner. Unlike other round robin variants, DRR provides deficit counters for individual traffic flows. A volume of service equal to its quantum size is provided to each active flow in every round. The remaining volumes of services are used in the next round for deficit counter. This means at the start of each round the volume of the deficit counter of a traffic flow is equal to the sum of its quantum size and the volume of its remaining service from the previous round. An active flow has the packets that can be served in each round are determined by the value of its deficit counter. The DRR scheduling mechanism initiates to attend the next active flow in each round under either of the two conditions. The present queue is either blank or there is not enough service volume left to serve packet in the current flow. There is a static quantum given to each flow to avoid misbehavior in traffic flow. The DRR can also eliminate the unfairness, caused in the changeability of packet sizes of different flows. This is truly the most important improvement of DRR compared to the original round robin scheme. It is clear that the service assigned to each flow depends only on its fixed quantum and any greedy traffic flows cannot take away the remaining service of the other flow [17]. Flow diagram of DRR is shown in Fig. 2.

B. Stochastic Fair Queue

Scheduling algorithm uses fair queuing technique. Fair queuing is used in a network scheduler. In every traffic flow, a separate data packet queue is used contrasting to the FIFO queue technique which uses a single queue for all data packet flows. Fairness is only accomplished when a small amount of resources is used. The fair queuing algorithm uses SFQ technique. Stochastic fair queue performance is not precise as compared to others but it is best according to fewer calculations. Conversation (or flow) is the main word which is often used in SFQ coincides with TCP session.

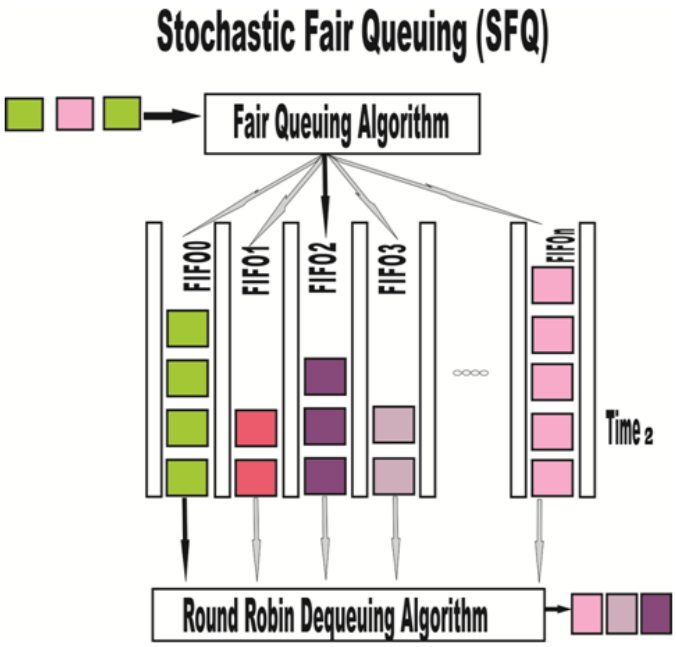


Fig. 3. Working diagram of stochastic fair queue.

When large amount of traffic is coming it is separated into FIFO queues using a single flow for each queue. Here a round robin technique is used for a huge traffic because every session has the equal chance to send their data. This is the most efficient process for every conversation and it does not drown the entire flow. SFQ uses an algorithm known as hashing algorithm, in this algorithm traffic is separated into a number of queues and does not assign queue in favor of every session. Fig. 3 shows stochastic fair queuing algorithm.

IV. Performance Evaluation

In this section, we describe simulation scenario with related parameters. The performance of DRR and SFQ are evaluated by a series of simulations using NS-2 tool [12]. In this research, file transfer protocol (FTP) is used at the application layer and transmission control protocol (TCP) is used at the transport layer. Destination sequenced distance vector (DSDV) routing protocol is used having simulation area of 250 x 250 m2 and simulation time is 300 seconds. Two numbers of wired nodes, five to fifteen mobile nodes with one base station (BS) are used. Random Way Point (RWP) is used as a mobility model. There is variation in time regarding mobility. DRR and SFQ are used as an AQM technique and the buffer capacity is 50 packets. The other parameters are listed in Table I.

TABLE I. SCENARIO CONFIGURATION PARAMETERS IN NS-2

Sr. No.	Parameters	Values
1	Application layer Protocol (ALP)	File Transfer Protocol (FTP)
2	Transport Layer Protocol	Transmission Control Protocol (TCP)
3	Number of Base Station	1 (Basestation)
4	Number of Wired Node	2 (Wired Nodes)
5	Number of Mobile Nodes	5, 10,15(Mobile Nodes)
6	Mobility Model	Random Way Point (RWP)
7	Mobility Speed	1.4, 3.3, 11.11 (m/Sec)
8	Active Queue Management Techniques	Deficit Round Robin (DRR) / Stochastic Fair Queue (SFQ)
9	General Buffer Capacity in nodes	50 Packets
10	Routing Protocol	Destination Sequence Distance Vector (DSDV)
11	Simulation Area	250 x 250 (m2)
12	Simulation Time	300 Seconds
13	Bandwidth between Wired Node / Base Station	5 Mbps
14	Delay between Wired Nodes/Base Station	2 ms Seconds

A. Simulation Parameters

In this section, we introduce some metrics that are necessary for performance evaluation. The parameter used for the QoS in the wired-cum-wireless network is average per packet delay and average packet dropped.

1) Average per packet delay: This parameter is used to find out the time consumed by the packet from one node to another node. The unit used in the end-to-end delay is a millisecond. Mathematical formula is as under.

$$Averageperpacketdelay = \frac{\sum_{i=1}^{n} EndtoEndDelay}{NumberofPackets} \quad (1)$$

2) Average packet drop rate: A number of packets dropped at a specific time is to find out by using packet dropped parameter. The unit for the packet dropped is packets per second. Mathematical formula for the packet dropped is as under:

$$conAveragepacketdropped = \frac{\sum_{i=1}^{n} DroppedPackets}{SimulationTime} \quad (2)$$

B. Simulation Scenarios

In this research study a simulation scenario is configured. On the basis of DRR and SQF different scenarios, i.e. increasing number of nodes, mobility and pause time are compared and analyzed to find out performance parameters, i.e. average per packet delay and average packet dropped.

C. Increasing Number of Nodes Scenario

DRR and SQF are the two AQM techniques which are observed by increasing the number of nodes from 5 to 15 while mobility and pause time remain constant.

1) Average per packet delay: The average packet delay is compared with DRR and SFQ, as shown in Fig. 4. Here it is clearly mentioned that DRR has the highest delay while SFQ has the lowest delay. As the number of nodes increases, packet delay in DRR also increases, on the other hand, using the same amount of nodes, SFQ gives the lowest delay. In short, SFQ outperformed DRR.

Using the scenario of DRR, when the numbers of nodes are less i.e. 5 nodes, the average per packet delay is less but as the number of nodes increases i.e. 15 nodes the average per packet delay also increases. It is because packets are organized in round robin. Using the scenario of SFQ, when the numbers of nodes are less i.e. 5 nodes, the average per packet delay is less but when the number of nodes increases i.e. 15 nodes, the average per packet delay is not that high as compared to DRR. It is because probability distribution and queues maintained statistically.

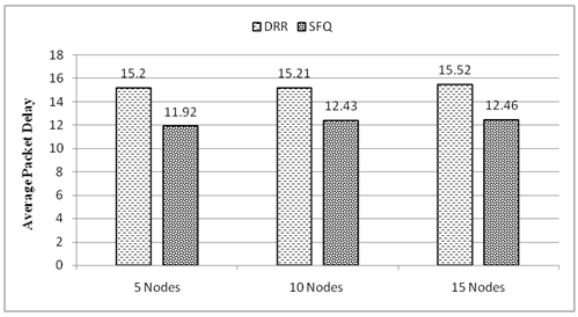


Fig. 4. Average per packet delay with respect to increasing number of nodes.

2) Average packet drop rate: Fig. 5 shows the average packet dropped between DRR and SFQ. Here, DRR has higher packet drop rate as the number of nodes are increasing from 5 to 15. When the number of nodes is 5 less packet are dropped in DRR, as the number of nodes increases the packet dropped ratio also increases. While on the other hand SFQ shows a slight difference in packet dropped rate. When the number of nodes is less i.e. 5 nodes, packet dropped ratio is high but as the number of nodes increases i.e. 15 nodes, the dropped ratio becomes low. In short, SFQ outperforms DRR with respect to packet dropped rate.

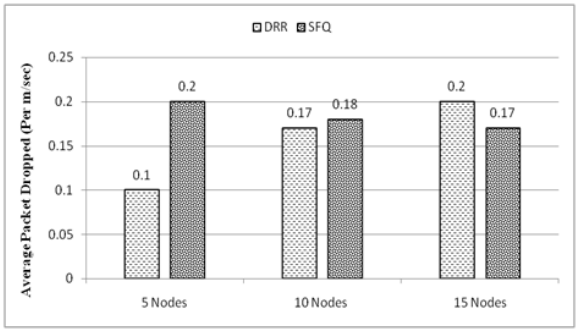


Fig. 5. Average packet drop with respect to increasing number of nodes.

D. Mobility Scenario

Two AQM techniques i.e. DRR and SFQ are used to compare mobility of nodes using different moving scenarios. Walking speed is 1.4 m/sec, running speed is 3.3 m/sec, and the speed of the vehicle is 11.11 m/sec. During mobility of nodes, other parameters like Pause time and number of nodes are kept constant.

1) Average per packet delay: Average packet delay of the AQM techniques i.e. DRR and SFQ are compared using multiple moving scenarios as shown in figure 6. It is clearly observed that walking speed, running speed and vehicle speed is kept changing like when a node is on walking speed delay is 15.21, when a node is running, the delay is lower i.e. 15.18, but when a node is in the vehicle delay is again rise to 15.23, the delay is almost same. This is in the case of sub-scenario of DRR, while in the case of SFQ packet delay in gradually increasing from walking speed to vehicle speed due to movement of nodes and links. Delay increases in SFQ because when nodes move distance between the nodes increasing and found problems in connectivity.

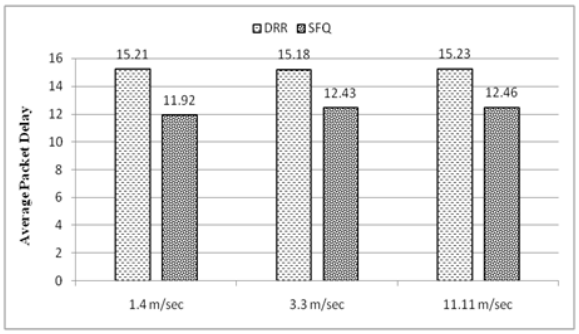


Fig. 6. Average per packet delay with respect to increasing mobility speed.

2) Average packet drop rate: Average packet dropped off the AQM techniques i.e. DRR and SFQ are compared using multiple moving sub-scenarios as shown in Fig. 7. It is clearly observed that average packet dropped rate is lower in walking speed (1.4 m/sec) for DRR while it is higher in SFQ. When the node is in running (3.3 m/sec) position average packet dropped rate is lower for DRR and higher for SFQ. Average packet dropped rate is lower in vehicle speed (11.11 m/sec) for DRR while it is higher in SFQ. Average packet dropped rate is gradually increased when mobility of nodes increases for DRR.

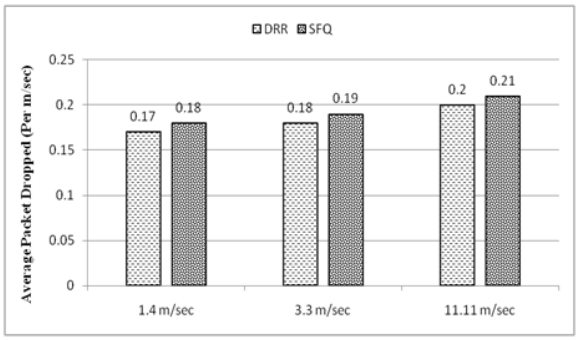


Fig. 7. Average packet dropped with respect to increasing mobility speed.

E. Pause Time Scenario

In this scenario number of nodes and mobility remain constant while pause time changes for DRR and SFQ.

1) Average per packet delay: Average packet delay is compared to AQM techniques i.e. DRR and SFQ in Fig. 8. It is observed that in each sub scenario pause time varies (between 2 to 10 seconds) for AQM techniques. In each sub-scenario, average packet delay is lowered both for DRR and SFQ.

When the pause time is 2 seconds, average per packet delay is highest for DRR and lowest for SFQ. When the pause time is 5 seconds, average per packet delay is highest for DRR and lowest for SFQ. When the pause time is 10 seconds, average per packet delay is highest for DRR and lowest for SFQ. It is observed from Fig. 4 and 5, as the time increases the node become too static, so the delay in the packet is lowered due to network link failure and stable network topology.

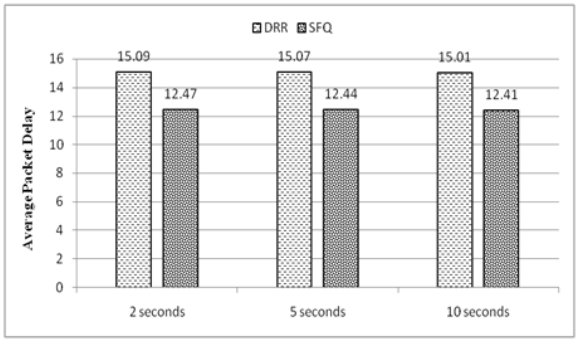


Fig. 8. Average per packet delay with respect to increasing pause time.

2) Average Packet Drop Rate: Average packet drop rate is gradually decreasing for AQM techniques i.e. DRR and SFQ by changing pause time. It is observed from Fig. 9 that in each sub scenario i.e. 2, 5 and 10 seconds pause time varies. Mostly average packet drop rate for DRR and SFQ is lowered due to low movement of nodes and no changes of intermediate nodes between sender and receiver. It uses the same route as established at the time of connection, no need to find a new route. As time increases the node come to static state, so no disconnection occurs.

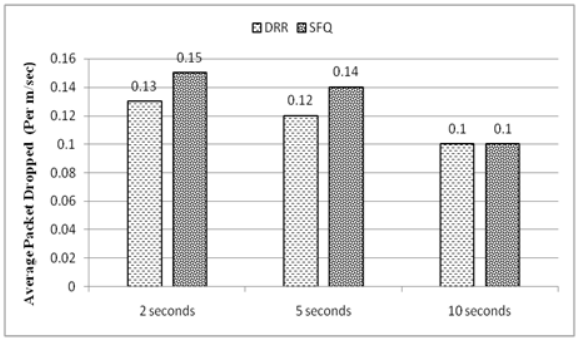


Fig. 9. Average packet drop with respect to increasing pause time.

V. Conclusion

This research paper focus on the performance evaluation of the DRR and SFQ using different scenarios such as increasing number of nodes, pause time and mobility scenario based on two parameters. It is observed the SFQ has outperformed than DRR by having comparatively low per packet delay. While on the other hand DRR has highest packet ratio as compare to

SFQ in case of increasing number of nodes. In mobility and pause time scenario, SFQ has less per packet delay while DRR has less packet dropped ratio. Considering all the evaluation and results, it is necessary to analyze AQM techniques with different traffic pattern in wired cum wireless network such as bursty traffic and constant bit rate. Furthermore, it is also necessary to modify SFQ to drop data equally from all flows instead of targeting single flow. DRR should be implemented with limited number of queues to improve its performance for future work.

REFERENCES

[1] Wang, X., Kar, K., and Low, S. H. (2009). *End-to-end fair rate optimization in wired-cum-wireless networks.* Ad Hoc Networks, 7(3), 473-485

[2] Rind, A. R., Shahzad, K., and Qadir, M. A. (2006). *Evaluation and comparison of TCP and UDP over Wired-cum-Wireless LAN.* In Multitopic Conference, 2006. INMIC'06. IEEE (pp. 337-342).

[3] Bitorika, A., Robin, M., Huggard, M., and Mc Goldrick, C. (2004). *A comparative study of active queue management schemes.* In Proc. of ICC.v2004

[4] Adms, R. (2013). *Active queue management: a survey.* IEEE communications surveys and tutorials, Vol. 15, Issue 3, pp-1425-1476.

[5] Ahmed, I., Badia, L., and Hussain, K. (2010). *Evaluation of deficit round robin queue discipline for real-time traffic management in an RTP/RTCP environment.* InFourth UKSim European Symposium on IEEE Computer Modeling and Simulation (EMS), pp. 484-489.

[6] Alsahag, A. M., Ali, B. M., Noordin, N. K., and Mohamad, H. (2014). *Fair uplink bandwidth allocation and latency guarantee for mobile WiMAX using fuzzy adaptive deficit round robin.* Journal of Network and Computer Applications, Vol.39, pp.17-25.

[7] Chitkara, M., and Ahmad, M. W. (2014). *Review on MANET: Characteristics, Challenges, Imperatives and Routing Protocols.* International Journal of Computer Science and Mobile Computing, 3 (2), 432-437.

[8] Chung, J., Claypool, M., and Kinicki, R. (2007,). *Stochastic Fair Traffic Management for Efficient and Robust IP Networking.* In Performance, Computing, and Communications Conference, 2007. IPCCC 2007. IEEE International (pp. 45-54).

[9] Cooper, N., and Meghanathan, N. (2010), *Impact of mobility models on multipath routing in mobile ad hoc networks,* AIRCC Int. J. Comput. Netw. Commun, 2, 174-185.

[10] Garg, P., Nagpal, C. K., and Bansal, S. (2013). *The impact of Random Waypoint Mobility Model on Hybrid Routing Protocols of Scalable Mobile Ad Hoc Network.* International Journal of Innovative Research and Development, 2(10).

[11] Goyal, E. H., and Kakkar, E. P. (2014,). *Performance Investigation of DYMO, DSR, AODV and LAR Routing Protocols using Different Mobility in MANETs.* In International Journal of Engineering Research and Technology (Vol. 2, No. 12 (December-2013)). ESRSA Publications.

[12] Issariyakul, T., and Hossain, E. (2011). *Introduction to network simulator NS2.* Springer Science and Business Media, pp 35-36

[13] Jonit, N. M., and Baba, M. D. (2011). *First in first out (FIFO) and Deficit Round Robin (DRR) scheduling analysis in WiMAX network.* In Control and System Graduate Research Colloquium (ICSGRC), 2011 IEEE (pp. 166-174).

[14] Jain, B., and Madan, S. *Comparative Analysis of various Active Queue Management Algorithms under Flooding-based LDDoS Attack.*

[15] Kortebi, A., Oueslati, S., and Roberts, J. (2005). *Implicit service differentiation using deficit round robin.* ITC19.

[16] Lin, D., and Hamdi, M. (2010,). *Two-stage fair queuing using budget round-robin. In Communications (ICC),* 2010 IEEE International Conference on (pp. 1-5).

[17] Liu, L., Jin, X., Min, G., and Li, K. (2010). *Performance modeling and analysis of Deficit Round Robin scheduling scheme with self-similar traffic. Concurrency and Computation: Practice and Experience,* 22(13), 1911-1926.

[18] Maan, F., and Mazhar, N. (2011). *MANET routing protocols vs. mobility models: A performance evaluation. In Ubiquitous and Future Networks (ICUFN),* 2011 Third International Conference on (pp. 179-184).

[19] Mahdipour, E., Rahmani, A. M., and Aminian, E. (2009). *Performance evaluation of destination-sequenced distance vector (DSDV) routing protocol.* In Future Networks, 2009 International Conference on (pp. 186-190).

[20] McKenney, P. E. (1990). *Stochastic Fairness is queuing.* In IEEE INFOCOM'90, Ninth Annual Joint Conference of the IEEE Computer and Communication Societies. The Multiple Facets of Integration. Proceedings, IEEE (pp. 733-740).

[21] Noon, A., Kalakech, A., and Kadry, S. (2011). *A new round-robin based scheduling algorithm for operating systems: dynamic quantum using the mean average.* arXiv preprint arXiv:1111.5348.

[22] Patel, S., Gupta, P. K., Garg, A., Mehrotra, P., and Chhabra, M. (2012). *Comparative analysis of congestion controls algorithms using ns-2.* arXiv preprint arXiv:1203.3654

[23] Radhakrishnan, S., Pan, R., Vahdat, A., and Varghese, G. (2012). *Netshare and stochastic net share: predictable bandwidth allocation for data centers.* ACM SIGCOMM Computer Communication Review, 42(3), 5-11.

[24] Shreedhar, M., and Varghese, G. (1996). *Efficient fair queuing using deficit round robin.* Networking, IEEE/ACM Transactions on, 4(3), 375-385.

[25] Khan, S. and Qadir, M. A. (2015). *Inter-path OOS packets differentiation based congestion control for simultaneous multipath transmission.* Int. Arab J. Inf. Technol. Vol.4, No.6, pp.907-913.

[26] Khan, S. and Qadir, M. A. (2017). *Deterministic Time Markov Chain Modelling of Simultaneous Multipath Transmission Schemes.* IEEE Access, Vol. 5, pp.8536-8544.

[27] Ali, H., Khan, S. and Quaid, M. (2015). *Comparative analysis of controlled delay (CoDel) with Deficit Round Robin (DRR) to overcome bufferbloat problem in wired network.* International Journal of Current Engineering and Technology, Vol.5, No. 5, pp. 3378-3386.

[28] Khan, F., Abbas, S. and Khan, S. (2016). *An Efficient and Reliable Core-Assisted Multicast Routing Protocol in Mobile Ad-Hoc Network.* International journal of advanced computer science and applications, Vol. 7, No. 5, pp. 231-242.

[29] Shakir, S. Khan,S. Hassain, L. and Matiullah (2017) *QoS Based Evaluation of Multipath Routing Protocols in Manets,* Advances in Networks. Vol. 5, No. 2, 2017, pp. 47-53. doi: 10.11648/j.net.20170502.13

[30] Khan, F., Abbas, S. and Khan, S. (2018). *Secure Core-Assisted Multicast Routing Protocol in Mobile Ad-Hoc Network*, Journal of Internet Technology, Article In Press.

Effect of Increasing Number of Nodes on Performance of SMAC, CSMA/CA and TDMA in MANETs

Samiullah Khan, Farooq Faisal,
Mahmood Nawaz, Farkhanda Javed
Institute of Business Management Sciences
The University of Agriculture Peshawar
Peshawar-Pakistan

Matiullah, Zia ullah, Muhammad Shoaib, Faqir Usman Masood
Department of Basic Sciences and Islamiat
Department of Computer System Engineering
University of Engineering and Technology
Computer science Department, Qurtuba University
Peshawar-Pakistan

Fawad Ali Khan, Rafidah MD Noor
Department of Computer System and Technology
University of Malaya, Kuala Lumpur
Malaysia

***Abstract*—The importance of Wireless Sensor Network (WSN) increases due to deployment for geographical, environmental and surveillance purpose in war fields. WSN facing several challenges due to its complex nature including key problems, such as routing and medium access control protocols. Several approaches were proposed for the performance evaluation of WSN on the basis of these issues due to the fact that MAC layer access protocols have a great impact on the performance of WSN. In this paper, we investigated the performance evaluation of three well known MAC Access protocols, i.e. sensor medium access control protocol (SMAC), carrier sense multiple access with collision avoidance (CSMA/CA), and time division multiple access (TDMA) over ad-hoc on demand distance vector (AODV) routing protocol. The number of simulation scenarios were carried out by using NS-2, the simulation metrics used are throughput, end-to-end delay and energy consumed. Simulation results showed that SMAC out perform CSMA/CA and TDMA by consuming less energy, less end to end delay and high throughput due to contention based approach to access the medium for transmission.**

Keywords*—Medium access control (MAC); sensor medium access control (SMAC); time division multiple access (TDMA); carrier sense multiple access with collision avoidance (CSMA/CA); ad-hoc on demand distance vector (AODV)

I. Introduction

The ad-hoc network is a network of self configuring mobile nodes that can communicate with each other without any centralized control. Intermediate nodes act as a router to forward data from source to destination. In ad-hoc networks, the nodes can move freely and the topology may change frequently due to nodes mobility. Every node in this network acts as host and router. Intermediate nodes act as a router to forward data from source to destination. Routing protocols are responsible for keeping track of the paths from source to destination in their routing tables. Routing tables are constructed by exchanging control messages to keep the information updated [1]. These messages causes extra overhead of the routing protocol and degrades the performance. An efficient routing protocol utilizes the network resources by minimizing delay and power consumption to provide better QOS and maximizing throughput, remains a challenging issue for the ad-hoc networks. Energy consumption is a major factor and have a great impact on the network performance in battery driven nodes in these networks [16].

Medium access control (MAC) plays an important role in the successful operation of the network. One fundamental task of MAC is to avoid collision, so that nodes in the same interference range do not transmit at the same time. The two main operations performed at medium access layer are controlling when to send and when to listen for a packet. There are many MAC protocols that are developed for wireless networks.Typical examples are TDMA, CDMA, CSMA/CA, SMAC, etc. The motivation for the selection of TDMA, CSMA/CA and SMAC in this research work were due to diverse type of access mechanism used by these protocols and have potential for further improvement in order to conserve the battery energy as mentioned in related work section [17]. The rest of the paper is organized as follows. Section II presents the overview of Medium Access protocols and the Routing protocol used and Section III describes the related work done in literature. Simulation environment and results are shown in Section IV and the paper is concluded in Section V.

A. Medium Access Control (MAC) Protocols

S-MAC uses three novel techniques to reduce energy consumption and support self-configuration as shown in Fig. 1. Nodes periodically sleeps by entering from listening to the idle channel to reduce energy consumption. During transmission of other nodes, SMAC sets the radio to sleep and it only uses in-channel signaling. An application that requires a store-and-forward processing as data move through the network message passing are applied by SMAC to reduce contentious latency of the network [18].

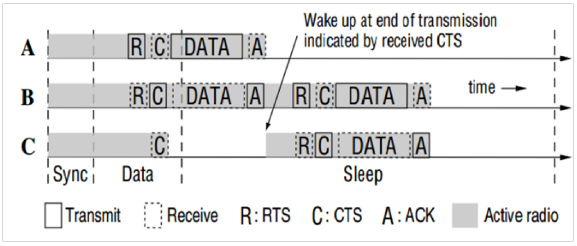


Fig. 1. S-MAC with adoptive listening [1].

CSMA/CA works on the principle of sensing before sending as shown in Fig. 2. As soon as a node receives a packet, it sense the medium and try to send it if the medium is clear or wait a randomly chosen period of time (2 RTT) if another transmission is going on and then checks again. This waiting time period is called back-off factor and is counted down by a back-off counter. If the back-off counter reaches zero and the channel is CSMA/CA will send the packets otherwise the back-off factor is set again and the process will be repeated [2].

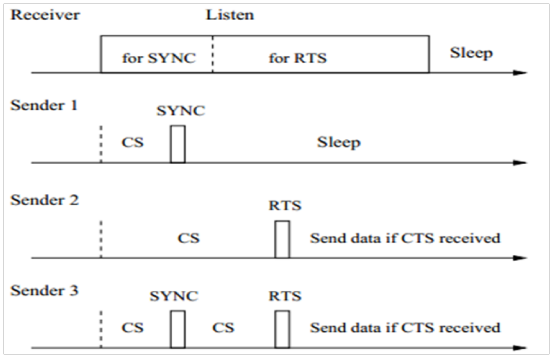


Fig. 2. Timing relationship between receiver and different senders [3].

TDMA shares a single carrier with several users to make use of non-overlapping time slots for each user as shown in Fig. 3. In TDMA, transmission of data is not continuous but occurs in bursts which results in low consumption of energy. For transmission and reception different, the time slots are used by TDMA, thus duplexes are not required. Bandwidth can be supplied on demand to different users by concatenating or reassigning of time slots based on priority.

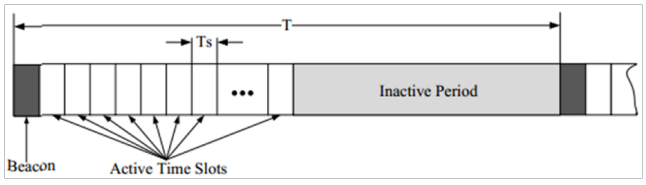


Fig. 3. TDMA frame structure [4].

This paper also investigates the behavior of the popular reactive routing protocol AODV over varying MAC protocols.

B. Routing in Ad-hoc Networks

Routing is the process of finding the shortest and optimal paths from source to destination in any network. Routing protocols for ad-hoc network can be classified into three main categories, i.e. Reactive, proactive and hybrid. Reactive routing protocols are more popular in ad-hoc networks because they are scalable and creating less overhead. In reactive routing protocols, The routes are established at the time when a node wants to communicate with another node in the network. There are no predefined route which results in reduction of overhead due to the fact that there is no exchange of information to keep track of the network topology [19].

Adhoc On-Demand Distance Vector (AODV) is a reactive routing protocol and works on an on-demand basis, i.e. a route is established only when it is desired by a source node. Routing table is maintained for every node in which the required information for communicating with other users in the network are stored. It reacts to the changes by maintaining only the active routes in the tables for a specified amount of expiration time [20]. AODV identify the most recent path by employing destination sequence numbers. The routes in AODV are expected to be available at a given instant. AODV uses three types of messages to be exchanged for establishing and maintaining routes i.e. route request (RREQ), route reply (RREP) and route error (RERR). When a node wants to communicate, it sends the RREQ as broadcast to its neighbor nodes. The node with the destination address will reply by sending RREP as unicast to the desired source node. It may be possible that a node loses connectivity due to some reason. The next hop nodes will notify all the nodes by sending RERR messages, to keep the information updated [5]. Due to this reactive approach, AODV was used as routing protocol in this study.

II. Related Work

Aggarwal et al. (2012) worked on the security challenges faced by ad-hoc routing protocols and proposed (AODVSEC) by enhancing the scope of AODV for the provision of security and compared it with (AODV) and SAODV in normal situation as well as in the presence of three concerned attacks i.e Resource Consumption (RC) attack, Route Disturb (RD) attack, Route Invasion (RI) attack and Black hole (BH) attack. The proposed scheme performed better than AODV and similar to SAODV but with less processing requirements which leads to save a lot of computational power [6].

Chen et al. (2012) studied the routing protocol for low power and lossy network (RPL) and the challenges faced by the protocol to find effective approaches to simulate and emulate the behavior and other extensions of applications. Their work has first provided a brief detail of the RPL protocol by considering two case studies (Contiki RPL and Tiny RPL) and initial experimental simulation results obtained from the COOJA simulator which is RPL capable. Secondly, their work was focused on the utilization of the RPL protocol in the agriculture sector by proposing dedicated hybrid network architecture [7].

Garg and Gupta (2012) studied the behavior of AODV routing protocols and highlighted the dynamic nature of MANETs such as limited power of nodes, bandwidth constrained, variable wireless links and dynamic topology that leads to performance degradation. They implement AODV protocol for optimizing the routing by reducing the loss rate in AODV.

NS-2 is used for simulating scenarios such as variation in the speed of nodes, packet transfer rates and number of nodes. this work concluded that AODV increases network performance by increasing the packet delivery ratio and reduced network delay [8].

Updhayayet et al. (2012) investigated the behavior of routing in ad-hoc network and the problem faced by the routing protocols by testing both reactive and pro-active protocols for efficiency under different circumstances. They first increase the density of nodes for analyzing the efficiency of the protocols. The metric used were packet delivery ratio, packet loss and end-to-end delay. The simulation showed that reactive routing protocol performed well in these circumstances as compared to proactive [9].

Taneja et al. (2010) reported the experimental analysis of the two prominent on-demand routing protocols by presenting their functionality with varying number of nodes. An effort was made by the author to perform analysis of a new random way point self created network scenarios. The performance parameter taken was the packet delivery ratio in the network scenario of changing speed of the nodes. The results reflected in graphs showed that both protocols are best in its own way, but AODV was emphasized best on the basis of well performing in dense condition [10].

Van Hoese et al. (2004) studied EMAC, a medium access protocol designed for WSN. The EMAC work on top of the effective TDMA scheme, in which each active node listens to the channel periodically and broadcast control messages for sending various types of information by utilizing low energy. The active node performs normal network functions while the passive nodes set aside to save energy. There are two operational modes, the monitoring mode in which the network merely keeps the network connected in a low duty cycle and communication mode in which the nodes are active and transfer data. Their simulation results showed that the presented approach is better as opposed to SMAC protocol in terms of power consumption [11].

Hang Su et al. (2009) proposed a cross layer based battery-aware medium access protocol based on TDMA for monitoring of body area network in wireless health care application. Their proposed approach provides a solution for electrochemical properties of the battery, the time varying wireless fading channels, and packet queuing by prolonged lifespan of battery with reliable and timely guaranteeing delivery of messages in WSN, which is the typical requirement of the patient monitoring network. Analytical and simulation results showed their proposed work performed better by consuming less energy as compared to the IEEE 802.15.4 and Bluetooth protocols [12].

Giuseppe Bianchi et al. (2001) studied the effect of medium access protocol on the performance of WSN by considering different MAC protocols. CSMA/CA is one of the medium access protocol for wireless sensor networks which decreases throughput as the number of nodes in the network increases, the throughput is much better in CSMA/CA with RTS/CTS. In these simulation scenarios, CSMA/CA was compared with TDMA and SMAC by analyzing the throughput, end to end delay and energy consumption [13].

Wei Y et al. (2002) presented a new type of mac protocol designed especially for wireless sensor networks. Energy conservation and self configuration of nodes is the primary goal of the protocol, while latency and per node fairness is less important. SMAC used three novel techniques to consume less energy and support for self configuration. To consume less amount of energy, nodes periodically sleeps to move from listening to idle state. Further, they reduced contention latency for sensor network application by message passing technique that required store-and-forward processing as data move across the network. The simulation showed that SMAC consumes 2 to 6 times less energy as opposed 802.11 like for traffic load with a message sent every 1-10 secs.

Demirkol et al. (2006) studied various aspects of the WSN to be kept in mind while developing a MAC layer protocol. Target detection, temperature judgment and hospitals are some of the areas in which different application of WSN is applicable and therefore hot areas for researchers [21]. MAC access protocols are the best choice for utilization of energy resources in WSN. A comparison of two MAC access protocol is made by considering the metric as throughput, delay and energy consumed. The simulation results show that both the protocol performed well in different condition and SMAC consumes less energy as compared to TDMA [14].

Anis Koubaa et al. (2006) showed the performance limitation of slotted CSMA/CA mechanism of sensor nodes by enabling beacon mode for broadcast transmission in WSNs. The work was conducted by modelling the CSMA/CA mechanism on top of realistic physical layer with respect to IEEE 802.15.4 standard specification. The parameter used for analysis are throughput and delay. The simulation was conducted using NS-2 by enabling beacon mode in the first case and disabling beacon mode in the second case. Beacon enabled mode showed efficient and good results as compared with the non-beacon mode, which provide synchronization and contention free period [15].

III. Result and Discussion

This paper investigates the performance evaluation of Medium access protocol, i.e. SMAC, CSMA/CA and TDMA over AODV routing protocol. The simulation has been performed in the Network Simulator-2 (NS-2), which provides a scalable environment for wireless sensor networks. The simulation area was 2000 X 2000 m square and the number of nodes used was 5, 10, and 20 and each node have 1000 joule of initial energy. The Constant Bit Rate (CBR) was used as the application layer traffic for all scenarios with an interval time of 1, 5, and 10 Seconds. The packet size, transport and routing protocol were 50 bytes, User Datagram protocol (UDP) and AODV, respectively. The simulation time was kept 1000 seconds. The configuration of simulation parameters are mentioned in Table I.

The simulation was performed using NS-2 which provides a scalable environment for WSNs. A number of simulation scenarios (Scalability and interval) were carried out by using performance analysis parameters i.e. throughput, end to end delay and energy consumed. The scalability scenario was designed with motivation to find out the performance efficiency of MAC layer protocols with increasing number of nodes various sub scenarios. The number of nodes used for scalability

TABLE I. GENERAL SIMULATION PARAMETERS CONFIGURATIONS

S/No	Attributes	Values
1	Simulator	NS-2 (Version 2.34)
2	Application Layer Protocol	Constant Bit Rate(CBR)
3	MAC Layer Protocol	SMAC, CSMA/CA and TDMA
4	Nodes Intial Energy Leve	1000 Joules
5	Packet Size	50 Bytes
6	Number of Nodes	5, 10 and 20 Nodes
7	Routing Protocol	AODV
8	Simulation Time	1000 Seconds
9	Area (m2)	2000 x 2000
10	Traffic Type	CBR/UDP
11	Data Interval	1, 5 and 10 mseconds

scenarios were 5, 10, and 20. The Interval scenarios described real behavior of the nodes in the WSNs in which do not communicates continuously. To simulate this behavior, three Interval scenarios were designed with an interval of 1, 5 and 10 ms.

A. Scalability Scenario

The SMAC has high throughput in all three sub scalability scenarios as compare to the CSMA/CA and TDMA. The CSMA/CA is a contention base protocol in which the number of collision increases as the number of nodes in the network increased. TDMA is conflict free protocol and contention is avoided by the use of reserving the slot for each transmitter. SMAC has comparatively higher throughput as shown in Fig. 4. SMAC has features of contention and scheduling. Throughput of SMAC rises with the increase in node density.

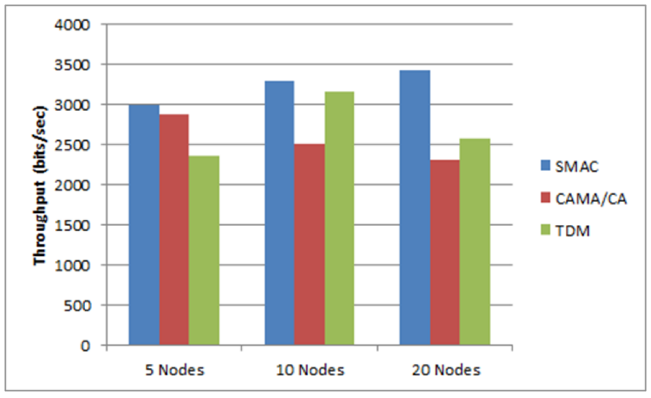


Fig. 4. Throughput of SMAC, CSMA/CA and TDMA in scalability scenario.

CSMA/CA has a comparatively higher delay (as shown in Fig. 5) with increase in the number of nodes in the network. This increase in delay of CSMA/CA is due to collisions and set back off the counter. This result in the comparatively less end to end delays, with the increase in the time interval. The simulation graph gives a clear picture of SMAC that have comparatively much lower delay as opposed to the other two MAC protocol.

The energy consumption of SMAC protocol is comparatively much lower than CSMA/CA and TDMA as shown in Fig. 6. SMAC used predictive techniques where nodes goes to idle state from communicating state due to periodic sleep.

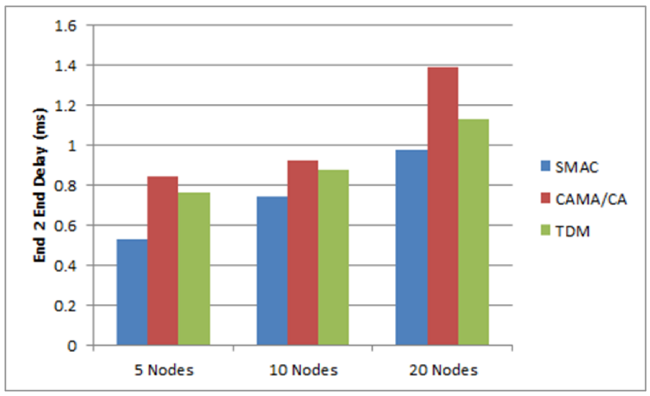


Fig. 5. End-2-end delay of SMAC, CSMA/CA and TDMA in scalability scenario.

This helps in reducing energy consumption of the protocol. The CSMA/CA senses the medium all the time for transmission and reception of data which increased the energy consumption of the protocol. Moreover, there is no predictive techniques used in TDMA which also consumes higher energy as compared to SMAC.

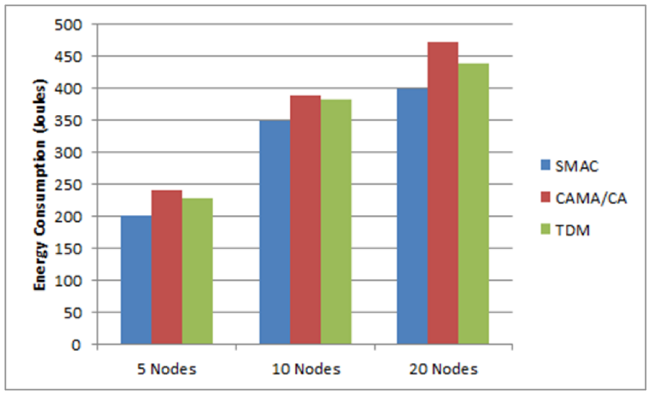


Fig. 6. Energy consumption of SMAC, CSMA/CA and TDMA in scalability scenario.

B. Interval or Data Rate Scenario

The number of nodes in interval scenarios are 20 and the interval time was kept 1, 5, and 10 ms. The throughput of all three MAC layer protocol decreases as the time interval increases as shown in Fig. 7. SMAC showed comparatively higher throughput as opposed to CSMA/CA and TDMA as shown in figure. This is due to the periodic sleeping techniques used by Sensor Medium Access protocol.

CSMA/CA and TDMA has comparatively higher delay as compared to SMAC as shown in Fig. 8. The simulation results showed that SMAC performed well and reduced the delay even when the time of interval was further increased. 1 ms interval corresponding to higher traffic while traffic generation decreased as the time interval was increased in the network.

The power consumption of TDMA and CSMA/CA is much higher than SMAC medium access protocol as shown in Fig. 9. The sleep delay prediction of SMAC gives the ability to protocol to consume less energy while TDMA and

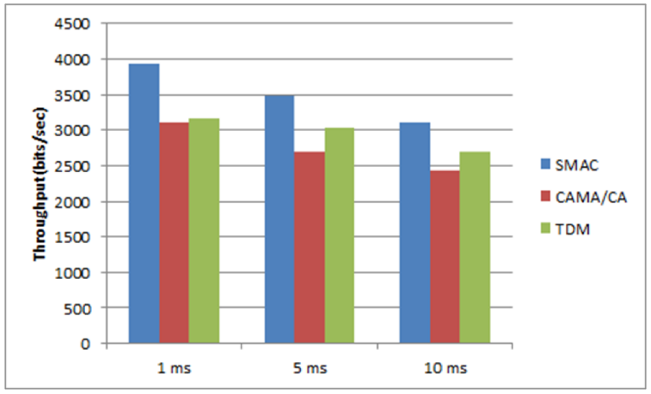


Fig. 7. Throughput of SMAC, CSMA/CA and TDMA in interval scenario.

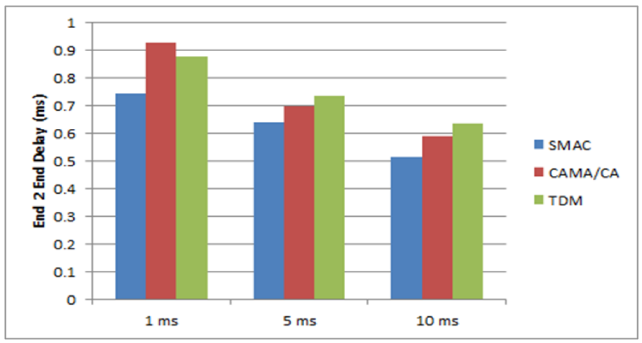


Fig. 8. End 2 end delay of SMAC, CSMA/CA and TDMA in interval scenario.

CSMA/CA was listening continuously to the medium which results in high consumption of energy.

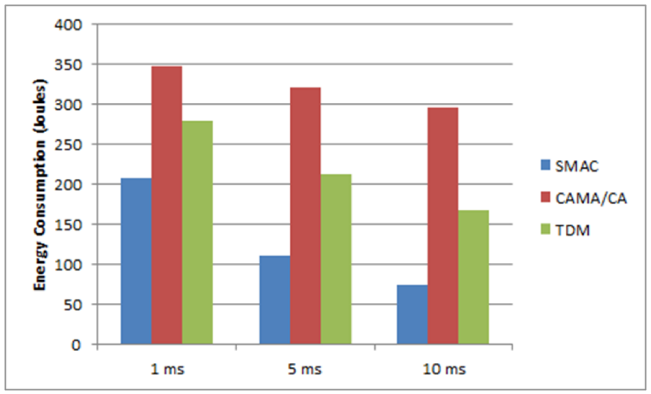


Fig. 9. Energy consumption of SMAC, CSMA/CA and TDMA in interval scenario.

IV. Conclusion

This paper examines the performance evaluation of three Medium Access Protocols, i.e. SMAC, TDMA and CSMA/CA over the AODV routing protocol for Wireless Sensor Networks (WSNs). The simulation results conclude that SMAC perform much better in both scalability and time interval scenarios as compared to the other medium access protocols, i.e. TDMA and CSMA/CA. The CSMA used contention to find out the collision by continually sensing the medium until getting free transmission medium. The conflict free MAC layer protocol, i.e. TDMA helped the nodes in avoiding the overlapping during transmission in WSNs. The SMAC was hybrid in nature as it have adopted the both the contention based mechanism of CSMA/CA to get the medium for transmission and conflict free nature of TDMA for dynamic allocation of medium or resources. It is concluded from the study and simulation results that SMAC performs better on AODV in WSNs in the given circumstances. Further research can be done in the same area by considering different simulation scenarios such as using different traffic pattern and routing protocols.

References

[1] A. Aggarwal and B. Garg (2012), *Survey on secure aodv for ad-hoc networks routing mechanism.*Int J Adv Res Comput Sci Softw Eng, vol. 2, no. 3, pp. 203-206.

[2] H. S. Bindra, S. K. Maakar, and A. Sangal (2010), *Performance evaluation of two reactive routing protocols of manet using group mobility model.* International Journal of Computer Science, vol. 7, no. 3, pp. 38-43.

[3] E. Ziouva and T. Antonakopoulos(2002) *Csma/ca performance under high traffic conditions: throughput and delay analysis* Computer communications, vol. 25, no. 3, pp. 313–321.

[4] B.-C. Kim (2001) *Method for forming frame structure for use in time division multiple access communication system.* uS Patent 6, pp.172-971.

[5] M. Alshowkan, E. A. Fattah, and A. Odeh,(2012) *Performance evaluation of dymo, aodv and dsr routing protocols in manet.* International Journal of Computer Applications, vol. 49, no. 11. pp. 176-191.

[6] A. Aggarwal, S. Gandhi, N. Chaubey, P. Shah, and M. Sadhwani (2012) *Aodvsec: A novel approach to secure ad hoc on-demand distance vector (aodv) routing protocol from insider attacks in manets.* arXiv preprint arXiv:1208.1959.

[7] Y. Chen, J.-P. Chanet, and K. M. Hou (2012) *Rpl routing protocol a case study: Precision agriculture.* First China-France Workshop on Future Computing Technology, pp. 6-17.

[8] A. Lal, S. Dubey, and B. Presswani (2012) *Reliability of manet through the performance evaluation of aodv, dsdv, dsr.* International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 5, pp. 125-134.

[9] S. Upadhyay, P. Joshi, N. Gandotra, and A. Kumari (2012) *Comparison and performance analysis of reactive type dsr, aodv and proactive type dsdv routing protocol for wireless mobile ad-hoc networking, using ns-2 simulator.* Journal of Engineering and Computer Innovations. vol. 2, no. 10, pp. 36–47.

[10] S. Taneja and A. Kush (2010) *A survey of routing protocols in mobile ad hoc networks.*International Journal of Innovation, Management and Technology, vol. 1, no. 3, pp. 279-291.

[11] L. Van Hoesel and P. J. Havinga (2004) *A tdma-based mac protocol for wsns* in Proceedings of the 2nd international conference on Embedded networked sensor systems. pp.303–304.

[12] H. Su and X. Zhang (2009) *Battery-dynamics driven tdma mac protocols for wireless body-area monitoring networks in healthcare applications.* Selected Areas in Communications, vol. 27, no. 4, pp. 424–434.

[13] Y. Tay and K. C. Chua (2001) *A capacity analysis for the ieee 802.11 mac protocol.* Wireless networks, vol. 7, no. 2, pp. 159–171.

[14] I. Demirkol, C. Ersoy, F. Alagoz (2006) *Mac protocols for wireless sensor networks: a survey* IEEE Communications Magazine, vol. 44, no. 4, pp. 115–121.

[15] A. Koubaa, M. Alves, and E. Tovar (2006) *A comprehensive simulation study of slotted csma/ca for ieee 802.15. 4 wireless sensor networks* in 5th IEEE International Workshop on Factory Communication Systems.pp. 183–192.

[16] Khan, S. and Qadir, M. A. (2015). *Inter-path OOS packets differentiation based congestion control for simultaneous multipath transmission.* Int. Arab J. Inf. Technol. Vol.4, No.6, pp.907-913.

[17] Khan, S. and Qadir, M. A. (2017). *Deterministic Time Markov Chain Modelling of Simultaneous Multipath Transmission Schemes.* IEEE Access, Vol. 5, pp.8536-8544.

[18] Ali, H., Khan, S. and Quaid, M. (2015). *Comparative analysis of controlled delay (CoDel) with Deficit Round Robin (DRR) to overcome bufferbloat problem in wired network.* International Journal of Current Engineering and Technology, Vol.5, No. 5, pp. 3378-3386.

[19] Khan, F., Abbas, S. and Khan, S. (2016). *An Efficient and Reliable Core-Assisted Multicast Routing Protocol in Mobile Ad-Hoc Network.* International journal of advanced computer science and applications, Vol. 7, No. 5, pp. 231-242.

[20] Shakir, S. Khan,S. Hassain, L. and Matiullah (2017) *QoS Based Evaluation of Multipath Routing Protocols in Manets,* Advances in Networks. Vol. 5, No. 2, 2017, pp. 47-53. doi: 10.11648/j.net.20170502.13

[21] Khan, F., Abbas, S. and Khan, S. (2018). *Secure Core-Assisted Multicast Routing Protocol in Mobile Ad-Hoc Network*, Journal of Internet Technology, Article In Press.

Dynamic Reconfiguration of LPWANs Pervasive System using Multi-agent Approach

Ghouti ABDELLAOUI, Fethi Tarik BENDIMERAD
University Abou Beker Belkaid,
Faculty of Technology
Tlemcen, Algeria

***Abstract*—The development of the Low Power Wide Area Network (LPWAN) has given new hope for the Internet of Things and M2M networks to become the most prevalent network type in industrial world in the near future. This type of network is designed to connect several entities in a radius that can reach up to 10 Km. This gain in scope is possible through a reduction in the amount of information exchanged. The latter will makes LPWANs Networks the most suitable for telemetry applications. This large network coverage offered by LPWAN gives the possibility to connect a large number of objects. On the other hand, it involves difficulties for the pervasive system associated with this kind of network to integrate dynamically all this objects, in which an automatic reconfiguration process becomes crucial to this kind of network. In this study, we propose the multi-agent systems as a solution to virtualize the heterogeneity of the peripherals and to facilitate their integration and their dynamic exploitation in the LPWAN system. That virtualization is possible due to the portability of the multi-agent systems and the standardization of the exploitation of the services offered by them.**

***Keywords*—Dynamic reconfiguration; pervasive system; multi-agent system; Low Power Wide Area Network (LPWAN)**

I. Introduction

The technology revolution has created a new world dedicated to information processing, where any real component is characterized by two sets (data and Handling). This can be seen as a global observing environment for several systems in different areas such as control of medical equipment, autonomous support, control and traffic safety, advanced automotive systems, process control, energy conservation, environmental control, control of critical infrastructure (electricity, communications system, e.g. tele-presence telemedicine), military systems, manufacturing systems, and any smart structures, etc. The majority of these systems are critical and vary chaotically in time, which require continuous changes of their representation in the information world.

The Low Power Wide Area Network (LPWAN) [1] such as LoRaWan [2] or SigFox [3] provides a solution to connect each component or object to the network and monitor their behaviour in real time through their pervasive system layer. For this purpose, we need to develop a mechanism for observing and dynamically updating the system, which can be considered as a dynamic reconfiguration for the initial state of the LPWAN pervasive system. This mechanism could be associated to the pervasive system layer of the LPWAN network, which is responsible for identifying and recognizing all the objects present in the network.

In this paper, to give pervasive system better perception of environment and make him more interacting "naturally" with the user, we propose a new mechanism improving the context-aware of this latter. For that, the systems must adapt to the different changes (adding and/or deleting a new subsystem) and dynamically reconfigure the changes that are imposed by the environment.

We will use the multi-agent systems approach [4], [5] to perform the dynamic reconfiguration. This chose is made, because they present better portability and flexibility by dint of the different behaviours offered by them, which gives a great capacity to improve context-aware of a pervasive system.

The rest of the paper is organized as follows. In the next one (Section II), we present the state of the art of the study. In this section, the dynamic reconfiguration and some reasons to do it with their approaches are given, in the second one, we presents the pervasive system and the most its important characteristics. After that, we talks about multi-agent system and what is the aim to use theme for the dynamic reconfiguration. This section will be end by the presentation of the multi-agent system as solution for our problematic. Section III gives the LPWAN networks like a case of use to implement our proposed approach in order to perform the dynamic reconfiguration of the pervasive system. Section IV presents our implementation for the proposed approach. Section V discuss the use of this approach and their advantages. Finally, Section VI draw the context of our work and future scope.

II. State of the Art

Along this section we give an overview of the dynamic reconfiguration and their different approaches and the multi-agent system as a solution to perform a dynamic reconfiguration on the LPWAN networks.

A. Dynamic Reconfiguration

Currently, many applications require a runtime dynamic reconfiguration. This requirement becomes critical when it is not possible to stop or restart a component or a service.

There are several examples in which the application of this practice is crucial. Especially in the industrial area where the embedded systems is complex, such as operating system of some modern cars control units requires operating systems while different modes with several possible configurations. In this case, we can said that the reconfiguration is the set of passage operations from one configuration to another, and we can take the following definition of the dynamic reconfiguration: "Dynamic reconfiguration is the ability of a software system to allow modification of a subset of the system during its execution, without interruption of service" [6].

In some areas, it is enough to have a finite set of configurations like the case of cars, but in others ones, we cannot anticipate all possible configurations. For this reason, the use of an approach based on configuration rules is more appropriate.

Among the applications which have trend towards dynamic reconfiguration, applications based on the agent approach are involved. This one becomes the most used method in the world of software development due its benefits to the use.

1) Reasons for the reconfiguration: Each system has its own reasons to be reconfigured according to the abnormalities, this may cause undesired operation. Some of ones, such as distributed system, risk of failures at server or become obsolete. Such a system requires a reconfiguration to correct this problem as in the case of zookeeper system [7]. In order to perform the composition of several web services and extend it more and more. This latte need to be changed as well performed and adapted to new ISC [8] specifications. In this approach the main element of the system is the service, since each one has its own specifications and works independently from the others.

The dynamic reconfiguration can be used to expand embedded systems with the reuse of hardware resources through a middle ware [9], or by the use of systems based on FPGAs [10], whose main characteristic is the dynamic reconfiguration.

The reconfiguration of FPGAs is possible through the ability to reprogram some blocks (CLB) [10] that are elected uncomfortable for the application. Knowing that each block or group of blocks can implement one or more spots in the application, this makes it modular and therefore the runtime partial maintenance is possible.

Also a dynamic reconfiguration can be done for several reasons (correction, adaptation, improvement); this involves the integration or deletion of system data, giving a new state for the system; this transition from one state to another gives a dynamic aspect to the system and therefore good reflection for the real system. The reconfiguration of one system will be the transition between two states.

2) The different approaches for the reconfiguration process: There are many approaches for the reconfiguration of systems due to their constraints and their constituents elements. The following paragraphs present some approaches for the reconfiguration:

a) Reconfiguration for the primary/backup system [7]: The reconfiguration in this approach is the transition between a source (S) and a target (S′) configuration, through the following steps:

- Persist information about S' on stable storage at a quorum of S.
- Deactivate S.
- Identify and transfer all committed S to S', persisting it on stable storage.
- Activate S'.

b) The reconfiguration steps to reuse embedded system resources [9]: The following steps show how to reuse embedded system resources or add embedded hardware device or add function to the middle ware:

- Analysing the embedded system resources.
- Dynamically loading them to the middle ware.
- Registering them on middle ware.

c) The reconfiguration steps for the component software approach [6]: Intuitively reconfiguration takes place as follows:

- Identify and locate the part of the system to be reconfigured.
- Suspend execution (to avoid corrupting the system).
- Change the configuration of the system (add, remove parts ...).
- Transfer the report to the new parts.
- Resume execution of the interrupted part of the system.

These three systems seem to be different, however their abstraction notice that they target the same thing (reconfigure), even with their different units of composition, but they share the same properties:

- Elemental.
- Independent.
- Connectible.
- Maintainable.

For this aim, our approach will be based on these properties to meet a large number of systems.
There are several technologies which can ensure these properties such as component programming (OSGIs) [11], web services [12], software agents, etc. all these technologies which are based on independent elementary entities, connectable and easily maintainable, except the software agents which has another feature that makes them closer to reality and allows us

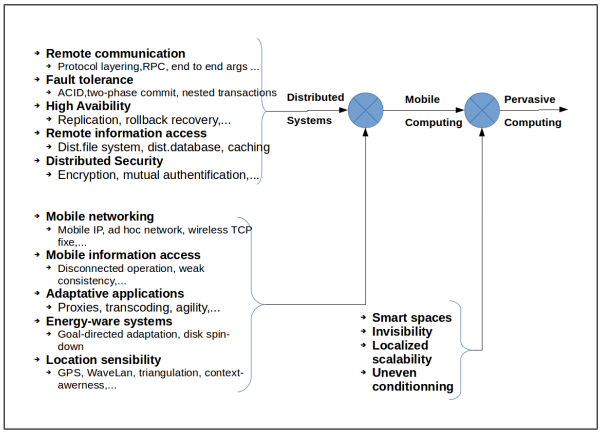


Fig. 1. Taxonomy of computer systems research problems in pervasive computing [14].

to create a copy of the real world in the virtual one (digital). This characteristic is offered by the behaviours of agent, which can effectively ensure the two main characteristics of a pervasive system (context-aware and ambient). Moreover, the software engineering for multi-agent systems can provide powerful techniques, methods and tools for the engineering of modelling and developing pervasive systems.

In the following sections, we present how we can use the software agents to perform dynamic reconfiguration of pervasive systems.

B. Pervasive System

In 1991, Mark Weiser [13] propose a new vision for computer systems; it was represented in the form of ubiquitous systems which will be developed later in pervasive systems. And his vision was defined as follow: "A new way of thinking about computers in the world, one that takes into account the natural human environment and allows the computers themselves to vanish in the background" [13].

In other words he wanted to create an environment with a computing capacity and communication. This vision has gone through various steps in order to arrive at the pervasive systems, which can be summarized as follows:

- Ubiquitously: Accessible anywhere over the network.
- Mobile: Integrate mobile terminals.
- Context-aware: Take the execution context into account.
- Ambient: Embedded in everyday objects.

So, pervasive systems [14] are therefore ubiquitous, mobile and context-aware systems as shown in Fig. 1. And the evolution of the pervasive systems depends directly on the evolution of the latter.

C. Multi-agents System

Multi-agent systems (MAS) provide abstractions that allow decomposing a system to a set of agents including distributed systems; Multi-agent systems provide flexibility for modelling more sophisticated, globally emergent behaviour; Multi-agent systems deploy services throughout the network; Multi-agent systems by their nature are powerful tools for modelling complex systems.

Moreover, multi-agent system can be easily reconfigured by addition or deleting agents from the platform, without any influence on the rest of the platform (the other active agents), which allow us to deploy a new agent with new services any time, this gives the global system new features without the need to shut down or restart the system. A second platform can be deployed to migrate or duplicate all agents for a possible maintenance or update of the main platform, in order to insure the non-stop feature of the system.

Multi-agent systems can be developed by JACK [15], JADE [16], JADEX [17], etc. And the interaction between the agents of the various platforms is possible through the standards imposed by the FIPA [18], for this reason, the use of agents for dynamic reconfiguration approach does not problem of heterogeneity because the communication is standardized by the FIPA-ACL language [18], hence virtualization of MAS development technologies. This gives developers the freedom to choose their own development platform while ensuring interaction with other agents in the system, even if they are developed by different technologies. Therefore, the possibility to integrate new subsystems in the overall system.

In MAS, agent can be a smallest independent entity of the system for the following reasons:

- Each agent can be deployed independently.
- Each agent has his own behaviour.
- Each agent offers one or more services.
- Each agent uses the FIPA-ACL standard to communicate with other agents, whatever their development platform.

So the dynamic reconfiguration of the system based agent can be:

- Adding a new agent.
- Deleting an agent.
- The deployment of the services offered by the agent.

III. Proposed Approach

LPWAN is one of the systems which they need to reconfigure dynamically their pervasive system, and since they cover a large area, a large number of connected terminals will be present on the network, and each terminal represent a sub-System and offer several types of information in order to design a real world projection in the virtual world. This information can be a temperature, a pressure, a humidity rate, a movement, a speed of an object, etc.

The dynamic reconfiguration of the LPWAN becomes a major priority when the number of terminals increases,

in particular when new terminals appear in the network, in order to deal with the full new services offered by them. This is ensured by the pervasive system layer as shown in the Fig. 2, that must manage all the information coming from the terminals (sensors) and automatically reconfigured whenever a new terminal appears in the environment, for this an automatic mechanism is necessary to perform the reconfiguration of the global network.

Multi-agent systems can handle the complexity of solutions through decomposition, modelling and organizing the interrelationships between terminals. That was when we have modelled the terminal by an agent which has an access to the information offered by the terminal sensor, also it can publish this information as services to other agents of the platform.

Therefore, we propose a new approach to perform the dynamic reconfiguration which is based on the following agents:

- Terminal agent: It is implemented at the terminal which has as tasks:
  - Discover the environment of the terminal and signal its presence in the network.
  - Publish the services offered by the terminal.
  - Make services network more accessible and easily taken in.
- Receptionist Agent: It is implemented at the main platform which has as tasks:
  - Receive the new terminal agents, and make their services exploitable across the platform by polling the new agent using a messaging system.
  - Reconfigure the platform to make these new services visible to users.
- Yellow Pages Agent: It is implemented at the platform and which register all the services offered by the other agents, in the case of the JADE technology this agent is already implemented and known under the name The Directory Facilitator (df) [19].

Based on these three types of agent we proposed the following approach for the dynamic reconfiguration:

- The terminal agent automatically searches and locates the main platform.
- The terminal agent automatically connects to the platform.
- The terminal agent publish its services at the “df” agent.
- The “df” agent inform the receptionist agent for the arrival of a new terminal agent and its services.
- The receptionist agent automatically reconfigures the platform in order to take into account the services offered by the new terminal agent.

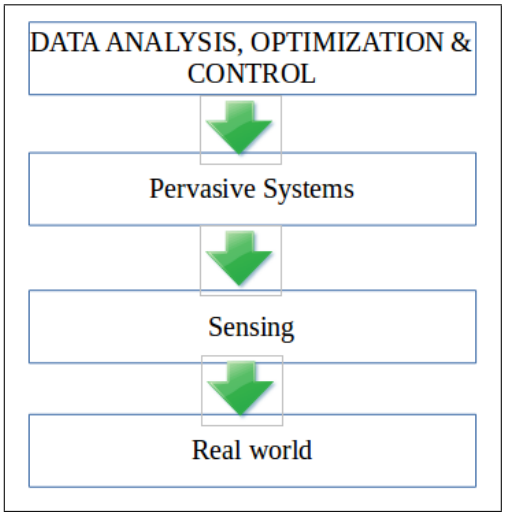


Fig. 2. The pervasive system layer in the data acquisition process.

IV. Implementation and Results

In the present approach, our solution is divided into two parts. The first one is implemented in the pervasive system layer of the LPWAN network (main platform) in order to get a better control of all the terminals that appears in the network, the second one is implemented in the terminal where will develop an agent (terminal agent) with the main behaviour is to locate the main platform in the network and publish its services.

Our solution is developed in JAVA using the JADE multi-agent systems development platform, the choice is made for the following reasons:

- JAVA: To ensure the portability of the solution on several kinds of terminals.
- JADE: It is a platform developed under JAVA, which implements yellow page agent which is called “df”.

The proposed solution is a kind of coordination between the receptionist agent of the main platform and the terminal agent; this is possible due to the different messages defined by the FIPA-ACL standard. These two agents plus the yellow page agent which is already defined in the JADE platform are cooperating in order to reconfigure the LPWAN pervasive system layer, by applying the proposed approach in the previous section and according to the scenario illustrated by Fig. 3:

A. Terminal Agent

This agent represents a node in the LPWAN network, physically it is a simple embedded system equipped with a sensor, an antenna to ensure the connection and a reduced operating system in which we will implement the terminal agent with access to the information provided by the sensor. The functioning of this terminal should consider the following steps:

- The connection to the network and obtaining IP address: This is possible because of the service offered by DHCP.

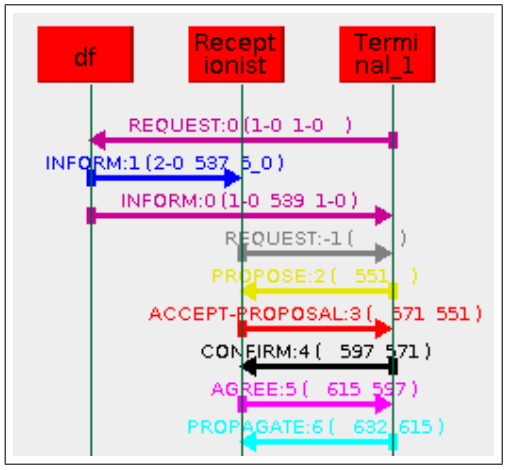


Fig. 3. Scenario for agent-based dynamic reconfiguration.

- Location and connection to the main platform.
- Services publication for the agent "df" of the main platform.
- Communication of the captured data by the sensor.

B. Receptionist Agent

This agent is implemented at the main platform, its main task is receiving the new terminal agents, this is possible through the "df" agent which will inform him with an ACL message (INFORM) every time an agent published its services in the platform, after the receptionist agent will interrogate the new terminal agent to get the different data captured by the latter and makes them accessible through the graphical interface included in the main platform.

C. Agents Communication

As shown in Fig. 3, the dynamic reconfiguration process is triggered by the arrival of a new terminal in the network, after that, a series of FIPA-ACL messages will be exchanged between the different agents defined by the approach according the following detailed scenario:

- The terminal agent sends a REQUEST message to the "df" to inform him of his presence and asks him to register his services in the platform.
- The agent "df" sends an INFORM message to the terminal agent to inform him that his request was accepted.
- The agent "df" sends another INFORM message to the receptionist agent to inform him of the presence of a new terminal agent and also communicates his services to him.
- The receptionist agent sends a REQUEST message to the terminal agent asking for the information offered by the services offered.
- The terminal agent sends a PROPOSE message to the receptionist agent for proposing the data format to send.
- The receptionist agent configures the main platform to receive the data from the new terminal agent and sends an ACCEPT-PROPOSAL message to the terminal agent informing him that his proposal was accepted.
- The terminal agent sends a CONFIRM message to the receptionist agent to indicate the start of sending data.
- The receptionist sends an AGREE message to the terminal agent to confirm that he is ready to receive the data.
- The terminal agent sends a PROPAGATE message to the receptionist agent to inform him of the sending of the first data and so on.

When the terminal agent disconnects the JADE platform automatically detects it and informs all agents through an INFORM message sent by The Agent Management System (AMS) [19], this allows the receptionist agent to reconfigure the system for the absence and breaking of data from the disconnected terminal agent.

V. Discussions

Comparing our approach to previous approaches such as primary/backup system, component software or reuse embedded system resources; we will notice that the main difference is how an element of the system is detected. In our approach, it is up to the terminal agent to discover its environment and look for the main platform to publish its services. However, the other approaches leave this task for the global system, which further complicates the search and the discovery of all subsystems present in environment, particularly where their number becomes important.

Knowing that the agent is an independent entity which means that the addition or deletion of an agent have no influence on the overall functioning of the system. This implies a continuity of operation of the system without the services provided by the agent or the disconnected terminal. therefore, the number of bugs in the application is minimized and the no-stop of the system is ensured, because the addition or deletion of the terminal agent do not need to restart the system.

After this comparison, we can say that the proposed approach allows us to develop an application with two main functionalities:

- Accept new services offered by terminal agents.
- Reconfigure the main platform (application) to make new services used by user.

These two features allow the application to expand and update its services with new services offered by all terminals agent, which is a remarkable addition to the LPWAN pervasive system.

VI. Conclusion and Future Scope

The application of dynamic reconfiguration to the LPWAN pervasive system layer offers more flexibility to the network

and a better perception of its environment since the terminals represent a set of sensors and a better adaptation the various changes by adding or removing a terminal. Our approach ensures the following characteristics:

- Flexibility.
- Independence.
- No-stop.
- Modularity.

These four characteristics can be deployed in several kind of systems such as sensor networks, smart homes or smart cities, the cyber physical systems [20] and any telemetric application where the global system has a pervasive system layer with several independent components.

Therefore, the future scope of the current proposed approach can be developed as a Framework. Then we can just use this Framework in various applications such as the addition of new devices in computer systems. In this case our approach can enable the operating system to take into account all new devices without the need to install drivers; this is possible by developing the main platform in the operating system, and terminal agent in the device that aims to publish the services offered by the hardware and how to use it. This can give a new vision for the development of the computer peripheral.

REFERENCES

[1] J. Petäjäjärvi, K. Mikhaylov, M. Hämäläinen, and J. Iinatti, "Evaluation of lora lpwan technology for remote health and wellbeing monitoring," in *Medical Information and Communication Technology (ISMICT), 2016 10th International Symposium on*. IEEE, 2016, pp. 1–5.

[2] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, "Understanding the limits of lorawan," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 34–40, 2017.

[3] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the iot and smart city scenarios," *IEEE Wireless Communications*, vol. 23, no. 5, pp. 60–67, 2016.

[4] C. A. Iglesias, M. Garijo, and J. C. González, "A survey of agent-oriented methodologies," in *International Workshop on Agent Theories, Architectures, and Languages*. Springer, 1998, pp. 317–330.

[5] W. Li, T. Logenthiran, and W. Woo, "Intelligent multi-agent system for smart home energy management," in *Innovative Smart Grid Technologies-Asia (ISGT ASIA), 2015 IEEE*. IEEE, 2015, pp. 1–6.

[6] J. POLAKOVIC, "Architecture logicielle et outils pour systmes dexploitation reconfigurables," Ph.D. dissertation, INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, 2011.

[7] A. Shraer, B. Reed, D. Malkhi, and F. P. Junqueira, "Dynamic reconfiguration of primary/backup clusters." in *USENIX Annual Technical Conference*, 2012, pp. 425–437.

[8] W.-T. Tsai, W. Song, R. Paul, Z. Cao, and H. Huang, "Services-oriented dynamic reconfiguration framework for dependable distributed computing," in *COMPSAC*, vol. 1, 2004, pp. 554–559.

[9] S. H. Moon and C. sick Lee, "Dynamic management software design in embedded system using middle," 2014.

[10] N. S. Voros and K. Masselos, *System level design of reconfigurable systems-on-chip*. Springer, 2005.

[11] A. Taherkordi, P. Herrmann, J. O. Blech, and A. Fernandez, "Service virtualization for self-adaptation in mobile cyber-physical systems," in *International Conference on Service-Oriented Computing*. Springer, 2016, pp. 56–68.

[12] H. Gao and H. Miao, "A quantitative model-based selection of web service reconfiguration," *2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 365–371, 2013.

[13] M. Weiser, "The computer for the 21 st century," *Scientific american*, vol. 265, no. 3, pp. 94–105, 1991.

[14] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Personal communications*, vol. 8, no. 4, pp. 10–17, 2001.

[15] M. Winikoff, "Jack intelligent agents: an industrial strength platform," in *Multi-Agent Programming*. Springer, 2005, pp. 175–193.

[16] F. L. Bellifemine, G. Caire, and D. Greenwood, *Developing multi-agent systems with JADE*. John Wiley & Sons, 2007, vol. 7.

[17] A. Pokahr, L. Braubach, and W. Lamersdorf, "Jadex: A bdi reasoning engine," in *Multi-agent programming*. Springer, 2005, pp. 149–174.

[18] A. Fipa, "Fipa acl message structure specification," *Foundation for Intelligent Physical Agents, http://www. fipa. org/specs/fipa00061/SC00061G. html (30.6. 2004)*, 2002.

[19] F. Bellifemine, A. Poggi, and G. Rimassa, "Jade–a fipa-compliant agent framework," in *Proceedings of PAAM*, vol. 99, no. 97-108. London, 1999, p. 33.

[20] S. Zanero, "Cyber-physical systems," *Computer*, vol. 50, no. 4, pp. 14–16, 2017.

Impact of Thyristor Controlled Series Capacitor on Voltage Profile of Transmission Lines using PSAT

Babar Noor[1], Muhammad Aamir Aman[1], Murad Ali[1], Sanaullah Ahmad[1], Fazal Wahab Karam.[2]

Electrical Engineering Department, IQRA National University, Peshawar, Pakistan[1],
Electrical Engineering Department, CIIT Abbottabad, Pakistan[2]

***Abstract*—In power system voltage stability is very important in order to maintain the voltage within the defined limits. The demand of electrical power increases in the last decade due to the lack of expansion in the generation and transmission network. The available transmission network is heavily loaded. This loading of transmission network cause the voltage instability. This instability on heavy loaded system creates voltage collapse which causes the power loss. Due to this phenomenon it is necessary to keep monitoring the voltage instability and to reduce voltage collapse. Voltage collapse mostly occurs due to the minimum availability of the reactive power. A power electronic device i.e. Flexible AC Transmission System (FACTS) is used to add the reactive power to the system which improve the voltage profile and minimizes the chances due to which the voltage collapse occur. In this paper FACTS series compensator, i.e. Thyristor Controlled Series Capacitors (TCSC) is injected between the two nodes of the IEEE 6 bus bar test system to check the voltage profile. PSAT (Power System Analysis Tool) tool, which is a new tool in MATLAB for study in power system analysis is used. IEEE-6 bus system is used as a test system for the effectiveness of the proposed method. The voltage profile with and without TCSC device is then compared to conclude the result.**

***Keywords*—Flexible AC Transmission System (FACTS); Thyristor Controlled Series Capacitors (TCSC); Power System Analysis Tool (PSAT) tool; voltage profile; voltage collapse**

I. Introduction

Power system is a vast and complex system in which large amount of generating units and substation are connected. The demand of power is increasing day by day and the system is overloaded because the system is already operated at its maximum capacity [1]. If contingency occur then the situation is not safe. These contingency or faults leads the system to voltage instability. Voltage instability happens in the system i.e. even a single disturbance fall the voltage down or rise to certain limits and at the end black outs and voltage collapse in the system occurs [2]. There is a gap between reactive power generation and demand, which causes problems in the voltage instability. Now a days two important factors are discussed in power transmission system, load and power quality. These relates to sudden increase in power demand to maintain the security of the system and a nonlinear load equipment such as machines and different electronics equipment's which are connected to transmission system. Due to the variation in the voltage profile, demand and generation causes voltage instability. This instability is due to the reactive power, which is unstable. Reactive power demand cannot be controlled because the load varies. To expand the transmission system it requires time, investment and right of way. Alternatively, FACTS device will be install instead of installing new transmission lines or replacement of transmission lines [3]. These devices are made from semiconductors, which are an ideal switch. FACTS devices are power electronic devices that are used to maintain operation of power system in a stable condition [4]. FACTS devices are used to solve many problems in power network and ensure secure operation of the power system. System flexibility is also provided by these devices. These devices improve the apparent power of the system that helps to enhance system operation and control over the system. SVC (Static Var Compensator), TCSC (Thyristor Controlled Series Capacitor), UPFC (Unified Power Flow Controller) and Static compensator, etc. are the several kinds of FACTS devices [5]. It is necessary to inject proper size FACTS device at optimal location because it causes voltage drop and too much heating due to huge amount of reactive power [2].

To prevent the system from voltage collapse, different methods for load flow analysis such as steady state analysis, continuation power flow method (CPF), optimal power flow (OPF), etc. are used. Reactive power is very important in power system consisting of different transformers, circuit breakers, transmission, distribution lines and different types of loads having different power factor. Due to their intrinsic characteristics these devices shift the current and voltage measures in VARs. Due to change in the reactive power volt-ampere reactive (VAR) variation in the voltage profile occurs. If high amount of VARs injected to the system it increases losses in the system and decrease power transfer capabilities and when amount of VARs are low so it leads to voltage sag. To make a system in a stable position an appropriate amount of reactive power will be inject in a system. The compensation of reactive power improves voltage profile, transmission capability and control power flow to operate the system with flexibility [6].

This paper comprises of six sections. In first section the compensation of reactive power is explained. In second section FACT device is implemented in the system to check the disturbance in power flow mathematically. Third section explains TCSC FACT device and this device is implemented on the IEEE 6-bus test system to check the voltage profile. In the fourth section comparison is done with and without FACT device. Result and future work is explained in fifth and sixth section.

II. Compensation of Reactive Power

The power network have a variable sending and receiving voltage due to the presence of different generation and utility appliances. These appliances have different magnitude and phase angles. If we want the voltage magnitude and phase at desire range between sender and receiver point or in stable position we have to compensate the system voltage. In order to compensate the system, different methods are used. For example, a capacitive load is add in parallel to produce reactive power that decrease the current drop in the line and voltage is improved. The second method to compensate a system is to install a compensator in the system that is use to mitigate voltage variations by controlling the reactive power automatically. These devices has the ability to inject or absorbed reactive power according to line requirement [7]. The equation of power flow in transmission line is shown as

$$P = \frac{V_s V_R}{X} \sin \delta \tag{1}$$

Where

V_s and V_R is sending and receiving voltages,

X is the impedance and δ is the power angle.

Now to increase power factor and reduce line impedance [7]. This power factor is increased by injecting the reactive power.

III. FACTS Devices

FACTS devices are used to improve power transfer capabilities and to improve system performance. These are solid-state devices consist of thyristor which do not contain any mechanical parts. It can be turn off and on with receiving signals. Different FACTS devices have multiple characteristics that are impedance, voltage and dynamic control of angle. The devices have very efficient control and quick response. Fig. 1 shows FACTS device injected between the sending and receiving end [7].

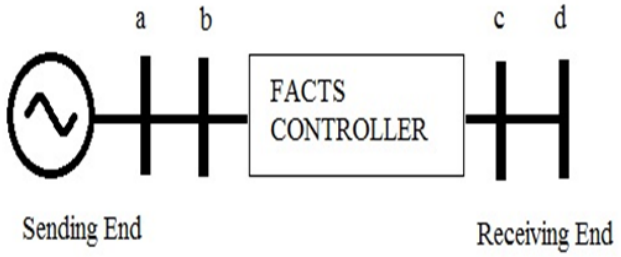


Fig. 1. FACTS device injected between the sending and receiving end [7].

IV. FACTS Device and Disturbance in Power Flow

Power flow and power quality is affected by the disturbance in the power system. The disturbances are usually voltage variation, transient instability and over voltages. These problems lead to overload the system and create power losses. Due to rapid increase in the population, the power demand is also increased. To cover the demand and supply gap, the generation must have to be increased but the capacity of existing transmission lines cannot be exceeded. As these transmission lines are overloaded and results in voltage reduction [8]. The possible solution of the problem is to extend the transmission lines capacity or to install new transmission lines to reduce the load on the existing transmission lines but the cost of transmission lines is very high. The alternative solution is to install FACTS device in our transmission system to prevent it from overloading condition and voltage collapse which causes the blackouts. This will also enhance the power quality, voltage stability and power transfer capability [3].

V. Thyristor Controlled Series Capacitor (TCSC)

Thyristor Controlled Series Capacitor (TCSC) is a FACTS device connected in series on the transmission lines to control the dynamic power flow and enhance the power quality. It consists of thyristor, reactor and capacitors in which the thyristor is connected with a reactor that is in parallel to the capacitor along with the transmission line. TCSC control reactance due to reactors with shunt connected capacitor. This will increases the efficiency of transmission line. The capacitive reactance of the controller can be varied by varying the firing angle of the Thyristor Controlled Reactor (TCR). The load ability of the system is increased by decreasing the line effective reactance due to the capacitive reactance connected in series with the line reactance [7]. TCSC having a LC circuit carry constant current as shown in Fig. 2. The impedance is changing from maximum to minimum and increase the minimum capacitance while the arrangement of TCSC is such that the reactor X_C is greater than the X_L [9].

$$X_L(\alpha) = X_L \frac{\pi}{\pi - 2\alpha - \sin 2\alpha} \tag{2}$$

Here

X_L = Reactance of the Reactor

α = Firing Angle

TCSC is connected between the transmission line nodes “i” and “j” as shown in Fig. 3.

From power flow equations, the reactive and real power flow between nodes “i” and “j” is given as:

$$Pij^n = Vi^2 Gij^n - ViVj(Gij^n \cos \delta_{ij} + Bij^n \sin \delta_{ij} \tag{3}$$

$$Qij^n = -Vi^2(Bij^n + B_{sh}) - V_i V_j(Gij^n \sin \delta_{ij} - Bij^n \cos \delta_{ij} \tag{4}$$

And the flow of power between node j to node i is

$$Pji^n = Vi^2 Gij^n - ViVj(Gij^n \cos \delta_{ij} + Bij^n \sin \delta_{ij} \tag{5}$$

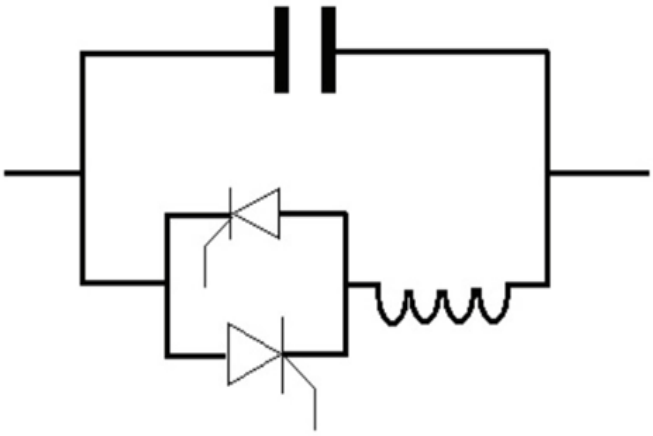


Fig. 2. Thyristor Controlled Series Capacitor (TCSC) [7].

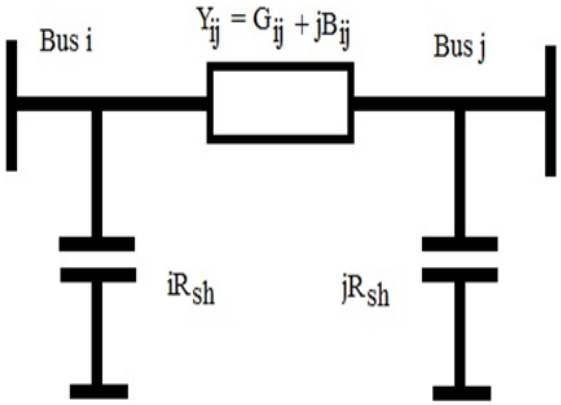


Fig. 3. Line representing between two buses [5].

$$Qji^n = -Vi^2(Bij^n + B_{sh}) - V_iV_j(Gij^n \sin\delta_{ij} - Bij^n \cos\delta_{ij}[5] \quad (6)$$

Where, G is the conductance and B is the line susceptance.

VI. TCSC Optimal Location

A. Sensitivity Factors

Sensitive lines are identified by sensitivity index based techniques to decrease reactive and active power loss and to improve voltage profile and stability by the placement of FACTS devices [5].

B. Line Voltage Stability Index (LVSI)

Voltage stability levels were predicted by using this LVSI values. These values will determine the weak transmission lines. Then according to their requirements these weak lines were provided proper amount of reactive power [5]. The line between the nodes "i" and "j" in the system is used to find the stability index of voltage.

LVSI is measured by the equation.

$$L_{ij} - \frac{4XQ}{(V_i \sin(\theta - \delta_{ij}))^2} \quad (7)$$

Where,

L_{ij} represent the transmission line voltage stability index between node-i and node-j

Q is the receiving end reactive power

X is the line reactance

θ is the line impedance

V_i shows the sending end voltage

δ_{ij} = δ_i - δ_j, δ_i and δ_j are the sending and receiving end angles of voltages.

When the line is stable, the value of line voltage stability index is always less than "l" [5].

VII. Cases

The IEEE-6 standard bus system is considered as a test system with power base of 100 MVA and frequency of 50 Hz. The analysis of power flow is done by using Power system Analysis Toolbox (PSAT) which is a new toolbox of power system analysis in MATLAB. Newton Raphson method is used to determine the result. TCSC FACTs device will be install in the system between node 1 and 6 and then analyze the system. The comparison of both the results, i.e. with and without FACTS device will be investigated and impact on voltage profile will be analyzed. Fig. 4 shows IEEE-6 bus system modeled in PSAT without TCSC controller. Fig. 5 shows the TCSC FACTS device that is installed between bus 1 and 6 in IEEE-6 bus test system.

The load flow analysis of the system is analyzed. This analysis will show the voltage level at each bus and power flow across the network.

A. System Parameters

IEEE-6 Bus bars System Frequency = 50 Hz Base MVA = 100 MVA Bus Voltage = 11 KV

B. Line Data

The line data is shown in Table I.

C. Bus Data

The bus data is shown in Table II.

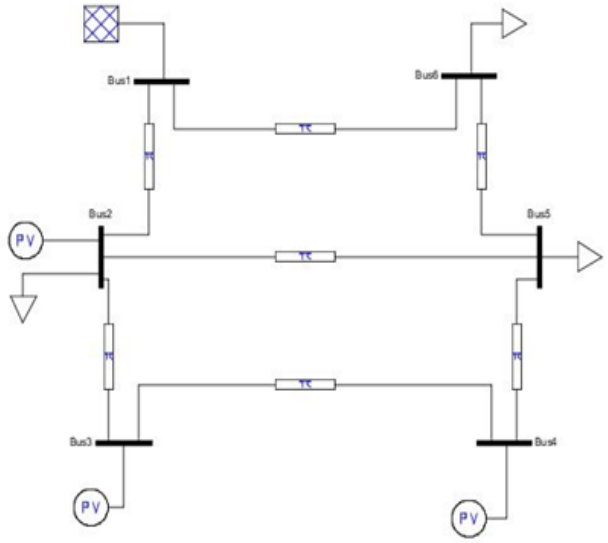


Fig. 4. IEEE 6-bus system without TCSC controller.

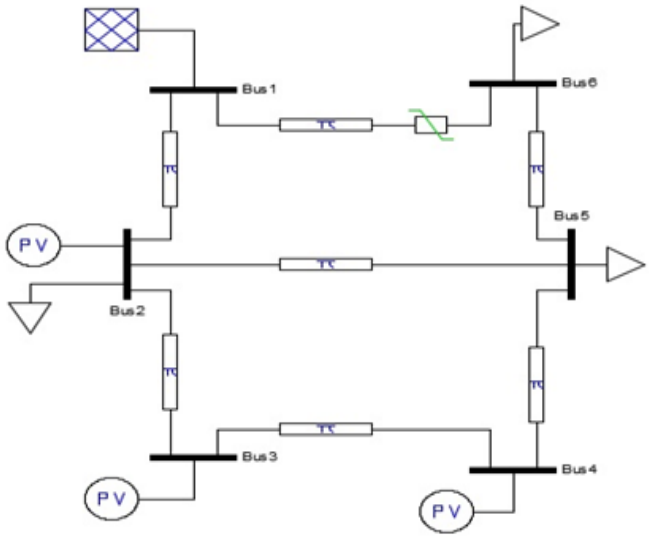


Fig. 5. IEEE 6-bus system with TCSC controller.

TABLE I. LINE DATA

Lines	Buses	Resistance P.U	Reactance P.U
1	1-2	0.05	0.20
2	2-3	0.10	0.50
3	3-4	0.20	0.80
4	4-5	0.10	0.30
5	5-6	0.20	0.40
6	6-1	0.10	0.15
7	2-5	0.20	0.50

TABLE II. GENERATION AND LOAD DATA

Buses No	Generation		Load	
	V P.u	MVA	MW	MVAR
1	1	100		
2	1.04	15	20	10
3	1.04	40		
4	1.07	30		
5			40	15
6			30	10

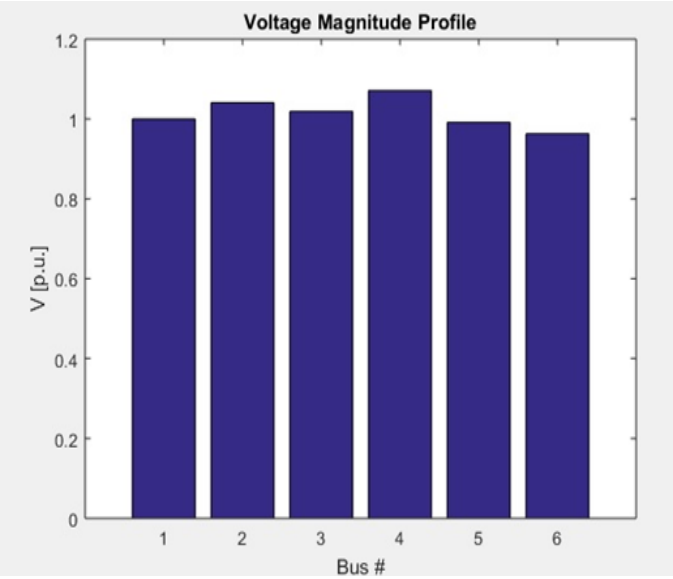


Fig. 6. Voltage profile of without TCSC controller.

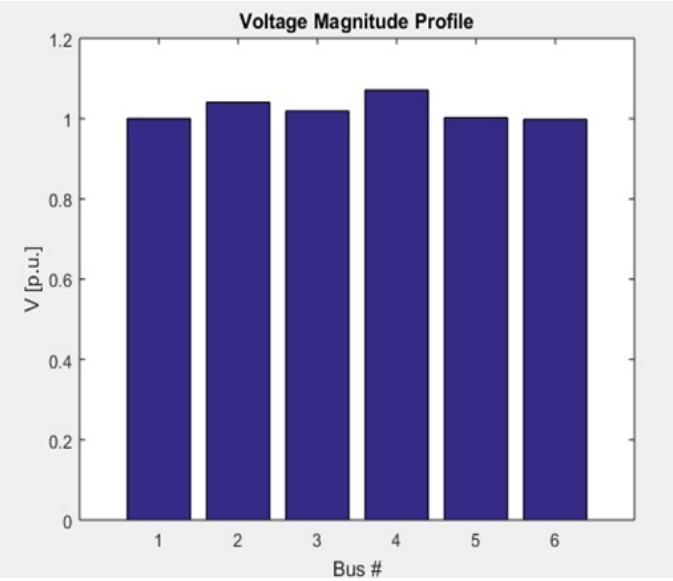


Fig. 7. Voltage profile with TCSC FACT device.

TABLE III. VOLTAGE COMPARISON WITH AND WITHOUT TCSC

	Without TCSC	With TCSC
Bus 1	1	1
Bus 2	1.041	1.041
Bus 3	1.019	1.019
Bus 4	1.071	1.071
Bus 5	0.991363	1.002456
Bus 6	0.962995	0.998168

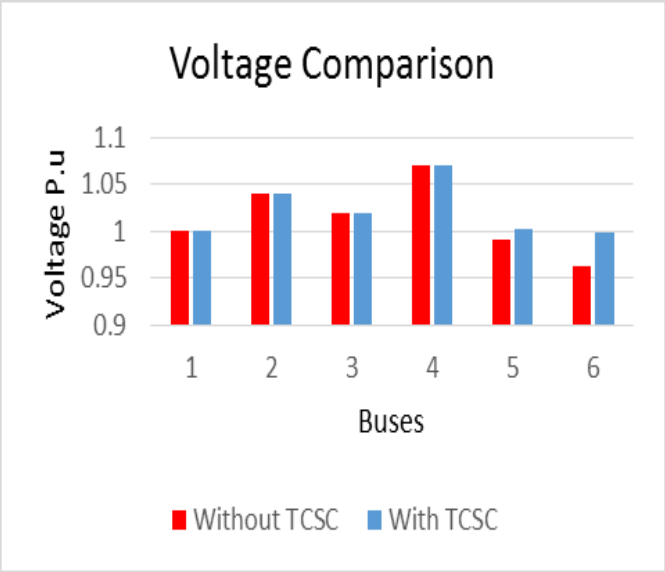


Fig. 8. Voltage comparison with and without TCSC device.

VIII. Results and Discussion

Fig. 6 shows the voltage magnitude of system without having TCSC device. Furthermore it was noticed that bus-6 has lower voltage as compare to other buses and this bus is consider as a weak bus in case of any fault or overloading. TCSC FACTS device is installed between bus 1 and 6 for improving voltage profile. Fig. 7 shows the voltage magnitude after connecting FACTS controller. It is observed that the voltage magnitude enhances from 0.962 to 0.998 which shows that by connecting TCSC device the system voltage profile is improved. Fig. 8 shows the comparison between voltage magnitude i.e. Without TCSC installation and after connecting the TCSC device in the system. This comparison is then tabulated in Table III.

IX. Conclusion and Future Work

In this paper the TCSC FACTS device is installed in the six bus bar IEEE test system and the voltage profile is analyzed using PSAT simulation tool. FACTS devices are used to improve the voltage profile of the network by injecting the reactive power to the system. Voltage level is analyzed at different buses and then identifies the weak voltage bus. The results are compared between the system without TCSC device and with TCSC device and concluded that by injecting the TCSC FACTS device between the node 1 and 6 of the IEEE-6 bus test system, the voltage level improves which directly reduces the power loss. This work can be implemented with other IEEE test systems and also with comparison to other FACTS devices to check the feasibility. Also it can be implemented in between transmission or distribution line to reduce the losses and improve voltage.

References

[1] M. A. Aman, S. Ahmad, and K. Mahmood, "Designing and strategic cost estimation of stand-alone hybrid renewable energy system," 2016.

[2] K. Hridya, V. Mini, R. Visakhan, and A. A. Kurian, "Analysis of voltage stability enhancement of a grid and loss reduction using series facts controllers," pp. 1–5, 2015.

[3] E. Yap, M. Al-Dabbagh, and P. Thum, "Applications of facts controller for improving power transmission capability," pp. 1–6, 2005.

[4] V. G. Mathad, B. F. Ronad, and S. H. Jangamshetti, "Review on comparison of facts controllers for power system stability enhancement," *International Journal of Scientific and Research Publications*, vol. 3, no. 3, pp. 2250–315, 2013.

[5] Y. Manganuri, P. Choudekar, D. Asija *et al.*, "Optimal location of tcsc using sensitivity and stability indices for reduction in losses and improving the voltage profile," pp. 1–4, 2016.

[6] J. Lakkireddy, R. Rastgoufard, I. Leevongwat, and P. Rastgoufard, "Steady state voltage stability enhancement using shunt and series facts devices," pp. 1–5, 2015.

[7] S. F. B. Shakil, N. Husain, M. D. Wasim, and S. Junaid, "Improving the voltage stability and performance of power networks using power electronics based facts controllers," pp. 1–6, 2014.

[8] M. A. Aman, S. Ahmad, A. ul Asar, and B. Noor, "Analyzing the diverse impacts of conventional distributed energy resources on distribution system," *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 8, no. 10, pp. 390–396, 2017.

[9] V. Yarlagadda, D. B. S. Ram, and D. K. Rao, "Voltage stability improvement using thyristor controlled series capacitor (tcsc) based on lmn and vcpi stability indices," *International Journal Of Scientific & Engineering Research*, vol. 3, no. 4, 2012.

Efficiency and Performance Analysis of a Sparse and Powerful Second Order SVM Based on LP and QP

Rezaul Karim, Amit Kumar Kundu
Department of Electrical and Electronic Engineering
Uttara University
Dhaka, Bangladesh

***Abstract*—Productivity analysis is done on the new algorithm "Second Order Support Vector Machine (SOSVM)", which could be thought as an offshoot of the popular SVM and based on its conventional QP version as well as the LP one. Our main goal is to produce a machine which is: 1) sparse & efficient; 2) powerful (kernel based) but not overfitted; 3) easily realizable. Experiments on benchmark data shows that to classify a new pattern, the proposed machine, SOSVM requires samples up to as little as 2.7% of original data set or 4.8% of conventional QP SVM or 48.3% of Vapnik's LP SVM, which is already sparse. Despite this heavy test cost reduction, its classification accuracy is very similar to the most powerful QP SVM while being very simple to be produced. Moreover, two new terms called "Generalization Failure Rate (GFR)" and "Machine-Accuracy-Cost (MAC)" are defined to measure generalization-deficiency and accuracy-cost of a detector, respectively and used to compare such among different machines. Results show that our machine possesses GFR up to as little as 1.4% of the QP SVM or 1.5% of Vapnik's LP SVM and MAC up to as little as 2.6% of the QP SVM or 35.9% of the Vapnik's sparse LP SVM. Finally, having only two types of parameters to tune, this machine is straight forward and cheaper to be produced compared to the most popular & state-of-the-art machines in this direction. These collectively fulfill the three key goals that the machine is built for.**

Keywords*—Generalization failure rate; Kernel machine; LP; QP; machine accuracy cost; Second Order Support Vector Machine; sparse

I. Introduction

Run time optimization of classifiers is a crucial issue for fast data classification. A prominent example is from Viola and Jones [1] on face detection based on a cascade of boosted weak classifiers. This framework is not efficiently applicable to kernel based classifiers like support vector machines (SVMs) [2], for instance, because boosting based on such strong classifiers as components is less effective. In many applications, the flexibility of such kernel machines is a real advantage. While SVM based classifiers play the leading role in pattern classification with highest accuracy, one of its key properties is that the learned classifier can be expressed in terms of only a subset of the training patterns, known as support vectors (SVs). But as the computational load of using such a classifier to classify a pattern is proportional to the number of SVs, SV sparsity is extremely important for large datasets. This is especially the case when the training is done once on powerful computers that can handle large data but the prediction is needed to be done multiple times possibly on a small low-powered devices in real time. This motivates to design kernel based classifiers maintaining the trade-off between accuracy and sparsity. Consequently, this problem has come to the center of main attention in research recently.

In this paper, we have proposed a new sparse algorithm "Second Order SVM (SOSVM)" and carried out experimental studies on it as well as standard QP SVM [2] and Vapnik's LP SVM [2] to analyze their performance & efficiency on the basis of computational cost and generalization ability. For simplicity in discussion, only Gaussian kernel is applied throughout the whole work. Standard machine learning benchmark data is used for experiment.

In Section II related works are discussed, in Section III we re-describe SVMs, in Section IV we explain our approach whereas Section V is for experiments and we use Section VI for conclusion and discussion including future work.

II. Related Works

Related work can be approximately, but not disconnectedly, classified

- into approaches [3]–[11] to the design of Reduced SVMs (RSVMs) that demand less computational loads than standard SVM for classifying a pattern;
- into approaches [7], [12]–[16] that exploit SVMs as components of a detector with structured architecture for classification
- into approaches [17]–[20] that develop SVM related cheaper classifiers, which are different from usual RSVMs;
- into approaches [21]–[31] that investigate ensemble-detector by boosting weak classifiers;
- into approaches [32]–[40] that improve one or more of the three variables cost, efficiency, & accuracy of a detector by applying different techniques on different hypothetical single classifiers using one of them or combining more of them considering un/balanced data.

Regarding the first class of approaches, RSVMs demand only a fraction of kernel evaluations to classify a pattern. Wavelet approximations of these latter vectors have also been investigated in [6] for an efficient evaluation of the arguments to which the kernel function is applied. However, while [4], [9], [11] proposed some smart iterative algorithms for reduced

SVMs with impressive results, [9] reported a memory run out from [4], [11] in case of their implementations on large dataset whereas [9] has a considerable practical variation with heavy parameter selection from its defined approach. The second class of approaches, in contrast to the first one, is focusing on structured SVM-based classification for pattern detection. Heisele et al. [13] studied a hierarchy of linear SVMs with a single nonlinear SVM at the end. Thresholds were tuned for optimizing classification performance and speed, followed by feature selection. Romdhani et al. [7] proposed a single chain of SVMs that is optimized also by threshold tuning, and by approximating a fully nonlinear SVM that has to be computed beforehand, whereas a decision tree with linear SVMs is suggested in [12]. Sahbi and Geman [14] presented a tree-structured hierarchy of SVMs that is optimized by reduced set technique in [7] and threshold selection, and operates on application specific partitioning of the space of patterns following different poses. Huo Chen [15] talked about numerical strategies for optimal cascade and checked three heuristics on synthetic data using binary SVM on each stage of a cascade. However, the third class of approaches, being a bit correlated to the second one as originated from the SVM principle, has reasonable discrepancy from that as well from the structural point of view. Maji et al. [17], [20], [36] showed that SVMs using histogram as well as additive kernels are faster and outperform linear SVM. Ladicky - Torr [18] proposed a novel locally linear SVM classifier with smooth decision boundary and bounded curvature while suggesting a trade-off the number of anchor points against the expressivity of the classifier in order to avoid overfitting and speed problem. Xu et al. [19] introduced a post-processing algorithm that compresses the learned SVM by further training on the SVs with adding few extra training parameters. Enthusiastically, the fourth class of approaches has a bit similarity to the second one from the construction principle as they both use a cascade like approach. Xiao et al. [21] used an idea named "Dynamic Cascade" as Face detector that is trained on large data set by dividing them into subsets and hence working on them while using "Bayesian Stump" as weak learners for boosting. Luo H. [22] designed optimization for cascaded classifier that finds the optimum thresholds of different stages for a fixed set up. Saberian et al. [23] introduced a mathematical model for a cascaded detector relating classification and complexity. Chen et al. [24] proposed an algorithm for a cascaded detector considering operational cost, accuracy, and feature extraction cost. Chen et al. [25] presented a general cascade framework that unifies detection learning and alignment for face detection. Li Zhang [26] offered a fast cascaded object detector having fewer stages and using logistic regression as weak learner, which emphasize on training efficiency. Raykar et al. [27] proposed a soft cascade where classifiers accept/reject patterns following probability distributions induced by the earlier stages' classifiers. Considering a fixed order of different stages in the cascade, they tried to find a trade off between accuracy and feature acquisition cost. Visentini et al. [28] devised an algorithm that dynamically builds a cascade of classifiers to speed-up the Online Boosting technique. The cascade explicitly considers the computational cost of the involved features to maintain real-time performance while its classifiers are automatically in tune balancing speed and accuracy. Saberian et al. [29] suggested a cascade boosting algorithm, fast cascade boosting (FCBoost) that minimizes Lagrangian risk while considering speed and accuracy. They introduced the concept of "neutral predictors" that robotically determines the cascade configuration such as number of cascade stage and number of weak learners in each stage. Xu et al. [30] offered a tree of classifiers to balance the test cost and accuracy while Xu et al. [31] analyzed the trade-off problem considering one more variable, feature orientation cost. At last, interestingly, the fifth class of approaches is quite diverge. Fu et al. [32] discussed a problem of combining linear SVMs to classify non-linear data set and claimed experimental results showing that their method can achieve the efficiency of LSVMs in the prediction phase while providing a classification performance comparable to nonlinear SVMs. Cheng- Jhan [33] proposed a pedestrian detector by cascading AdaBoost and SVM classifiers in different stages. A classifier for digit recognition was proposed by Maji et al. [34] that poses reduced operational cost with improved features. It also claimed to have the best result in all three aspects like accuracy, train-cost, and test-cost while using histogram-gradient features and intersection kernel SVM. Gu - Han [35] designed a Clustered Support Vector Machine (CSVM), by weighted combination of linear SVMs (LSVM) trained on the clustered subsets of the training data to separate the data locally. These combined LSVMs are regularized globally to leverage the inter cluster information and avoid over-fitting in each cluster. They derive a data-dependent generalization error bound for CSVM, which explains the advantage of CSVM over linear SVM. Sharma et al. [37] offered an approach for learning non-linear SVM at reduced computational cost in the test phase and empirically analyzed the tradeoff between encoder and classifier complexity and strength. Osadchy et al. [38] proposed a so called hybrid classifier to tackle the problem with data set having high asymmetry as the large portion of the pattern space belongs to the negative class; their kernel hybrid classifier is for further efficiency than SVM while having similar accuracy [39]. Vedaldi et al. [40] offered a three-stage classifier combining linear, quasi-linear, and non-linear kernel SVMs. They showed that increasing the non-linearity of the kernels increases their discriminative power at the cost of an increased computational complexity. Nevertheless, their three stage cascade to overcome the complexity cost has resulted in quite a 'heavy' algorithm in both training and testing.

III. Support Vector Machine (SVM)

Support Vector Machines (SVMs) is a state-of-the-art and popular machine learning technique that has been confirmed as a very powerful tool for Supervised Classification. In this part, we re-describe SVM with its two main variants; one is the standard & most common method using the quadratic programming (QP), we call it QPSVM, while the other one is the Vapnik's linear programming SVM, we call it VLPSVM. We also make a mild comparison between these two.

A. Quadratic Programming SVM (QPSVM)

Here, we briefly review the basic learning algorithm of the QP based Support Vector Machine (SVM) using margin maximization between two classes, which consists in finding the separating hyperplane that is furthest from the closest object; a detailed introduction could be found in [2].

Consider a binary classification problem of dataset where a set of training patterns $\{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^d$ and

$y_i \in \{-1, 1\}$ is given. As the objective of the SVM algorithm is to find the optimal separating hyperplane that skillfully separates these patterns into two classes, it offers a classifier using a decision function (for the input pattern x) of the form $f(x) = w \cdot \phi(x) + b$ leading to $class(x) = sgn(f(x))$, where $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ is a kernel function and the parameters w and b are found from a series of calculations starting from the following QP problem:

$$\min_{w,b,\zeta} f_P(w) \;=\; \frac{1}{2}\|w\|^2 \;+\; C\sum_{i=1}^{N}\zeta_i \tag{1}$$

$$s.t.\;\; y_i\big(w \cdot \phi(x_i) + b\big) \;\geq\; 1 - \zeta_i \tag{2}$$

$$\zeta_i \;\geq\; 0;\; i = 1, 2, ..., N \tag{3}$$

Where the set of constraints (2) implies that the decision function should classify correctly all patterns from the given training set up to some tolerable errors, the slack variables $\zeta_i > 0$ hold for margin-outward-deviated patterns (that is, patterns staying outwards from their class-margins) and $C > 0$ is a parameter of the classifier that controls the trade off between two main goals of the objective function in (1): one is to optimally maximize the margin between the two classes and another is to minimize the number of misclassifcations on the training patterns.

Common practice to realize a solution for this problem is to solve its dual problem, developed by introducing a Lagrangian and the Lagrangian of the problem form (1)-(3) is

$$L_P(w, b, \zeta, \alpha, \gamma) = \frac{1}{2}\|w\|^2 \;+ C\sum_{i=1}^{N}\zeta_i$$

$$-\sum_{i=1}^{N}\alpha_i\Big(y_i\big(w \cdot \phi(x_i) + b\big) \;- 1 + \zeta_i\Big) \;-\sum_{i=1}^{N}\gamma_i\zeta_i \tag{4}$$

$$\alpha_i, \gamma_i \;\geq\; 0;\; i = 1, 2, ..., N \tag{5}$$

where α_i are Lagrange multipliers and we get the corresponding dual problem as

$$\max_{\alpha} f_D(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j) + \sum_{i=1}^{N}\alpha_i \tag{6}$$

$$s.t \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{7}$$

$$0 \leq \alpha_i \leq C; \quad i = 1, 2, ..., N \tag{8}$$

This is also a QP problem and optimum values of its variable α are used to find the primal variables w, b as $w = \sum \alpha_i y_i \phi(x_i)$ and $b = y_s - w \cdot \phi(x_s)$ where s is an index of any pattern for which $0 < \alpha_s < C$. One of the KKT conditions for the problem (1)-(3) is $\alpha_i\Big(y_i\big(w \cdot \phi(x_i) + b\big) \;- 1 + \zeta_i\Big) = 0$ from which for $\alpha_i \neq 0$, we get $y_i\big(w \cdot \phi(x_i) + b\big) \;- 1 + \zeta_i = 0$. These patterns having $\alpha_i > 0$ are support vectors (SVs), which are usually far less (depending on the data set) in number compared to the total training set size that proves QPSVM to be sparse. From these SVs, $\alpha_i < C$ (are those having $\zeta_i = 0$) patterns lie on the margin of own class whereas $\alpha_i = C$ (are those having $\zeta_i > 0$) patterns stay outwards from their corresponding margins. Interestingly, the constraint (7) that is $\sum_{i=1}^{N}\alpha_i y_i = 0$ ensures that in this QPSVM, SVs set must have members from both classes. In QPSVM model, SVs are the only training patterns that contribute in designing an optimal classifier.

B. Vapnik's LP SVM (VLPSVM)

Here, we concisely go through the linear programming approach proposed by Vapnik to find a separating hyperplane that is very similar to that of the QPSVM one but demands comparatively less computation to classify a pattern. More elaborate could be found in [2].

Inferring that the classifier has the same form of kernel expansion using the SVs in the QPSVM, Vapnik used a linear objective function to minimize the sum of all the coefficients used in the kernel expansion. Each coefficient is associated with its corresponding KCV (kernel computing vector) in the expansion. For better clarification, we name these vectors of this machine as "Expansion Vector (EV)", which is similar to SVs in QPSVM.

Considering that the decision function preserves exactly the same form of kernel expansion as the QPSVM and the error constraints of the QPSVM also remain almost the same, Vapnik proposed this VLPSVM focusing at minimizing the number of kernel computation by reducing EVs of the separating hyperplane that has the weight vector w_V of the decision function $f_V(x) = w_V \cdot \phi(x) + b_V$ leading to $class(x) = sgn(f_V(x))$. For this purpose, he formed a linear objective function using the coefficients of the EVs directly and coupling the error penalty on top of the error constraints as below:

$$\min_{\lambda,\xi,b_V} \quad \sum_{i=1}^{N}\lambda_i \;+\; C_V\sum_{i=1}^{N}\xi_i \tag{9}$$

$$s.t.\;\; y_i\Big(\sum_{j=1}^{N}\lambda_j y_j \phi(x_j) \cdot \phi(x_i) + b_V\Big) \;\geq\; 1 - \xi_i \tag{10}$$

$$\lambda_j \;\geq\; 0;\; j = 1, 2, ..., N \tag{11}$$

$$\xi_i \;\geq\; 0;\; i = 1, 2, ..., N \tag{12}$$

Alike the QPSVM, the set of constraints (10) implies that the decision function should classify correctly all patterns from the given training set up to some tolerable errors, the slack variables $\xi_i > 0$ hold for absolute-unity-outward-deviated patterns (that is, training patterns having $ClassLabel(x_i) \cdot f_v(x_i) < 1$) and $C_V > 0$ is a parameter of the classifier that controls the trade off between data learning and overfitting.

For this machine, the bias term, b_V of the decision function is an optimization variable of the main problem ((9)-(12)) and found along with other optimization variables λ and ξ. The optimum values of λ are used to find the weight vector w_V as $w_V = \sum \lambda_j y_j \phi(x_j)$. Training patterns, x_j with coefficients $\lambda_j > 0$ are the EVs, which are usually very much smaller (depending on the data set) in number compared to the total training set size showing VLPSVM to be sparser. Unlike QPSVM, here we have no constraints that forces the machine to have KCV from both classes.

C. Comparison between VLPSVM & QPSVM

Unlike QPSVM, we directly implement the primal to be optimized in case of VLPSVM and get the optimal value of the bias along with the optimal co-efficient set of the expansion vector. Interestingly, in case of this machine, values of the absolute-unity-outward-deviation parameters ξ_i are also found as a subset of the optimum variable by solving the LP.

Although QPSVM & VLPSVM do not have the same constraint set fully, part of the constraints they use are almost the same. For example, they use nearly the same error constraint & the non-negative lower bounds of the variables (KCV-expansion co-efficients α, λ, margin/absolute-unity outward deviation ζ, ξ).

Both machines are heavily influenced by the error penalty parameter (on top of kernel parameter) where the QPSVM introduces this parameter, C through the constraint but the VLPSVM uses this parameter, C_V by directly optimizing the primal cost function that includes it, which gives it a chance to have further significance.

For the QPSVM, maximizing the margin and minimizing the error are the two basic modules in its mathematical modelling while finding the best trade off between minimizing the error and minimizing the sum of the KCV coefficients is the main theme of VLPSVM. To do so, in its mathematical formulation, VLPSVM directly puts the sum of the non-negative coefficients of KCVs in the objective function to be minimized that gives a sparser solution. Amazingly, with this direct involvement of non negative coefficients of KCVs (by a non-negative summation) in the cost function, VLPSVM very closely replicates the QPSVM. However, by so far, compared to the QPSVM, VLPSVM misses many of the interesting properties that make QPSVM academically richer such as a concrete and validated theoretical base with the efficient dual transformation that couples the kernel functions in the simplest and productive fashion. Still, our experiments on benchmark data as well as other reports [41] show that considering classification performance with generalization, VLPSVM is quite competent like QPSVM while being more efficient proving LPSVM to be empirically richer and more productive. Apparently, it appears to be paradoxical to the basic principle of SVM that a machine with less number of KCVs poses similar generalization performance to that of a machine with more KVCs, but the key point here is that the KCVs in VLPSVM do not have exactly the same topoloical and geometric interpretation as that in QPSVM despite the fact that they are being extensively called by the same name (SVs) in many literature. Further in the same path, unlike QPSVM (due to its constraint $\sum_{i=1}^{N} \alpha_i y_i = 0$) there is no condition in VLPSVM that the KCV set contains training patterns from both classes, which may help it to reduce the number of KCVs in the decision function. Furthermore, as L_1 norm is usually more intending to sparser solution compared to L_2norm [42], by formulating an indirect L_1 norm in VLPSVM cost function, it leads to further sparsity compared to QPSVM that uses the L_2 norm for such. Another concern related to the larger number of KCV of the QPSVM (compared to VLPSVM) is its KKT condition $\alpha_i\Big(y_i\big(w \cdot \phi(x_i) + b\big) \ - 1 + \zeta_i\Big) = 0$ that forces all the training patterns staying on the class-margin of own class or outwards to be KCV. That means, some sorts of training patterns must be included in the KCVs set in the QPSVM and this becomes specially serious in case of large and noisy datasets as they contain such overlapping and non-separable examples with a big portion. On contrary, VLPSVM has no such condition, which gives it flexibility to chose KCVs from more scattered pattern space following the demand of the stochastic and topological property of data-patterns leading to pick up few but crucial patterns that are perfect to be KCVs for a very sparse but powerful classifier with strong generalization capability.

IV. Proposed Method: SOSVM

While a powerful classifier is essential to handle with the difficulties from large and noisy data, controlling the classifier-complexity is also important to achieve better generalization. Additionally, considering both cost and accuracy, the best classifier is the most sparse one, having the highest generalization capability, posing least test error. To serve this purpose, we try a novel algorithm by applying both QP and LP in a structured sequence.

As we discussed earlier that although both QPSVM and LPSVM are sparse, VLPSVM produces sparser solution than QPSVM while posing very similar accuracy. Still, this sparsity form VLPSVM is not sufficient for large data set. So, we look for a machine that is even further sparse and faster but more generalized and powerful aiming at real-time classification on very large and complex data.

We know that the sparseness of SVMs heavily depends on the noise and complexity of the data. When the data set is very noisy, a good generalized QPSVM may get more outliers, which will be included into the SVs set in addition to the patterns that are just on the margins. So, number of SVs will soar while generalization capability of the machine will also rise and this SVs set is one of the best representative sets of the whole data. Moreover, as the SVs set from QPSVM are sufficient to represent the discrimination between the classes, we consider only this SVs patterns (who also mostly stay around the discrimination boundary using margin maximization concept) for next manipulation in order to produce our efficient classifier by further sparsification without losing generalization ability. We then run LP (in VLPSVM fashion) on this SVs set as this LP will impose a co-efficient vector carrying weights of these SVs patterns to minimize the objective function while maintaining classification accuracy and generalization potential. Hence, this weight vector will have updated co-efficient values (from the QPSVM SVs-coefficients) being further (2nd time) sparse and will promote to an extensive computational reduction by enabling much smaller number of KCVs (after throwing a large part of the SVs) to be involved in the final decision function.

This gives us twofold benefits: one, training set is condensed by a pure filtration picking only the significant patterns that are already bases of a theoretically solid and powerful classifier and are sole representer of the data. By this, we also abstain from the computation of novel representatives of SVs as this relies upon complex optimization problems that are susceptible to initialization, step sizes, etc. Second, we take advantage from the sparser, and flexible pattern selecting capability of VLPSVM for KCV from a scattered and random

Our Algorithm: SOSVM

1: **Input:** A training set $\left(x_i, y_i\right)_{i=1}^{N}$
2: **Output:** A discriminator $f_{SOSVM}(\cdot)$
3: Select two of the best pairs of $(Penalty\ parameter, Kernel\ parameter) \equiv (C, \sigma), (C_V, \sigma_V)$ for two stages
4: Run QPSVM on the training set solving the following following problem:

$$\min_{\alpha} f(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(x_i)\cdot\phi(x_j) - \sum_{i=1}^{N}\alpha_i$$
$$s.t \ \ \sum_{i=1}^{N}\alpha_i y_i = 0$$
$$0 \le \alpha_i \le C; \ \ i = 1, 2, ..., N$$

5: Extract SVs with their labels $\left(x_l, y_l\right)_{l=1}^{M}$ from the QPSVM using $\alpha_i > 0$
6: Run LP in VLPSVM fashion on these SVs patterns solving the following following problem:

$$\min_{\lambda,\xi,b_V} \ \ \sum_{l=1}^{M}\lambda_l \ + \ C_V\sum_{i=1}^{M}\xi_l$$
$$s.t. \ \ y_i\Big(\sum_{l=1}^{M}\lambda_l y_l \phi(x_l)\cdot\phi(x_i) + b_V\Big) \ \ge \ 1-\xi_i$$
$$\lambda_l \ \ge \ 0; \ l = 1, 2, ..., M$$
$$\xi_i \ \ge \ 0; \ i = 1, 2, ..., M$$

7: Extract EVs with their labels $\left(x_m, y_m\right)_{m=1}^{P}$ using $\lambda_m > 0$ and the bias b_V from the LP
8: $w_{SOSVM} \leftarrow \sum_{m=1}^{P}\lambda_m y_m \phi(x_m)$; $b_{SOSVM} \leftarrow b_V$
9: Return $f_{SOSVM}(\cdot) = w_{SOSVM}\cdot\phi(\cdot) + b_{SOSVM}$

pattern space. Additionally, the sequential training of two inductive submachines where the second one is denser being truncated and a function of the first one leads to a resultant final one higher ordered function. So, our discriminator virtually plays the functional role of a second ordered decision function, which echos the name "Second Order SVM (SOSVM)" of our algorithm while this higher ordered nature better handles the random and nonlinear behavior of the data.

Patterns having the non-zero co-efficient values from final output are the KCVs of our SOSVM and we call them "Machine Vectors(MVs)" whereas the bias of this SOSVM comes from the solution of the final optimization problem; these components construct the SOSVM using corresponding patterns and labels. It is worth mentioning here that while the optimizations in the two stages are done in sequence, their supporting parameters are found simultaneously by a joint search using a modified cross validation technique, which is in accordance with the overall system as a single machine.

Fig. 1 shows the decision boundary with number of Kernel computing vectors (KCVs), and training error rates from QPSVM, VLPSVM, and SOSVM on Banana data set from machine learning benchmark. Fig. 1(a) shows the decision boundary for QPSVM with $C = 4096$ and $\sigma = 2$, Fig. 1(b) shows the decision boundary for VLPSVM with $C_V = 4$ and $\sigma_V = 1$ and the decision boundary for SOSVM with $C = 8, \sigma = 0.125, C_V = 4$ and $\sigma_V = 1$ is shown in Fig. 1(c). Banana is a two dimensional data with 400 training patterns from which QPSVM uses 94 KCVs, VLPSVM uses 14 whereas our SOSVM uses only 11 while posing training error rates 6.75%, 8.75%, and 8.75%, respectively. Therefore, to classify a single pattern, SOSVM demands only 11.7% kernel execution of QPSVM (which is sparse) and 78.6% that of the VLPSVM (which is sparser) while offering very similar accuracy!

V. Experiments and Results

A. Key Terms to Analyze Machine's Perfection

So far, there is a convention to find the test error rate of a classifier to realize its generalization-quality. In fact, it tells about the classifier's performance on test data, which is a must to know but may not give complete info about machine's bridging capability between the training and test data. So, we define a novel term called "Generalization Failure Rate (GFR)" that includes machine's performances on the training set, test set and their difference. To evaluate the classifier's deficiency, GFR is based on the two coupled info: 1) How much the classifier intends to overfit; and 2) How bad it performs on the test set.

Further, to terminate the confusion between the usefulness of an expensive machine with highest accuracy and a cheaper machine with acceptable accuracy, another new term "Machine-Accuracy-Cost(MAC)", expressing cost per accuracy is defined.

1) Generalization Failure Rate (GFR): The main goal of a classification algorithm is to discover a discriminating function basing on the training set (input patterns and the corresponding labels) that will generalize well by classifying the novel patterns with the least possible errors. However, to make it real time applicable, the secondary objective is that the classifier should be as sparse as possible, that is, in case of basis-vector based machine, it should have as few basis as possible. But this basis-set (hence its size) radically influences the machine's generalization performance.

If this basis-set lead towards a very simple model, it fails to learn the data-complexity and thus poses poor performance on both the training and test set by underfitting.

On contrary, if this basis-set lead towards a very complex model, it learns the irrelevant detail and noise in the training dataset (and weakening the general model) leading to the decrement of the training error by overfitting and increment of the test error with generalization-failure. Thus measuring the generalization quality of a classifier is really indispensable. However, although both overfitting and underfitting can lead to model's performance failure, the most frequent problem in machine learning is overfitting. So, we start by focusing on it and define a term called "Overfitting Tendencey(OT)" as the difference between Test Error Rate and Train Error Rate per Train Error Rate; mathematically, $OT = \frac{TestErrorRate - TrainErrorRate}{TrainErrorRate}$. Hence, OT gets higher for a higher value of $\frac{TestErrorRate}{TrainErrorRate}$, which increases for lower Train Error and higher Test Error. Further, to include the loss done by underfitting, we divide this OT by Test Accuracy and define the term as the GFR that measures the overall Generalization deficiency of the model, that is $GFR = \frac{OT}{TestAccuracy}$. It is quite clear that GFR gets higher either by increasing in OT or by decreasing the Test Accuracy or by both. It is to note that the term "GFR" is defined here on the assumption that $TestErrorRate, TrainErrorRate \in (0, 100)\%$ and $TestErrorRate > TrainErrorRate$, which is the usual case.

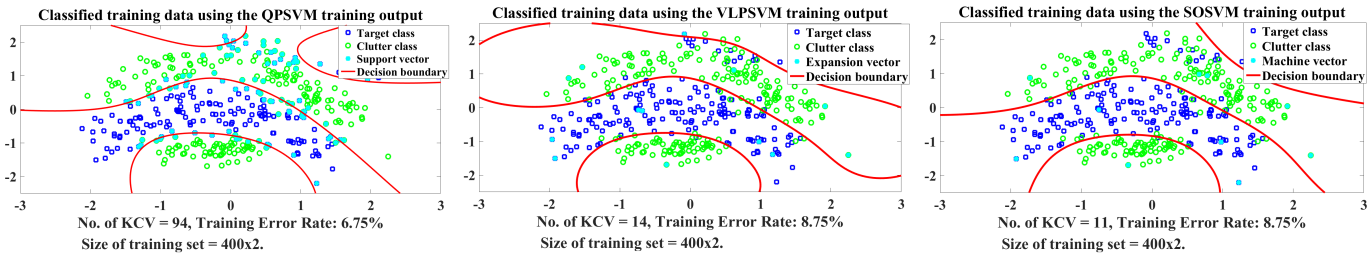


Fig. 1. Decision boundaries, number of Kernel Computing Vector (KCV), & training error rates from (a) QPSVM, (b) VLPSVM, & (c) SOSVM on Banana data.

2) Machine-Accuracy-Cost (MAC): In almost all cases, it is desired to have an efficient machine, that is a machine with less computational cost but having high performance. Therefore, we need to build a machine demanding less kernel execution (to classify a test pattern), hence with less number of Kernel Computing Vectors (Support Vectors or Basis Vectors or Expansion Vectors or Machine Vectors, however it is called by different authors for different machines, we, here, call those "Kernel Computing Vectors(KCVs)" that involve kernel evaluation) and high test accuracy. To measure the achievement of such property by a machine, we define a term "Machine-Accuracy-Cost(MAC)" as $MAC = \frac{Number\ of\ KCVs(\#KCVs)}{Test\ Accuracy(TeA)}$. So, a machine with the highest number of KCVs which increases for lower Train Error and hig lowest test accuracy will have the maximum MAC(which is never desired), whereas a machine with the lowest number of KCVs giving the highest test accuracy will have the minimum MAC (which is always desired).

B. Experimental Set-up and Results

In this section, the experimental results obtained from the proposed method are presented to show the efficiency of the SOSVM and to compare with QPSVM and VLPSVM. The experiments were performed on six benchmark machine learning datasets [43] namely Banana, Diabetics, Heart, Thyroid, Titanic and Twonorm as listed in Table I. For all three machines Gaussian kernel is used throughout the whole experiment. In case of QPSVM and VLPSVM, the penalty parameters C, C_V and the kernel parameters σ, σ_V are chosen based on the lowest crossvalidation error rate for each dataset using five-fold crossvalidation scheme. This two cases are experimented with $C \in \{2^{-2}, 2^0, 2^2, ..., 2^{12}\}$ and $\sigma \in \{2^{-2}, 2^0, ..., 2^6\}$. For SOSVM, there are C, σ in the first stage and C_V, σ_V in the second stage. To figure out the best C, σ, C_V and σ_V, a modified scheme of five-fold crossvalidation is implemented. In this scheme, 4 folds of randomly chosen training data are used to feed in the first stage for a particular value of C and σ. The KCVs from the first stage are used to feed as training set with a particular value of $C_V$and σ_V in the second stage. The returned KCVs of the second stage VLPSVM are treated as the overall KCVs of SOSVM and this classifier is used to test the remaining one fold of training data. The parameters C, σ, C_V, σ_V with the lowest crossvalidation error rate are chosen as the best ones. This approach is experimented with $C \in \{2^{-2}, 2^0, 2^2, ..., 2^{12}\}, \sigma \in \{2^{-2}, 2^0, ..., 2^6\}, C_V \in C \times \{2^{-2}, 2^{-1}, 2^0, ..., 2^5\}$ and $\sigma_V \in \sigma \times \{2^{-2}, 2^{-1}, 2^0, ..., 2^5\}$.

To evaluate the quality of the result obtained by the proposed SOSVM, it is compared with the results obtained by QPSVM and VLPSVM in Table I. Table I presents the number of training and testing patterns of different dataset along with the average number of KCVs and average test error rate. From this Table, it is observed that the proposed method (SOSVM) results in a lower number of KCV than QPSVM in all cases and also lower than VLPSVM in most cases. In case of QPSVM the average number of KCVs for all the datasets is 154.15 (SD 8.19) whereas for VLPSVM it is 29.08 (SD 8.16). But in case of our proposed SOSVM the average number of KCVs is 18.29 (SD 8.34) which is $1/9^{th}$ of QPSVM and $2/3^{rd}$ of VLPSVM. Therefore, it is clear that in most cases SOSVM results in a reduction in the number of KCVs and in some cases substantial reduction.

Moreover, in order to show how well our machine is capable of generalizing than the traditional machines, the performance of the proposed method along with the traditional QPSVM and VLPSVM is compared in terms of Generalization Failure Rate (GFR). In Table II, the GFRs of traditional QPSVM, VLPSVM and the proposed SOSVM are listed for various dataset. Also, the ratios of GFRs of the proposed SOSVM with respect to traditional machines are listed to show how well SOSVM performs to generalize the training data. It can be seen from the table that the average GFR value of our machine as small as 30% of the most powerful classifier like the QP SVM and 2% of the LP SVM.

Finally, in order to judge the proposed in terms of Machine Accuracy Cost (MAC), the MAC values of the proposed method along with the traditional QPSVM and VLPSVM is listed Table III. Also, the ratios of MACs of the proposed SOSVM with respect to traditional machines are listed to show how well our machine minimizes the cost. It can be seen from the table that the average MAC value of our machine as small as 16% of the most the most powerful classifier like the QPSVM and 80% of the sparse LP SVM.

VI. Conclusion and Future Work

We have developed a fast but powerful classifier by using sequential optimization that is supported with simultaneous parameter search. We have also defined two new terms "GFR" and "MAC" that can be directly and easily measured to verify a detector's perfection. Our classifier is very much straight forward using least effort to train. Compared to the state-of-the-art sparse classifiers, it is more efficient, hence, posing average MAC value as small as 16% of the standard QP

TABLE I. NO. OF KCVS (KERNEL COMPUTING VECTORS), WHICH INCREASES FOR LOWER TRAIN ERROR AND HIGH TEST ERROR RATE ON BENCHMARK DATA FOR DIFFERENT MACHINES

Dataset	No. of train pattern	No. of test pattern	Data Dimension	QPSVM mean SVs (SD)	QPSVM mean TeER (SD)	VLPSVM mean EVs (SD)	VLPSVM mean TeER (SD)	SOSVM mean MVs (SD)	SOSVM mean TeER (SD)	$\frac{\#MVs}{\#SVs}$	$\frac{\#MVs}{\#EVs}$
BANANA	400	4900	2	102.26 (14.09)	10.61 (0.53)	15.08 (1.30)	10.75 (0.51)	15.23 (1.75)	10.91 (0.52)	0.1489	1.0099
DIABETIS	468	300	8	263.22 (15.63)	23.28 (1.70)	12.58 (1.94)	23.40 (1.77)	12.56 (1.93)	23.43 (1.78)	**0.0477**	0.9984
HEART	170	100	13	68.23 (6.01)	16.60 (3.05)	21.94 (2.61)	17.44 (3.49)	10.60 (9.47)	15.55 (3.20)	0.1554	**0.4831**
THYROID	140	75	5	43.51 (3.10)	5.20 (2.08)	8.97 (1.59)	5.09 (2.11)	4.34 (1.27)	7.77 (8.14)	0.0997	0.4838
TITANIC	150	2051	3	148.50 (3.33)	22.69 (0.86)	83.91 (36.52)	22.91 (0.60)	48.48 (34.06)	23.34 (1.39)	0.3265	0.5778
TWONORM	400	7000	20	299.18 (7.00)	2.42 (0.14)	32.00 (5.03)	3.71 (0.55)	18.50 (1.55)	3.38 (0.38)	0.0618	0.5781
Average	288	2404.33	8.50	154.15 (8.19)	13.47 (1.39)	29.08 (8.16)	13.88 (1.50)	18.29 (8.34)	14.06 (2.57)	–	–

TABLE II. GENERALIZATION FAILURE RATE (GFR) FOR DIFFERENT MACHINES

Dataset	QPSVM	VLPSVM	SOSVM	$\frac{SOSVM}{QPSVM}$	$\frac{SOSVM}{VLPSVM}$
BANANA	0.0031	0.0026	0.0019	0.6265	0.7618
DIABETIS	0.0010	0.0014	0.0014	1.3083	1.0100
HEART	0.0039	0.0060	0.0012	0.3073	0.2018
THYROID	0.0918	0.0109	0.0013	**0.0142**	0.1194
TITANIC	0.0017	0.0015	0.0010	0.6021	0.6975
TWONORM	0.0017	1.7005	0.0246	14.2239	**0.0145**
Average	0.0172	0.2871	0.0052	–	–

TABLE III. MACHINE ACCURACY COST (MAC) FOR DIFFERENT MACHINES

Dataset	QPSVM	VLPSVM	SOSVM	$\frac{SOSVM}{QPSVM}$	$\frac{SOSVM}{VLPSVM}$
BANANA	1.1439	0.1690	0.1520	0.1329	0.8996
DIABETIS	3.4311	0.1642	0.0876	**0.0255**	0.5337
HEART	0.8181	0.2657	0.0955	0.1168	**0.3594**
THYROID	0.4590	0.0945	0.0945	0.2059	1.0000
TITANIC	1.9208	1.0885	1.0666	0.5553	0.9799
TWONORM	3.0661	0.3323	0.2001	0.0653	0.6021
Average	1.8065	0.3524	0.2827	–	–

SVM, 80% of the sparse LP SVM by Vapnik. Moreover, being optimally complex and powerful, its overfitting tendency is really low, which leads it to offer average GFR value as small as 30% of the most powerful classifier like the standard QP SVM and 2% of Vapnik's LP SVM.

Due to this exceptionally good performance, one question pops up about how it manages to perform better than Vapnik's LP SVM (VLPSVM) or the standard QP SVM (QPSVM). We do not have any theoretical explanation for it now (and left for future work) but one plausible explanation for it could be that by the second layered training from the sequential combination of the two sub-machines generated by QP and LP (using corresponding parameters by a joint and simultaneous search) respectively, we get a machine vectors set being second ordered filtered and scaled (hence learned) having stochastic and topological properties complex and sophisticated than Support Vectors (in case of QPSVM) or Expansion Vectors (in case of VLPSVM) while working in the similar method for the discriminator. Thus, our algorithm produces a hybrid and unconventional hyperplane, based on a compact second ordered representer set coupled with corresponding co-efficent vector and bias that collectively adopts the statistical and geometric properties of training data very skillfully and generalization is boosted.

Anyway, while our classifier has consistently over performed state of the art complex and sparse classifiers with respect to computational cost and accuracy, we are considering some further manipulation with it where parts are given below:

1) An extension from two-stages including further stages.

2) A deep theoretical analysis relating the data characteristics, components of the sub-machines as well as their sequential behavior and pattern-space sharing including low GFR and MAC would be interesting.

At last, an efficient and accurate classifier like our SOSVM is very much essential. For example, our classifier is indispensable in real life, where one may have more time and resources to train but very less to test.

REFERENCES

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[2] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[3] F. Vojtěch and V. Hlaváč, "Greedy algorithm for a training set reduction in the kernel methods," in *Proc. of the International Conference on Computer Analysis of Images and Patterns*. Springer, 2003, pp. 426–433.

[4] S. S. Keerthi, O. Chapelle, and D. DeCoste, "Building support vector machines with reduced classifier complexity," *Journal of Machine Learning Research, (JMLR)*, vol. 7, no. Jul, pp. 1493–1515, 2006.

[5] E. Osuna and F. Girosi, "Reducing the run-time complexity of support vector machines," in *Proc. of the International Conference on Pattern Recognition (ICPR)*, 1998, pp. 1–10.

[6] M. Rätsch, S. Romdhani, G. Teschke, and T. Vetter, "Over-complete wavelet approximation of a support vector machine for efficient classification," in *Proc. of the Joint Pattern Recognition Symposium*. Springer, 2005, pp. 351–360.

[7] S. Romdhani, P. Torr, B. Schölkopf, and A. Blake, "Efficient face detection by a cascaded support–vector machine expansion," in *Proc. of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 460, no. 2051, 2004, pp. 3283–3297.

[8] M. Wu, B. Schölkopf, and G. Bakır, "A direct method for building sparse kernel learning algorithms," *Journal of Machine Learning Research (JMLR)*, vol. 7, no. Apr, pp. 603–624, 2006.

[9] A. Cotter, S. Shalev-Shwartz, and N. Srebro, "Learning optimally sparse support vector machines," in *Proc. of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 266–274.

[10] Y. J. Lee and O. L. Mangasarian, "RSVM: reduced support vector machines," in *Proc. of the First SIAM International Conference on Data Mining (SDM)*, 2001, pp. 1–17.

[11] T. Joachims and C. N. J. Yu, "Sparse kernel svms via cutting-plane training," in *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML)*, 2009, p. 8.

[12] K. Z. Arreola, J. Fehr, and H. Burkhardt, "Fast support vector machine classification using linear SVMs," in *Proc. of the 18th International Conference on Pattern Recognition (ICPR)*, vol. 3, 2006, pp. 366–369.

[13] B. Heisele, T. Serre, S. Prentice, and T. Poggio, "Hierarchical classification and feature reduction for fast face detection with support vector machines," *Pattern Recognition*, vol. 36, no. 9, pp. 2007–2017, 2003.

[14] H. Sahbi and D. Geman, "A hierarchy of support vector machines for pattern detection," *Journal of Machine Learning Research (JMLR)*, vol. 7, no. Oct, pp. 2087–2123, 2006.

[15] X. Huo and J. Chen, "Building a cascade detector and applications in automatic target detection," *Applied Optics: Information Processing*, vol. 43, no. 2, pp. 1–47, 2003.

[16] R. Karim, M. Bergtholdt, J. H. Kappes, and C. Schnörr, "Greedy-based design of sparse two-stage svms for fast classification," in *Proc. of the 29th DAGM Symposium on Pattern Recognition*, 2007, pp. 395–404.

[17] S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel svms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 66–77, 2013.

[18] L. Ladicky and P. H. S. Torr, "Locally linear support vector machines," in *Proc. of the 28th International Conference on Machine Learning, (ICML)*, 2011, pp. 985–992.

[19] Z. E. Xu, J. R. Gardner, S. Tyree, and K. Q. Weinberger, "Compressed support vector machines," *arXiv preprint arXiv:1501.06478*, 2015.

[20] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[21] R. Xiao, H. Zhu, H. Sun, and X. Tang, "Dynamic cascades for face detection," in *Proc. of the IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.

[22] H. Luo, "Optimization design of cascaded classifiers," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 480–485.

[23] M. J. Saberian and N. Vasconcelos, "Boosting classifier cascades," in *Proc. of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*, 2010, pp. 2047–2055.

[24] M. Chen, Z. E. Xu, K. Q. Weinberger, O. Chapelle, and D. Kedem, "Classifier cascade for minimizing feature evaluation cost," in *Proc. of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 218–226.

[25] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. of the 13th European Conference on Computer Vision (ECCV)*, 2014, pp. 109–122.

[26] J. Li and Y. Zhang, "Learning SURF cascade for fast and accurate object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3468–3475.

[27] V. C. Raykar, B. Krishnapuram, and S. Yu, "Designing efficient cascaded classifiers: tradeoff between accuracy and cost," in *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 853–860.

[28] I. Visentini, L. Snidaro, and G. L. Foresti, "Cascaded online boosting," *J. Real-Time Image Processing*, vol. 5, no. 4, pp. 245–257, 2010.

[29] M. J. Saberian and N. Vasconcelos, "Boosting algorithms for detector cascade learning," *Journal of Machine Learning Research (JMLR)*, vol. 15, no. 1, pp. 2569–2605, 2014.

[30] Z. E. Xu, M. J. Kusner, K. Q. Weinberger, and M. Chen, "Cost-sensitive tree of classifiers," in *International Conference on Machine Learning*, 2013, pp. 133–141.

[31] Z. E. Xu, M. J. Kusner, K. Q. Weinberger, M. Chen, and O. Chapelle, "Classifier cascades and trees for minimizing feature evaluation cost," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2113–2144, 2014.

[32] Z. Fu, A. R. Kelly, and J. Zhou, "Mixing linear svms for nonlinear classification," *IEEE Trans. Neural Networks*, vol. 21, no. 12, pp. 1963–1975, 2010.

[33] W. C. Cheng and D. M. Jhan, "A cascade classifier using adaboost algorithm and support vector machine for pedestrian detection," in *Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, 2011, pp. 1430–1435.

[34] S. Maji and J. Malik, "Fast and accurate digit classification," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-159*, 2009.

[35] Q. Gu and J. Han, "Clustered support vector machines," in *Proc. of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013, pp. 307–315.

[36] S. Maji and A. C. Berg, "Max-margin additive classifiers for detection," in *Proc. of the IEEE 12th International Conference on Computer Vision (ICCV)*, 2009, pp. 40–47.

[37] G. Sharma, F. Jurie, and P. Perez, "Learning non-linear SVM in input space for image classification," Ph.D. dissertation, GREYC CNRS UMR 6072, Universite de Caen, 2014.

[38] M. Osadchy, D. Keren, and B. F. Specktor, "Hybrid classifiers for object classification with a rich background," in *Proc. of the 12th European Conference on Computer Vision (ECCV)*, 2012, pp. 284–297.

[39] M. Osadchy, D. Keren, and D. Raviv, "Recognition using hybrid classifiers," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 38, no. 4, pp. 759–771, 2016.

[40] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. of the IEEE 12th International Conference on Computer Vision (ICCV)*, 2009, pp. 606–613.

[41] A. Nefedov, J. Ye, C. Kulikowski, I. Muchnik, and K. Morgan, "Experimental study of support vector machines based on linear and quadratic optimization criteria," *DIMACS Technical Report 2009-18*, 2009.

[42] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[43] G. Rätsch. Benchmark data sets. [Online]. Available: http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

A Fuzzy based Soft Computing Technique to Predict the Movement of the Price of a Stock

Ashit Kumar Dutta
Department of Computer Science and Information System
Shaqra University
Kingdom of Saudi Arabia

***Abstract*—Soft computing is a part of an artificial intelligence, and fuzzy logic is the study of fuzziness on data. The combination of these two techniques can provide an intelligent system with more ability and flexibility. The nature of data in the stock/capital market is more complex and challenging to predict the movement of the price of the stock. The study has combined both fuzzy c-means and neural network technique for the prediction of the price of the stock. The research finds an optimum solution to predict the future price of a stock. The comparison of time and space complexity proved that the proposed method is better than the existing methods.**

Keywords*—Soft computing; fuzzy logic; stock recommendation; fuzzy based soft computing, soft computing systems

I. Introduction

A wide variety of engineering control processes are using the concept of fuzzy logic (FL). There is a need of expert knowledge to develop an intelligent system. FL concepts were evolved in late 60's from crisp fuzzy sets and reached a high level of perceptions in recent years [6], [7]. A researcher can apply every kind of possibilities ranges from approximate, linguistics to perception, computing with words, natural language processing, computational theory of perceptions, and perception-based probability theory. Logical, relational, set-theoretic and epistemic are the four facets of FL [8], [9]. Sigmoid functions can be used to define fuzzy sets. Professor Zadeh has found many operators in FL and completely replaced the ordinary Boolean logic. Fuzzification and defuzzification are the functions similar to encode and decode functions. Many expert systems were implemented with the concept of FL.

Soft computing (SC) is related to FL used to develop intelligent systems. The purity of thinking, machine intelligence, and fuzziness processing capabilities are the attributes of SC. The SC aims to exploit tolerance for imprecision, uncertainty, and fuzziness to achieve low-cost computation. It can learn from experience, universalize domain, simulation of biological processes, and faster computations [10], [11]. It has the critical role in science and engineering applications [20]-[22]. It is widely used in machine learning, belief networks, chaos theory, wisdom based expert system and probability reasoning. It has feature of extraction of knowledge/information from inaccurate and uncertain data [12], [13] and vast influence of artificial neural network (ANN), FL and genetic algorithm [14], [15]. The combination of FL and SC is used to develop an accurate expert system [23]-[25].

Stock/capital market is familiar to investors due to its high returns. The non – linear, chaotic, noisy behavior nature of data in the market made difficult to predict the price of a stock [1]. Technical, fundamental, and macroeconomics are the popular methods to analyze the market and predict the movement of a stock. A technical analysis is used to find the price of the stock in a period. Mathematical methods are used in the technical analysis to indicate the buy/sell signals [2], [3]. Fundamental analysis is the technique to analyze the company performance by calculating profit and liabilities. A macroeconomic analysis is a technique to analyze the recent decision taken by the company and its impact on the market [4], [5]. The aim of the research is to build an intelligent system illustrated in Fig. 1 to predict the price of a stock in Indian capital market.

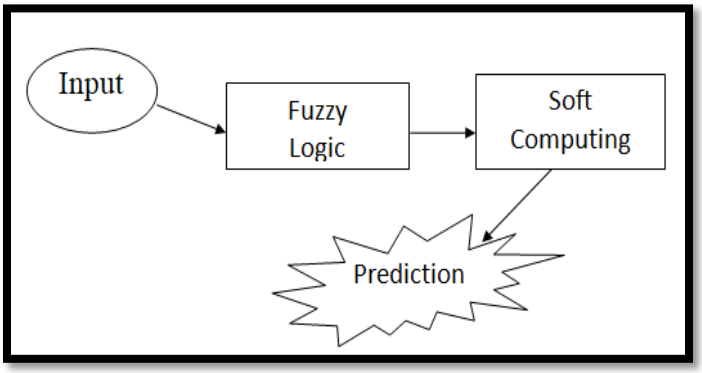


Fig. 1. Stock recommendation systems.

Current and previous day price of the stock, Relative strength index (RSI), the volume of the stock and 30 days average moving price of the stock are the inputs required for the system and FL – SC combined proposed method will predict the future price of the stock [16], [17]. The Nifty – 50 and Pharmacy stocks used in the research. Support vector machine (SVM), Neural networks (NN) are the existing methods used in the research to show the effectiveness of the proposed method. Section 2 elaborates the review of the literature, Section 3 define the methodology of the research and experiment and results provided in Section 4. Finally, Section 5 concludes the research.

II. Review of Literature

Ibrahim M. Hamed et al. [1], proposed an intelligent system to recommend the price of the stock. They have used multilayer perceptron-based ANN for the research. The model proposed in their research had several stages. Input selection, pre-processing of data and signal detection were some of the critical stages in the model. The algorithm was used in the model update the weights between the neural network neurons to minimise the error of the prediction results. Blind source separation techniques are used in the research for the signal detection and forward to the final stage for the user. ANN is a slow learner, and moreover, the multilayer perceptron leads to taking more time for the prediction of results.

Monica Tirea and Viorel Negra [2] have developed an intelligent stock market analysis system for the prediction of the stock price. The study has taken many attributes and constructed a multi-agent prediction system. Economic, industrial and behavioural are the types of approaches in fundamental analysis. The macro – economical analysis have taken interest rates, earning of the company, profitability and growth rate. The overall structure of the system had three essential parts. They are multi-agent stock trading system, news information system and portfolio optimisation system. The portfolio system has data interpretation level, behavioural level, risk detection level and portfolio optimisation level. The multi-agent stock trading system has used the numerical data with multi-agents. Text mining agent is used to crawl the recent news and forwards to the portfolio optimisation system and modify signals according to the news. The system took more time and space to predict the price and consume more memory comparing to the other new methods.

Sunil Kumar Khatri et al. [3] have proposed a sentiment analysis to predict Bombay stock exchange using ANN. Many forums are there for the investors to participate and share their views related to the market. Data of blogs/forums and E – media were the inputs for the system and pre-processed into four different moods like happy, hope, sad and disappointing. The ANN model was trained with sample data from a detailed set and produces the final prediction from the test data. The output of the ANN model prediction will be positive, negative/neutral. A maximum of 15 hidden layers were used for the prediction of the price of the stock based on the sentiments. The research entirely based on the sentiment of the users in the forum and there was no relation with the real price of the stock. The accuracy is the primary concern of the research as it depends on the sentiment shared in the forum.

Qiaghua wen et al. [4], have proposed an intelligent stock recommendation system using SVM techniques and box theory of stock. The SVM uses statistical learning theory and structural risk minimisation principle. The principle behind the SVM is to map the inputs into a high – dimension space by a non – linear mapping with kernel function to do a linear regression. In this research, the radial basis function is acted as the kernel function of SVM estimators. The selection of the user-prescribed is a vital role in overall performance. The sliding window method is employed in the research to train SVM estimators. The accuracy and computation time of SVM was more compared to other existing methods.

Chin – Cheng Tseng [5] have proposed a portfolio management using hybrid recommendation system. The research has deployed a group of agents with different functionality and groups them under case – based situations. Agents are used to gather, filter, and process various cases in the systems. It does not have any mechanism to deal with uncertainty and complex data. It will help an individual to select a stock based on historical data and not on real-time data. Influence diagram module, decision tree system module, web information agent module, and interface module were the components of the hybrid recommendation architecture. Decision tree is one of the old methods and if the number of data increased then the computation time will be more and create problems in the process of making the decision.

Lei wang et al. [6], have developed a stock recommendation system based on a method called HLP. The method was used to extract data with different frequency and amplitude. The data were extracted from the server will describe the nature of a price of stock and its movement. ANN modules are used to forecast the movement of the stock. The following Table I shows the type of technique and year used in the research.

TABLE. I. Techniques used in the Research

S.No.	Technique	Year	Authors
1	Decision Tree	2004	Chiu –Che et.al. ,
2	SVM	2009	QinghuaWen et. al. ,
3	ANN	2011	Ibrahim M. Hamed et.al.,
4	Decision Tree	2013	Zhen Hu et.al.,
5	ANN	2014	Monica Tirea et. al. ,
6	ANN	2014	Sunil Kumar Khatri et. al. ,
7	ANN	2014	Neelima Budhani et.al.,
8	ANN	2016	Amin Hedayati Moghaddam et. al.,

III. Proposed Method

The proposed method uses fuzzy C-means and ANN model to build an intelligent system to predict the movement of the price of the stock. Fuzzy c-means is similar to K – means algorithm. In this study, Fuzzy C – means and NN are combined into the Fuzzy logic neural network (FL-N) to predict the price of the stock. The fuzzy algorithm used to cluster stock market data, clusters the stock data extracted from National Stock Exchange (NSE) portal. The clusters are given as input to NN. The NN has two input and one output layers. The historical data will be pre-processed for the model for the prediction of its future [18], [19]. Let $\sum D$ be the price data and $f(\sum D)$is the fuzzification of data. ANN model will take the fuzzified data and compute the future price of the data. Current and previous day price of the stock, Relative strength index (RSI), the volume of the stock and 30 days average moving price of the stock are the inputs. A total of 15 hidden layers were used in NN to produce results. The NN used in the research is the feed forward, back propagation network. It can learn from the failures and fault tolerance. The following algorithm will describe the process of the computation in detail.

Step 1: Start
Step 2: Historic data downloaded from the Nifty 50 and Pharma

Step 3: Initial pre – process of data to normalize it for the fuzzification
Step 4: Fuzzy C – means clusters the data.
Step 5: The clusters transformed into NN inputs.
Step 6: Multi – Layer perceptron NN model receives the transformed input for the training phase.
Step 7: Learnt NN produces results
Step 8: Goto step 3.
Step 9: End

Fig. 2 shows the screen of NSEIndia (www.nseindia.com) website having details about NIFTY 50 stocks. Fig. 3 shows the details of NIFTY Pharma stocks. Fig. 4 shows the pre – processing of data extracted from NSEINDIA. The data were normalized by fuzzy plug-ins for Microsoft Excel 2007.

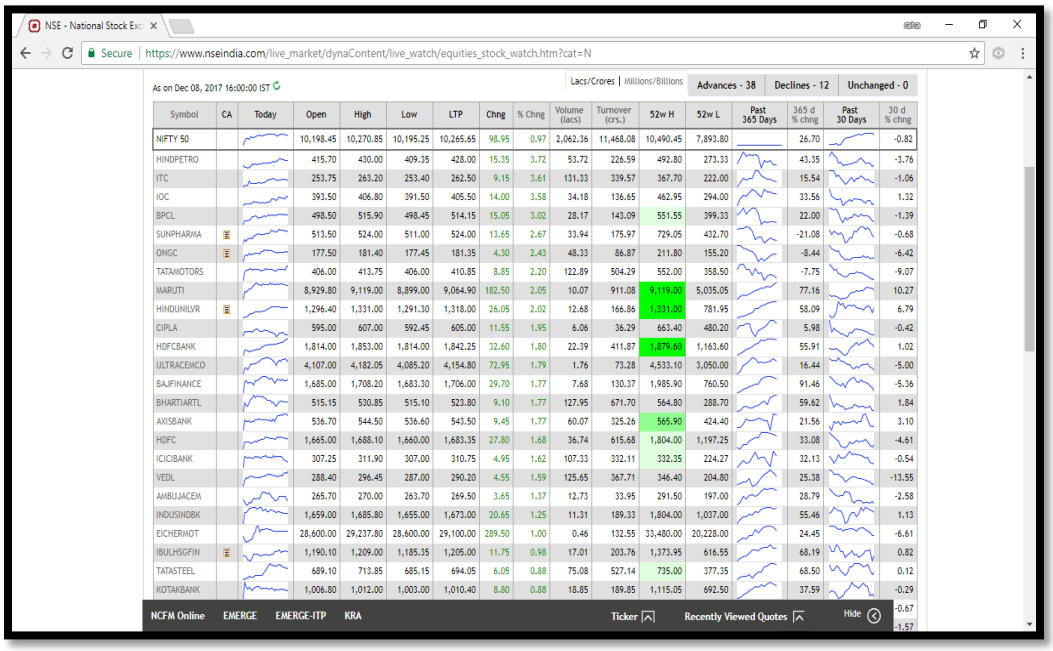


Fig. 2. NSE India – NIFTY 50 Stocks.

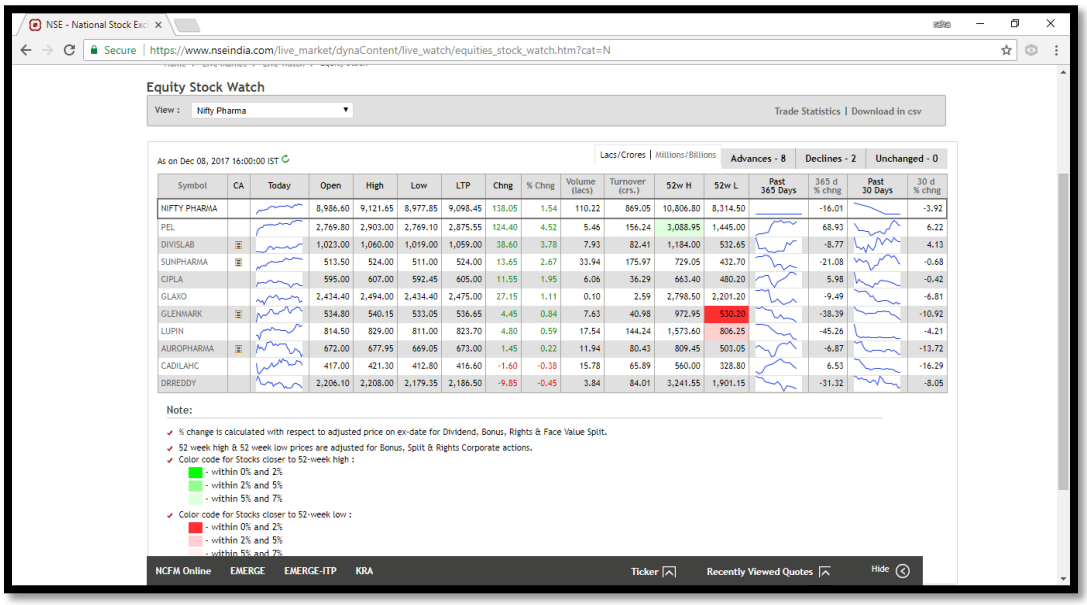
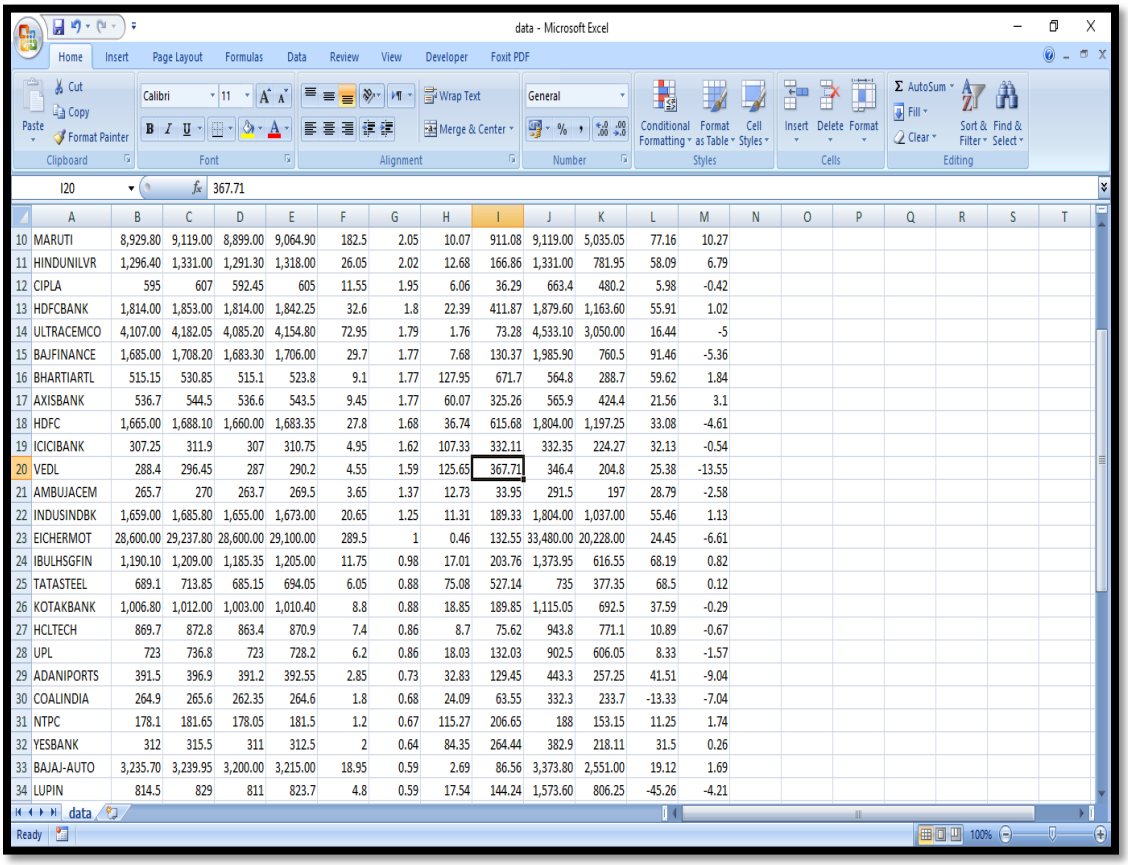


Fig. 3. NSE India – NIFTY PHARMA Stocks.



	A	B	C	D	E	F	G	H	I	J	K	L	M
10	MARUTI	8,929.80	9,119.00	8,899.00	9,064.90	182.5	2.05	10.07	911.08	9,119.00	5,035.05	77.16	10.27
11	HINDUNILVR	1,296.40	1,331.00	1,291.30	1,318.00	26.05	2.02	12.68	166.86	1,331.00	781.95	58.09	6.79
12	CIPLA	595	607	592.45	605	11.55	1.95	6.06	36.29	663.4	480.2	5.98	-0.42
13	HDFCBANK	1,814.00	1,853.00	1,814.00	1,842.25	32.6	1.8	22.39	411.87	1,879.60	1,163.60	55.91	1.02
14	ULTRACEMCO	4,107.00	4,182.05	4,085.20	4,154.80	72.95	1.79	1.76	73.28	4,533.10	3,050.00	16.44	-5
15	BAJFINANCE	1,685.00	1,708.20	1,683.30	1,706.00	29.7	1.77	7.68	130.37	1,985.90	760.5	91.46	-5.36
16	BHARTIARTL	515.15	530.85	515.1	523.8	9.1	1.77	127.95	671.7	564.8	288.7	59.62	1.84
17	AXISBANK	536.7	544.5	536.6	543.5	9.45	1.77	60.07	325.26	565.9	424.4	21.56	3.1
18	HDFC	1,665.00	1,688.10	1,660.00	1,683.35	27.8	1.68	36.74	615.68	1,804.00	1,197.25	33.08	-4.61
19	ICICIBANK	307.25	311.9	307	310.75	4.95	1.62	107.33	332.11	332.35	224.27	32.13	-0.54
20	VEDL	288.4	296.45	287	290.2	4.55	1.59	125.65	367.71	346.4	204.8	25.38	-13.55
21	AMBUJACEM	265.7	270	263.7	269.5	3.65	1.37	12.73	33.95	291.5	197	28.79	-2.58
22	INDUSINDBK	1,659.00	1,685.80	1,655.00	1,673.00	20.65	1.25	11.31	189.33	1,804.00	1,037.00	55.46	1.13
23	EICHERMOT	28,600.00	29,237.80	28,600.00	29,100.00	289.5	1	0.46	132.55	33,480.00	20,228.00	24.45	-6.61
24	IBULHSGFIN	1,190.10	1,209.00	1,185.35	1,205.00	11.75	0.98	17.01	203.76	1,373.95	616.55	68.19	0.82
25	TATASTEEL	689.1	713.85	685.15	694.05	6.05	0.88	75.08	527.14	735	377.35	68.5	0.12
26	KOTAKBANK	1,006.80	1,012.00	1,003.00	1,010.40	8.8	0.88	18.85	189.85	1,115.05	692.5	37.59	-0.29
27	HCLTECH	869.7	872.8	863.4	870.9	7.4	0.86	8.7	75.62	943.8	771.1	10.89	-0.67
28	UPL	723	736.8	723	728.2	6.2	0.86	18.03	132.03	902.5	606.05	8.33	-1.57
29	ADANIPORTS	391.5	396.9	391.2	392.55	2.85	0.73	32.83	129.45	443.3	257.25	41.51	-9.04
30	COALINDIA	264.9	265.6	262.35	264.6	1.8	0.68	24.09	63.55	332.3	233.7	-13.33	-7.04
31	NTPC	178.1	181.65	178.05	181.5	1.2	0.67	115.27	206.65	188	153.15	11.25	1.74
32	YESBANK	312	315.5	311	312.5	2	0.64	84.35	264.44	382.9	218.11	31.5	0.26
33	BAJAJ-AUTO	3,235.70	3,239.95	3,200.00	3,215.00	18.95	0.59	2.69	86.56	3,373.80	2,551.00	19.12	1.69
34	LUPIN	814.5	829	811	823.7	4.8	0.59	17.54	144.24	1,573.60	806.25	-45.26	-4.21

Fig. 4. Pre-processing of data for FL-N.

IV. Results and Discussion

The research used Windows 10 operating system with I7 processor for the pre – process of data extracted from NSEINDIA. Table II shows the training time of the methods used for 60 days from the data. Usually training time will be more comparing to testing time; every method will take time to learn about the environment. The training data has to be unique and should not be used in testing phase. Comparing to others methods, the proposed method have taken less time in the training phase. Fig. 5 shows the relevant graph of the training phase. Days indicate the data collected from the range of 10 to 60 days. The process of data collection begun from 01-02-2017 to 01-12-2017.

TABLE. II. Training Time (in Seconds) of Methods

Days /Methods	**Days**					
	10	**20**	**30**	**40**	**50**	**60**
ANN	0.432	0.326	0.423	0.521	0.563	0.452
SVM	0.621	0.568	0.528	0.612	0.589	0.632
DT	0.745	0.695	0.485	0.569	0.541	0.458
FL - N	0.385	0.215	0.412	0.235	0.346	0.295

Table III shows the testing time taken by the methods and it is evidence that the proposed method have taken less time to produce the results. The FL-N has taken only 0.098 seconds to produce results from 50 days data. Fig. 6 displays the relevant graph of Table III.

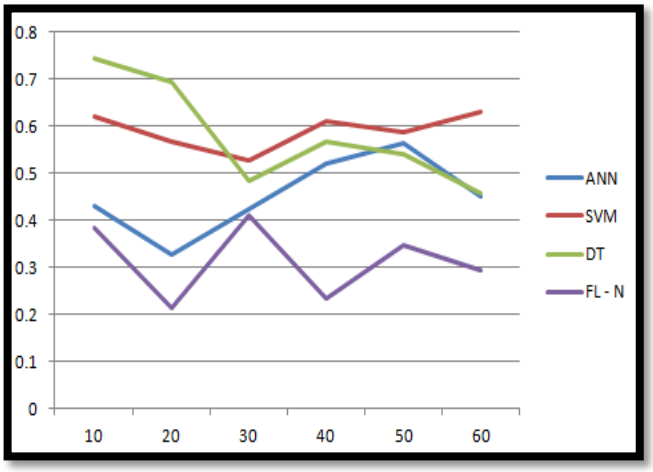


Fig. 5. Training Phase (in seconds).

TABLE. III. Testing Time (in Seconds)

Days /Methods	**Days**					
	10	**20**	**30**	**40**	**50**	**60**
ANN	0.234	0.256	0.185	0.215	0.265	0.278
SVM	0.285	0.312	0.165	0.178	0.195	0.174
DT	0.195	0.215	0.175	0.185	0.139	0.149
FL - N	0.124	0.138	0.135	0.152	0.098	0.112

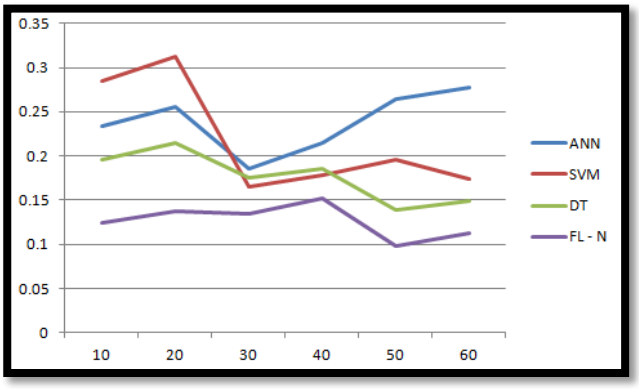


Fig. 6. Testing Phase (in seconds).

Tables IV and V show the accuracy for total of 60 days data during training time. It is necessary to measure the accuracy during the training time. It will be useful for the research to find any deviation between training and testing phase. Table IV shows the percentage of accuracy of the methods used in the research. Fig. 7 shows the relevant graph of Table IV. The proposed method has reached an average of 88% for all predictions. DT has also reached the better accuracy comparing to other two methods.

TABLE. IV. ACCURACY (ROUNDED OFF PERCENTAGE) (TRAINING TIME)

Days /Methods	**Days**					
	10	**20**	**30**	**40**	**50**	**60**
ANN	87	86	85	87	85	86
SVM	89	84	88	87	85	86
DT	85	87	86	88	89	87
FL - N	91	89	87	89	90	86

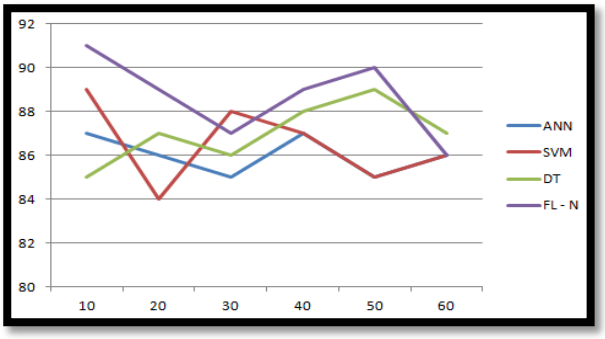


Fig. 7. Accuracy of methods employed in the research (Training time).

Table V shows the percentage of accuracy of the methods used in the research during testing phase. The proposed method has reached an average of 93% for all predictions. DT has also reached an average accuracy of 93%comparing to other two methods but took more time to produce it. Fig. 8 shows the relevant graph of Table V.

TABLE. V. ACCURACY (ROUNDED OFF PERCENTAGE) (TESTING TIME)

Days /Methods	**Days**					
	10	20	30	40	50	60
ANN	91	94	93	96	97	89
SVM	88	92	91	93	89	94
DT	89	91	90	94	87	87
FL - N	93	92	94	94	95	92

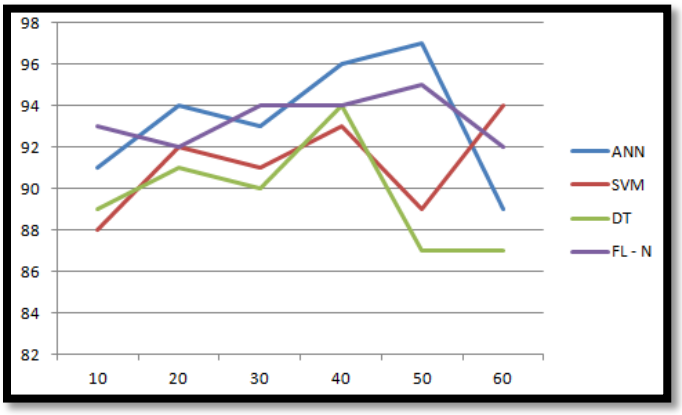


Fig. 8. Accuracy of methods employed in the research (Testing Time).

Precision, Recall and F1 measure are the metrics used to indicate the performance of an algorithm and Tables VI, VII and VIII shows the performance of the proposed and other methods. The proposed method has better score than other methods. FL-N has the ability to perform well in complex situation. Even the web data is hard and complex; the FL – N has learnt well about the environment and produced good results. Fig. 9 shows the relevant graph of F1 measure. FL-N and SVM have performed well comparing to DT and ANN.

TABLE. VI. PRECISION

Clusters /Methods	**Days**						**Overall Average**
	10	**20**	**30**	**40**	**50**	**60**	
ANN	78.5	82.3	84.6	79.8	81.3	83.4	81.65
SVM	82.3	86.4	86.3	83.4	82.3	84.3	84.17
DT	76.5	79.6	79.4	76.2	73.4	79.5	77.43
FL - N	89.3	84.3	87.3	86.3	84.9	85.7	86.3

TABLE. VII. RECALL

Clusters /Methods	**Days**						**Overall Average**
	10	**20**	**30**	**40**	**50**	**60**	
ANN	79.6	78.6	81.2	75.6	76.4	77.4	78.13
SVM	86.3	85.3	85.3	86.3	85.3	86.3	85.8
DT	79.5	79.4	76.3	78.3	79.6	84.3	79.57
FL - N	85.9	86.3	88.2	87.9	88.6	87.5	87.4

TABLE. VIII. PRECISION, RECALL AND F1 MEASURE

Clusters /Methods	**Precision**	**Recall**	**F1 Measure**
ANN	81.65	78.13	79.85
SVM	84.17	85.8	84.98
DT	77.43	79.57	78.49
FL - N	86.3	87.4	86.85

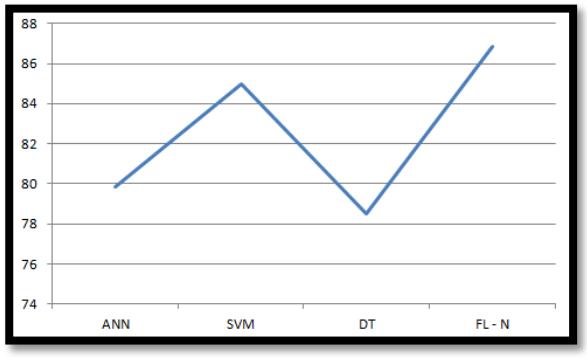


Fig. 9. F1- measure of methods.

The research has measured the overall memory to produce the results during testing time. Training time will take more memory as all methods have to learn the environment. The proposed method has taken 0.3 MB memory to produce the results. DT has taken a maximum of 2 MB to produce the result whereas SVM has occupied 0.7 MB and ANN has taken 0.5 MB. The performance of proposed method is far better than the existing method. ANN and SVM lacks in performance due the nature of consuming more time and memory. DT has the scalability issue so could not produce better results comparing to the proposed method.

V. CONCLUSION

The objective of the research is to predict the movement of the price of a stock. The research has extracted data from NSEINDIA. NIFTY – 50 and PHARMA data were taken for the purpose of experiment. The experimental results proved that the proposed method has an average accuracy of 88% to predict the stock price. Testing and Training time results shows that the time taken by the proposed method is better than the other existing methods. The overall memory for the prediction is 0.3 MB for the proposed method. The future work of the research is to predict the movement of stocks in Bombay stock exchange.

REFERENCES

[1] M. Hamed, A. S. Hussein and M. F. Tolba, "An intelligent model for stock market prediction," The 2011 International Conference on Computer Engineering & Systems, Cairo, 2011, pp. 105-110. doi: 10.1109/ICCES.2011.6141021.

[2] M. Tirea and V. Negru, "Intelligent Stock Market Analysis System - A Fundamental and Macro-economical Analysis Approach," 2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, 2014, pp. 519-526. doi: 10.1109/SYNASC.2014.75

[3] S. K. Khatri, H. Singhal and P. Johri, "Sentiment analysis to predict Bombay stock exchange using artificial neural network," Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, Noida, 2014, pp. 1-5.

[4] Q. Wen, Z. Yang, Y. Song and Peifa Jia, "Intelligent stock trading system based on SVM algorithm and oscillation box prediction," 2009 International Joint Conference on Neural Networks, Atlanta, GA, 2009, pp. 3341-3347.

[5] C. C. Tseng, "Portfolio management using hybrid recommendation system," e-Technology, e-Commerce and e-Service, 2004. EEE '04. 2004 IEEE International Conference on, 2004, pp. 202-206.

[6] L. Wang and Q. Wang, "Stock Market Prediction Using Artificial Neural Networks Based on HLP," 2011 Third International Conference on Intelligent Human-Machine Systems and Cybernetics, Zhejiang, 2011, pp. 116-119.

[7] Amin Hedayati Mogaddam, Moein Hedayati Moghaddam, Morteza Esfandyari," Stock market index prediction using artificial neural network", Journal of economics, finance and administrative science, vol.21, issue 41, December 2016, pages 89 - 93.

[8] Zhen hu, Jie zhu, Ken tse," Stock market prediction using support vector machine", 6th International conference on information managment, innovation management and industrial engineering, China, 23 - 24 November 2013.

[9] Thira Chavarnakul David Enke "Intelligent technical analysis based equivolume charting for stock trading using neural networks" Expert Systems with Applicationsvol. 34 pp. 1004-1017 2008.

[10] David Enke Suraphan Thawornwong "The use of data mining and neural networks for forecasting stock market returns" Expert Systems with Applicationsvol. 29 pp. 927-940 2005.

[11] Chen An-Sing T. Leungb Mark Daoukc Hazem "Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index" Computers & Operations Researchvol. 30 pp. 901-923 2003.

[12] Kim Kyoung-jae Boo Lee Won "Stock market prediction using artificial neural networks with optimal feature transformation" Neural Comput & Applicvol. 13 pp. 255-260 2004.

[13] F. Ham I. Kostanic,"Principles of neuron-computing for science and engineering", New York NY:McGraw-Hill 2001.

[14] K. Hornik M. Stinchcombe H. Wite "Multilayer Feedforward Networks Are Universal Approximators" Neural Networks vol. 2 pp. 359-366 1989.

[15] Fredric M. Ham "Ivica Kostanic" in Principles of Neurocomputing for Science and Engineering New York NY:McGraw-Hill Science 2000.

[16] P. H. Algoet and T. M. Cover. "Asymptotic optimality and asymptotic equipartition properties of log-optimum investment". The Annals of Probability, pages 876-898, 1991.

[17] T. M. Cover. "Universal portfolios". Mathematical Finance, vol. 1, pages 1-29, 1991.

[18] T. M. Cover and E. Ordentlich. "Universal portfolio with side information". IEEE Transactions on Information Theory, pages 348-363, March 1996.

[19] G. Gorry and G. Barnett, "Experience with a model of sequential diagnosis", Computers and Biomedical Research, 1968.

[20] J. Grass and S. Zilberstein, "Value-Driven Information Gathering", AAAI 97 Building Resource-Bounded Reasoning Systems Workshop, 1997.

[21] P. J. Gmytrasiewicz and E. H. Durfee, "Elements of a Utilitarian Theory of Knowledge and Action", IJCAI, 1993, pp. 396-402.

[22] D. Heckerman, E. Horvitz and B. Middleton, "An Approximate Nonmyopic Computation for Value of Information", IEEE Transaction of Pattern Analysis and Machine Intelligence, 1993.

[23] E. Horvitz and M. Barry, "Display of Information for Time-Critical Decision Making", Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence, 1995.

[24] R. A. Howard, "Information value theory", IEEE Transactions on Systems Science and Cybernetics, 1966.

[25] R. A. Howard, "From Influence to Relevance to Knowledge", Influence Diagrams, Belief Nets and Decision Analysis, 1990, pp. 3-23.

A Compact Modified Square Printed Planar Antenna for UWB Microwave Imaging Applications

Djamila Ziani, Sidi Mohammed Meriah, Lotfi Merad
Faculty of technology, University of Tlemcen
Laboratory of Telecommunications Tlemcen (LTT)
Tlemcen, Algeria

***Abstract*—In this paper, both frequency and time domain performances of a new compact planar antenna for the ultra-wideband (UWB) applications are fully investigated. The proposed antenna has the size of 12x18 mm² providing a fractional bandwidth more than 128123% (3.057 GHz to 13.98 GHz, S11<-10 dB). The results show that the proposed antenna performances in term of wide bandwidth, small size, gain and radiation pattern, transmission coefficient and system fidelity factor are very satisfactory. Moreover, by fabricating and testing the proposed antenna, the simulation results are fairly verified.**

Keywords*—Modified square planar antenna; high bandwidth; return loss; UWB antenna; low profile; frequency and time domain analysis

I. INTRODUCTION

The increased needs of surveillance in different fields led to make the humankind more curious about the unseen and the unknown. In this context, detection through obstacles such as walls, ground or even human tissue has been of a strong interest for the research community. Microwave imaging science has enabled and helped us to resolve these problems. It is actually a good alternative for the other methods of detection [1]. Thanks to their advantages, ultra-wideband is one of the widely used technologies in the microwave imaging systems since the Federal Communication Commission (FCC) allocated the spectrum from 3.1 GHz up to 10.6 GHz in 2002 [2] for this technology communications. It can be defined as any technology using signals that have a bandwidth (BW) greater than 500 MHz or have a fractional bandwidth (FBW) greater than 20% of the center frequency, as described in (1) and (2) [3].

$$BW = f_h - f_l \quad (1)$$

$$FBW\ \% = 2 * \frac{f_h - f_l}{f_h + f_l} * 100 \quad (2)$$

Where f_h and f_l are the higher and the lower frequencies of the operating bandwidth respectively.

As the UWB antenna is the key issue of such systems, plenty of UWB antennas are introduced such as TEM horn, bowtie, spiral antennas and so on [4]-[8]. However, they present some drawbacks in term of bandwidth, size and pulse distortion. For that reason, the necessity of achieving the main requirements of Simplicity, low weight, low cost of fabrication and a wide fractional bandwidth is crucial [9].

Recently micro-strip printed planar antennas attract tremendously many antenna's designers thanks to their interesting characteristics meeting the above criteria, which allow them to be used in emerging UWB applications. This type of antennas consist of a radiating element (patch) which can take several shapes , substrate and ground plane that can be either full or partial. Whereas, they present some limitations in term of the bandwidth, therefore many techniques of bandwidth and impedance matching enhancement are investigated.

In this paper, we propose an UWB small planar antenna for UWB applications that meet the previously mentioned criteria and covers the UWB range. Numerical simulations are conducted to evaluate the antenna performance in both frequency and time domain.

A. Related Work

The design of an UWB planar antenna for imaging system presents a real challenge. The main problem of this types of antennas is their operational bandwidth and their size. In this fact, a trade-off between the antenna size and the bandwidth is needed.

In the literature, there are fully examples of micro-strip patch antennas such as antenna having a semicircular slotted ground plane [10], new UWB monopole antenna with dual band notched [11], diamond antenna [12], antenna which have a rose leaf structure [13], bow-tie slot antenna for breast cancer detection [14], rectangular ultra-wideband antenna for UWB and different wireless applications [15].

Several techniques are studied and exploited in order to broaden the operating bandwidth, including modifications in ground plane [16], slot-loading techniques [17]. In our work, our purpose lies on designing a small planar antenna which have a wide operational bandwidth relying on the advantage of both partial ground plane and the insertion of slots are exploited.

B. Contributions

The high bandwidth is an important factor in UWB applications particularly in microwave imaging. In our case, the antenna is intended for array-based imaging system in medical application. A high bandwidth will ensures a good images resolution [18].

Thus, our contributions are:

- High bandwidth is achieved.

- The proposed antenna provides a good performances in term of gain and the radiation patterns are stable in the desired frequency band.
- The designed antenna has a small size, compact, simple and easy to fabricate which allow it to be integrated easily in array system for medical applications.

This paper is organized as follows: Section II discusses our contributions, describes the antenna design and shows the simulation and experimental results. Sections III and IV are respectively devoted for the frequency and time domain analysis. Finally, Section V concludes the findings of this paper.

II. Antenna Design and Analysis

Fig. 1 presents the geometry of the proposed antenna showing the top and the bottom views of the exponential slot antenna.

The proposed antenna of dimensions (12x18) mm² is designed on the FR-4 substrate with a dielectric permittivity of εr= 3.38 and thickness h = 1.524 mm, with two metallization faces representing the radiating element and a partial ground plane which has a thickness of t = 0.035 mm. The antenna will be fed by a 50 Ohm micro-strip line.

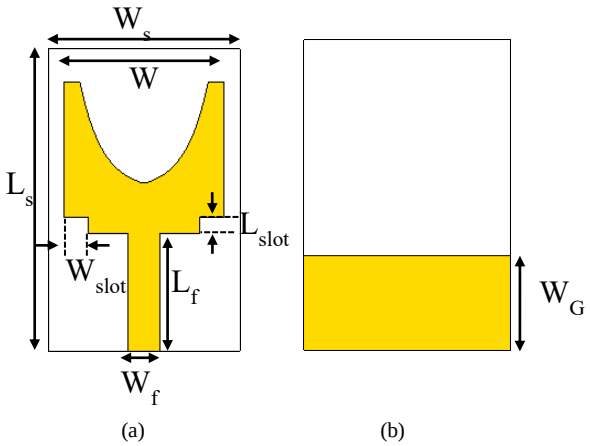


Fig. 1. Geometry of the proposed antenna. (a) Bottom view, (b) Top view.

- The basic antenna design consists of a simple square patch. First of all, a parametric study was conducted on the length of the ground (W_G parameter) of the ordinary square antenna varying from 3.5 mm to 5.5 mm, the best results are obtained for W_G=4.5mm as shown in Fig. 2.

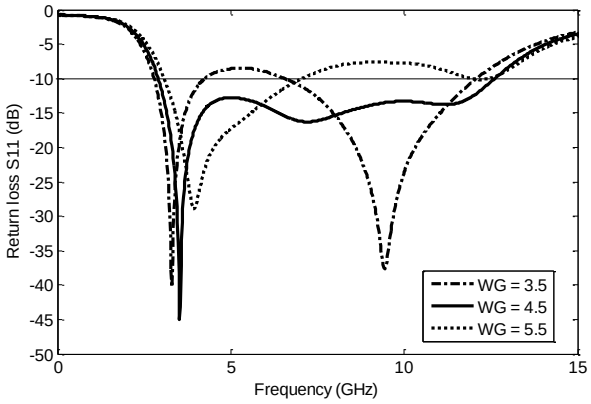


Fig. 2. Return loss parameter for the different values of the ground plane length.

To further enhance the antenna impedance matching and increase the operating bandwidth, we use four rectangular slots which are placed in the edges of the patch and another slot which has an exponential form made with (3). The progress of the design of the antenna, starting from the ordinary square patch to the modified square planar antenna is presented in Fig. 3.

$$y = c_1 . e^{Rx} + c_2 \quad (3)$$

An additional path for surface current is created with the presence of these slots, which leads to producing another resonance, and as a result, increase the bandwidth [19], [20]. The final antenna parameters are illustrated in Table I.

Fig. 4 presents the effect of each variation reaching to the final antenna. It is noticed from the figure that the exponential slot antenna provides a wide fractional bandwidth of more than 128% (3.057-13.98 GHz).

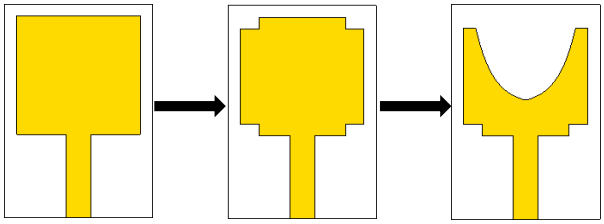


Fig. 3. Progress of the design of the proposed antenna. (Ant. 1) the ordinary square antenna , (Ant. 2) square antenna with rectangular slots and (Ant. 3) final antenna with rectangular and exponential slot.

TABLE I. DIMENSIONS OF THE DESIGNED ANTENNA

Variable	Dimension (mm)
Ws	12
Ls	18
W	10
Wf	2
Lf	7
W_G	4.5
Lslot	1
Wslot	1.5
c1	0.698
c2	9.302
R	0.7

The structure is constructed and numerically simulated with the CST MWS software (Computer Simulation Technology-Microwaves Studio) [21] and fabricated in our laboratory (Fig. 5). The measured and simulated reflection coefficient of the antenna are plotted in Fig. 6, where we observe evidence of wideband characteristic of the design, defined at threshold limit of −10 dB or less. A satisfactory agreement between the measured results and the ones achieved with numerical simulation with small differences, which are probably caused by losses due to the SMA connector used for the measurement taking in account the miniature size of the propose antenna.

The proposed antenna gain is shown in Fig. 7, and as it is apparent the antenna achieves a good gain values in the desired operating frequency band. The maximum gain achieved is over 4dB.

Fig. 8 shows the radiation patterns of the proposed antenna at different frequencies (5, 8.5 and 11.5 GHz) for both E and H planes. The antenna presents acceptable quasi omnidirectional pattern required to receive information signals from all directions.

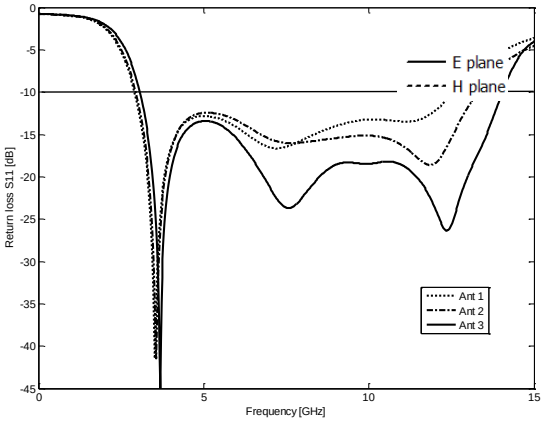


Fig. 4. Return loss parameter for each antenna.

Fig. 5. Fabricated prototype of the antenna (*top and bottom view*).

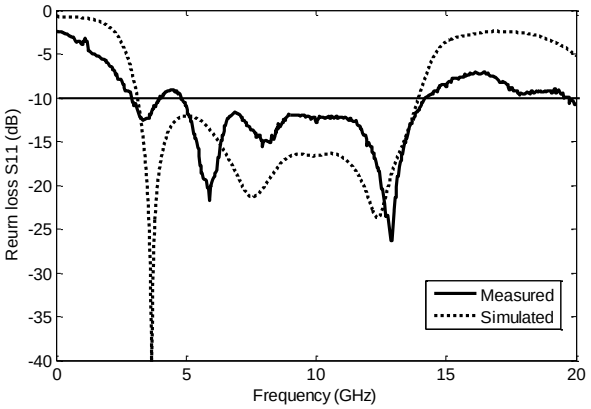


Fig. 6. Measured and simulated return loss of the antenna.

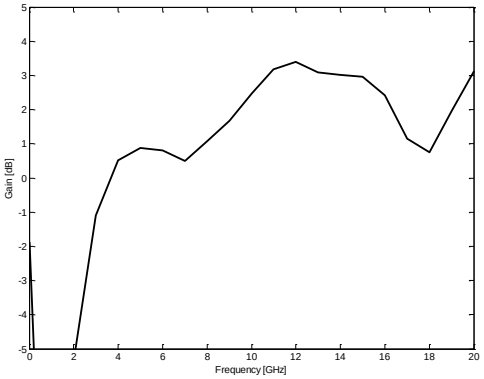
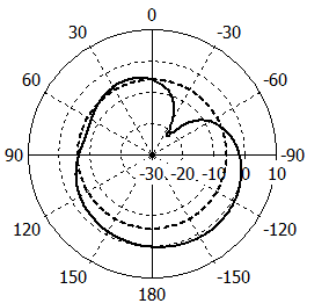
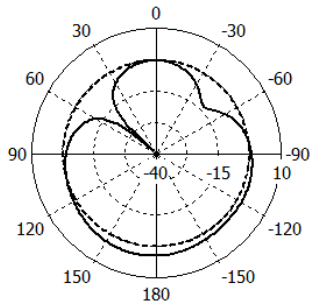


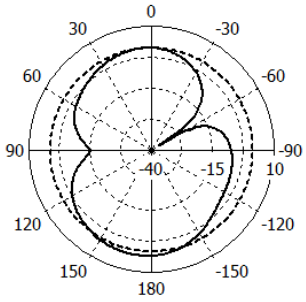
Fig. 7. Antenna gain vs frequency.



(a)



(b)



(c)

Fig. 8. 2-D radiation pattern for E an H planes at 5, 8.5 and 11.5 GHz.

III. Frequency-Domain Analysis

The antenna frequency domain characteristics are summarized in the transfer function (transmission parameter S_{21}) and the group delay [22]. The transmission coefficient's magnitude and phase should be respectively flat and linear and any distortion will cause signal dispersion. We can evaluate the phase distortion by determining the group delay parameter with the relation

$$Group\ delay\ = -d\varphi/df \qquad (4)$$

Where φ is S_{21} phase. A constant group delay results to in linear S21 phase.

For that reason, a system of two identical antennas is considered. The antennas are separated by d=10 cm from each other (to have each antenna in the far-field of the other one) and placed in two different orientations, the face to face and side by side orientations (Fig. 9); one of the antennas is the transmitter and the other one act as the receiver.

Fig. 10 presents the simulated transfer function and the group delay for each orientation. It can be seen that the transfer function is flat and the phase is fairly linear over the entire frequency band.

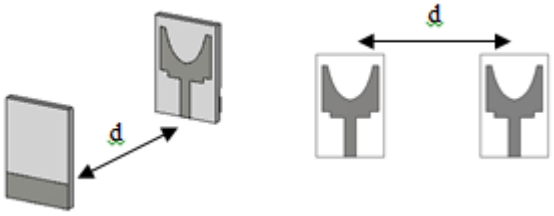
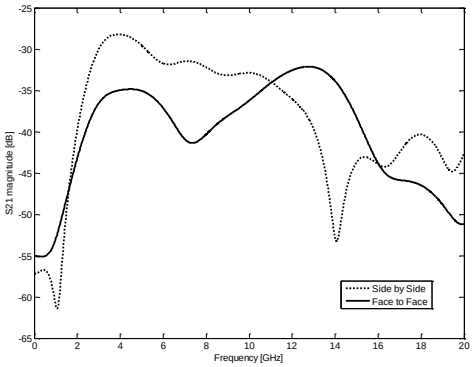
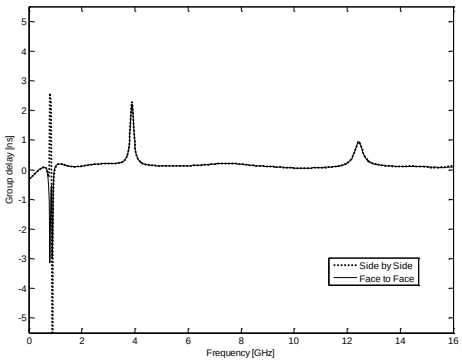


Fig. 9. Antennas orientations (a) face to face (b) side by side.



(a)



(b)

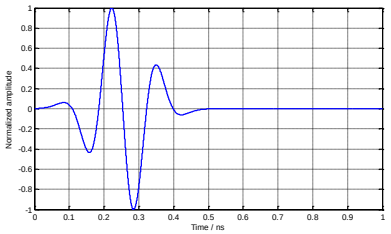
Fig. 10. (a) Magnitude of S21, (b) Group delay.

IV. Time-Domain Analysis

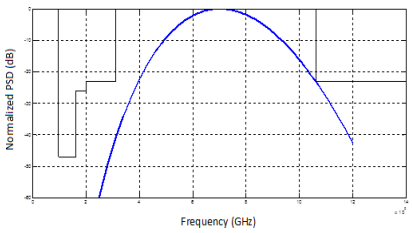
For ultra-wideband antennas, evaluating the time domain behavior and studying the dispersion is very important [11]. For that, a reference signal which meets the FCC mask covering the range from 3.1 up to 10.6 GHz, should be applied at the antenna input [9]. The optimal pulse is the one suggested in [23] which has a good corresponding to the FCC mask and it can be obtained by the Gaussian 5th derivative (Fig. 11). Mathematically it can be given by (5):

$$y(t) = A\left(-\frac{t^5}{\sqrt{2\pi}\sigma^{11}} + \frac{10t^3}{\sqrt{2\pi}\sigma^{9}} - \frac{15t}{\sqrt{2\pi}\sigma^{7}}\right) exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad (5)$$

Where A is a constant chosen to meet the limitations set by the FCC. To cover the right frequency band, the value of σ was set at 50.788 ps.



(a)



(b)

Fig. 11. The representation of the Gaussian 5th derivative (a) in time domain, (b) spectrum compared with the FCC mask.

The system fidelity factor is one of the most important factors to evaluate the antenna performance [24], [25]. This factor can be defined as the maximum magnitude of the correlation coefficient between the received and the transmitted signals, in fact it can judge the similarity between the two signals. In general, a value less than 0.5 (50%) of the SFF will deliver a high distortion. The system fidelity factor can be evaluated using (6) [26]:

$$SFF = \max_{\tau} \frac{\int_{-\infty}^{\infty} s_t(t) s_r(t-\tau) dt}{\sqrt{\int_{-\infty}^{\infty} |s_t(t)|^2 dt . \int_{-\infty}^{\infty} |s_r(t)|^2 dt}} \quad (6)$$

Where $s_t(t)$ and $s_r(t)$ are the transmitted and received signals respectively. The normalized transmitted and received signals in the time domain are presented in Fig. 12. System fidelity factor is also calculated according to (6) using Matlab software for the two proposed orientations, and the results are summarized in Table II. We can see that a better system fidelity factor SFF is achieved (more than 92%). Therefore, the antenna has no any considerable dispersion.

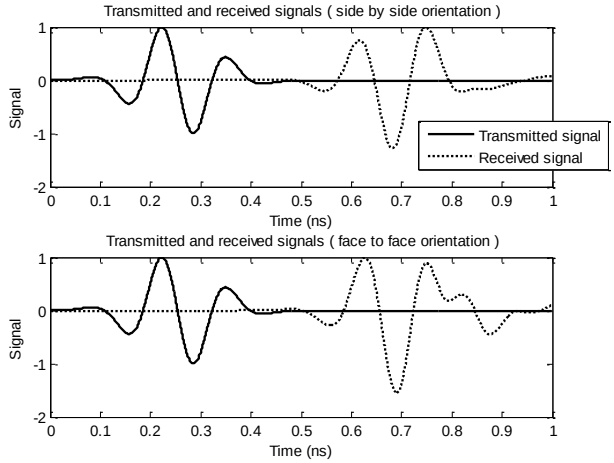


Fig. 12. Normalized transmitted and received signals.

TABLE II. SYSTEM FIDELITY FACTOR OF THE TWO IDENTICAL ANTENNAS BETWEEN THE TRANSMITTED AND RECEIVED SIGNALS

Variable	Dimension (mm)
Side by Side	0.9525
Face to Face	0.9246

V. CONCLUSION AND FUTURE WORK

In this letter, a compact printed planar UWB antenna for radar and microwave imaging is proposed. The simulation results show that the performance of this antenna meet the desired requirement in terms of return loss parameter, gain, wide bandwidth and radiation pattern. The proposed antenna was fabricated and measured. Good agreement was obtained between simulated and measured data, covering the range from 3.057 GHz to 13.98 GHz (128%). The antenna presents a good performance in both time domain and frequency domain, where it exhibits a good system fidelity factor, a flat transfer function and a linear group delay, moreover, it has a low profile (< 2 cm) that can be easily integrated in arrays reducing their dimensions. The results of this antenna confirm that it is suitable for UWB applications. In future work, we aim to use the proposed antenna to design a Radar system for medical imaging applications able to offer a good images resolution benefiting of the UWB characteristis of the our antenna.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge Pr. SALAH Belkhoudja from Telecommunications and Digital Signal Processing Laboratory, University of Sidi Belabbes for his technical support in providing the experimental data.

REFERENCES

[1] S. M. Chouiti, L. Merad, S. M. Meriah, X. Raimundo, A. Taleb-Ahmed, “An Efficient Image Reconstruction Method for Breast Cancer Detection Using an Ultra-Wideband Microwave Imaging System,” Electromagnetics, Vol. 36, No. 4, pp. 225-235, 2016.

[2] Federal Communication Commission, “Revision of Part 15 of the Commission's Rules RegardingUltra-Wide-Band Transmission SystemsFirst Report and Order,” FCC 02-V48.Tech.Rep, Washington,DC, USA, 2002.

[3] Z. N. C. a. M. Y. W. Chia, “Broadband Planar Antennas, “Design and Applications,” John Wiley & Sons, Ltd, pp. 180-190, 2006.

[4] K. L. Shlager, G. S. Smith, and J. G. Maloney, “Accurate analysis of TEM horn antennas for pulse radiation,” IEEE Trans. Electromagn. Compat, Vol. 38, No. 3, 1996.

[5] S. C. Hagness, A. Taflove, and J. E. Bridges, “Wideband ultra low reverberation antenna for biological sensing,” Electronics Letters, vol. 33, no. 19, pp. 1594–1595, 1997.

[6] G. E. Atteia and A. A. Shaalan, “Wideband partially-covered bowtie antenna for ground-penetrating-radars,” Progress In Electromagnetics Research, PIER 71, 211–226, 2007.

[7] T. W. Hertel and G. S. Smith, “Analysis and design of conical spiral antennas using the FDTD method,” *Proc. of IEEE Antennas and Propagation Society Int. Symp*., Vol. 3, 1540–1543,2000.

[8] J. Ali, N. Abdullah, M. Y. Ismail, E. Mohd and S. Mohd Shah “Ultra Wideband Antenna Design for GPR Applications: A Review,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8,No. 7, pp. 392–400,2017.

[9] M. Ershadh, P. Krishna, Bhagyaveni., S. Subramanian, “Design of a Novel Antenna and Its Characterization in Frequency and Time Domains for Ultra Wide Band Applications,” Progress In Electromagnetics Research, PIER, Vol. 48., 69-76, 2014.

[10] N. Seladji, F. Z. Marouf, L. Merad, S. M. Meriah, F. T. Bendimerad, M. Bousahla, N. Benahmed, “Antenne Microruban Miniature Ultra Large Bande ULB pour Imagerie Micro-onde”, Mediterranean Telecommunication Journal, Vol. 3, No. 1, pp. 21-25, february 2013.

[11] M. Mighani and M. Akbari, “New UWB Monopole Antenna with Dual Band Notched,” Progress In Electromagnetics Research C, Vol. 52., 153-162, 2014.

[12] D. Ziani Kerarti, S. M. Meriah, "New Monopole Antenna for Ultra Wideband Applications", International Journal of Computer Applications, Vol. 47, No. 11, pp. 40-43, 2012.

[13] A.A Lotfi Neyestanak "Ultra WideBand Rose Leaf Microstrip Patch Antenna" Progress In Electromagnetics Research, PIERS 86,pp. 155-168, 2008

[14] N. Seladji-Hassaine, L. Merad, S. M. Meriah, F. T. Bendimerad, "UWB Bowtie Slot Antenna for Breast Cancer", World Academy of Science, Engineering and Technology, International Journal of Biomedical and Biological Engineering Vol. 6, No. 11, pp. 571-574, 2012.

[15] Z. Ul Abedin and Z. Ullah, "Design of a Microstrip Patch Antenna with High Bandwidth and High Gain for UWB and Different Wireless Applications", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8,No. 10, pp. 379–382,2017.

[16] N. PrombutrH, HP. KirawanichH, and P. Akkaraekthalin "Bandwidth Enhancement of UWB Microstrip Antenna with a Modified Ground Plane" International Journal of Microwave Science and Technology, Volume 2009 (2009),

[17] F. Amini and M.N. Azarmanesh, and M. Ojaroudi, Small semicircle-like slot antenna for ultra-wideband applications, Prog Electromagn Res C 13 (2010), 149–158.

[18] A.G. Yarovoy, L.P; Lighart, "UWB Radars: Recent Technological Advances and Applications", IEEE radar conference, pp. 43-48, April 2007.

[19] S. Maci and G. B. Gentili, "Dual frequency patch antennas," IEEE Antennas Propag. Mag., vol. 39, no. 6, pp. 13–20, Dec. 1997.

[20] S. Maci and P. Piazzesi, "Dual-band slot-loaded patch antenna," Proc. Antenna Propag., vol. 142, no. 3, pp. 225–232, Jun. 1995.

[21] CST Microwave Studio, CST Inc .,2014.

[22] A. Mehdipour, K. Mohammadpour-Aghdam, R. Faraji-Dana, "Complete Dispersion Analysis of Vivaldi Antenna for Ultra Wideband Applications," Progress In Electromagnetics Research, 85-96, 2007.

[23] H. sheng, P. Orlik, A.M. Haimovich, L.J. Cimini, Jr, J. Zhang, "On the Spectral and Power Requirements for Ultra-Wideband Transmission," IEEE International Conference on Communications (ICC), December 2003.

[24] G. Quintero, J.F. Zürcher, A.K. Shrivervik, "System Fidelity Factor: A New Method for Comparing UWB Antennas," IEEE transaction on antennas and propagation, Vol. 59, July 2011.

[25] M. Koohestani, A. A. Moreira, A. K. Skrivervik, "Fidelity Concepts Used in UWB Systems" *IEEE Antennas and Propagation Society Int. Symp*,2014.

[26] L. Akhoondzadeh-Asl, M. Fardis, A. Abolghasemi, G. Dadashzadeh, "Frequency and Time Domain Characteristics of a Novel Notch Frequency UWB Antenna," Progress In Electromagnetics Research, 337-348, 2008.

Time-Dependence in Multi-Agent MDP Applied to Gate Assignment Problem

Oussama AOUN, Abdellatif EL AFIA
Operations Research and Logistics Team
ENSIAS Mohammed V University
Rabat, Morocco

***Abstract*—Many disturbances can impact gate assignments in daily operations of an airport. Gate Assignment Problem (GAP) is the main task of an airport to ensure smooth flight-to-Gate assignment managing all disturbances. Or, flights schedule often undergoes some unplanned disruptions, such as weather conditions, gate availability or simply a delay that usually arises. A good plan to GAP should manage as possible stochastic events and include all in the planning of assignment. To build a robust model taking in account eventual planning disorder, a dynamic stochastic vision based on Markov Decision Process theory is designed. In this approach, gates are perceived as collaborative agents seeking to accomplish a specific set of flights assignment tasks as provided by a centralized controller. Multi-agent reasoning is then coupled with time dependence aptitude with both time-dependent action durations and stochastic state transitions. This reflection will enable setting up a new model for the GAP powered by a Time-dependent Multi-Agent Markov Decision Processes (TMMDP). The use of this model can provide to controllers at the airport a robust prior solution in every time sequence rather than bringing a risk of online schedule adjustments to handle uncertainty. The solution of this model is a set of optimal decisions time valuated to be made in each case of traffic disruption and at every moment.**

Keywords*—*Time-dependent Multi-Agent Markov Decision Processes; stochastic programming; flight delays; Gate Assignment Problem

I. INTRODUCTION

More interest in recent years is allowed to providing advanced techniques in the air traffic framework. This is resulting from the increase of air transport traffic [1]. The main objectives are best allocation and management of airport and airline resources in the best way effectively and efficiently. Caused by the dynamic stochastic operational environment conditions of air transport, the scheduling problems currently confronted by the airport and airline managers are leading to challenging and complex planning problems that involve innovative models and solutions. This is triggered by the significant diversity of resource segments that have to be regarded including terminals, flights, crews, baggage …, and most are interdependent. In fact, stochastic disruptions in air traffic transport raised the complexity of the resolution models. This is progressively more taken under consideration in most recent studies.

The main target of an airport is to guarantee a fluent flights traffic. Optimal assignment of aircraft guaranteed to make available over time the proper gates. If an aircraft is not assigned, it will be forced to wait on the ramp very well as in the air; This type of scenarios are quite undesirable on account of time wasting and let to flight delays. Also, ramps and airspace are as well resources with a limited capacity.

Gate flight assignment is an essential task of an airport; it is the primary activity in airline traffic transport management [2]. Moreover, several airports today have severe capacity constraints resulting from the increase in air traffic volume. The GAP can be regarded as such a problem of constraint resource assignment, in which gates represent resources and aircraft considered as resource consumers.

Furthermore, GAP is thought to be a challenging problem [3] since it includes very inter-dependent resources integrating aircraft, crews, and gates. Therefore, severe disruptions in the airport manifested as flight delays are caused by inadequate assignment, which reduces the customer services and produces inefficient use of gate services and conflicting flights.

Various circumstances could potentially cause stochastic disruptions in gate assignment; it can possibly be interrelated to possible gate dysfunction, a flight delay or earliness, extreme weather conditions, or for any more causes. This type of daily disturbances might reduce the overall performance of the currently assigned gates once associated with actual operations. Therefore, even a unique variation in a single flight plan could engender a series of disturbances for additional aircraft, which have been designated to the same gate. This sort of phenomenon is very unwanted in airline operations due to its noticeable costly impact.

Various GAP models and techniques are identified from the literature. Static as well as stochastic models are developed. Working with methods with an exact solution can be obviously more suitable. However, [4] states that these kinds of exact methods are actually ineffective to resolve real problems. This is because flights in static models are allocated to gates depending on the expected flight schedule using fixed parameters. Nonetheless, in real operations, stochastic disruptions occur frequently, leading to real-time adjustments of gate assignments and flight delays. Consequently, stochastic methods have been widely motivated in recent researches.

Consequently, to build a significantly better gate flight assignment approach, it has to include in the model the possibilities of stochastic flight delays that may arise in real operations.

When it comes to stochastic environments, Markov Decision Processes (MDPs) [5] have confirmed to be effective in optimal decision making. A derived version of MDPs called multi-agent Markov decision process [6] was developed to manage some challenges in the standard GAP based MDP firstly introduced. In this work, collaborative multi-agent based MDPs is built, which are composed of multiple agents attempting to produce the best allocation of aircraft to gates. A new methodology for GAP is provided and regarded as a multi-agent problem that includes robustness for a stochastic disturbance. The GAP is therefore designed as a Multi-agent MDP that is intended to resolve within the assumption of environment uncertainty the GAP. Then, incorporating Time dependence to the developed Multi-Agent model enables further stochastic planning ability. In this method, the stochastic feature is considered caused by flight delays with the flexibility to consider additional constraints for the constructed model. Gates are designed as agents having a centralized controller. Consequently, individual agent or gate possesses full visibility of airport operations and so can be aware of flights allocated to every gate at the time of planning horizon. Built policies take into consideration the time dimension. Time-dependent Multi-agent Markov Decision Processes allows more real illustration of the Gate problem with rewards and transitions varying with time. So TMMDP includes Multi agent aspect coupled with a real-valued time component.

This paper is structured as follows: the next section provides a literature review. Section 3 shows the MMDP model for the GAP. Section 4 presents the model of GAP in stochastic circumstances with time dependence. Furthermore, Section 5 will provide experimentations, and at last, conclusion and perspective are dressed.

Airport Gate Assignment Problem is referred to as setting an appropriate gate for each arriving aircraft to the airport until the time of its departure. It is one among the primary components in what concerns the management of airport resources. Gates, being a resource, is subject to the next two groups of constraints as categorized in literature: strict and soft constraints (see [3]).

The first category is obligatory to represent the problem of gate assignment. It comprises these constraints:

- Single: Each aircraft have to be assigned just to one gate.
- Feasible: A single gate could be assigned to one single aircraft simultaneously.

Soft constraints are various and can possibly be related to either airlines or airports. Mainly common among constraints in the literature is about to minimize the total walking distances within passenger transfer. (e.g. [7]), or just like assigning aircraft to some specified gates, also taking into consideration the size aircraft for allocating the gate [8]. It can also be minimizing the number of aircraft obligated to wait for a gate.

There are several objective functions in GAP models. Some notable ones of literature are cited. These functions are like minimizing the total walking distance [7] or the total waiting time for passengers such as in [9] or also minimizing the number of un-gated aircraft in [2]. Others like minimizing the current schedule modification from an initial schedule, or also maximizing the preferences of assigning particular aircraft to individual gates (e.g. [10]) and minimizing gate conflict in [2]. In this paper, the stochastic model will implement particularly the last one to minimize conflicting assignment due to flights disruptions.

The GAP formulation is classified into two main types: deterministic and stochastic models. In the first kind, just static parameters are regarded (including passengers, gates, number of flights…); due to stochastic perturbations in real-world operations deterministic models becomes infeasible. Stochastic GAP models have been investigated to consider those disruptions in air traffic into concern such as flight delays or some sever weather conditions.

Deterministic models are more a lot discussed in the literature, such as [7]. Most have as an objective the minimization of the total passenger-walking distance. Lately, stochastic and robust models are more reviewed assisting operators to act in response to possible uncertain events.

To illustrate stochastic and robust GAP resolutions in literature, [11] displays that having a planned buffer time into the flight schedule can increase schedule punctuality. In [12] and [9], they use in their GAP a fixed buffer time among two consecutive flights assigned to the same gate in order to absorb the possible stochastic flight delays. In [12] author produces a multi-commodity network flow approach as well as in [13]. In [14], author builds up a heuristic approach sensitive to stochastic flight delays in a framework that consists of three components, a stochastic gate assignment model, then a real-time assignment rule, in addition to two penalty correction methods.

In [2], GAP is modelled as a stochastic programming model and altered it into a binary programming model; the resolution contains hybrid meta-heuristic, a tabu search, and a local search. Also, an ant system combined with a local search in [15] has been used to an over-constrained airport Gate Assignment Problem with the interest of choosing and allocating aircraft to the gates minimizing the total passenger interconnection.

Recently, a model based heuristics of Mixed Integer Programming in [16] has been presented, it has been confirmed to be more efficient when compared to the linearized models, and more robust. Likewise, a multi-objective optimization model of GAP has been offered in [17], a particle swarm algorithm for resolution is used for resolution, which gives an improved comprehensive service of gate assignment regarding robustness. Applying also a metaheuristic for resolution, authors in [18] designed a three-objective problem to the GAP and using a non-dominated sorting genetic algorithm for resolution.

Markov process theory, in general, has been proven for application in airline transport like in [19]. Notably, the use of MDP model for GAP has been applied in [20] to deal with gate disturbances with consideration of aircraft size in the assignment, where neighboring gates can just only accept

aircraft of a specified size or are possibly blocked. A most recent robust GAP with multi-agent MDP model has been provided in [21].

In a similar idea of incorporating stochastic disturbance for establishing gate assignment, a multi-agent system with time dependence for modeling with time dependence is used in this paper. Multi-agent systems (SMA) are a part of Distributed Artificial Intelligence. Their applications are large: game theory, humanities, economics, and other real-world applications including air traffic control, robotics, and networking. SMA methods are interested in connections between independent entities. This circumstance is mainly examined in SMA as the cooperation that requires complex components.

In planning with multi-agent systems, it is commonly supposed to possess are some number of agents, each one with their individual group of actions, and a provided tasks to be solved recognizing that interaction with the other agents is essential. Reinforcement learning has been a practical methodology to construct coexisting agents (e.g. [22]) as well as Markov games (see, e.g. [23]). In general, each agent may possess its personal goals. In this paper, the concern is given to the case of fully cooperative agents; where all of the agents have a shared similar goal to maximize the total expected reward. In particular, where agents are autonomous and distributed, a local Markov Decision Process (MDP's [5]) is used to express every single agent's state and actions space. Therefore, the utility of any given system state is similar for all agents, and with models of uncertainty and general utility, Multi-agent Markov decision process (MMDP) is developed by [24] to incorporate such numerous adaptive agents that interact to compute some given goals. MMDP has been applied in various domains as well as in the air transportation (see [25]).

MMDP is the basis of full observability of the global state by every single agent; it is designed as a set of interacting learners agents, which are autonomous. These agents have to learn in order to cooperate and obtain their assigned goal. It can also either centralized or decentralized in term of decision-making main feature [26]. Hence, this paper incorporates Markov decision processes as a formalism in the multi-agent structure (e.g. [24]). It supposes having a centralized controller knowing all information regarding the system (Fig. 1), including actions, the global state of the system, and rewards; thus the controller possesses the decision authority and keeps information distributed among agents.

Multi-Agent notion can as well be combined with real-time valued to include time evolution into the multi-agent system dynamics. A Time-dependent Markov Decision Process (TMDP) is provided by [27] to give this extension. This model is composed of stochastic state transitions and as well as stochastic time-dependent action durations. The actions in TMDP model are stochastic and time-varying:

$$a(t) \sim policy(s, a(t)) \quad (1)$$

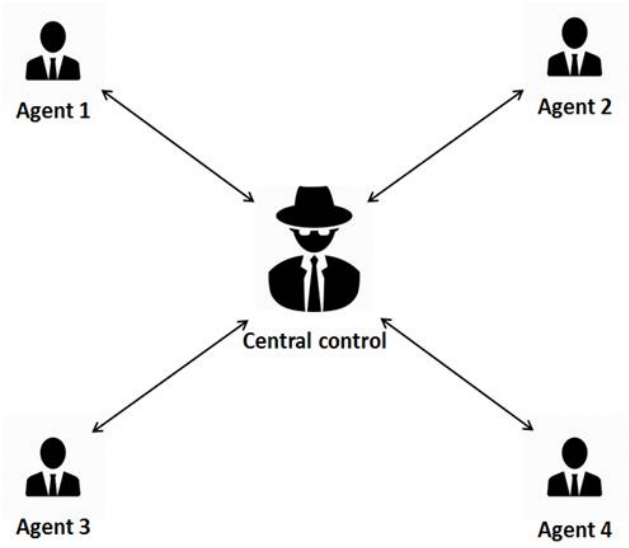


Fig. 1. Centralized control in MMDP.

Resulting policies are actions to be performed by agents in every single time sequence. Then, the real planning window can be widespread to problems under uncertainty changing with time.

So, in this formulation as in [28], first, MMDPs consider an assignment centered decomposition approach, which is intermediate between the join MDP method and the method of independent agents. The centralized controller is adopted having the complete relevant information regarding the states of all agents to allocate jobs and assign jobs and resources to agents determined by a task level value functions associated with agents. After the jobs are allocated to agents, the particular lower level actions of agents are driven by the task level value functions till the primary controller reassigns jobs. Then, adding time dependence behavior will give a more realistic representation of the Gate Assignment Problem, inspired by TMDP and coupled with the MMDP approach providing a new formalism of time-dependent Multi agent MDP. This method will help us to have real-time policies to apply in every case of disturbance for the GAP problem.

II. Theoretical Background

Giving the theoretical knowledge, Markov Decision Processes (MDPs) are defined (see [5]), and then generalized to multi-agent settings. Then, the basic model of Time-Dependent Markov Decision Process (TMDP) (given by [27]) is provided to finally conclude a new extension of MMDP depending on time and formalize the Time-Dependent Multi-Agent Markov Decision Process (TMMDP).

A. Standard Markov Decision Process

Considerably, more research interested in problems having uncertainty in the planning with possibly conflicting objectives. As a tool of artificial intelligence (AI) planning, decision-theoretic dress those challenges, especially, Markov Decision Processes Theory (MDPs). It finds a significant attractiveness in recent researches equally as a computational and conceptual model. MDP is defined by a tuple $< S, A, P, R >$ where s is a finite set of states S describing

systematic interests, a finite set of actions A featured to the agent, and then a reward function R. When an action can take an agent from one state to a second one, the results of actions is uncertainty described by the probability P considered as transition model. A mapping $\pi : S \rightarrow A$ defines a policy. The objective is to identify the optimal policy $\pi*$ maximizing per each state the expected discounted future reward. MDP is considered in this paper to possess an infinite horizon with exponentially discounted future rewards by a discount factor $\gamma \in [0, 1)$.

B. Multi-Agent Markov Decision Processes

The MDP model can be extended to multi-agent systems to define The Markov Decision Processes Multi-agent or MMDP as in [6]. In this formalism, the same goal of maximizing the total expected reward is shared for all agents having the same joint utility function. MMDP can be viewed as a generalization of MDP with a single agent; Or, but also a special case of Markov games [29] where the payoff function is identical for all agents. Let define first the MMDP formalism before offering it as a useful framework to constitute a new GAP model.

A MMDP is identified via a tuple <n, S, A, P, R>. Where each one action is identified by the set of actions of all single agents, it constitutes a joint action. Each element is defined as :

- n: the total number of agents in the system.
- S: refers to the set of states S.
- $A = A_1 \times \cdots \times A_n$: identifies the set of joint actions of all agents, A_{g_i} defines the set of local actions designed for the agent i.
- P defines the transition function; it provides the probability $P(s, a, s')$ of the system moves from a state s into a state s' once agents run the joint action $a \in A$.
- R identifies the reward function. $R(s, a, s')$ is the reward received after moving from a state s to a state s' performing an action a.

Solving a MMDP is about determining a joint policy $\pi = < \pi_1, \ldots, \pi_n >$. Where π_i corresponds to the policy of a local agent. It identifies a function $\pi_i : S \rightarrow A_i$ that gives a mapping to any system state to the action of agent i. The joint policy will be computed applying the standard algorithm the Value Iteration (continue to operating in the general situation of decentralized agents, see [18]).

C. Time-Dependent Markov Decision Processes

In standard previously defined MDPs, transitions and rewards are thought to be stationary functions; they do not undergo any change during decision epochs. In literature, some approaches like [30] define Stochastic Time-Dependent Network where stochastic transition durations are included, but transition outcomes are deterministic. A model given by [27] is one of the first models to focus on time as an independent observable state variable; it is named as Time-dependent Markov Decision Process.

Time-dependent Markov Decision Process extends the Markov decision process model where a continuous observable time dimension is contained in the state space. The added time variable allows more real representation of large problems with transitions or rewards time-varying. So TMDP includes problems with following properties:

- State transitions are stochastic;
- Time-dependent action durations are stochastic.
- Rewards are Time-dependent.

In the TMDP model, each transition, which arises from making an action, is decomposed into a set of possible outcomes $\{\mu\}$. Every single outcome identifies both a transition duration and a resulting state.

The TMDP model decomposes each transition resulting from the application of action into a set of possible outcomes $\{\mu\}$. Each outcome describes a resulting state and transition duration.

Formally, the TMDP is defined as in [27] by:

- S: Discrete space state.
- A: Discrete action space.
- M: Discrete set of outcomes, of the form $\mu = (s'_\mu, T_\mu, P_\mu)$:
  - $s'_\mu \in$ S: is the resulting space
  - $T_\mu, \in$ {ABS, REL}: identifies the type of the resulting time distribution (if it is absolute or relative)
  - $P_\mu(t')$ (If T μ = ABS): probability density function (pdf) over absolute arrival times of μ
  - $P_\mu(\delta)$ (If T μ = REL): probability density function over durations of μ
- L: $L(\mu|s, t, a)$ is the likelihood of outcome μ given action a, state s, and time t
- R: $R(\mu, t, \delta)$ is the reward associated to outcome μ at time t with a duration δ

In the figure below (Fig. 2), it shows a simple graphic representation of TMDP evolution.

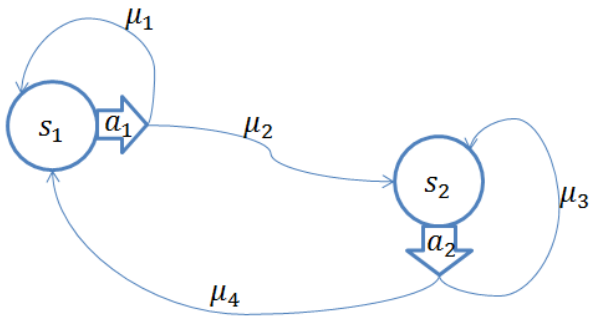


Fig. 2. Elementary example of TMDP.

In TDMDP and at time t, if in a state s_1 agent executes an action a_1 , it will be generated outcome $\mu 1$ by certain probability $L(\mu_1|s_1, t, a_1)$ and an another outcome μ_2 by a probability $L(\mu_2|s_2, t, a_2)$. μ_2 represents the transition to s_2 and P_{μ_2} gives the transition absolute arrival time, while μ_1

represents the return to s_1 (failure to leave s_1) with a duration P_{μ_1}. Implicitly, a waiting time is inserted before each action in the model.

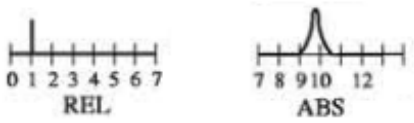


Fig. 3. Representation of pdf types.

The likelihood functions L governs possible outcomes in the model. Time distributions in a TMDP could be either "relative" (REL) or "absolute" (ABS) as shown as an example in Fig. 3.

The TMDP model can be represented by the Bellman equations below:

$$V(s,t) = \max_{a \in A} Q(s,t,a)$$

$$Q(s,t,a) = \sum_{\mu \in M} L(\mu|s,a,t).U(\mu,t)$$

$$U(\mu,t) = \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t'-t) + V(s'_\mu,t')]dt' \quad (2)$$

$$(if\ T_\mu = ABS)$$

$$U(\mu,t) = \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t'-t) + V(s'_\mu,t')]dt'$$

$$(if\ T_\mu = REL)$$

Where

$U(\mu,t)$: Utility associated to the outcome μ in time t

$V(s,t)$: Time-value function of the immediate action

$Q(s,t,a)$: Expected Q time-value through outcomes.

The resolution of this model is performed using Bellman equations, (2) representing an undiscounted continuous-time MDP. At each state, the optimal time-value function is a piecewise linear function of time, which could be precisely calculated by value iteration [27]. The TMDP model is more general than semi-Markov decision processes [31] that have no notion of absolute time. With absolute time included in the state space, comprehensive set of domain objectives can be modeled beyond the objective to minimize expected time, like for example the probability of designing a deadline. Actually, the variable time dimension may represent further quantities; it can consider planning with the non-linear utilities, or also with continuous resources.

D. Time-Dependent Multi-Agent Markov Decision Processes

Based on the two previous definitions of MMDP and TMDP, a new formalism is defined combining between those approaches. So, it is called Time-Dependent Multi-Agent Markov decision process TMMDP. This is a MMDP seen as cooperative multi-agent systems as in [6] or associated with a time dependence capabilities as defined by [27]. MMDP is then extended to take a continuous observable time dimension contained in the state space. Supposing time variable is common between agents, a global time is associated to all agents.

A TMMDP is defined by:

- n : Number of agents.
- S : refers to the set of states s
- $A = A^1 \times \cdots \times A^n$: The set of joint actions for the agents i is the set of local actions of the agent A_{g_i}.
- M: Discrete set of outcomes, of the form $\mu = (s'_\mu, T_\mu, P_\mu)$:
  - $s'_\mu \in$ S: the resulting space
  - $T_\mu, \in$ {ABS, REL}: identifies the type of the resulting time distribution (absolute or relative)
  - $P_\mu(t')$ (If T μ = ABS): pdf (probability density function) over absolute arrival times of μ
  - $P_\mu(\delta)$ (If T μ = REL): pdf over durations of μ
- L: $L(\mu|s,t,a)$ is the likelihood of outcome μ given join state s, time t and join action $a = (a_1, \ldots, a_n)$.
- R: $R(\mu,t,\delta)$ Reward attached to outcome μ at time t for all agents with duration δ.

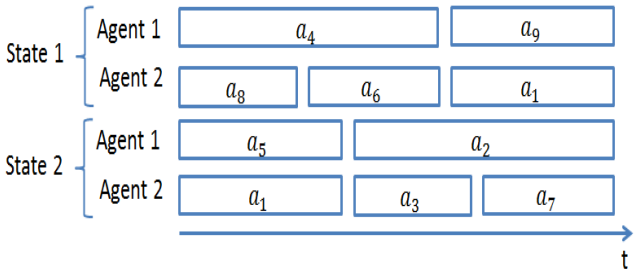


Fig. 4. TMMDP policy representation

The aim of defining TMMDP formalism is to model and solve large real problems of planning under uncertainty taking into account either cooperative agent property and time evolution. Resulting policies are actions to be performed by agents in every time sequence (see Fig. 4).

III. The Proposed Approach

A. Multi-Agent Reasoning

Various efforts made in the literature to manage uncertainty (see Section 2). With the Objectif to build a robust, a multi-Agent based method is selected to develop a solution that can resist the most to flights disturbances. The choice for this specialized background to model the problem. MAS methods are getting large approval being an effective instrument to solve more complex problems and then designs a promising alternative. As well, many advantages related to multi-agent reasoning such as the distribution of processing, which made some type of problems more simple in conception. Additionally, it provides an intelligent alternative to complex problems and logical approach of decomposing into individual agents that cooperate.

In this paper, MAS is considered to be managed by a centralized controller, and the solution is composed of all possible decisions that could be taken within the planning horizon of gates to flights assignment. Therefore, This approach supposes there is no need to take real-time optimization since it is predetermined the solutions for all likely case of disturbances. Hence, for any provided gate assignment combination, the solution offers the best decision of gate allocation to make.

B. Time-Dependence Behavior

The real interest is given to sequential decision problems. Theoretical aspect based on MDPs gives a best well-known tool to model and solve them, giving optimal results. However, real-world problems have additional and specific behavior, which is time dependence. MDP reflects only fixed time steps between decision epochs, which can be easily modeled as iteration steps. This property does not reflect the real evolution of problems like the subject of gate assignment. To bypass this limitation, Time-dependent MDP (TMDP) has been proposed in those models (see the previous section), the transition between states is not instantaneous but proceeds in specific time t. Also in TMDP, the time is always observable, optimal policies give to the agent the best moment to make a decision or execute an action due to the state of the system.

Inspired by other occurrences like the truck dispatching system where decisions about truck assignments and destinations are made in real-time [32], choosing to benefit from temporal aspect and to project it to Gate Assignment Problem. Therefore, the rewards associated with action outcomes in the time-dependent frameworks will be represented as time-dependent functions including more real evolution information of the problem.

C. Multi-Agent Model for the GAP:

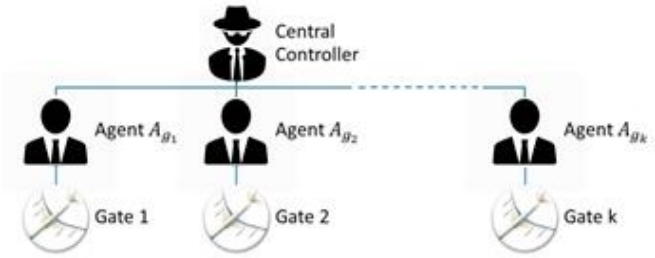


Fig. 5. Agents representation

Before extending the model of GAP to be time-dependent, an earlier formulation like in [21] of the Gate Assignment Problem with MMDP is presented. The model is given by is a tuple <K,S, A, P, R> as a follow (see Fig. 5):

The State $S = S_1 \times \cdots \times S_K$ is a vector giving the diverse feasible combinations of flights indexed by its assignment position $S_i = (s_1, \ldots, s_k)$, where k is number of gates and $s_i \in V$. V represents the set of flights to be allocated to gates during the planning horizon (one day in general).

The set of actions $A = A_1 \times \cdots \times A_K$ describes the set of joint actions for the agents, A_i gives the set of local actions of the agent i. For each single agent, performing $a \in A_i$, will match an action of allocation a flight $a \in V$ to the gate i.

Therefore, each agent is in charge of handling a particular gate, and $a \in A_i$ for agent i considers that there is a set of feasible flights to be affected to the gate i. $A_i \subset V$ that are appropriated to be allocated to gate i. This supposition regarded as a feasibility constraint that describes the possible assignment.

Defining: $A_{i,t}$ set of feasible flights for the gate i at a discrete time t. Then:

$$A_i = \sum_t A_{i,t} \quad , \quad (i \le k) \tag{3}$$

$P(s, s', a)$ gives the probability of transition as :

$$P: \ S \times S \times A \to [0, \quad 1]$$

It represents the probability of the going from state s into another state s' when agents perform a joint action $\in A$. This probability is views as the possibility of modifying assignment combination from s to s' resulting from executing a re-assignment action.

The probability P is integrating the complete stochastic information about assignment of gates including stochastic delays as well as additional disturbances that impact gate assignment and computed as a probability of occurrence. This probability utilizes other estimation techniques to build the probabilistic model of GAP under possible disruptions.

The way how transition probabilities are defined is essential for building the robustness of the GAP based MMDP model. The state transition stochastic matrix P defines all likely possible state transition probabilities (p_{ij}):

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{11} & p_{11} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{4}$$

Where:

$$\sum_{j=1}^{n} p_{ij} = 1 \ (i = 1,2, \ldots, n), p_{ij} \ge 0 (i, j = 1,2, \ldots, n)$$

Various statistical estimating methods could be applied to calculate state transition probabilities described above. The method as in [33] is applied using statistical data of state transition. Actions corresponding to flights combination are identified, and the arising states are collected from data. The collected values from observed data, $k_1(a)$ corresponds to the case without disruption on state s_1 performing action a, and k_{12} (a) is the case of disruption observed between state s_1 and state s_2 performing action a. therefore the transition probability between s_1 and s_2 performing an action a is estimated from observed data as :

$$p(s_1, s_2, a) = {k_{12}(a)}/{k_1(a)} \tag{5}$$

R (s, a, s') corresponds to the reward acquired once transiting from a state s to a state s' performing an action a. This involves costs as negative reward or positive reward as benefits of each reassignment.

R is thought as : $R: S \times S \times A \rightarrow R$

Where its function is defined as:

$$R(s,a,s') = -\lambda\delta_{ss'}p(s,a,s') + \gamma(1-\delta_{ss'})p(s,a,s') \quad (6)$$

having:

- $\delta_{ss'} = 1$ if $s = s'$ and 0 otherwise
- λ Penalty unit
- γ Recompense unit

The main task of a decision maker is to compute a policy as:

$$\pi: S \rightarrow A$$

A state-action sequence of decisions that maximize the expected total reward is denoted as π^*, and corresponds to the policy optimal. V*(s) gives the maximum cumulative reward attained by the optimal policy beginning with states. Therefore, the optimal decision in a state s is to choose an action a maximizing the sum of the immediate reward $R(s,a,s')$ and the value V* of the immediate successor state, discounted by γ ($0 \leq \gamma < 1$):

$$\pi^*(s) = arg\ \max_a[R(s,a) + \gamma V^*(p(s,a))] \quad (7)$$

The solution concerns obtaining an optimal stationary policy π ∗ that maximize for each state s and for all agents the expected discounted future reward. π ∗ contains the optimal decisions to make in every gate considering the assignment state.

MMDP model representing the GAP problem is solved using the value iteration function determined by Howard algorithm (see [5]) and provided as follows:

Policy Iteration Algorithm [34]:

- π any policy
- While $\pi \neq \pi'$
  - (1) $\pi := \pi'$
  - (2) For all $s \in S$
    Compute $V_\pi(s)$ by solving the system of $|S|$ unknowns given by Eq (1)
  - (3) For all $s \in S$
    If there exists an action $a \in A$ such that:
    $$R(s,a) + \gamma \sum_{s' \in S} P(s,s',a)V_\pi(s') > V_\pi(s)$$
    Then π′(s): = a
    Else π′(s): = π(s)
- Return π

This algorithm is assured to converge (as in [5]).

D. Multi-Agent with Time Dependence Model:

Based on the same approach as the previous model, this paper presents another model with Multi-Agent reasoning but including the time evolution aspect of the Gate Assignment Problem. The considered Time-dependent Multi–agent Markov Decision Problem illustrated in Fig. 6.

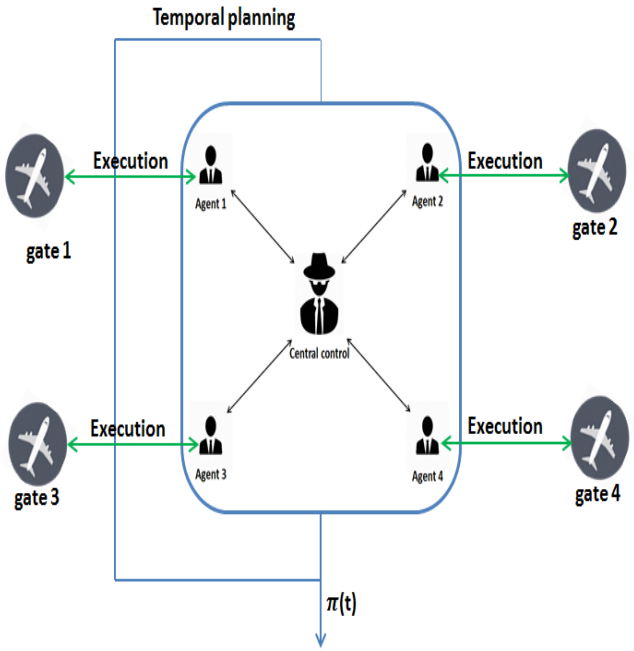


Fig. 6. Agents distribution and temporal planning

Let K be the Number of agents; also correspond to the number of gates. Taking the same actions definition from the previous model, the set of actions $A = A_1 \times \cdots \times A_K$ defines the set of joint actions of agents being also for every agent i assigning a flight $a \in V$ to a gate I, V the set of flights.

Additional temporal information will be included first in the Discrete set M set of outcomes, of the form $\mu = (s'_\mu, T_\mu, P_\mu)$:

- $s'_\mu \in S$: the resulting state space
- $S = S_1 \times \cdots \times S_K$ gives different possible combinations of flights $a \in V$.
- T_μ, ∈ {ABS, REL}: Type of the time distribution (absolute or relative).
  - If $T_\mu = ABS, P_\mu(t')$ will be a pdf over absolute arrival times of μ and corresponds to distribution time associated to some gates assignment configuration action.
  - If T_μ = REL, $P_\mu(\delta)$: pdf will be over durations of μ that corresponds to the duration needed to establish the assignment configuration action.
- L: $L(\mu|s,t,a)$ is the likelihood of outcome μ given state of gate assignment s, time t and action of next assignment to execute $a = (a_1, \ldots, a_n)$, $a_i \in V$.
- R: $R(\mu,t,\delta)$ is the reward for the outcome μ at time t with duration δ, corresponding to reward of spending δ duration at time t with airport assignment action μ. The

reward includes as the previous model two components :

- A benefit from the gate assignment outcome μ.
- A penalty to assignment outcomes μ that causing a possible disturbance at time t and with duration δ.

IV. Experiment

A. Multi-Agent Model experiment:

Computational analysis is done to test the efficiency of the used Multi-Agent MDP approach, and utilizing a simple data example to conduct experimentations.

For simplification, data includes two gates and three aircraft to allocate in a discrete window of time between T_0 and T_3.

$V_i \in V$, is set of flights and for $i = 0$ it match a vacant assignment gate.

As a sample, in this experimental instance exist three possible states:

$s_1 = (V_1, V_2)$, $s_2 = (V_3, V_0)$, $s_3 = (V_0, V_3)$. Two agents are affiliated to the two gates, therefore actions are: $a_1 = (V_1, V_2)$, $a_2 = (V_3, V_0)$, $a_3 = (V_0, V_3)$.

As an initial policy, the solution provided first by a deterministic approach to the problem from literature is used. Simple values are used as input parameters only for simulation. The preliminary policy is as follows: $\pi_0 = (a_2, a_1, a_1)$:

TABLE I. Initial Policy without Disruption

		T_0	T_1	T_2
Agent 1	**Gate 1**	V_1	V_3	
Agent 2	**Gate 2**	V_2		

It is designed regarding observations, transition probabilities and rewards are shown in Fig. 7.

```
>> P

P(:,:,1) =

     1     0     0
     1     0     0
     1     0     0


P(:,:,2) =

    0.7500    0.2500         0
         0    1.0000         0
         0         0    1.0000


P(:,:,3) =

    0.0500         0    0.9500
         0    1.0000         0
         0         0    1.0000
```

```
>> R

R(:,:,1) =

  -100     0     0
   100     0     0
   100     0     0


R(:,:,2) =

   -75    25     0
     0  -100     0
     0     0  -100


R(:,:,3) =

    -5     0    95
     0  -100     0
     0     0  -100
```

Fig. 7. Transitions and rewards matrixes

With $\lambda = \gamma = 1$.

Like in Table I, $p(s_1, s_1, a_2) = 75\%$ expresses a probability of disruption performing action a_2 on s_1, which corresponds to the situation in Table II (V_1 is delayed and still allocated to gate 1 that V_3 cannot be re-assigned, which results in conflict):

TABLE II. Conflicting Assignment in Initial Policy Due to Delay

		T_0	T_1	T_2
Agent 1	**Gate 1**	V_1 Conflict		V_3
Agent 2	**Gate 2**	V_2		

A simple experimentation is done to demonstrate the feasibility of the suggested resolution method.

The initial policy is not possible as a result of delay of the flight V_1 (Table II), which causes a conflict in gate allocating. Therefore this solution is used as initial policy in the policy iteration algorithm then the algorithm is performed.

After execution of value iteration algorithm in MatLab, the provided solution offers another order in the gate assignment; optimal policy is $\pi^* = (a_3, a_1, a_1)$ identified as in Table III:

TABLE III. Optimal Policy

		T_0	T_1	T_2
Agent 1	**Gate 1**	V_1		
Agent 2	**Gate 2**	V_2	V_3	

Table III shows that the proposed approach can give a solution that is more robust to delays. Compared with the sample agent MDP in [21], this approach is more representative of the problem structure because of the Multi-agent distribution of processing, that simplify its conception. Also, MMDP gives gate assignment configurations in multi-dimensional policies instead of having in MDP a single gate to flight assignment.

However, MMDP model gives only fixed time steps between decision epochs (iteration steps), that does not reveal the real evolution of gate assignment witch time is different from iteration step and always observable. Next paragraph gives an experiment with time dependence.

B. Time-Dependent Multi-Agent Model Experiment:

In this paragraph, it is conducted an experiment of Time-Dependent Multi-Agent MDP modeled earlier.

For simplification, every action possesses a single outcome. Hence actions and outcomes can be directly recognized ($a_i \leftrightarrow \mu_i$) and actions thought to be deterministic with regard to the discrete component of the state. This is expressed as:

$\forall i$ Such that a_i is feasible in state s, $L(\mu_i | s, t, a_i) = 1$

It is used a real data from six flights of Hong Kong international airport as in Table IV; tree gates are dedicated to those flights.

A Gate conflict is detected between flights LH738/739 and SQ862/861 due to some disturbance.

Starting with a specific state of the system s_1 corresponding to the airport gate assignment:

s_1 =(CA101/102, LH738/739, TG600/601)

TABLE IV. DATA FROM HONG KONG INTERNATIONAL AIRPORT [33]

Flight	Arrival	Departure	Route	Airline
CA101/102	11:25	12:45	Beijing-Hong Kong-Beijing	Air China
LH738/739	11:30	13:10	Frankfurt-Hong Kong-Frankfurt	Lufthansa
TG600/601	11:45	12:45	Bangkok-Hong Kong-Bangkok	Thai Airway
JL710/702	13:15	15:00	Osaka-Hong Kong-Osaka	Japan Airlines
BR869/870	14:25	15:30	Taipei-Hong Kong-Taipei	EVA Air
SQ862/861	14:20	16:00	Singapore-Hong Kong-Singapore	Singapore Airlines

Moreover, exploiting other possible actions is done to apply adapting assignment to arriving flights representing a change in gate configuration.

- a_1 =(BR869/870, JL710/702, SQ862/861)
- a_2 =(BR869/870, SQ862/861, JL710/702)
- a_3 =(JL710/702, SQ862/861, JL710/702)

Fig. 8 below shows the state transition corresponding diagram.

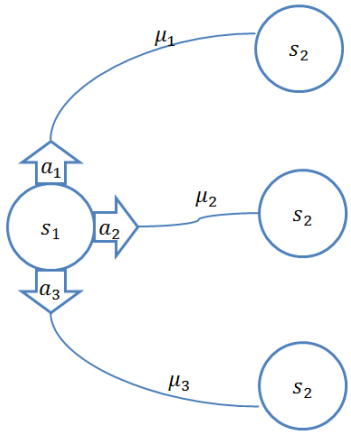


Fig. 8. State transition diagram.

Just for simplification, all outcomes have parameter $T_\mu = ABS$, so outcomes with durations are not considered. The probability density functions P_μ are the defined for every outcome see as example Fig. 9.

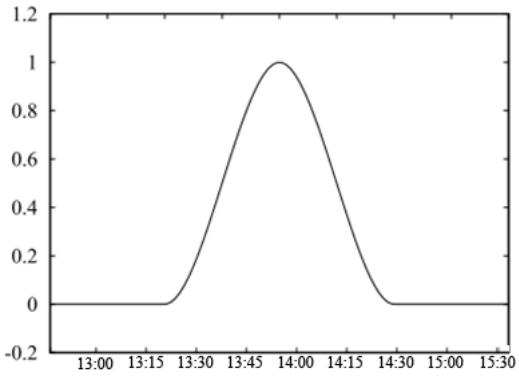


Fig. 9. Probability density functions of μ_2.

This probability includes stochastic information related to action execution. Rewards are given in a way to score every action of assignment in the airport.

So, implementing the resolution algorithm, the value iteration algorithm gives an exact resolution [27]. The given solution consists of time-dependent policy choosing outcome μ_2 that avoid the disturbance situation. Then, the solution given by this approach is robust and handles flight delays. The fact of including the information about the possible disturbances improves more the GAP solution quality.

V. CONCLUSION AND PERSPECTIVE

In this work, A new approach has been formulated for the Gate Assignment Problem (GAP) powered by Time-dependent Multi-Agent Markov Decision Processes (TMMDP). This method aims to constitute a robust mechanism that will give a time valuated approach dealing with disturbances in every time sequence. The provided solution is all of the decisions at every time that could be performed at the time of the planning horizon of flights assignment. This kind of model takes into account real-time optimization because it assumes to have a solution at every time which manages disturbances.

Experimentations on this approach using a real sample data by simulation of the associated value iteration algorithm provides a best feasible solution that the deterministic model.

The aim behind this reflection is to offer to controllers at the airport a robust time valuated solution take in consideration possibilities of gate conflict, even if may take more time to resolution, it can manage well risks in gate assignment.

As perspective, this reflection about this type of model can be more extended to take into account as possible other real constraints of gate assignment.

REFERENCES

[1] B. Pearce, "The state of air transport markets and the airline industry after the great recession," *Journal of Air Transport Management*, vol. 21, pp. 3–9, 2012.

[2] A. Lim and F. Wang, "Robust airport gate assignment," in *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on*, Nov 2005, pp. 8 pp.–81.

[3] C.-H. Cheng, S. C. Ho, and C.-L. Kwan, "The use of meta-heuristics for airport gate assignment," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12430–12437, Nov. 2012.

[4] T. Obata, "Quadratic assignment problem: evaluation of exact and heuristic algorithms," 1979.

[5] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press, 1960.

[6] C. Boutilier, "Planning, learning and coordination in multiagent decision processes," in *In Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge (TARK96*, 1996, pp. 195–210.

[7] J. Xu and G. Bailey, "The airport gate assignment problem: Mathematical model and a tabu search algorithm," in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 3 - Volume 3*, ser. HICSS '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 3032–.

[8] U. Dorndorf, F. Jaehn, and E. Pesch, "Modelling robust flight-gate scheduling as a clique partitioning problem," *Transportation Science*, vol. 42, no. 3, pp. 292–301, Aug. 2008.

[9] S. Yan and C.-M. Huo, "Optimization of multiple objective gate assignments," *Transportation Research Part A: Policy and Practice*, vol. 35, no. 5, pp. 413 – 432, 2001.

[10] A. Drexl and Y. Nikulin, "Multicriteria airport gate assignment and pareto simulated annealing," *IIE Transactions*, vol. 40, no. 4, pp. 385–397, 2008.

[11] M. Hassounah and G. N. Steuart, "Demand for aircraft gates," *ransportation Research Record, n 1423, pp.26-33*, 1993.

[12] S. Yan and C.-M. Chang, "A network model for gate assignment," *Journal of Advanced Transportation*, vol. 32, no. 2, pp. 176–189, 1998.

[13] B. Maharjan and T. I. Matis, "Multi-commodity flow network model of the flight gate assignment problem," *Computers & Industrial Engineering*, vol. 63, no. 4, pp. 1135 – 1144, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835212001799

[14] S. Yan and C.-H. Tang, "A heuristic approach for airport gate assignments for stochastic flight delays," *European Journal of Operational Research*, vol. 180, no. 2, pp. 547 – 567, 2007.

[15] C.-M. Pintea, P. Pop, C. Chira, and D. Dumitrescu, "A hybrid ant-based system for gate assignment problem," in *Hybrid Artificial Intelligence Systems*, ser. Lecture Notes in Computer Science, E. Corchado, A. Abraham, and W. Pedrycz, Eds. Springer Berlin Heidelberg, 2008, vol. 5271, pp. 273–280. [Online]. Available: http://dx.doi.org/10.1007/-978-3-540-87656-4_34

[16] C. Yu, D. Zhang, and H. Lau, "Mip-based heuristics for solving robust gate assignment problems," *Computers & Industrial Engineering*, vol. 93, pp. 171–191, 2016.

[17] W. Deng, H. Zhao, X. Yang, J. Xiong, M. Sun, and B. Li, "Study on an improved adaptive pso algorithm for solving multi-objective gate assignment," *Applied Soft Computing*, vol. 59, pp. 288–302, 2017.

[18] S. Mokhtarimousavi, D. Talebi, and H. Asgari, "A nondominated sorting genetic algorithm approach for optimization of multiobjective airport gate assignment problem," Tech. Rep., 2018.

[19] O. Aoun, M. Sarhani, and A. E. Afia, "Investigation of hidden markov model for the tuning of metaheuristics in airline scheduling problems," *IFAC-PapersOnLine*, vol. 49, no. 3, pp. 347 – 352, 2016, 14th IFAC Symposium on Control in Transportation SystemsCTS 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-S2405896316302543

[20] O. Aoun and A. El Afia, "Using markov decision processes to solve stochastic gate assignment problem," in *Logistics and Operations Management (GOL), 2014 International Conference on*, June 2014, pp. 42–47.

[21] ——, "Application of multi-agent markov decision processes to gate assignment problem," in *Information Science and Technology (CIST), 2014 Third IEEE International Colloquium in*, Oct 2014, pp. 196–201.

[22] H. Yanco and L. A. Stein, "An adaptive communication protocol for cooperating mobile robots," in *Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. MIT Press, 1993, pp. 478–485.

[23] H. W. Kuhn and A. W. Tucker, "John von neumann's work in the theory of games and mathematical economics," *Bulletin of the American Mathematical Society*, vol. 64, pp. 100–122, 05 1958.

[24] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *IN PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*. Morgan Kaufmann, 1994, pp. 157–163.

[25] O. Aoun and A. El Afia, "A robust crew pairing based on multi-agent markov decision processes," in *Complex Systems (WCCS), 2014 Second World Conference on*, Nov 2014, pp. 762–768.

[26] P. Stone and M. Veloso, "EnglishMultiagent systems: A survey from a machine learning perspective," *EnglishAutonomous Robots*, vol. 8, no. 3, pp. 345–383, 2000. [Online]. Available: http://dx.doi.org/10.1023/-A%3A1008942012299

[27] J. A. Boyan and M. L. Littman, "Exact solutions to time-dependent mdps," in *in Advances in Neural Information Processing Systems*. MIT Press, 2000, pp. 1026–1032.

[28] S. Proper and P. Tadepalli, "Solving multiagent assignment markov decision processes," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, ser. AAMAS '09. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 681–688. [Online]. Available: http://dl.acm.org/citation.cfm?id=1558013.1558107

[29] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 3, pp. 387–434, Nov. 2005.

[30] M. P. Wellman, K. Larson, M. Ford, and P. R. Wurman, "Path planning under time-dependent uncertainty," in *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1995, pp. 532–539.

[31] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, ser. Wiley Series in Probability and Statistics. Wiley, 1994. [Online]. Available: https://books.google.co.ma/-books?id=tsiiQgAACAAJ

[32] G. S. Bastos, L. E. Souza, F. T. Ramos, and C. H. Ribeiro, "A single-dependent agent approach for stochastic time-dependent truck dispatching in open-pit mining," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. IEEE, 2011, pp. 1057–1062.

[33] J. Wang, S. Hou, Y. Su, J. Du, and W. Wang, "Markov decision process based multi-agent system applied to aeroengine maintenance policy optimization," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, vol. 3, Oct 2008, pp. 401–408.

[34] J.-Y. Greff, L. Idoumghar, and R. Schott, "AnglaisApplication of markov decision processes to the frequency assignment problem," *AnglaisJournal on Applied Artificial Intelligence*, vol. 18, no. 8, pp. 761–773, 2004.

A Novel DDoS Floods Detection and Testing Approaches for Network Traffic based on Linux Techniques

Muhammad Tahir*[1], Mingchu Li[1], Naeem Ayoub[2], Usman Shehzaib[3], Atif Wagan[4]
[1]School of Software Technology, Dalian University of Technology, (DUT), Dalian, Post (116621), P.R. China
[2]School of Computer Science & Application Technology, Dalian University of Technology, Dalian, P.R. China
[3]Dept. Of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan
[4]School of Computer Science & Eng., Nanjing University of Science & Technology, Nanjing, P.R. China

***Abstract*—In Today's Digital World, the continuous interruption of users has affected Web Servers (WSVRs), through Distributed Denial-of-Service (DDoS) attacks. These attacks always remain a massive warning to the World Wide Web (WWW). These warnings can interrupt the accessibility of WSVRs, completely by disturbing each data processing before intercommunication properties over pure dimensions of Data-Driven Networks (DDN), management and cooperative communities on the Internet technology. The purpose of this research is to find, describe and test existing tools and features available in Linux-based solution lab design Availability Protection System (Linux-APS), for filtering malicious traffic flow of DDoS attacks. As source of malicious traffic flow takes most widely used DDoS attacks, targeting WSVRs. Synchronize (SYN), User Datagram Protocol (UDP) and Internet Control Message Protocol (ICMP) Flooding attacks are described and different variants of the mitigation techniques are explained. Available cooperative tools for manipulating with network traffic, *like;* Ebtables and Iptables tools are compared, based on each type of attacks. Specially created experimental network was used for testing purposes, configured filters servers and bridge. Inspected packets flow through Linux-kernel network stack along with tuning options serving for increasing filter server traffic throughput. In the part of contribution as an outcomes, Ebtables tool appears to be most productive, due to less resources it needed to process each packet (*frame*). It is pointed out that separate detecting system is needed for this tool, in order to provide further filtering methods with data. As main conclusion, Linux-APS, solutions provide full functionality for filtering malicious traffic flow of DDoS attacks either in stand-alone state or combined with detecting systems.**

Keywords*—DDoS attacks; floods detection; Linux-APS architecture; mitigation techniques; network traffic; Netfilter; testing approaches

I. Motivation and Introduction

Through the development of information and communication technology (ICT), our societies become global information societies with all-around smart computing environments, but unfortunately, the-security systems and policies that regulate this environment are not accelerated as needed. Attackers do not use the quarry in various vulnerabilities found in applications running on systems.

Among the various cyber-attacks, (such as: SQL Injection (SQLi), session recording, internal site scripting attack, Denial-of-Service (DoS) attack, have become the most threatening attacks so far, as very few methods have managed to mitigate these attacks problems). Distributed Denial-of-Service (DDoS) attacks, aggressors do not use a particular crowd on behalf of their attacks however a cluster of numerous loads or uniform hundreds of central processing units (CPUs), towards fixing an SYN (synchronized) attacks. After the creation of development solutions towards resolution the incidence of attacks stimulated the development of the attacks themselves. Currently, DoS and DDoS attacks have been outdated via DDoS attacks. The World Wide Web (WWW) Safety FAQs, scheduled DDoS declares that [1].

"A DDoS attack, usages numerous supercomputers towards promotion of a synchronized DoS attack beside one or further goals. By client and server machinery, the criminal is intelligent to increase the efficiency of DoS, pointedly by connecting the properties of numerous unknowing accessory supercomputers which help as attack stands. Classically a DDoS, main suite is connected the one (CPU, storage, memory, by a filched version). The main suite, on a selected time, before transfers to some digits of ***'proxy'*** plans, connected on CPUs, everywhere on the Internet. The proxies, once they obtain the -command, recruit the attack. By client and server know-how, the main suite can pledge hundreds or else level thousands of proxy series inside ages."

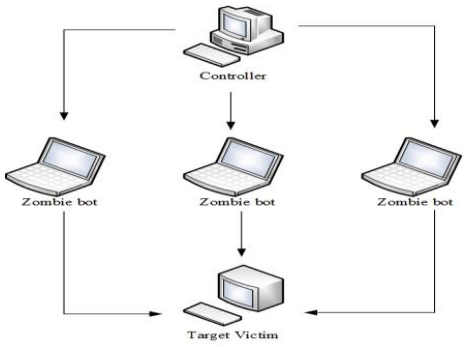


Fig. 1. Provides a clear idea about DDoS attacks.

Fig. 1 proves two main aims that make DDoS attacks, eye-catching to intruders. Firstly, there are effective automatic-tools to attack all victims [2]. That is, experience is not necessarily required. Secondly, that one is typically dreadful toward find an aggressor lacking general communication with a person or else lacking novel roles fashionable utmost the Internet routers [3].

According to *Akamai-state* of the Internet security reports that the frequency of DDoS attacks has been increased by **131%** Worldwide in **2017**. Along with recent event, involving the *WannaCry ransom-ware* attacker, (also known as *WanaCrypt0r & W-Cry Ransom ware*) malicious software spreading security and protection questions for Internet users are becoming more important than ever before.

The importance of this research work is to find out and testing most efficient and reliable tools existing in Linux-based systems for filtering and aggregation DDoS attacks. Linux-based solutions are considered as bit and easy to configured systems among available competitors.

Most common WSVRs, focused DDoS attacks, will be taken into consideration such as SYN, UDP and ICMP floods. By configuring filter servers and applying suitable setup, most efficient and reliable solution will be chosen. Aggregation of data traffic flow will be considered from a point of impact on filtering productivity.

- *This research work, is divided into following four parts:*
- ***Firstly,*** the overview of the DDoS attack is provided.
- ***Secondly,*** literature analysis & background gives a brief survey of research going in the area of most common types and methods of mitigations of DDoS attacks and available Linux-based solutions for data traffic filtering and aggregation. After going through the literature analysis and its background.
- ***Thirdly,*** the problem statement and related work has been identified and will describe experimental network components and kernel tuning.
- ***Finally,*** it focuses on the conclusion of the work that will include implementation of the selected solutions on data filter servers, differentiated by installed hardware.

II. Literature Analysis and Backgroud

DDoS attacks have become more common and fashionable in recent years. Large-scale systems of septic computers ***'zombie/bot'*** trust their processing plus processing capabilities to overload public service and deny it to authentic users.

The attacks on major *e-commerce locations* in February **2000** and attacks on origin domain name system (DNS), service **2003** and **2007** made community devotion towards the-problematic of DDoS attacks [4]. Nowadays, medium-volume Web pages are generally condemned through crooks toward getting defense from their venders. These are also deprived of appealing [5]. Additionally, Internet Service Providers (ISPs), must address the problem that DDoS traffic and increases the bandwidth of the communication channel.

The first tools such as Tribe Flood Network (TFN) [6], Stacheldraht, Trinoo or Mstream [7], have used communication structures without cryptography and were organized hierarchically. Best of these implements recycled TCP, SYN, UDP, and ICMP floods using potentially recognizable limits.

As several of these attacks have been positively relieved, an innovative formation of robots takes developed Spartan-dominion Robot (SDBot), Agobot, commonly used Phatbot and well-known representatives using Internet Relay Chat (IRC), as a secure link [8]-[10]. These tools also include distribution methods besides take additional refined attack algorithms that can stay updated concluded the Internet.

Sudden mitigating techniques of DDoS attacks, on the Internet sources or on kernel seems impossible owing towards the circulated and authoritative environment of the Internet protocol IP, based system network. Several methods were suggested to search for the original IP address, of the attacker using filtering mechanisms.

Internet Engineering Task Force (IETF) documented filter entry approach is defined in request for comments **(RFC: 2827)** [11]. And assistances mitigate DDoS attacks, through IP spoofing, which indirectly deal with different types of network misuse, causes Internet traffic to control source. Network filter-is a ***'good neighbor'***. This policy is based on mutual cooperation between ISPs, for their common benefits.

In order to avoid manipulation of IP addresses, Park et al, [12] Proposed packet filters distributed on standalone systems over the Internet are to be stopped packets of counterfeit IP addresses.

Suspenseful Savage et al. [13] is recommended that IP Trace back find the basis of fake IP addresses using probabilistic labeling packages. Song et al. [14] provides an extended scheme for probabilistic packet selection to reduce the frequency of false positives to restore the attack path. Another improved scheme for probabilistic labeling of packages was proposed by Bellovin et al. [15] to reduce the cost of calculation. This ICMP is a tracking system similar to a probabilistic scheme for labeling packages.

In this system, the routers generate ICMP, packets at the Low-probability destination. For a significant data traffic flow, the recipient can gradually restore the route made by the packets in the leak.

Mahajan et al. [16] provide a system where routers learn fixed costs to provide good traffic from bad traffic. Siris et al. [17] represents the variance finding algorithms on behalf of identifying TCP and synchronized SYN attacks.

Modifying threshold algorithm and a specific request of the total amount of algorithm is for finding of a switching point. Modifying threshold algorithm associates with number of the SYN packets established over an estimated number of predetermined intervals, founded on the new dimensions. Towards promotion of increases, a panic inception should remain overdone in series. The, Cumulative Sum Control Chart (CUSUM), algorithm usages the change among the sum of SYN packet per time pause besides the number valued on behalf of the similar range such as ***'Gaussian'*** random variable. Then, flood attack SYN, is sensed by consuming the total amount built on the probability ***like;*** instantaneous significance due to the change in the average velocity of the circulation.

Several authors, including Wang et al., [18] has proposed a method for detecting SYN flood attacks. These are built and going on the construction of TCP, SYN, FIN flags and Rapid Spanning Tree Protocol (RSTP). Protocols on leaflets connect the final nodes to the Internet. There is a change in the sum of SYN, FIN packets perceived rest (RST) flag set and CUSUM, algorithm is used to detect the switch point. Towards decrease the influence of changed entree designs at diverse locations, the change among the amount of: (SYNs), (FINs) and (RSTs) remains controlled with the calculated typical by (FINs) and (RSTs).

Luo et al. and others [19] Also proposed a system for detecting Pulsing Distributed Denial-of-Service (PDoS) and DoS attacks on expressive target's WSVRs, with **(t\¥0)** variances produced by PDoS attacks, specifically jitter in inbound documents traffic flow and reduction of outward---bound TCP, and acknowledgement (ACK) data networks traffic flow.

Cabrera et al. [20]. Detecting DDoS using Management Information Base (MIB), traffic flow valuables on behalf of the aggressor and destination. Appropriate autographs were identified towards detect attacks from known attacks. On the side of the attacker, the DDoS attacks must be detected before it is launched by identifying precursors based on MIB.

In addition to previous work, intrusion detection systems and firewalls are one of the best security systems that work by pairing packets with predefined rules and filtering them accordingly. Linux-kernels had a **1.1** packet filter. At the end of **1994**, kernel-hacker Alan Cox carried the IP firewall from Berkeley Software Distribution (BSD), to Linux. In mid-**1998**, Rusty Russell et al. [21] and others revised most of the networks under the Linux kernel of the **2.1** development series and introduced the IP chains user space tool.

The Linux-kernel that preceded this had some very serious disadvantages with some main working functionality. The old Linux-firewall code does not apply to fragments. The **32-bit**-counter (*at least on Intel*), does not allow to enter other protocols than TCP, UDP and ICMP.

It cannot make major atomic changes that can not specify ***'Inverse Functions'***. It has little and can be tricky to handle (*which makes it prone to user failure*).

Russell's work redefined radically the Linux-network layer and enabled filtering of kernel-packets in user-space (*which simplified usage, configuration, and security*).

Finally, the next advanced generation of tools, ***"Ebtables-Iptables"*** and another core transcript took place in mid-**1999** for Linux **2.4**. The enhancement initiated in **2.2** continued, and the extensive Linux arsenal of networking tools. The core is rewritten as ***'Netfilter'*** [22].

In this research article, we proposed different Linux-based resolution approaches and testing simulations for network traffic, filtering data packets to protect against DDoS attacks using cooperative ***"Ebtables"*** and ***"Iptables"*** tools, mitigation techniques, and Linux-based resolution lab design firewall architecture.

III. Problem Statement and Related Work

- *Problem No.1.* The DDoS attacks, stands a rigid challenge towards mark and connected packages inaccessible near all users, frequently via briefly disturbing before interrupting sudden services from their domain server. Originally based on target service resources limitation, DDoS attacks can be done by either spoofing attacker IP address or using so called: ***'Botnet'***. Botnet is a network of hacked devices, connected to global network which control is gained by third party. Compromise devices can send data traffic to target services which makes DDoS mitigation complex. As it is hard to distinguish authentic malicious traffic flow from the DDoS attacks.

- *Problem No.2.* The DDoS attacks, stance has a main warning to the internet. This must be reduced the class from the internet service. This is significant because the internet has now become a critical resource whose violation has financial consequences or even terrible consequences for human security. More and more critical services use the internet for daily work. DDoS attacks do not just mean losing the latest game results of the environment. This could malicious traffic flow from DDoS attacks losing an effort on the item you want to buy or lose to your customers for a day or two under attack. So this one is significant in the direction of consume funds towards stop or else moderate the system.

Smart breaks support unbroken competition among the aggressors and the protectors. Passing of these remain a real ramparts beside a confident category of attack. The aggressors' revolution devices finding an approach toward avoid this defense.

In addition, since absolute protection is not feasible, developing effective defense framework involves an often complicated set of Trade-offs:

For this purpose, numerous solutions have been proposed as discussed in literature analysis background. ***'Firewall'*** is one of them. By using a dedicated firewall we can block all the IP addresses that are flood the server to consume its resources.

According to Radware's: **(2016-2017)** Global application and network security report [23]. The most common types of attacks were **2016** (e.g., UDP, SYN and ICMP Flood).

A. Distributed Dinel-of-Service (DDoS) Flooding Detection - Tools & It's Mitigation Techniques

1) The Synchronized (SYN) Flood Detection:

Attacks of this kind targeting three-way TCP, link mechanism that sends connection requests faster than the target computer that can handle them, causing network capacity [24].

In common situations, the client method is activated through distribution of SYN junk messages toward affecting server.

The affecting server formerly confirms effective SYN junk messages by distribution and SYN acknowledgement (ACK) messages toward sudden client.

Affecting client server formerly complements impressive configuration of the linking, responds by an ACK message.

This opens powerful link among the client and server then provides capacity information container to be swapped among sudden client plus server. **Fig. 2(a)** displays three-way appearances regarding junk message flow:

A potential weak point is that when the server system sends the confirmation of (SYN-ACK) towards the client, however it does not consume all data so far to acknowledge (ACK-Message). It is called a partial built-up linking.

The SYN server made a data structure (Queue-SYN) in system memory, relating all incomplete connections [25].

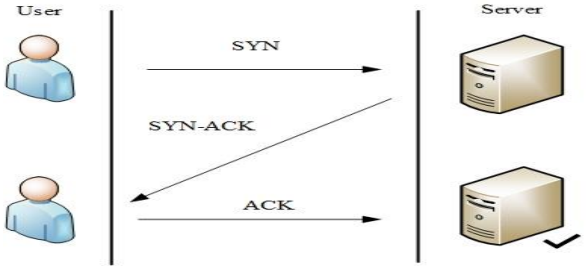


Fig. 2. (a) Three-way handshake mechanism.

This data structure has a problem dimension and it can be transmitted through purposefully generating as well numerous moderately sweeping networks. Generating sweeping networks remains calm by IP spoofing. Suspenseful attacker drives (SYN- Junk Messages) toward the quarry's server systems. They seem genuine, but they actually refer to a client system-that cannot reply to (SYN-ACK-Messages). These resources are used in order to terminate (ACK-Messages) spirit.

It is not once being directed near the quarry's server systems, as per displayed in **Fig. 2(b).** The figure displays norms of SYN flood attacks:

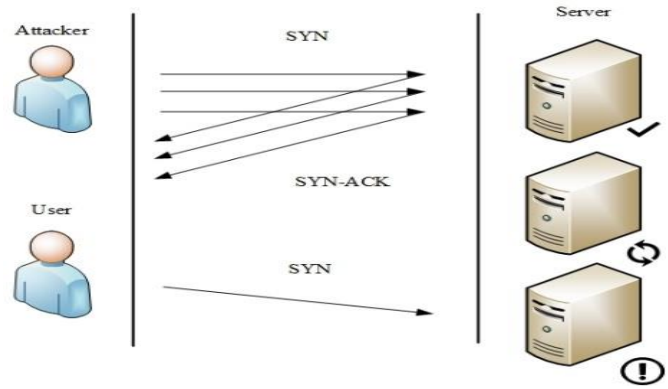


Fig. 2. (b) Norms of SYN flood attacks.

Among possible solutions for mitigation, these kinds of attacks are using firewalls with (SYN-Cookie). This cookie feature enabled, filtering by limitation of possible (SYN-Packets) per second passing accepted by the server and blocking attacker source IP addresses.

The (SYN-Cookies) is the technique by which the original (TCP-Sequence) numbers will be selected by the (TCP-Servers). These are designed to mitigating the SYN flood attacks.

The main differences are between the original sequence number that created by the server and client are:

- ***Highest 05 bits:*** T mod 32, wherever T stands a 32 bit device hostage, increasing each 64 seconds;
- ***The following Subsequent 3-bits:*** Is a programming of a ministry of state security (MSS) designated via the client server cutting-edge reply near client MSS;
- ***Lowest 24 bits:*** Sudden server-designated classified purpose regarding sudden client IP- address also port no. of sudden server IP address and port no. and **T**.

Therefore, sudden client server using SYN cookies, files should no more delete networks after the situation SYN queue is full. This one will return SYN+ACK, by way of the SYN queue was higher.

After tense server obtains an ACK message, its authorizations whether the Top-secret function workings aimed at the last significance of **T**. Besides reconstructs of the SYN queue has been scheduled to encrypt Maximum Segment Size (MSS).

Restrictions can be done based on exact server statistics, including average traffic rate, connections per second during specific time.

2) The User Datagram Protocol (UDP) Flood Detection:

The UDP, is an offline network protocol that provides data integrity check numbers and port numbers for addressing functions [26].

If there is no first handshake, there is no guarantee of data transfer, sorting or duplication of protection to establish a valid connection.

Therefore, a lot of best-staked traffic can be sent over UDP channels to any host without built-in security to limit UDP, DDoS throughput.

This means that not only UDP, thread attacks are very effective, but they can also be done with relatively few resources.

Through UDP flood attack, attackers direct a huge amount of UDP packets towards the offer system, resulting in network saturation and bandwidth reduction available for authentic quote requests. Once ***'rib-tickling'*** offer structure obtains the UDP packets, it controls which request is to come for electrifying endpoint docks.

After the server decides so that the application is not spoofing to the application, it will generate an: ***"Unavailable"*** ICMP packets for the fake home address. Unknown sufficient UDP packets are brought toward the victim's docks. Sudden system resolve will be continued to reduce.

Another way to perform an attack is to send huge amount of UDP packets to certain the ***'Opened-Ports'***, leading to link bandwidth exhaustion.

Among known option of mitigating is to close all unused port on server and filter all incoming traffic that is destined to target server, except Domain Name Servers (DNS).

3) The Internet Control Message Protocol (ICMP) Flood Detection:

The ICMP, is recycled by devices, including the router, to send operating evidence messages to support networks for example, diagnostic or control drives [27].

An attack using ICMP Flood can be performed in the ***"ping of dead"*** mode and sends victims a large number of the ***"ICMP_ECHO_ REQUEST"*** messages that cause the target server to respond.

Thereby it leads to saturation of the network connection with the victim's behavior.

During the ICMP flood attacks, the foundation IP address might be manipulated. Possible solutions for mitigations are limit size of ping requests as well as the rate at which they can be accepted and denied all ***"Icmp_Echo_Requests"*** from all source IP addresses, except local network.

B. Linux-Based Resolution Availability Protection System (Linux-APS) Architecture

In this section, we present the Linux-based resolution lab design architecture for DDoS network control.

While deciding which best resolution is to choose for traffic filtering and DDoS attacks mitigations.

Linux-kernel based products can become a preferable for considering price tag and user friendly environment.

There is lot of tools and tuning options for working with traffic, on which we need to, have a closer look:

Linux distributions are Operating Systems (OS), created from a collection of software [28].

These are based on the Linux-kernel and packet management system.

The Linux kernel is a CPU suite and it is the essential for the OS, through full controller, especially utilized in our given method to be able to apply the suitable techniques for filtering traffic.

It is a crucial to know, how incoming and outgoing packets of Linux network stack are being handled by Linux Protection System (LPS).

Linux-based resolution lab design architecture provides clients and users with a chance to discover the LPS.

This lab design architecture enables clients and users to deliver demos to execute their own-due meticulousness in our given proposed testing methods of surroundings.

With our demonstration tool end-users will gain: additional information of the Linux-based availability protection system (Linux-APS) and it also help users to fast-track safe data on WWW.

In this connection, **Fig. 3** provides protection solution to assist sense and mitigation techniques against DDoS attacks and cutting-edge malicious threats.

Linux-based (APS) mitigates accessibility threats. *Such as:* web application-layer DDoS attacks previously they influence WSVRs, accessibility.

So we compromises a lab setting to simulate a tier **1** web application and DDoS attacks, which can be mitigated by the Linux-based (APS) system.

The lab enables web domain partners to fully understand the APS, system and provide informative demonstrations for their clients and users.

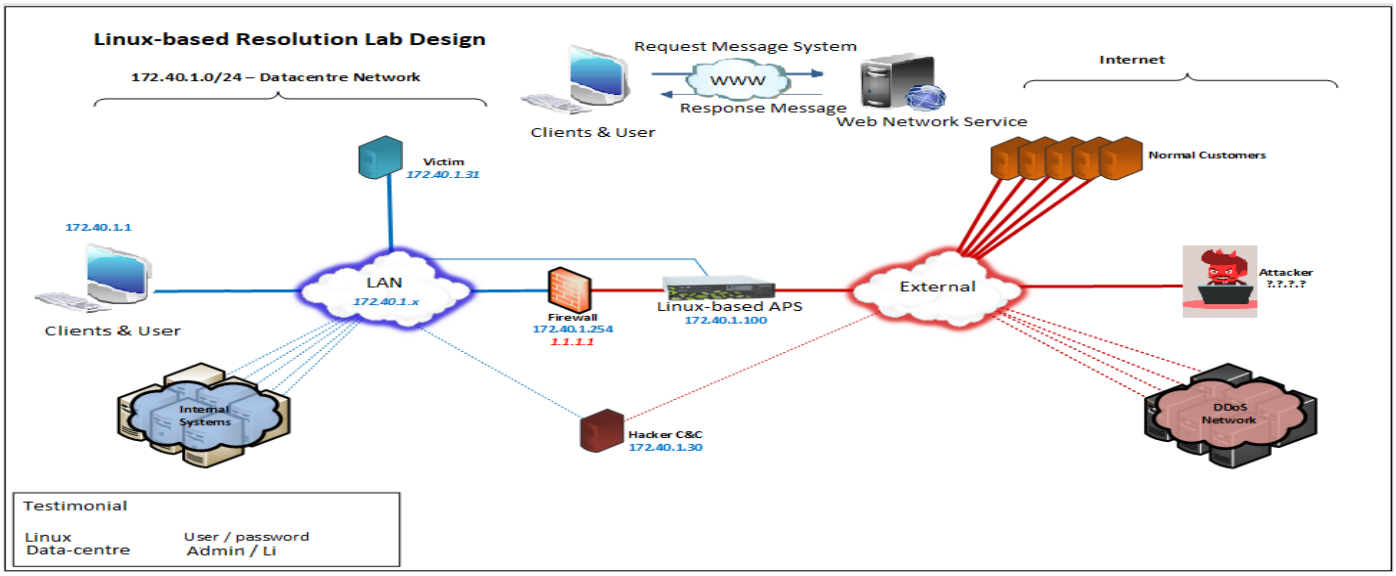


Fig. 3. Linux-based resolution lab design architecture.

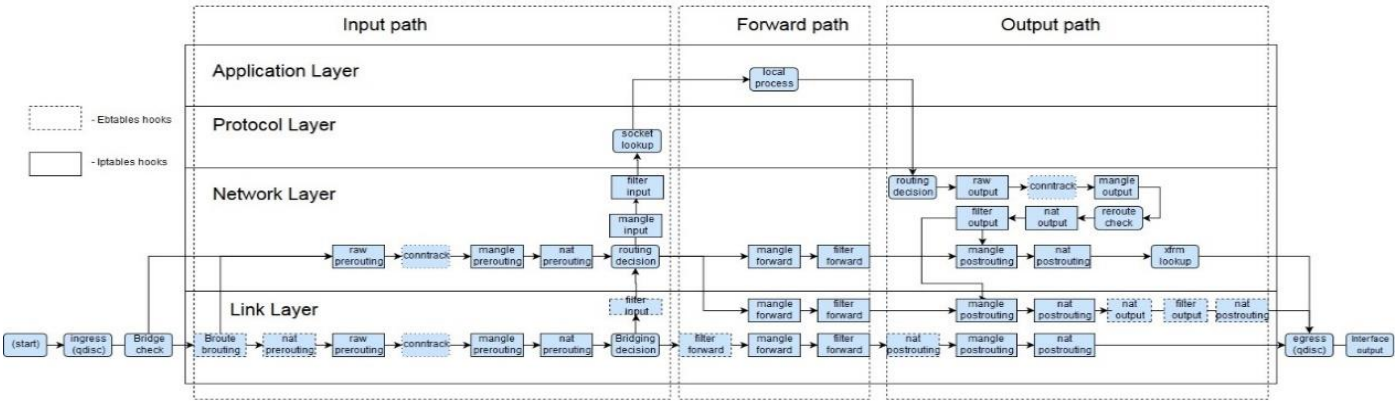


Fig. 4. Data packets flow over Netfilter.

C. The Netfilter Data Packet's Flow

Disreputable for all manipulating with network stack in Linux kernel is Netfilter. To be able to apply the suitable techniques for filtering traffic, it is crucial to know, how incoming and outgoing packets of Linux network stack are being handled by Netfilter. Netfilter is the infrastructure that is provided by the Linux-kernel and it is also a set of hooks inside the, enabling single core modules to record call-back features with the core network stack [29]. Which you can see in a general packets flow path through Netfilters hooks. It is displayed in **Fig. 4.**

Now, let's have closer look at the most essential functions of our testing purposes:

- ***Classless Queuing Disciplines (qdiscs):*** It is the scheduler and the major building block in Linux - traffic-control process. It queues all packets based on appropriate queuing discipline and transmits the packet as soon as it can. There is ingress the (*inbound traffic*) and egress the (*outbound traffic*) *qdiscs*. The default queuing discipline for all interfaces under Linux is ***'pfifo_fast'***. It is based on a conventional first in, first out (FIFO) qdisc and provides prioritization. There are three different bands for separating traffic. The highest priority traffic is placed into the band **0** and is always serviced first.
- ***Bridge Check:*** It is simply checked, if the interface, from which packet was received, that may belong to bridge or not. If so, frame will not be processed at this point and will be sent to bridging decision function.
- ***Conntrack Tools:*** It is the connection tracking subsystem. It supplies evidence around the formal of a-linking, including foundation and endpoint IP addresses, docks digit sets, procedure forms, public, and recreation in structured memory [30].
- ***Bridging Decision:*** At this point frame is being investigated whether its destination is local process or its endpoint media access control (MAC) address is-located on extra sideways of the bond. This can be done with four different frame things:

1. Bridge it.

2. Flooding it over, if the endpoint media access control (MAC) address is unidentified to the bond.

3. Permit this to developed procedure encryption (IP-code).

4. Disregard it, condition an endpoint media access control (MAC) address stands on the similar cross of the bond.

- ***Routing Decision:*** Based on IP address it decides if packet destination is local process or it should be forwarded. Packet will be sent through bridge interface at this point, if forwarded.

D. Proposed Linux-based Resolution Tools and Mitigation Techniques for Web Servers (WSVR's)

a) Ebtables Tool:

The Linux kernel is an integrated filtering tool, starting with core version 2.2 [31], [32] which allows you to configure and to maintain control tables that control Ethernet frames. This is similar to iptables tool, but with less functionality, because the Ethernet frame header is less complex than that of the IP packet header. Ebtables rules are working only with bridged frames and compare to iptables, the frame is processed earlier in the stack, consuming less resources.

There are three tables named ***filter***, ***nat*** and ***broute*** [33]. Syntax for managing with ebtables rules is the same as with iptables rules:

- The ***filter*** is the default table and contains three embedded strings: INPUT (for frames - intended for the bridge), OUTPUT (for locally generated frames or (b) and FORWARD (for frames sent by the source).
- The ***nat*** is usually used to change media access control MAC addresses and contains three - embedded strings: PREROUTING (to change frames as soon as they arrive), OUTPUT (to - change locally- generated or (b) routed to their bridge) and POSTROUTING they will leave.
- The ***broute*** table has a built-in chain: BROUTING. The goals for DROP and ACCEPT are of particular importance in the broute table. DROP means that the frame must be routed, while ACCEPT means that the

frame must be locked. The BROUTING chain goes very early. However, it only passes through the frames entering the bridge port, which is in the forward-looking state.

b) Iptables Tool:

Netfilter iptables is a user-space command line utility to configure packet filtering rules [34].

It's the default firewall management utility on Linux-systems.

Iptables exists recycled to configure, preserve plus validate the IPv6 packet refine control tables cutting-edge the Linux-built kernel.

There are four different tables ***filter***, ***nat***, ***mangle*** and ***raw.*** All tables' covers a digit of integrated-chains and might similarly cover well user-defined-chains.

All new chains are lean of comments that may correspond to a fixed of packets. Also every instruction sets a goal: [35].Which means action with matching packets.

- ***Filter:*** Standard tables, contains built-in INPUT chains (for local-purpose plugs), FORWARD (on behalf of packets sent over sudden field), and the OUTPUT (On behalf of nearby produced- packets).
- ***Nat:*** The nat-table, while a packet is found so that makes an original linking. It involves like three diverse integrated modules: PRE-ROUTING (To modification packets when they will- enter), OUT-PUT (To change packets created locally before routing) and POSTROUTING (To change packets when they-expire).
- ***Mangle:*** The mangle-table remains recycled on behalf of particular data-packets changes. From the **02.04.19** kernel, ternary added embedded chains are moreover maintained: IN-PUT (On behalf of-incoming packets to powerful actual block), FORWARD (To change-packets that pass finished the-packet), and POST-ROUTING (To change packets once- drive-towards departure), PRE-ROUTING (Changes trendy for incoming pre-routing- packets) and OUTPUT (To change packets that are generated locally before routing).
- ***Raw:*** This table is primarily used to configure connection tracking exception along with the NOTRACK target. It is registered in higher priority Netfilter hooks, then stands so-called formerly an ***'ip_conntrack'***, before several further IP tables. Tendency delivers suspenseful next integrated new chains: PRE-ROUTING (On behalf of packets coming by some system edge) OUT-PUT (On behalf of packets produced via native methods).

c) ***The Link Aggregation:*** Link aggregation contains several methods for combining numerous system networks in similar to growth bandwidth.–Such as a linking can support then deliver termination cutting-edge situation of failure of one of the links [36]. Link Aggregation Group (LAG), syndicates a range of physical ports to create an only high-bandwidth-information channels near performance capacity distribution of traffic - between member ports and improve connection reliability. The channel aggregation principle is displayed in **Fig. 5.**

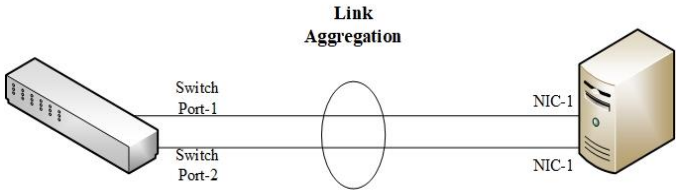


Fig. 5. The appearances of link aggregation.

Following are the most important benefits of link aggregation:

1) ***Multiple connections with little speed loss:*** Single file transfers or one-at-a-time transfers do not get much benefit from link aggregation, but multiple connections and file transfers do.

Some transfer rate increase might be apparent, but link aggregation perfectly works with multiple simultaneous transfers where several clients connect and download concurrently. Having more network lanes available allows all clients to encounter faster download speeds. Examples include a personal media server or network attached storage where multiple devices or users connect.

2) ***Redundancy:*** The physical links can be spread across multiple switches. If one switch fails or a cable is torn or disconnected, the transfer continues but at slower speed until the issue is resolved.

3) ***Load Balancing:*** This balances the network load across multiple network cards for more performance and better throughput. Rather than making one card do most of the work, let the other cards distribute the workload among them.

The Linux-built bonding connection delivers a technique on behalf of combining manifold system edges ***(Slaves)*** keen on a unique logically connected edge ***(Bond)***. Linux-built kernel provisions two bonding methods [37].

- The ***IEEE 802.3ad*** link-aggregation mode, and that tolerates single or else extra associates toward stay combined composed near method a link aggregation group (LAG), supposed that a MAC, client is able to delicacy the link-aggregation cluster by way of unknown draw stood a solitary relation.
- The ***balance-xor*** mode, somewhere the bonding of slave interfaces are static and fully slave interfaces are dynamic for load balancing and fault tolerance devotions.

IV. Discussion and Filter Testing

A. The Spirental Avalanche Commander Filter Testing Tool

Nowadays, it's essential that the performance of our ***'Network Infrastructure'***, ***'Security Systems'***, and ***'A-Web Applications'*** are carefully tested to ensure that performance goals are met. Heavy demands are placed on the- network by

the emerging combination of voice, video, and data traffic, creating new challenges for client users and Digital Information Technology (DIT), staff and Web Networks.

The Avalanche can quickly identify potential points of failures by ***'stress-testing'*** the infrastructure. Large quantities of highly-realistic simulated user and network traffic can be generated, according to a wide range of real-world loading scenarios.

This proactive approach enables you to correct trouble spots and bottlenecks before network slowdowns or costly outages occurs [38].

All tests and measurements carried out on a specifically design and built network, equipped with network tester the ***"Spirent Avalanche 3100B"***, which allows generating traffic by **1Gbit/s** links. General view of network is displayed in **Fig. 6.**

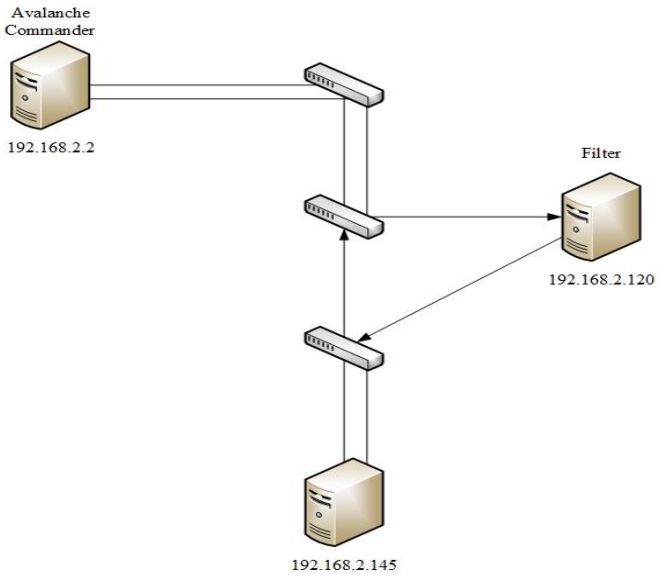


Fig. 6. Demonstrations of new testing network infrastructure.

The new testing filters **V1**, **V2** and **V3** are presented in the form of three separated server machines which are connected parallel. Same filtering tools will be used on all machines to compare hardware influence and link aggregation on filtering process. Hardware specifications for new testing filters are displayed in **Table I.**

TABLE I. THE NEW TESTING FILTERS SPECIFICATION

Requirements	The New Testing Filter V1	The New Testing Filter V2	The New Testing Filter V3
Number of Cores	4	16	2
Clock Speed(min/max)	2800/3600	1600/2200	3400
Network Interface Controller (NIC)	2x NC7782 Gigabit Server Adapter	2x 10-Gigabit X540-AT2	4x 82546EB Gigabit Ethernet
NIC Drivers	E1000	Ixgbe	E1000
Bus	PCI-X (66MHz)	PCIe(16)	PCI-X

- ***PCI-X:*** 64-bit parallel computer bus with theoretical maximum of the 1.06GB/s data exchange speed between computer processor and peripherals.
- ***PCIe:*** Is a serial point-to-point (P2P) connection bus with possible 4GB/s bandwidth in each direction.
- ***Ixgbe:*** Network interface card (NIC) driver with abilities to reduce the number of queues per interface-direction to the number of logical central process units (LCPUs).

The reasoning for this reduction is that each queue requires some over headed, and there is no advantage in maintaining for more queues between designed CPUs.

Because the filter software is used in one of the Linux-based distributions, ***'Debian (OS)'*** since it is entirely free software, most are licensed under the ***'GNU operating-system'*** and a (*Linux-distribution-General Public- License free software*).

In order to have traffic passing through filtering server, first bridge is needed to be configured. The configuration file for all interfaces is located at the */etc./network/interfaces*.

There are three network interfaces on current filtering server. One is serving for virtual private network (VPN) connection, remote control (RC), and other two services for carrying traffic.

B. The Link Aggregation and Interface Bonding

The bonding in place of link-aggregation essential exists maintained through in cooperation bottom line. Dual Linux-based machinery linked by edge chains can proceeds improvement of link-aggregation.

A lone device linked by dual physical chains near a switch whichever provisions port **'*trunking*'** know how to usage link-aggregation to the switch. First straight switch determination develops deeply disordered via a hardware-address looking scheduled numerous ports concurrently [39].

The bonding provision in Linux remains portion like a great accessibility resolution. Designed aimed at an entrance idea keen on the complexity of great convenience cutting-edge combination by Linux [40].

Affecting term concerning the edge is able to state in the user.

The situation remains usually bond **0** or else somewhat parallel. Equally a common-sense edge, that one is able to use in routing-tables together through **'*tcpdump*'**. Link aggregation and bonding-interface setup, is displayed below in **Fig. 7.**

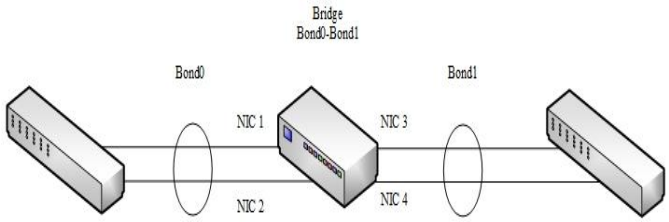


Fig. 7. Shows the links aggregation and interface bonding

To be able to affect bridged frames, we needed to install ebtables tool and bridge-netfilter infrastructure. Also for tracking bandwidth and CPUs utilization, tools such as **'nload'** and **'htop'** will be installed on all servers [41].

That is it exist to console displays web network traffic flow and bandwidth procedure in physical time and **'htop'** powerful main excessive object. Nearby *htop* remains so to resolve display you practice apiece CPU, to boot expressive manuscript diagram of your memory and switch norm accurate on the maximum energy remain installed on all servers. Below code show the creation of headers for installation bridge-utilization, tools and run all servers.

- #apt-get install bridge-utilization
- #apt-get install ebtables
- #apt-get install htop
- #apt-get install nload

C. Data Packets Response Time Testing

Data packets response time testing, satisfaction increased traffic data of network interface controller (NIC) and Linux kernel. We need to improve packet reception process and all filters. **Fig. 8(a), (b)** shows the main check multi queue mode and ***rx_queues*** settings and buffers size enabling, if it's possible on data packets receiving network devices. **Fig. 8(b)** received frames on ***eth3*** will be processed in **63** queues, which increase possible amount of filtering the traffic of Cumulative NIC ***"ring buffers size"*** mentioned in above **Table I.**

```
root@debian:~# ethtool -l eth3
Channel parameters for eth3:
Pre-set maximums:
RX:             0
TX:             0
Other:          1
Combined:       63
Current hardware settings:
RX:             0
TX:             0
Other:          1
Combined:       16

root@debian:~# ethtool -L eth3 combined 63
```

(a)

```
root@debian:~# ethtool -g eth3
Ring parameters for eth3:
Pre-set maximums:
RX:             4096
RX Mini:        0
RX Jumbo:       0
TX:             4096
Current hardware settings:
RX:             512
RX Mini:        0
RX Jumbo:       0
TX:             512

root@debian:~# ethtool -G eth3 rx 4096  tx 4096
```

(b)

Fig. 8. (a) Shows the increasing number of *rx_queues* example. (b) Shows the ring buffers size.

- ***Supporting Receive Packet Routing:*** Therefore it will help prevent network data drops at the NIC during - periods, when large numbers of data frames are received we use this query such as: *(#echo 1 > /sys/class/net/eth3/queues/rx-0/rps_cpus).*Checks the hardware queue of a single NIC, from becoming a bottleneck by creating the hash from the IP addresses and Port-numbers, which uses to determine the CPU near send the data packets.
- ***Hash Function Usage:*** The use of the hash function ensures that packets for the same stream of data are sent to the same CPU, which helps to increase performance.

V. Implementation and Experimental Testing Results

To make testing consistent, we will start tests from layer Two open systems interconnections (OSI) model. This means frames filtering with using only bridge code in Linux and then go up to network layer with filtering packets. For all DDoS types we will use: ***'ebtables rules approach'*** to filter Ethernet frames and ***'iptables rules approach'*** to filter IP packets as they are well-known and reliable.

Authentic traffic will be represented as **500** simulated users are sending ***'HTTP get'*** request to web server port **80** each second for **3** minutes. The output of such request and working preconfigured bridge is shown in **Table II and Fig. 9(a).**

First big spike corresponding is to unsuccessful transactions tab which is related to ***'Avalanche-Commander'*** specific functioning. Amount of traffic generated by authentic users on the '***Web server'*** is shown in **Fig. 9(b).**

A. Testing Result's of SYN (Synchronize) Flood Detection

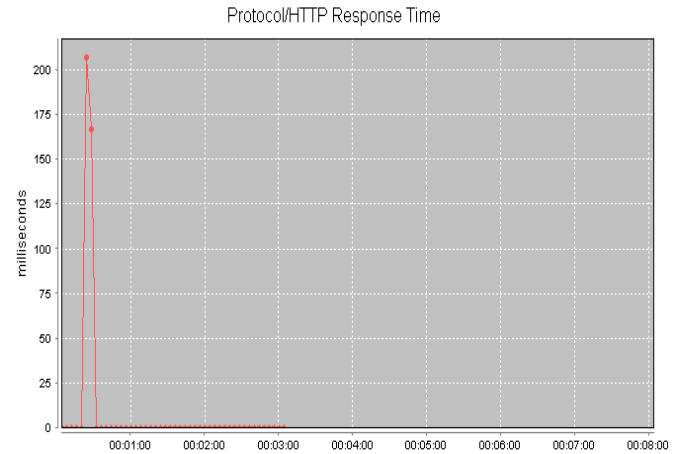
In this test ***'Avalanche Commander'*** sending packets with SYN flags set from IP addresses range **192.168.2.163** to **192.168.2.167**. By generating a big amount of SYN floods packets along with authentic traffic flow it is able to increased server response time besides execution script packet suite, which is shown in **Fig. 10(a).**

Mostly, there is no way to distinguish malicious packets traffic flow with SYN flag set on the Ethernet layer, so for filtering needed detections system which will provide us with malicious traffic flow from DDoS attacks using IP addresses:

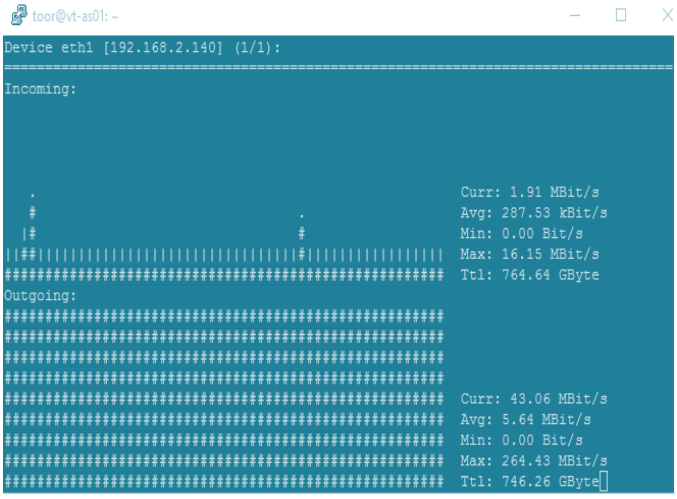
- #ebatables -f
- #ebtables -n syn_flooding
- #ebtables -a forward -p ipv6 --ip-proto tcp --ip-dport 80 -j syn_flooding
- #ebtables -a forward -p ipv6 --ip-proto tcp --ip-dport 443 -j syn_flooding
- #ebtables -a forward -p ipv6 --ip-proto tcp -j drop
- #ebtables -a syn_flooding -p ipv6 --among-src-file data -j drop
- #ebtables --atomic-file syn_flooding -t filter --atomic-save

TABLE II. USERS TRAFFIC FLOW ON WEB SERVERS (WSVRS)

Relations			Time (m s)					TCP Connections	
Total		Rate Per Second		Page Response	URL Response	To TCP SYN/ACK	To First Data Byte	Est. Server Response	Total	
Attempted	91879	491	**Minimum**	0	0	0.094	0.225	0	**Attempted**	91879
Successful	86031	460	**Maximum**	6017	6017	13999.418	6017.389	1996.095	**Established**	87494
Unsuccessful	5848	31	**Average**	14.329	14.329	210.003	14.594	3.928		
Aborted	0	0								



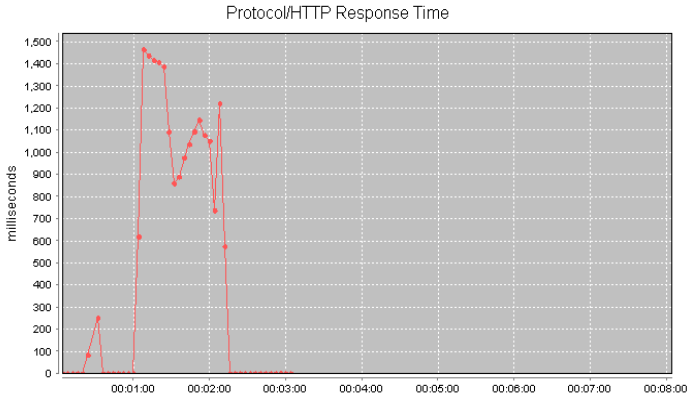
(a)



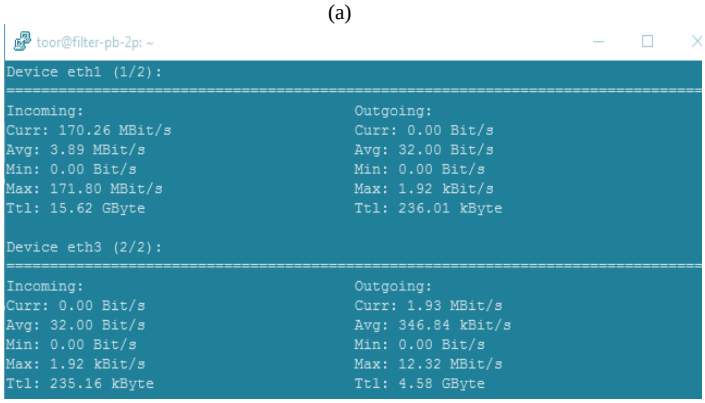
(b)

Fig. 9. (a) Shows specified milliseconds response time for HTTP/Protocol request. (b) Shows authentic user's incoming and outgoing network traffic flow generation on the web server.

In the file ***'data'*** we specified the range of MAC/IP-addresses. To test applied rule we will generate maximum available SYN flood speed and see how much of it is coming from filter server and how much is dropped. Using above IPv6 testing headers, mitigation techniques and ***'ebtables-tool'*** we get the incoming and outgoing SYN flood filter speed which is display in **Fig. 10(b).**



(a)



(b)

Fig. 10. (a) Shows authentic network traffic flow in SYN flood attacks. (b) Shows SYN flood incoming and outgoing filtering speed via ebtables.

As we see only user's traffic is passed through the server. There was no big additional CPU usage, corresponding to frame blocking. The maximum speed of incoming frames which kernel was able to filter is **170-180Mbit/s (~400000pps)**, including users and malicious traffic flow from DDoS - attacks, which corresponds to maximum throughput of NIC. Another way to filter SYN flood is limiting passing traffic flow which is based on '***packets/s'***, to decrease some load from target servers. This resolution is affecting user traffic flow and also it is not preferable.

By default, only ebtables code is able to process bridged frames, so to let ***iptables rules*** receive traffic flow from bridged ports. So we need to enable ***bridgen-of-call-iptables*** feature. With the help of below IPv6 addresses, testing headers, mitigation techniques and ***'iptables-tool'***, we compare filtering flow with iptables rules utilizing following headers and result is display in **Fig. 10(c).**

- #iptables –f
- #iptables –p forward accept
- #iptables -n syn_flooding
- #iptables -a forward -p tcp –m tcp --syn -j syn_flooding
- #iptables -a syn_flooding -m iprange --src-range
- 192.168.2.163- 162.168.2.167 -j drop

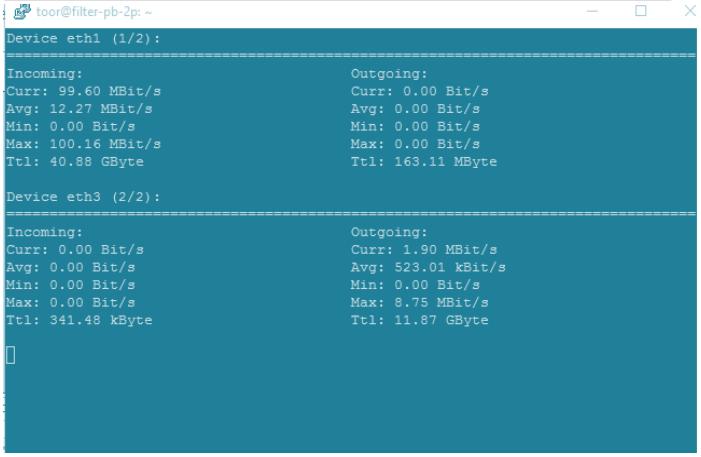
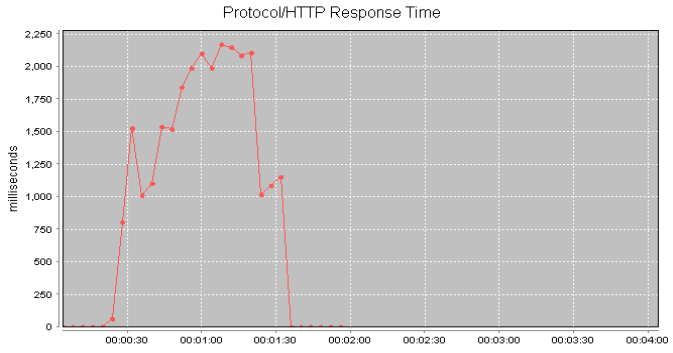
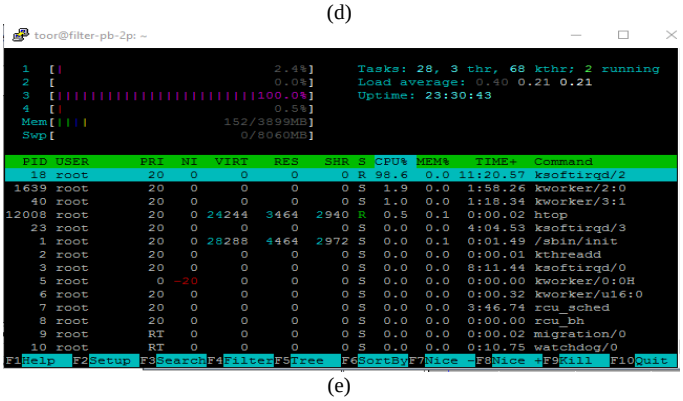


Fig. 10. (c) Shows SYN flood incoming and outgoing attacks via iptables filtering.

Although we are able to filter traffic on speed of **90-100Mbit/s**, the web server response time is still greatly increased due to CPU overloading, which is made by *Ksoftirqd-Per-CPUs* built kernel cord in order that turns while the device is below full Soft-Interrupt load [42]. Increased response time and CPUs utilization are shown in **Fig. 10(d) and (e).**



(d)



(e)

Fig. 10. (d) Shows milliseconds response time in SYN flood attacks. (e) Shows results of CPUs utilization.

Previously we disabled ***'rpc_cpus'*** feature to test CPUs utilization. Now we enable it back a run same test again. The result is shown in **Fig. 10(f).**

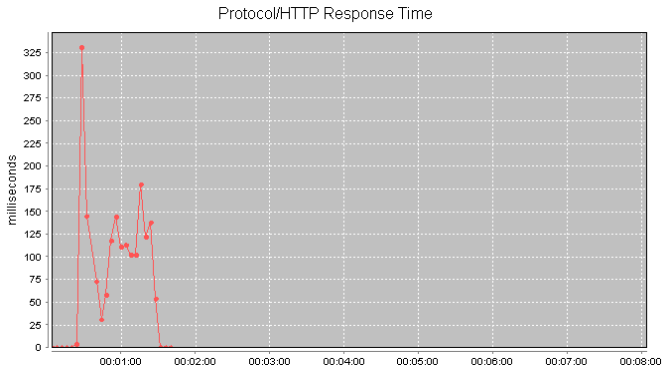


Fig. 10. (f) Shows milliseconds SYN flood attacks response time in *'smp_affinnity'*.

Only **3** out of **4** cores are loaded with processing packets, which decreasing response time to the acceptable range. However maximum amount of possible filtered traffic is increased from the **90Mbit/s** to **130Mbit/s**. Assuming using more complex iptables rules with same user's data traffic will lead to DDoS attacks. So we utilize more complex rules using headers and ***'Conntrack-tool'*** Linux kernel as follows:

These rules set the limit on amount connections per second coming from one IP address, assuming all **500** users will start simultaneously sending connection requests. The data traffic filtering speed, in this case, is around: **100Mbit/s** with enabled load spread. Now we compare same new tests to those which were conducted on all filter servers and compared results are shown in **Table III**.

- #iptables -f
- #iptables -n syn_flooding
- #iptables -a forward -p tcp -m state –state-new -j syn_flooding
- #iptables -a syn_flooding -m connlimit --connlimit-above 500 -j drop

TABLE III. SYN (SYNCHRONIZE) FLOOD ATTACKS FILTERING

SYN Flood Mbit/s	The New Testing Filter V1	The New Testing Filter V2	The New Testing Filter V3
Ebtables	182	178	174
Iptables V1	130	177	102
Iptables V2	100	184	83

As we can see, ebtables successfully allows filtering traffic on desired load. However, using ebtables can only be related with existence of any kind of detection system, which supplies filter with needed data. Using iptables rules we are able to filter in more stand-alone way but it requires more hardware resources to use in order to filter the same amount of data – traffic per second. Filter with aggregated links can benefit from having 1Gbps links instead of one, since the *'bottleneck'* is not in filtering. More of that due to less computing resources it show lower filtering throughput.

B. *Testing Result's of User Datagram Protocol (UDP) Flood Detection*

The User datagram protocol (UDP) flood attacks testing in our network based on sending as many packets as possible on web server port No.**80** with spoofed source IP addresses.

The main goal is to utilize all data filter servers for CPUs usage. Effect from generating UDP flood attacks can be seen in **Fig. 11(a) and (b).**

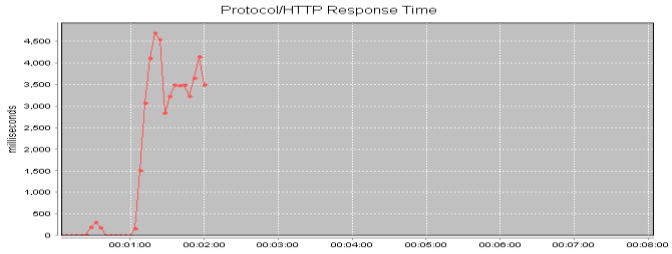


Fig. 11. (a) Shows Web Servers milliseconds response time in UDP flood attacks.

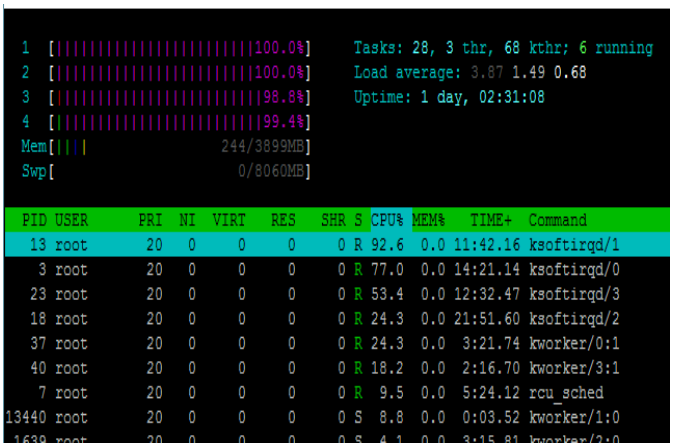


Fig. 11. (b) Shows results of CPUs utilization by UDP flood attacks.

On link layer to filter UDP flood attacks we able to set following headers and mitigation technique rules: UDP data packets which are destined to domain name servers (DNS) server on port 53 will pass, all others UDP data packets should be dropped. Results are shown in **Fig. 11(c) and (d).**

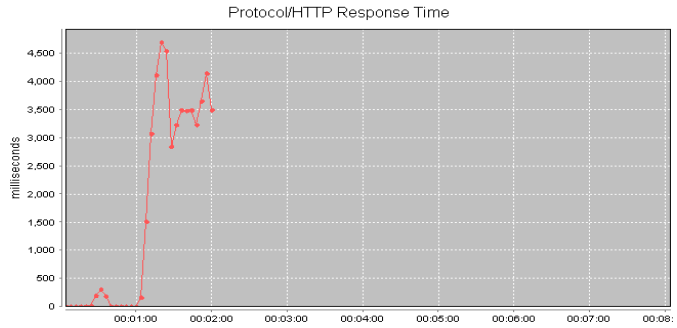


Fig. 11. (c) Shows Web Server milliseconds response time by data filtering.

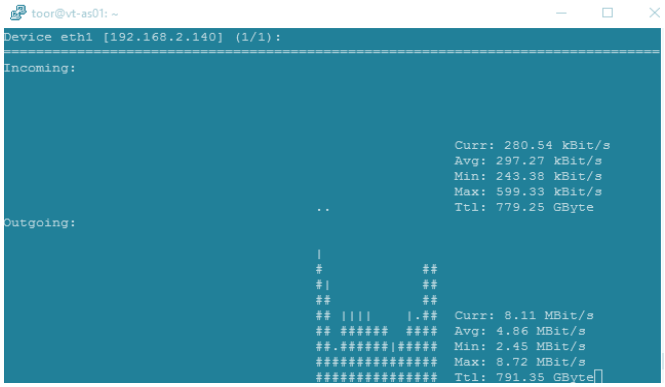


Fig. 11. (d) Shows results of Web Server incoming and outgoing users load.

- #ebtbales - f
- #ebtables - n udp_flooding
- #ebtables - p udp_flooding drop
- #ebtables - a forward -p ipv6 --ip-proto udp! --ip-dport 53 -j udp_flooding

As it can be seen on Web server data traffic that is still struggle to pass through filtering. All **4** filter cores server are loaded with 'Ksoftirqd-Program' and filtering is possible but web server has still big response delay.

To reduce CPUs utilization on filter, it's possible to apply dropping packets even before them being processed by kernel. Among ebtables hooks there is nat table with: the PREROUTING chain which is logically located between kernel network stack and NIC.

Applying same rules in table nat PREROUTING chain results for better performance, which can be shown in **Fig. 11(e) and (f).**

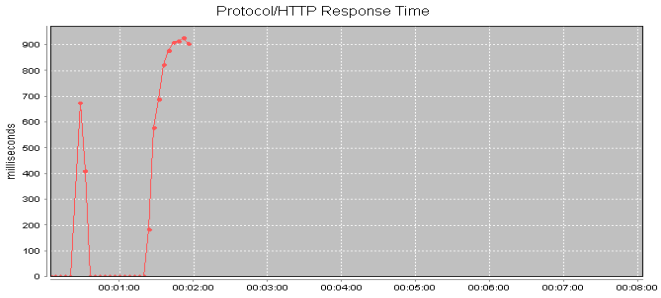


Fig. 11. (e) Shows users traffic milliseconds response time.

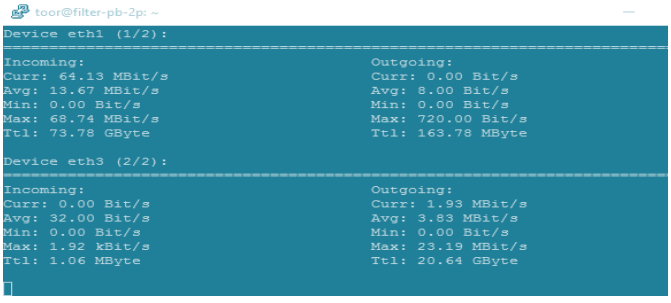


Fig. 11. (f) Shows results of filtering incoming and outgoing UDP flood attacks with ebtables.

In the part of user datagram protocol (UDP) Flood was decreased along with 'http-response time'. Although it stays in acceptable range under the **1** second further countermeasures in network are required.

Considering ebtables experience, iptables rules have to be applied on corresponding netfilter hook to have better result. So iptables table raw with: the PREROUTING chain should be configured as follows:

- #iptables - f
- #iptables – t raw –a pre – routing –p udp –dport 53 –d 192.168.2.145 -j
- Accept
- #iptables - t raw -a pre – routing –p udp –d 192.168.2.145 –j drop

For all passing user datagram protocol (UDP) packets going to web server should be dropped, except destined to domain name servers (DNS), we set the above headers. The Results output is shown in **Fig. 11(g) and (h).**

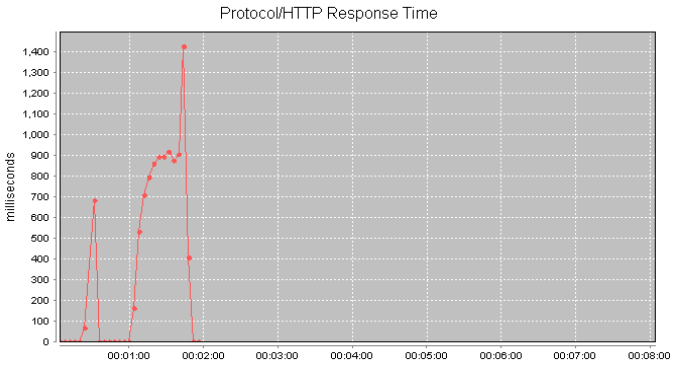


Fig. 11. (g) Shows milliseconds response time of the UDP flood with iptables.

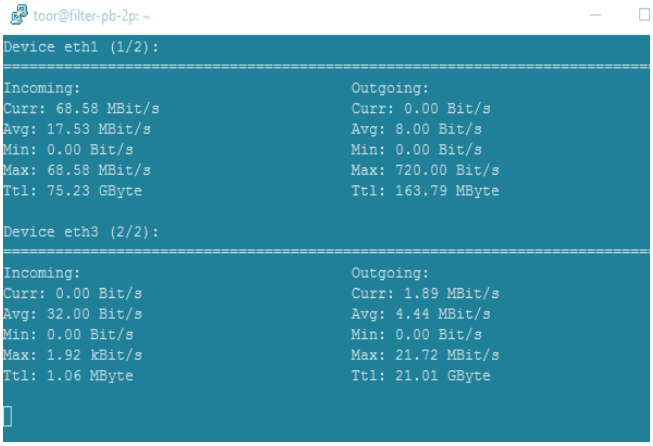


Fig. 11. (h) Shows results of maximum incoming and outgoing UDP flood traffic through output.

How can be seen both ebtables and iptables are able to filter UDP, flood attacks. However, using iptables gives us litter bigger delay while communicating with web server. That is consequences of that iptables uses more code to process each packet, so it need more calculating time for the rest of server. Same tests were conducted and results are combined in **Table IV.**

TABLE IV. CONDUCTED REST OF SERVERS UDP FLOOD FILTERING TESTS RESULTS

UDP FLOOD (Mbit/s)	The New Testing Filter V1	The New Testing Filter V2	The New Testing Filter V3
Ebtables	67	67	67
Iptables	68	64	65

As this attack doesn't consume many resources, each server was able to completely filter UDP flood on maximum available any through output.

C. *Testing Result's of Internet Control Message Protocol (ICMP) Flood Detection*

Internet control message protocol (ICMP) flood attacks rely on constantly sending 'echo_request' to force our web server to respond and consume additional resources. Generating attacks along with sending users data traffic flow also consumes, as shown in **Fig. 12 (a) and (b).**

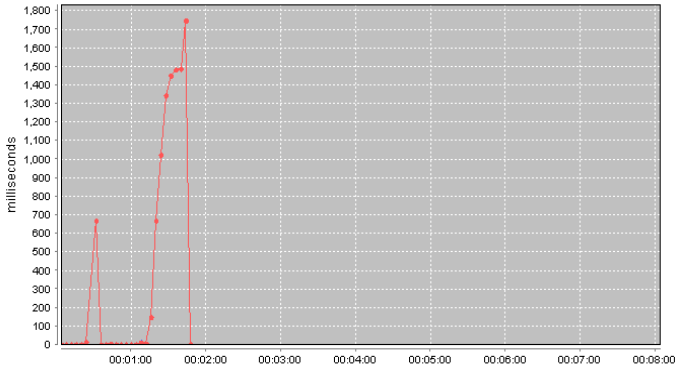


Fig. 12. (a) Shows HTTP network traffic milliseconds response time.

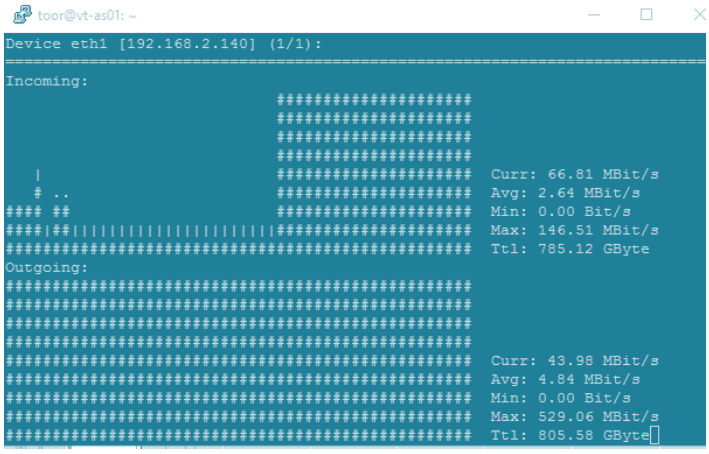


Fig. 12. (b) Shows results of Web Server incoming and outgoing load in ICMP flood attacks.

ICMP, flood data traffic flow is crucial for network control and it cannot be dropped at all. The best way is to filter data flow using following headers and ***'link-layer'*** mitigating technique for set limits:

As a standard icmp_echo request rate is 12 packet per 10 second, our limit is letting users to ping servers and describing filtering process which is shown in **Fig. 12 (c) and (d)**.

- #iptables -n icmp_flooding
- #iptables -a forward -p icmp -j icmp_flooding
- #iptables -a icmp_flooding -p icmp --icmp-type echo-request -s 192.168.2.0/24 -j Accept
- # Iptables -a icmp_flooding -j drop

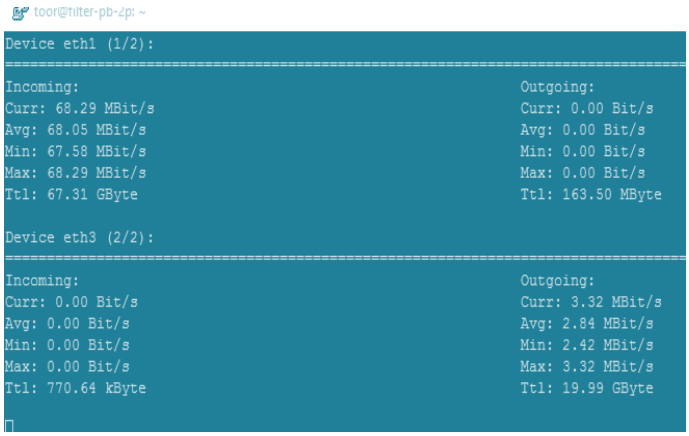


Fig. 12. (c) Shows ICMP flood incoming and outgoing filtering using ebtables.

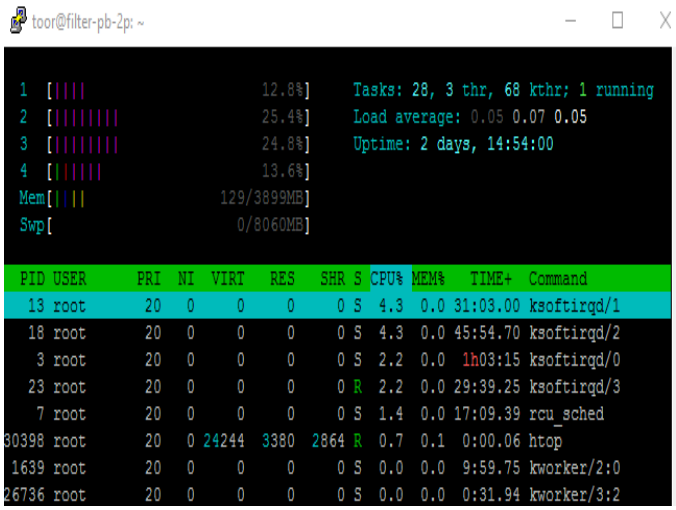


Fig. 12. (d) Shows results of CPUs load in ICMP filtering.

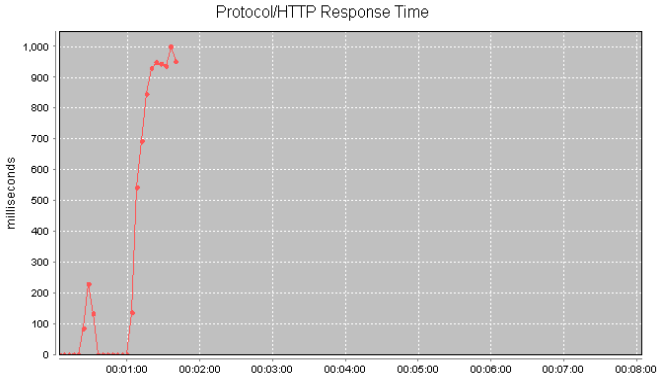


Fig. 12. (e) Shows milliseconds response time of ICMP flood filtering with iptables.

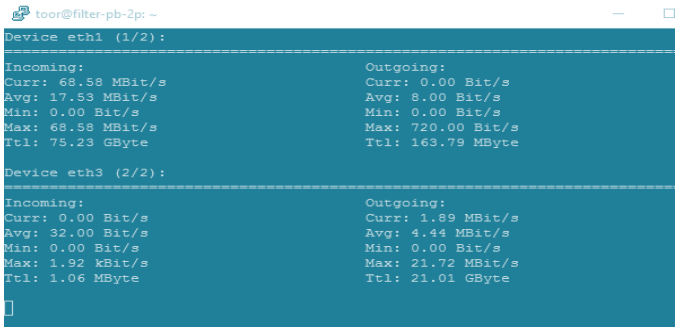


Fig. 12. (f) Shows results of the incoming and outgoing response time in iptables filtering.

As can be seen there is no restriction from hardware on filtering side, but still there is a small delay related with packets processing. On network layer it's possible to deny any: echo_request packets from outside of our network, since we want to leave troubleshooting options for network administrator. For that purpose we set following headers and need to follow below rules and regulations. Desired results are described in **Fig. 12 (e) and (f)**.

As expected all malicious internet control message protocol (ICMP) traffic was filtered by adding a small delay in response time. The results for rest of the servers are described in **Table V.**

TABLE V. ICMP FLOOD NEW TESTING RESULTS

ICMP FLOOD (Mbit/s)	The New Testing Filter V1	The New Testing Filter V2	The New Testing Filter V3
Ebtables	68	70	69
Iptables	68	68	68

A tiny part of computing resources is required for processing internet control message protocol (ICMP) flood, which makes this type of attacks filtering less complex and available across networks worldwide.

VI. CONCLUSION AND OUTLOOKS

In this paper, we discussed how Linux-based resolution approaches for Web Servers WSYRs, tools and mitigation technique principles can be applied to effectively control DoS and DDoS attacks.

These approaches can give several benefits to WWW, users. The proposed approaches pose a number of challenges for DDN, management and cooperative communities on the internet technology losses.

Although DDoS attacks are effectively achieving such goals, you can find reasonable protection against certain types of DoS and DDoS attacks. In other words, the frequency, power, harshness of bandwidth, processor speed, and the number of available systems that can be attacked and compromised will continue to grow, as well as attack tools complexity to compromise with computers and their usage for attacks.

To mitigate DoS and DDoS attacks usually requires many different security mechanisms. However, the situation remains

no more advisable toward softly picking a huge usually like protection devices in contradiction of DoS and DDoS attacks.

This smart-stated research work focus was to test the correct tools to filter data traffic flow. Maximum throughput achieved with SYN (Synchronize) filtering flood attack is **187 Mbit/s,** user datagram protocol (UDP) flood max. Performance is **67 Mbit/s,** and internet control message protocol (ICMP) flood is filtered to the maximum **71 Mbit/s**.

In view of the major consideration, we can say in this study that the Linux-based resolution approaches provide the complete set of tools needed to filter traffic flow. For example, the SYN Flood DDoS attack can be reduced by packet filtering using iptables without additional hardware or software.

"But by installing additional detection systems on the network, the best filtering performance can be achieved with the ebtables tool."

The difference lies in the packet flow within the central network stack.

DDoS attacks on the UDP, thread can be alleviated with all available tools, but additional kernel configuration is required. The ability to add aggregates with Linux-based software can be used in cases where a high-voltage network is used.

An extra speed can be achieved for users who use so-called: **'bonding'**.

In real time, proposed methods to mitigate DDoS should be considered temporary measures due to limited resources in the provided systems.

The latest DDoS speeds in the world outperform the possibilities of any filtering equipment or software. Therefore, good cooperation between all parties involved in global WWW, work is required to successfully filter malicious traffic.

VII. Future Scope

Expressive netfilter or iptables coordination stays perfect model for Linux-based method proprietors, system proprietors. It may also be suitable for all users which need towards design firewalls, allowing near their explicit requirements. In the direction of excluding change arranged firewall resolutions, besides entire device packets IP filtering.

Although this research paper provides, adequate protection against the refusal of attacks, many areas remain under the protection mechanism used for further investigation that can be expanded by this work by using iptables advanced features such as network address translation (NAT), IP masking, redirect IP Tables packets.

These have not only the ability to redirect packets, such as IP chains, but also have a destination NAT (DNAT) allocation, which allows to freely change the distribution of the IP address and port-numbers.

So you can actually hide web address where the service packets go through DNAT. Additional dynamic exploration part could be the development of policy scripts in the ***'Bro-language'*** to detect DoS attacks using ***'Bro-IDS'*** and develop effective against Anti DDoS ***'Snort-IDS'*** rules.

Acknowledgment

The Publication of this research was support by Nature Science Foundation of China under Grant No. 61572095.

Abbreviations

The following abbreviations are used in this research manuscript:

ACK	Acknowledgement Data Networks
LAPS	Linux - Availability Protection System
BSD	Berkeley Software Distribution
CPUs	Central Processing Units
CUSUM	Cumulative Sum Control Chart
DoS	Denial-of-Service
DDoS	Distributed Denial-of-Service
DDN	Data-Driven Networks
DNS	Domain Name Servers
DIT	Digital Information Technology
DNAT	Destination Network Address Translation
FIFO	First In, First Out
ICMP	Internet Control Message Protocol
ISPs	Internet Service Providers
IRC	Internet Relay Chat
IP	Internet Protocol
IETF	Internet Engineering Task Force
LPS	Linux Protection System
LAG	Linux Aggregation Group
LCPUs	Logical Central Processing Units
MIB	Management Information Base
MSS	Ministry of State Security
MSS	Maximum Segment Size
MAC	Media Access Protocol
NIC	Network Interface Card
NAT	Network Address Translation
OS	Operating System
OSI	Open System Interconnections
PDOS	Pulsing Distributed Denial-of-Service
P2P	Peer-to-Peer
QDISCS	Queuing Disciplines
RSTP	Rapid Spanning Tree Protocol
RFC	Request for Comments
RC	Remote Control
SYN	Synchronize
SDBot	Spartan-Dominion Robot
SQLi	SQL Injection
TFN	Tribe Flood Network
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VPN	Virtual Private Network
WSVRs	Web Servers
WWW	World Wide Web

References

[1] Cobb, S., things to know about the October 21 IoT DDos attacks,". WeLiveSecurity, In press release. 24.

[2] Wu, L., Designing Effective Security and Privacy Schemes for Wireless Mobile Devices. 2017, Temple University.

[3] Bonguet, A. and M. Bellaiche, A Survey of Denial-of-Service and Distributed Denial of Service Attacks and Defenses in Cloud Computing. Future Internet, 2017. 9(3): p. 43.

[4] Saleh, M.A. and A. Abdul Manaf, A novel protective framework for defeating http-based denial of service and distributed denial of service attacks. The Scientific World Journal, 2015. 2015.

[5] Tahir, M., et al., The Novelty of A-Web based Adaptive Data-Driven Networks (DDN) Management & Cooperative Communities on the Internet Technology. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 2017. 8(5): p. 16-24.

[6] Alam, M.F., Application layer ddos a practical approach & mitigation techniques. bdHUB Limited, 2014.

[7] Hadi, A.D.A., F.H. Azmat, and F.H.M. Ali. IDS Using Mitigation Rules Approach to Mitigate ICMP Attacks. in Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on. 2013. IEEE.

[8] Zargar, S.T., J. Joshi, and D. Tipper, A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. IEEE communications surveys & tutorials, 2013. 15(4): p. 2046-2069.

[9] Bhuyan, M.H., D.K. Bhattacharyya, and J.K. Kalita, Network anomaly detection: methods, systems and tools. Ieee communications surveys & tutorials, 2014. 16(1): p. 303-336.

[10] François, J., I. Aib, and R. Boutaba, FireCol: a collaborative protection network for the detection of flooding DDoS attacks. IEEE/ACM Transactions on Networking (TON), 2012. 20(6): p. 1828-1841.

[11] Senie, D. and P. Ferguson, Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing. Network, 1998.

[12] Park, K. and H. Lee. On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets. in ACM SIGCOMM computer communication review. 2001. ACM.

[13] Savage, S., et al., Network support for IP traceback. IEEE/ACM transactions on networking, 2001. 9(3): p. 226-237.

[14] Song, D.X. and A. Perrig. Advanced and authenticated marking schemes for IP traceback. in INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. 2001. IEEE.

[15] Bellovin, S.M., M. Leech, and T. Taylor, ICMP traceback messages. 2003.

[16] Mahajan, R., et al., Controlling high bandwidth aggregates in the network. ACM SIGCOMM Computer Communication Review, 2002. 32(3): p. 62-73.

[17] Siris, V.A. and F. Papagalou. Application of anomaly detection algorithms for detecting SYN flooding attacks. in Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE. 2004. IEEE.

[18] Wang, H., D. Zhang, and K.G. Shin. Detecting SYN flooding attacks. in INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. 2002. IEEE.

[19] Luo, X. and R.K. Chang. On a New Class of Pulsing Denial-of-Service Attacks and the Defense. in NDSS. 2005.

[20] Cabrera, J.B., et al. Proactive detection of distributed denial of service attacks using mib traffic variables-a feasibility study. in Integrated Network Management Proceedings, 2001 IEEE/IFIP International Symposium on. 2001. IEEE.

[21] Russell, R., Equity in Eden: can environmental protection and affordable housing comfortably cohabit in suburbia. BC Envtl. Aff. L. Rev., 2002. 30: p. 437.

[22] Russell, R. and H. Welte, Linux netfilter hacking howto. Disponıvel em http://www. netfilter. org/documentation/HOWTO//netfilter-hacking-HOWTO. letter. ps (Junho de 2005), 2002.

[23] Fouda, K.M., Payload based signature generation for DDoS attacks. 2017, University of Twente.

[24] Lau, F., et al. Distributed denial of service attacks. in Systems, Man, and Cybernetics, 2000 IEEE International Conference on. 2000. IEEE.

[25] Tunc, C. and S. Hariri, CLaaS: Cybersecurity Lab as a Service. J. Internet Serv. Inf. Secur., 2015. 5(4): p. 41-59.

[26] Xu, R., W.-l. Ma, and W.-l. Zheng. Defending against UDP flooding by negative selection algorithm based on eigenvalue sets. in Information Assurance and Security, 2009. IAS'09. Fifth International Conference on. 2009. IEEE.

[27] Douligeris, C. and A. Mitrokotsa, DDoS attacks and defense mechanisms: classification and state-of-the-art. Computer Networks, 2004. 44(5): p. 643-666.

[28] Bitzer, J. and P.J. Srhroder, Linux vs. Windows: A Comparison of Application and Platform Innovation Incentives for Open Source and Proprietary Software Platforms "published in Elsevier BV. 2006.

[29] Chakeres, I.D. and E.M. Belding-Royer. AODV routing protocol implementation design. in Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference on. 2004. IEEE.

[30] Rivero de la Cruz, A., High available GNU/Linux systems. 2014, Universitat Oberta de Catalunya.

[31] Radhakrishnan, S., Linux–advanced networking overview version 1. 1999.

[32] de Schuymer, B. and N. Fedchik, Ebtables/Iptables interaction on a Linux-based Bridge. 2003.

[33] Hoffman, D., D. Prabhakar, and P. Strooper. Testing iptables. in Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research. 2003. IBM Press.

[34] Andreasson, O., Iptables Tutorial 1.1. 19. Internet article at: http://people. unix-fu. org/andreasson/iptables-tutorial/iptables-tutorial. html, 2001.

[35] Diekmann, C., et al. Verified iptables firewall analysis. in IFIP Networking Conference (IFIP Networking) and Workshops, 2016. 2016. IEEE.

[36] Shannon, F.M., et al., Link aggregation protection. 2012, Google Patents.

[37] Williams, M., Linux ethernet bonding driver HOWTO. 2006, April.

[38] Chen, S.-J., et al., A Behavior-Based Web Application Firewall Total Solution to Detect Various Web Application Service Attacks.

[39] Perloff, R.S., A.H. Frederik, and T.J. Christian, Multi-device link aggregation. 2005, Google Patents.

[40] Dow, E., S. Loveland, and G. Markos, A reference implementation architecture for deploying a highly-available networking infrastructure for cloud computing and virtual environments using OSPF. IBM Platform Test-z Systems library, 2009.

[41] Bowen, J.L., et al., NLOAD: AN INTERACTIVE, WEB- BASED MODELING TOOL FOR NITROGEN MANAGEMENT IN ESTUARIES. Ecological Applications, 2007. 17(sp5).

[42] Aust, S., et al. Evaluation of Linux bonding features. in Communication Technology, 2006. ICCT'06. International Conference on. 2006. IEEE.

AUTHORS' PROFILES

Muhammad Tahir: Was born in 1987. He received his Bachelor of Science BS Degree in Software Engineering from University of Sindh, Jamshoro, Pakistan in 2008. And Master of Engineering Degree in Software Engineering from CHONGQING University, Chongqing P.R. China in 2014. Currently he is a Ph.D. Research Scholar in School of Software Technology, Dalian University of Technology, P.R. China. His research interest includes: Cooperative Game Theory, Web based Software Defect Prediction Models, Web-based Security & its Testing, IOT, Cloud & Fog Computing, Software Defined Networks (SDN) & Networking.

Mingchu Li: Received the BS Degree in Mathematics, from Jiangxi Normal University, P.R. China and the MS Degree in Applied Science from University of Science and Technology, Beijing, P.R. China. He worked as an Associate Professor from 1989 to 1994. He worked for School of Software, Tianjin University, as a full Professor, (from 2002 to 2004) & from (2004 to now), for School of Software Technology, Dalian University of Technology as a full Professor, and Vice Dean. His main research interests include Theoretical Computer Science & Information Security, and Trust Models and Cooperative Game theory.

Naeem Ayoub: Became the student member of IEEE in 2017. He is pursuing his Ph.D. Degree at School of Computer Science and Application Technology, Dalian University of Technology, P.R. China. He received his Master's Degree in information technology from University of Education Lahore, Pakistan in 2013. His research interest includes: Computer

Vision, Image Processing, Saliency Detection, Iris Recognition, IOT and Wireless Sensor Networks (WSNs).

Usman Shehzaib: Is an Assistant Professor of Computer Science, in the Department of Computer Science, COMSATS Institute of Information Technology Lahore, Pakistan. His research interest includes: Databases, Data Mining and Big Data.

Atif Wagan: Is a Ph.D. Research Scholar in the School of Computer Science and Engineering, Nanjing University of Science & Technology, (NUST), Nanjing, P.R. China. His research interest includes: Software Testing, Big Data, Artificial Intelligence, Machine Learning and Deep Learning.

Choice of Knowledge Representation Model for Development of Knowledge Base: Possible Solutions

Sabina Katalnikova, Leonids Novickis
Faculty of Computer Science and Information Technology
Riga Technical University
Riga, Latvia

***Abstract*—In current society knowledge, information and intelligent computer systems based on knowledge base play a great role. The ability of an intelligent system to efficiently implement its functions depends on the efficiency of organizing knowledge base, and on the fact whether the applied knowledge representation models comply with the set requirements. The article is devoted to the research of the problem of choosing the knowledge representation models. Based on the requirement analysis for knowledge representation models, one of the solutions for the researched problem shown is application of extended semantic networks. Analysis of extended semantic networks' properties is carried out, as well as relevant examples of representing knowledge of extended semantic networks' application for various spheres offered.**

***Keywords*—Extended semantic networks; knowledge base; knowledge representation model; semantic networks**

I. Introduction

The end of 20th and the beginning of 21st century can be characterized by transfer from industrial to the so-called information society the peculiarity of which is a significant increase of the role of knowledge and information. Transfer from economy which was dominated by such traditional factors as land, labour and capital to knowledge based economy marked a new approach to the concept of economic efficiency. In the circumstances of the new reality society's interest in information and knowledge grew both in theoretical and practical aspect as apparently the lack of innovation, innovative products and services significantly decreases economic efficiency.

Computer systems which use data bases to solve several typical formalized tasks are based on developed rules, models and algorithms. On the other hand, the arisen complicated tasks are not always solvable with formalized rules and algorithms. In order to solve new problems, knowledge rooted intelligent systems are applied, which are based on knowledge bases and where the main focus is processing knowledge.

One of the central and most complex problems solved in development of knowledge-based systems is the problem of knowledge representation and processing: the efficiency of system being developed and the correctness of solutions obtained by its means depend on success of this problem's solution [1].

Many articles in the world are dedicated to the problem of knowledge representation [2]-[4]. Success of adjusting knowledge representation models directly depends on the fact whether the applied models comply with the set requirements. Problem of choosing knowledge representation model and processing methods can be defined in the following way: how to represent knowledge structures from such sources as special literature and knowledge of highly qualified professionals (namely, to choose knowledge representation model), so that their automatized processing could allow efficient solutions of domain tasks and gain positive results.

The paper is organized as follows. Section II describes requirements for knowledge representation model. Section III outlines one of the knowledge representation models - semantic networks and their advantages. Section IV presents a general description of extended semantic networks, while Section V discusses in detail their possibilities. Section VI concludes the paper by briefly discussing the future direction of research.

II. Requirements for Knowledge Representation Model

The basis for intelligent system is knowledge base which comprises all the information that the system uses in a systemized way. Thus, all the used and workable information within intelligent systems is presented in a semantically structured unified knowledge base which represents a notion of a wholesome world where this system "lives".

The aim of knowledge representation is the organization of necessary information is such a form that the intelligent system would be able to apply it for decision making, planning, analysis, judgment output and other function implementation process. In order for the knowledge representation model to be applicable in development of an intelligent system, it has to ensure representation of all the knowledge types necessary for the operation of the intelligent system.

Firstly, the knowledge base of the intelligent system must contain knowledge on the domain where this system will function. That is knowledge on basic units (concepts and objects) within the domain, as well as knowledge on how these units are related. Such knowledge includes relations that directly connect concepts as well as more sophisticated types of knowledge that represent several types of dependency among the domain concepts (logical and functional).

The majority of intelligent system knowledge bases also contain particular knowledge on the domain (subject

knowledge) that is represented as concept specimens (in a form of particular objects) and the relation between them – in a form of relations' or restrictions' specimens.

Another important type of knowledge necessary to be represented in an intelligent system is knowledge on problems and solutions within the modelled domain (methods and algorithms). This knowledge characterizes problem environment of the intelligent system. Such knowledge can be declarative and procedural. Declarative knowledge describes division of tasks in subtasks and their link to the solution methods. Such information is knowledge received from the user. This information is included in the system, it constantly changes and determines the system solutions. Procedural knowledge is task solving methods and particular algorithms. Such knowledge is developed once in such a way that the setup of the system is done based only on declarative knowledge.

In addition to knowledge included in the knowledge base, it is necessary that the knowledge describing a fragment of reality (situation) that defines the context and entry data for tasks solved by the intelligent system are represented in the intelligent system. Such knowledge, similarly to subject knowledge, usually is given as concept and relation and/or restriction specimen type.

There are many requirements submitted to a knowledge representation model. By analyzing these requirements, it is possible to define a requirement cluster for knowledge representation model in intelligent systems, namely, [5]-[8]:

- representation of knowledge meaning; acquiring a unified character of knowledge to be represented with an intention to comply with all the substantial objects from the viewpoint of the solvable task, their characteristics and relations, and ignore the irrelevant ones;
- representation of knowledge within concepts of natural language of the domain to be studied; clearness of development and representation of logical links and semantic relations of domain to be studied;
- preservation of initial information and acquisition of new information;
- representation of hierarchical structure of knowledge;
- possibilities of representing fuzzy knowledge,
- representation of both declarative and procedural knowledge;
- representation possibilities of logical operations and quantifiers;
- representation possibilities of intentional and extensional;
- possibilities of recognizing contradiction in knowledge to be represented;
- model uniformity;
- provision of integrity of knowledge to be represented;
- possibility of merging the knowledge structures.

III. SEMANTIC NETWORKS

One of the knowledge representation models is net-type model where the domain is examined as a body of objects and their binding relations (for instance, semantic networks, conceptual graphs). Knowledge representation in network models is the closest to knowledge in natural language texts. Initially semantic network was made as representation model of long-term memory structure in psychology, but later it became one of the basic types of knowledge representation. The task of semantic networks is representation of concept clusters, namely, establishment of basic organisation of domain notions. The necessities for development of a semantic network [9]:

- analysis of structural interworking of content to be researched;
- exhaustive description of concepts and their relations;
- thorough processing of knowledge;
- the link between the new concepts and the existing concepts and notions.

Knowledge representation concept in the semantic network type is based on the idea that all the knowledge can be represented as a cluster of objects (concepts) and links (relation) between them. Semantic network possesses such characteristics from the viewpoint of the requirements mentioned in the previous section: knowledge representation in natural language notions, declarative knowledge representation, domain semantic link representation, clearness of knowledge description, integrity of knowledge structure representation. Thus, it is possible to conclude that a necessary model can be continuation of semantic network model adding to its logic and computing property [1].

In the example of the semantic network it is possible to establish the difference of data base (working memory) and knowledge base. Domain is a cluster of possible conditions of its entities. This cluster which is represented through common terminology, concepts, relations and laws creates knowledge base as an intentional semantic network. But in every particular situation characteristics of this domain entities have particular values. This particular data is represented in extensional semantic network (data base or working memory). Working memory is used to store temporary data. Information on aims, current tasks, finished tasks, incoming and outgoing messages and short-term connections are located here.

Advantages of semantic models' information processing [10], [11]:

- similarity of semantic network structure to semantic structure of natural language phrases;
- clearness of knowledge representation model;
- allness that is achieved as a result of choosing the appropriate relation cluster;
- knowledge representation with semantic networks allows significant simplification of knowledge

integration process that is implemented as identification and pasting of synonymic elements of integrated semantic networks;

- properly developed intelligent system's knowledge base as a semantic network completely eliminates doubling of information within such knowledge base;
- knowledge representation as a semantic network simplifies the associative access for various types of knowledge based fragments;
- knowledge processing semantic models are well suited for parallel asyncronic processing of information.

IV. Extended Semantic Networks

As it was said, knowledge base is a depository of various types of knowledge that constantly, without any restrictions has to store, change, and adjust. It is possible if knowledge representation model allows rather arbitrary modification of its constructions. Thus, it is preferable for the inner language to include simple, uniform structures, which could be removed and added. On basis of such considerations conception of uniformity developed. Semantic networks comply with requirements of uniformity, but are restricted, for instance, from the viewpoint of generalized information, relations between situations or relation representation. In relation to this, extended semantic networks were developed in which node-concept necessary clusters and special complex elements – link nodes were introduced [12]-[15]. Extended semantic networks can be used for representation of sophisticated objects, logical, generalized information, different requirements and so on.

In extended semantic networks the nodes correspond not only to objects or concepts, but also to relations, logical components of information (truth or untruth facts), complex objects and so on. All that could be regarded as an independent unit must correspond to a separate node. In such networks instead of semantic network edges the so-called link nodes are used. This node does not correspond to any object or relation, it is used only for indicating the link and ensures unified significance for nodes that correspond to separate components or information units. As a result a fragment that corresponds to the elementary situation forms. From such fragments the networks are composed.

There are also special constructions that are called semantic graphs. With their help it is additionally listed which component should be searched first, which – afterwards and so on. Every graph gives its own operations that are carried out on the network and leads to finding or distributing to network nodes. Network can be regarded as a special case of graph, namely, network is a graph where the processing direction is not set. Networks and graphs are composed of uniform fragments, every fragment can be removed or added to the network without damaging the correctness of syntax or semantics.

A special complex element – link node – connects with the help of marked edges to node-relation and nodes-objects, as a result a fragment is made that corresponds to an elementary situation, that is, objects related to a certain relation. This fragment is called an elementary fragment in the following way: D_0 (D_1, D_2, ... , D_k / D_{k+1}), where D_0 is the relation word, D_1, D_2, ... , D_k – objects participation in relation, D_{k+1} – link node that describes the whole cluster participating in relation, D_0, D_1, D_2, ... , $D_{k+1} \in D$, D – node cluster, $\kappa>0$. Extended semantic networks are regarded as a final cluster of elementary fragments. With the help of semantic networks relation cluster, various situations, scenarios are represented. Every elementary fragment has its own link node that describes its fragment [14].

Formal description of extended semantic network [14]:

1) if $\{D_0, D_1, D_2, ... , D_k, D_{k+1}\} \subseteq D$, $k>0$, then D_0 (D_1, D_2, ... , D_k / D_{k+1}) = T_0;

2) every T_k is extended semantic network;

3) if T_1 and T_2 – extended semantic networks, then compositions T_1T_2 and T_1T_2 are also extended semantic networks, moreover, $T_1T_2 \equiv T_2T_1$.

In an extended semantic network a rather free placement of nodes in various positions is allowed. Any node that stands in a position of any elementary fragment (and represents a particular object or object cluster, or relation word) can be placed in a different position of a different fragment. As a result, it is possible to represent a situation where relation words have the role of objects and close their relations. Link node of an elementary fragment can also be included in different elementary fragments but in a different role. With the help of such elementary fragments representation of such cases when some cluster is reviewed as a complex object that, for its turn, comes into relation is ensured. It defines the model's uniformity.

For processing of extended networks comparison by example principle is applied using a method of two network overlay. This principle is based on identification rules that allow linking the nodes and compare the networks on the basis of logic laws [10].

V. Review of Extended Semantic Networks

We shall review the compliance of extended semantic networks to a requirement cluster that is proposed for knowledge representation model in intelligent systems. Extended semantic network model characteristics with informative examples are summarized in Table I.

By analyzing the examples and explanations offered in the table, it is possible to conclude that the extended semantic networks comply with the requirements of knowledge representation model in intelligent systems and can be used for development of knowledge base.

TABLE I. EXTENDED SEMANTIC NETWORK MODEL POSSIBILITIES/OPTIONS

Model Properties	Explanation
Knowledge meaning representation, acquisition of uniform character of representable knowledge	Semantic network by definition is a knowledge system with a definite meaning in a network, the nodes of which correspond to concepts and objects, and edges – to concepts and object relations, in an integrated character type [16]. Implementation of knowledge meaning representation requirement is achieved by including in the network structure relations that exist among the object of domain to be studied.
Knowledge representation in natural language notions, knowledge description clearness	For example, natural language expression "Cranberry – red, sour berry, which grows in a bog" is depicted in the following way [13]: SUB('berry','cranberry') COLOUR('cranberry','red') TASTE('cranberry','sour') GROWS_IN('cranberry','bog') SUB('cranberry',' cranberry1') SUB('cranberry',' cranberry2').., where relation SUB('cranberry',' cranberry1') SUB('cranberry',' cranberry2')..., renders particular objects (for instance, those could be cranberries bought in a market or picked somewhere).
Knowledge hierarchical structure representation	Hierarchical structures are the knowledge basis on which catalogues, explanatory dictionaries, etc. are constructed. An example of representing such structures [17]: SUB('human','man') SUB('human','woman') SUB('man','Janis1') SUB('man','Karlis1') SUB('woman','Mara1') SUB('woman','Una1') For class relations node SUB is used. Here it is depicted that a human – it is men and women. In the lower level there are nodes representing particular humans. Each such node can have its own connection that represents characteristics and/or relations. Characteristics (relations) can also be for nodes-classes, and they are referable to all the class elements. SUB branches must form a tree (it cannot be a cycle). It is necessary from the viewpoint of characteristic's succession – every high level node characteristics are inherent to all the lower level nodes. This principle allows significant decrease of the knowledge amount, using object class characteristics (relations) and automatically distributing them to particular objects.
Fuzzy knowledge representation	Expression "A1 is not very little" looks like this [12]: LENGTH(a1, x11) 'NOT'(x21,x11) EVALUATION(x21, 'little') 'VERY'(x21)
Declarative knowledge representation	An example of declarative knowledge representation [14]: network SUB(human, man) SUB(man,a1) SUB(man,a2) NAME(Jānis,a1) NAME(Jānis,a2) FATHER(a1,a2) represents that there are two men, who are human, who are named Jānis, one of them is the father of the other man.
Procedural knowledge representation	An example of procedural knowledge representation [13]: It is necessary to calculate the value of variable B3 by the following formula: B3:=(B1+25)*B2 The network looks like this: VALUE('B1', $x1^1$) VALUE ('B2', $x2^1$) +($x1^1$,'25',$x3^1$) *($x2^1$,$x3^1$,$x4^1$) ?($x4^1$) VALUE ('B3', $x4^1$), where the relation "VALUE" means that the node $x1^1$ corresponds to the value of variable B1; upper index 1 means that x^1 corresponds to some one single variable.
Logical operation representation	For example, the expression ¬(P1∧P2) ∨ (P2∧1) = 0 is depicted in the following way [14]: ∧(x1,x2,x3) ¬(x3, x4) ∧(x2,t,x5) ∨(x4,x5,f), where the variables P1 and P2 correspond to nodes x1 and x2, constants 1 and 0 correspond to nodes t and f.
Quantifier representation options	Predicate expression (∀X1∈M1)R1(X1) is depicted in the following way [14]: ∈(x1, m1) ∀(x1,x2) r1(x2), where node x1 corresponds to a cluster, node x2 corresponds to cluster representatives which are all inherent to common (additional) characteristics or relations r1. In the case of quantifier ∀ node x2 corresponds to the whole cluster, quantifier ∃ - sub-cluster, quantifier $\exists_1$ – one cluster element.
Preservation of initial information and acquisition of new information	Outer world is dynamic, although something is continuously changing. As it was mentioned before, special type of extended semantic networks are used – productions which have an important role in representing different types of time and cause-consequence dependence, definitions, explanations, etc. For example, we shall review a production which depicts the following changes – at first the object X1 possessed the characteristic R2, afterwards – R3. We are talking about an object that has relation R1 with A1 [13]: R1($x1^1$, a1) R2($x1^1$/γ1) IN BEGINNIN-THEN(γ1, γ2) R1($x1^1$, a1) R3($x1^1$/γ2), where γ1 and γ2 mean nodes, in the cluster of the corresponding situation, γ1 – the initial situation, γ2 – the end situation.
Intensional and extensional representation options	A possibility to represent extensional and intensional means a possibility to represent in a model a cluster of objects that is described with a particular word (word extensional) and cluster or object characteristics that is described with a particular word (word intensional) [7]. We shall review the following example [18]. Study object is any material that can be used for teaching. When designing a study object the following main characteristics must be provided: • they are little subject quants that last for 2 to 15 minutes; • they are closed, that is, they can be used separately; • repeatedly usable; • can be aggregated in one group; • marked with metadata. Study objects can have different types: theoretical information; explanation; example; question; task; commentary etc. Such definition can be depicted in the following way: SUB('study material', 'study object') MIN_DURATION('study object', '2 min') MAX_DURATION('study object', '15 min') CHARACTERISTIC1('study object', 'closed') CHARACTERISTIC2('study object', 'repeatedly usable') CHARACTERISTIC3('repeatedly usable', 'can be aggregated with others') CHARACTERISTIC4('study object', 'marked with metadata') SUB('study object', 'theoretical information') SUB('study object', 'explanation') SUB('study object', 'example') SUB('study object', 'question') SUB('study object', 'commentary') SUB('study object', 'task'). Here six subtypes of a study object are established (theoretical information, explanation etc.) that are node words and make extensional of the word 'study subject'. The words 'completed', 'repeatedly applicable' etc. make intensional of the word 'study object'.

TABLE I EXTENDED SEMANTIC NETWORK MODEL POSSIBILITIES/OPTIONS (CONTINUATION)

Model Properties	Explanation
Contradiction recognition	When a new fragment of knowledge is received, it must be established whether it contradicts a fragment known before. For example, the system knows that *Janis* (who is our focus) *is the only child in the family*. Then the expression: *his brother's friend is definitely married* should form actions that are related both to outer actions (repeated question; statement that it cannot be etc.) and inner actions (change of notions) [13].
Knowledge structure representation integrity	Every complicated object or situation can be regarded as a unified whole. Then it is allocated its own node (link node).
Model uniformity	Network is constructed from uniform fragments, every fragment can be removed or added to the network without damaging the correctness of syntax or semantic.
Option to merge knowledge structures	Possibility to merge knowledge structures means the possibility of the knowledge representation model to merge different networks structure which can contain, for example, nodes that represent words with the same meaning. Expediency of network structure merging is connected with the necessity to merge knowledge structures that describe domains with an intention to acquire a possibility to make decision within the framework of integrated domain. Before merging of knowledge structures it is necessary to check implementation of several provisions [13]: • there should be no alternatives; if we talk about separate objects, they are to be found, if several objects exist, they must be specified; • all the search possibilities must be used, including those that are connected with equivalent transformations, logical conclusion elements; search can be ineffective, for example, because there are relations in the expression that are depicted differently within knowledge; then reassessment of notions is necessary; • it is necessary to find out whether merging causes a contradiction to the information available before.

In the end we shall consider an informative example of application of extended semantic networks.

The most important attributes of the study course are its name, laboriousness, sections (course content), the competences formed in the process of studying the course, keywords. The study course is studied at a certain stage, is part of a study program, is completed with certain report, requirements of employers were taken into account, etc. Fragment of study course description with means of an extended semantic network is shown in Fig. 1.

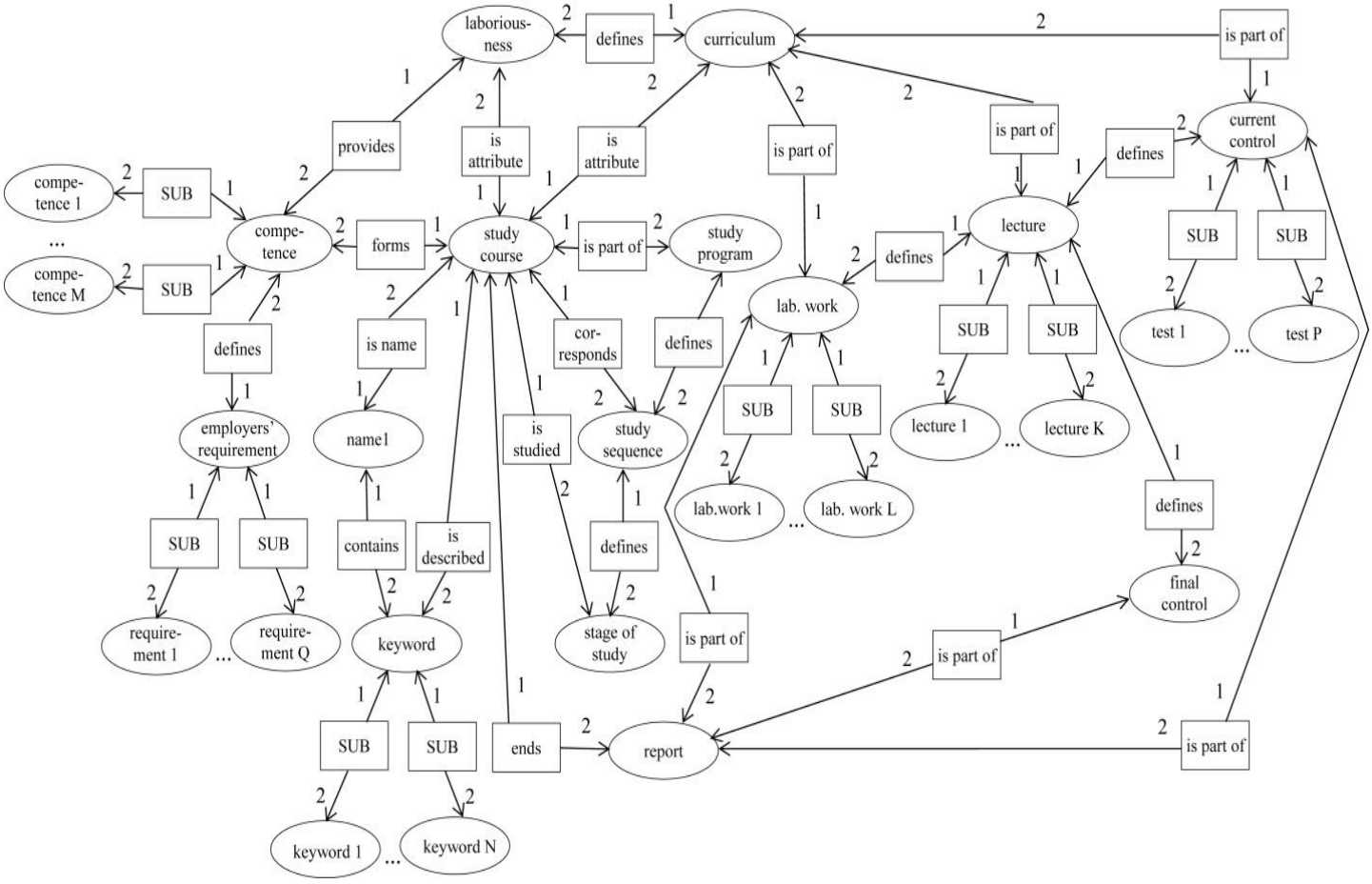


Fig. 1. Fragment of study course description.

This network is described as follows:
IS PART OF('study course', 'study program')
CORRESPONDS('study course', 'study sequence')
IS NAME('name' 'study course')
IS DESCRIBED('study course', 'keyword')
FORMS('study course', 'competence')
……. etc.

VI. CONCLUSION

Topicality of knowledge representation problem is evident. Choice of knowledge representation model is one of the main problems in intelligent system's development, and its essence is to choose such a model that would satisfy the set requirements.

Software developers frequently try to describe complicated domains, where informative sophisticated tasks are solved, by using monotone regular structures that are too primitive for representation of the whole variety of domain's meaning nuances, even though they are convenient for further processing [18].

The article examines body of requirements for a knowledge representation model in intelligent systems, it is offered to use extended semantic networks for knowledge representation and with examples it is shown that this model complies with the aforementioned body of requirement. The offered material could be useful for developers of intelligent systems and other researchers to continue the work in knowledge representation and processing problem solution.

The authors believe that use of extended semantic networks in their work is expedient. Future work will focus on the evolution of extended semantic networks model for its use in development of collaborative intelligent educational systems.

REFERENCES

[1] A. L. Yalovets (2011). "Representation and Processing of Knowledge from the Point of View of Mathematical Modeling". Naukova Dumka, Kiev [In Russian].

[2] J. Bentahar, B. Moulin, M. Bélanger (2010). "A taxonomy of argumentation models used for knowledge representation" // Artificial Intelligence Review, vol. 33 (3), pp. 211-259. doi: 10.1007/s10462-010-9154-1

[3] M.M Abdul Jalil, C.P. Ling, N.M. Mohamad Noor, F. Mohd (2017). "Knowledge Representation Model for Crime Analysis" // Procedia Computer Science, vol. 116, pp. 484-491

[4] I.N. Gluhih, R.K. Akhmadulin (2017). "Problem-Oriented Corporate Knowledge Base Models on the Case-Based Reasoning Approach Basis" // IOP Conference Series: Materials Science and Engineering, vol. 221, issue 1, article number 012025

[5] P. Clark (2007). "Requirements for a Knowledge Representation System Working", Note 10, pp. 1-10.

[6] R. Davis, H. Shrobe, P. Szolovits (1993). "What Is a Knowledge Representation?" In: AI Magazin, Vol. 14(1), pp. 17-33.

[7] H. Ueno, M. Ishizuka (1989). "Knowledge Representation and Usage'. Moscow, Mir [In Russian].

[8] D. H. Jonassen (1995). "Computers as Cognitive Tools: Learning with Technology, not from Technology, in: Journal of Computing in Higher Education", Spring, Vol. 6(2), pp. 40-73.

[9] A. Bashmakov (2005). "Intelligent Information Technology". MSIT, Moskow [In Russian].

[10] V. Golenkov, N. Guliakina (2014). "Project of Open Semantic Technology of the Componential Design of Intelligent System"s. Part 1: The Principles of Creation, in "Ontology of Designing" Scientific Journal, 1(11), pp. 42-64 [In Russian].

[11] E. V. Zolotov, I. P. Kuznetso (1982). "Expanding Systems of Active Dialogue". Nauka, Moscow [In Russian].

[12] I. P. Kuznetsov (1986). "Semantic Representation". Nauka, Mockow [In Russian].

[13] I. P. Kuznetsov (1989). "The System of Knowledge Processing on Extended Semantic Network"s (in the form of a scientific report), Moscow [In Russian].

[14] E. B. Kozerenko, I.P. Kuznetsov (2010). "Evolution of Linguistic Semantic Presentations in the Intelligent Systems Based on the Extended Semantic Networks", in: Computer Linguistics and Intelligent Technologies, Vol. 9, pp. 205-211

[15] S. P. Habarov (2017). "Knowledge Representation in Information Systems", http://www.habarov.spb.ru/bz/index.htm (accessed: October 04, 2017)

[16] I. P. Kuznecov, A.G.Matskevich (2006). "Semantic-Oriented Systems Based on Knowledge Bases", MTUSI, Moscow [In Russian].

[17] J. Bule (2011). "A Set of Models for Managing Adaptive Computer-Based Learning", Doctoral Thesis, Riga, RTU [In Latvian].

[18] A. Novikov (2016). "Problem Definition of Representation of Knowledges in Computer Systems" // Applied Informatics, vol. 11, Nr. 2(62), pp. 107-143, ISSN: 1993-8314 [In Russian]

Behavior of the Minimum Euclidean Distance Optimization Precoders with Soft Maximum Likelihood Detector for High Data Rate MIMO Transmission

MAHI Sarra, BOUACHA Abdelhafid
Faculty of technology, University of Tlemcen,
Laboratory of Telecommunication of Tlemcen (LTT),
Tlemcen, Algeria

***Abstract*—The linear closed loop Multiple-input Multiple-output (CL-MIMO) precoding techniques characterized by the channel state information knowledge (CSI), at both sides of the link, aims to improve information throughput and reduce the bit error rate in the communication system. The processing involves multiplying a signal by a precoding matrix, computing from the CSI with some optimized criteria. In this paper, we proposed a new concatenation of the precoders optimizing the minimal Euclidean distance with soft Maximum Likelihood (soft-ML) detection. We analyze the performance in terms of bit error rate (BER) for the proposed association with the three well-known quantized precoders: Maximum of minimum Euclidean distance (Max-dmin) precoder, Orthogonalized Spatial Multiplexing precoder (POSM), and Orthogonalized Spatial Multiplexing (OSM) based on the same criteria, in coded MIMO system over a Rayleigh fading channel, using Quadrature Amplitude Modulation (QAM). Simulations show the interest of the proposed association of the dmin-based precoder with a soft - Ml detector, and the best result is achieved for Max-dmin precoder.**

Keywords—MIMO; max-dmin; POSM; singular values decomposition (SVD); soft-ML detector

I. INTRODUCTION

Modern wireless communications [1], especially fifth-generation (5G) cellular networks [2], [3], require high data throughput with low transmission latency. For example, high-speed coding, high-order modulation, and Multiple-Input Multiple-Output (MIMO) technology are essential tools for achieving high data rates. MIMO technology not only offer the diversity and capacity gains, but also achieve higher link reliability comparable with single antenna systems (SISO). The advantages of using Multiple antennas at the transmitter and receiver of the wireless MIMO system have been well exploited in the recent years [4]. The benefits of MIMO communication are generally ensured by both open-loop and closed-loop MIMO techniques. Open loop techniques, such as spatial coding (STC) and spatial multiplexing (SM) [5], [6], are used without the need for Channel State Information (CSI) at the transmitter. In order to overcome the multipath effect and to improve the robustness of spatial multiplexing systems, closed loop linear pre-coding techniques [7] may be used at the transmitter. The principle of precoding techniques is that, when the channel knowledge is available to the transmitter, the transmit signal is pre-multiplied by a precoding matrix so that the inter-symbol interference (ISI) in the receiver is greatly reduced. This knowledge of the characteristics of the channel makes it possible to anticipate any damage caused by the propagation, in order to obtain a "global" transmission channel favorable to communication. This technique is used in particular in WLAN networks (IEEE 802.11n standard) and mobile networks (LTE standard of the 3GPP project) [8]. Several types of precoders have been proposed in the literature, they have been designed according to various criteria. We can sites output capacity maximization [9], the BER minimization [10], signal-to noise ratio (SNR) maximization [11], mean square error (MSE) minimization [12] and minimum singular value maximization [13], this provides diagonal precoders and focus on power allocation schemes. Recently others precoders completely optimize the precoding matrix for a very specific purpose such as maximizing the minimal Euclidean distance between the constellations received referred to as Max-dmin precoder [14], whose principle is basing on the maximization of the minimum Euclidean distance between the received symbols. It has demonstrated its ability to improve both the spectral efficiency and the robustness of the transmission, and outperforms other kinds of MIMO precoders in terms of BER performance [15], particularly in correlated propagation scenarios, Like the Max-SNR, channel information is required for transmission.

In the literature, MIMO pre-encoding (or pre-equalization, or channel formation) consists of pre-mixing the signals prior to the channel, choosing the precoder according to the available CSIT, so to obtain a "global" transmission channel favorable to communication. Research conducted at Bretagne Telecom has made it possible to determine the pre-mix for globally optimizing a MIMO / OFDM system using criteria based on the minimum distance (Dmin) [15], [16].

The results make it possible to obtain a robust transmission with respect to the fading of the channel, which is the main cause of error in the estimation of the symbols. In addition, the proposed MIMO precoding in [15]-[17] has the advantage of preserving a maximum transmission rate comparable to the spatial multiplexing technique where the data is transmitting

independently over several transmitting antennas and in the same frequency band. This excellent flow / performance compromise is a crucial asset for future MIMO systems.

MIMO systems can make the most of the useful information available in the CSI but with CSIT uncertainty robustness. The difficulty lies in the fact that the conventional singular value decomposition (SVD) and water-filling (WF) techniques are sensitive to the CSIT error, while other alternatives available, for example spatio-temporal coding, cannot fully exploit the advantages offered by the CSIT available. The problem remains to design transmission schemes for MIMO channels that can fully exploit the available CSIT benefits while being robust against CSIT uncertainty. In addition, convolutive error correcting codes and soft ML detection is not taken into account for this precoders.

In this context, we investigate the precoder design for convolutive encoded MIMO systems by assuming Soft detection at the receiver. In this paper, we propose firstly a new concatenation of the precoder optimizing the minimal Euclidean distance (Max dmin precoder) with convolutive error correcting code and soft Maximum Likelihood (soft-ML) detection. Thus, as a second contribution, and to see the effective method for our proposed system, we compare the performance of Max-dmin, POSM and OSM procedures using Soft-ML detection. These three techniques optimize the same criterion (minimum Euclidean distance), but each precoder uses a different method, the SDV, the antenna selection and Coding-Orthogonalization respectively.

. We simulate the three techniques for a Rayleigh channel model and perfect CSI-T (channel state information at the transmitter). The performance is based on the evaluation of the BER (Binary Error Rate) for different number of antennas at both sides of the link and several modulation profiles. We use the classical Spatial Multiplexing technique as reference for our results.

Section 2 of this paper gives the MIMO close-loop system model (precoders, channel and detection techniques). Section 3 presents the simulation results, and finally Section 4 concludes the paper.

II. System Model

Let us consider a MIMO system with n_T transmit and n_R receive antennas, i.e. a (n_T , n_R) MIMO system, and b independent data-streams over a Rayleigh fading channel. The basic system model is defined by:

$$y = GHFs + Gn \quad (1)$$

Where y is the $b \times 1$ received symbol vector, G is the $b \times n_R$ linear decoder matrix, H is the $n_R \times n_T$ channel matrix, F is the $n_T \times b$ linear precoder matrix, s the $b \times 1$ transmitted symbol vector, n is the $n_r \times 1$ additive Gaussian noise vector, We assume that $b \leq r = rank(H) \leq min(n_T , n_R)$ and I_b (denotes the $b \times b$ identity matrix).

$E[ss^*] = I_b, E[sn^*] = 0$ *and* $E[nn^*] = R$, With R the noise covariance matrix, and superscript * stands for conjugate transpose. Under the perfect CSI condition at both the transmitter and the receiver, the channel can be diagonalized by using the virtual transformation (Fig. 1) and is decomposing in three steps: noise whitening, channel diagonalization and dimensionality reduction.

. Firstly, the precoding and decoding matrix can be written as $F = F_v F_d$ and $G = G_v G_d$. Then, the new decomposition of F_v and G_v matrices into the product of three matrices are considering:

$$F_v = F_1F_2F_3 \text{ and } G_v = G_1G_2G_3 \quad (2)$$

Where the (F_i, G_i) perform the particular operations: noise whitening, channel diagonalization and dimensionality reduction.

Therefore, the input-output relation (1) will be:

$$y = G_d H_v F_d s + G_d n_v \quad (3)$$

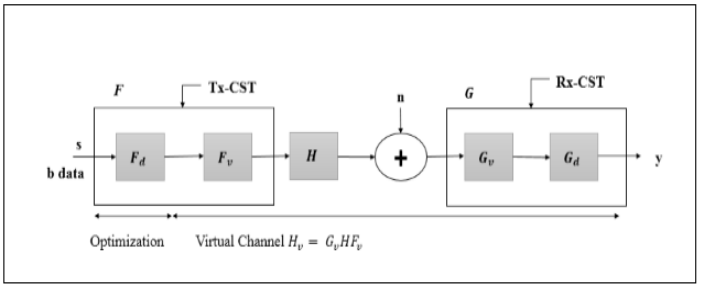


Fig. 1. MIMO channel precoded in virtual channel.

The decoding matrix G_d has no effect on the performance when the ML detection is considered. Therefore, we adopt in this paper that G_d is an $b \times b$ identity matrix.

H_v is the $b \times b$ virtual channel matrix, written as:

$$H_v = diag(\sigma_1, \dots, \sigma_b) \quad (4)$$

Were σ_i are entries with: $\sigma_1 > \sigma_2 > \cdots > \sigma_i$ and $i = 1, \dots, b$. $n_v = G_v n$ is the $b \times 1$ transformed additive Gaussian noise vector. Now we can write the virtual system model as:

$$y = H_v F_d s + n_v \quad (5)$$

A. The Max-dmin Precoder

The Max-dmin is non-diagonal precoder, which maximize the minimal distance between received constellations [14] affects the system performances, especially with the ML detecto. The value of the minimal distance between received constellations is denoted d_{min} and given by:

$$d_{min}(F_d) = min_{(s_k, s_l) \epsilon C^b, s_k \neq s_l} \|H_v F_d (s_k - s_l)\| \quad (6)$$

Were C represents the set of complex symbols of the constellation.

x_k and x_l are two transmit signals, and S is the set of all these possible transmit vectors. Let us define $\check{x}$ a difference vector as $\check{x} = s_k - s_l$, with $s_k \neq s_l$.The Max-dmin solution consists to find the F_d matrix coefficients, which maximize the minimum distance of the received constellation:

$$F_d = \arg\max_{F_d} \{d_{min}(F_d)\} \quad (7)$$

with

$$trace(F_d F_d^*) = P_0 \quad (8)$$

This problem may be particularly difficult to solve, because the distance expression takes account of several parameters: the transmission channel H_v , the digital modulation and the number of considers channels b. In this paper a solution is found for $b = 2$ and BPSK, QAM-4 and QAM-16 modulations. The optimization for the virtual sub-channels proposed in [18] is obtained by a variable change of two channels eigenvalues denoted σ_1 and σ_2. It is a simple change of Cartesian coordinates into polar coordinates. The new variables are:

$$\begin{cases}\sigma_1 = \rho\cos\gamma \\ \sigma_2 = \rho\sin\gamma\end{cases} \quad \text{and} \quad \begin{cases}\gamma = \arctan\frac{\sigma_2}{\sigma_1} \\ \rho = \sqrt{\sigma_1^2\sigma_2^2}\end{cases} \qquad (9)$$

Where ρ and γ represent the channel gain and channel angle, respectively. The virtual channel is then given by:

$$H_v = \begin{pmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{pmatrix} = \rho\begin{pmatrix}cos\,\gamma & 0 \\ 0 & sin\,\gamma\end{pmatrix} \qquad (10)$$

Note that $\sigma_1 \geq \sigma_2 > 0$, so we have $0 < \gamma \leq \pi/4$.

We give here the precoding matrix F_d for a 4-QAM modulation; this solution is relatively simple with two forms of precoder:

- If $0 < \gamma \leq \gamma_0$

$$F_d = F_{r1} = \sqrt{E_s}\begin{pmatrix}\sqrt{\frac{3+\sqrt{3}}{6}} & \sqrt{\frac{3-\sqrt{3}}{6}}e^{i\frac{\pi}{12}} \\ 0 & 0\end{pmatrix} \qquad (11)$$

- If $\gamma_0 \leq \gamma \leq \frac{\pi}{4}$

$$F_d = F_{octa} = \sqrt{\frac{E_s}{2}}\begin{pmatrix}cos\Psi & 0 \\ 0 & sin\Psi\end{pmatrix}\begin{pmatrix}1 & e^{i\frac{\pi}{4}} \\ -1 & e^{i\frac{\pi}{4}}\end{pmatrix} \qquad (11)$$

Where:

$$\begin{cases}\Psi = arctan\frac{\sqrt{2}-1}{tan\gamma} \\ \gamma_0 = arctan\sqrt{\frac{3\sqrt{3}-2\sqrt{6}+2\sqrt{2}-3}{3\sqrt{3}-2\sqrt{6}+1}} = 17{,}28^{\circ}\end{cases} \qquad (12)$$

The parameter Ψ in relations is the power allocation on each sub-channel, and the constant threshold γ_0 allows the precoder to use one or two sub-channels. The value of γ_0 is obtained when considering that the two precoders give the same minimum Euclidean distance d_{min}. This one depends on ρ and γ and is expressed in [18].

$$d_{min} = \begin{cases}\sqrt{E_s}\,\rho\sqrt{1-\frac{1}{\sqrt{3}}}cos\gamma & if\ 0<\gamma\leq\gamma_0 \\ \sqrt{E_s}\,\rho\sqrt{\frac{(4-2\sqrt{2})cos^2\gamma sin^2\gamma}{1+(2-2\sqrt{2})cos^2\gamma}} & if\ \gamma_0<\gamma\leq\frac{\pi}{4}\end{cases} \qquad (13)$$

B. The POSM Precoder

The technique known as Orthogonalized Spatial Multiplexing (OSM) proposed for closed-loop MIMO systems, associates a symbols coding with an orthogonalization by rotation (Coding-Orthogonalization). In addition, Y. Kim and al propose in [19] a precoding technique for the OSM system, this is referred as the P-OSM precoder. Like the Max-dmin, the P-OSM optimize minimal Euclidean distance of the received constellation. The OSM system transmits $b = 2$ independent informations channels on $n_t = 2$ transmission antennas and if $n_t > 2$, an antenna selection method must be associated. The principle of the OSM consists to precode the transmitted symbols x_1 and x_2 as:

$$F(x,\theta) = \begin{bmatrix}1 & 0 \\ 0 & e^{j\theta}\end{bmatrix} s(x) \quad (14)$$

Where θ is the phase rotation angle applied to the second antenna and s(x):

$$s(x) = \begin{bmatrix}Re[x_1] + jRe[x_2] \\ Im[x_1] + jIm[x_2]\end{bmatrix} \qquad (15)$$

In a real representation, the P-OSM system model can be written as:

$$y_r = H_r^{\theta}s_r(x) + n_r = \left[h_1^{\theta}\,h_2^{\theta}\,h_3^{\theta}\,h_4^{\theta}\right]s_r(x) + n_r \qquad (16)$$

The real vector h_i of length $2n_r$ represents the ith column of the real matrix H_r^{θ}. The columns h_1^{θ} and h_2^{θ} are respectively orthogonal to h_3^{θ} and h_4^{θ}. The angle of rotation θ necessary to guarantee the orthogonality between h_1 and h_4 or h_2 and h_3 is calculated from the original channel matrix as follows:

$$\theta = tan^{-1}\left(\frac{B}{A}\right) \pm \frac{\pi}{2} \qquad (17)$$

Were $A = \sum_{m=1}^{n_r}\left|\dot{h_{m1}}\right|\left|\dot{h_{m2}}\right|\sin(\angle\dot{h_{m2}} - \angle\dot{h_{m1}})$
and $B = \sum_{m=1}^{n_r}\left|\dot{h_{m1}}\right|\left|\dot{h_{m2}}\right|\cos(\angle\dot{h_{m2}} - \angle\dot{h_{m1}})$.

$\angle$ is the argument.

At this stage, the OSM system do not optimize any criterion, but enable the orthogonalization of the received symbols. Thus, the P-OSM precoder, which maximizes d_{min}, is also given in an actual representation by:

$$y_r = H_r^{\theta}\bar{P}s_r(x) + n_r = H_r^{\theta}\begin{bmatrix}P & 0 \\ 0 & P\end{bmatrix}s_r(x) + n_r \quad (18)$$

Where $\bar{P}$ is an actual precoding matrix of size 2×2 and the power constraint is:

$$trace(Ps_r(x)s_r(x)^*P^*) = P_0 \quad (19)$$

For simplicity, the precoder is not optimal. So, the matrix $\bar{P}$ is decomposed as [19]:

$$\bar{P} = R_{\theta 1}\begin{bmatrix}p & 0 \\ 0 & \sqrt{2-p^2}\end{bmatrix}R_{\theta 2} \qquad (20)$$

With

$$R_{\theta i} = \begin{bmatrix}cos\theta_i & -sin\theta_i \\ sin\theta_i & cos\theta_i\end{bmatrix} \qquad (21)$$

The angle θ_1 is directly computed from the matrix H_r^{θ}.

$$\theta_1 = tan^{-1}\left(\frac{C+\sqrt{C^2+4D^2}}{2D}\right) \qquad (22)$$

With

$$\begin{cases}C = \left\|h_2^{\theta}\right\|^2 - \left\|h_1^{\theta}\right\|^2 \\ D = (h_1^{\theta}).h_2^{\theta}\end{cases} \qquad (23)$$

θ_2 and p are chosen according to the modulation under consideration and a parameter k defined as:

$$k = \frac{\left\|h_1^{\theta_1}\right\|}{\left\|h_2^{\theta_1}\right\|} \qquad (24)$$

$$H_r^{\theta 1} = H_r^{\theta}R_{\theta 1} = \left[h_1^{\theta}\,h_2^{\theta}\,h_3^{\theta}\,h_4^{\theta}\right] \quad (25)$$

$\|h_1^{\theta 1}\|$ and $\|h_1^{\theta 1}\|$ represent respectively the first and the second singular value (σ_1 and σ_2) of the channel matrix H. The solution for a 4-QAM is [17]:

- If $1 \le k < 7$ $p = \sqrt{\frac{6}{(k+3)}} \approx 1$ and $\theta_2 = 45^\circ$
- If $k \ge 7$ $p = \sqrt{2}$ and $\theta_2 = 26.5^\circ$

Fig. 2 shows the entire system block diagram to simulate.

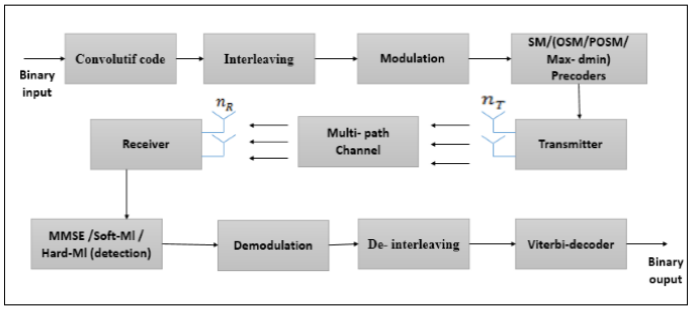


Fig. 2. Block diagram of precoded MIMO system.

C. Detection Techniques

At receiver, the MIMO detection consists to estimate the symbols generated at the transmitter before the coding channel processes. Several methods are available in the literature for the MIMO systems [20]. The most common approach is the maximum likelihood (ML) detection [21], which achieve optimal performances at the expense of computational complexity. Besides the ML criterion, the zero forcing (ZF) and minimum mean square error (MMSE) [22] detections are linear equalization-based methods. They are low-complexity and simple to implement, but with lower performances. In this work, we compare the ML hard and soft detection performances for different precoding techniques.

In the MIMO case, ML-hard detection consists of finding the most likely transmit symbols vector:

$$\check{s} = arg\ \max_{s \epsilon S} f\left(\frac{y}{s}\right) \quad (26)$$

Where s is the set of all possible M^{n_T} transmit symbol vector candidates with the modulation order M.

$$f(y/s) = \frac{1}{\pi^{n_R}} exp(-\|y - Hs\|^2) \quad (27)$$

So, the ML hard detection is equivalent to finding the transmit symbol vector that minimizes the Euclidean Distance (ED) $\|y - Hs\|^2$ between y and Hs :

$$\check{s} = \arg\min_{s \epsilon S} \|y - Hs\|^2 \quad (28)$$

In this detection technique, there are low-complexity algorithms that try to find optimum s without calculating all the EDs for all transmit symbols vector. In this case, hard detection ignores a large part of the information contained in the receive vector y.

In ML-soft detection case [23], the information about the decision and its reliability are usually delivered jointly for every bit $b_{x,n}$ using log likelihood ratios (LLRs) [24]. The n-th bit of x-th stream:

$$L(b_{x,n}) = log\left(\frac{P\{y|b_{x,n}=1\}}{P\{y|b_{x,n}=0\}}\right) \quad (29)$$

Using Max-Log-MAP approximation, we calculate the approximate LLR:

$$\tilde{L}(b_{x,n}) \triangleq min_{s \in S_{x,n}{}^{(0)}} \|y - Hs\|^2 - \min_{s \in S_{x,n}{}^{(1)}} \|y - Hs\|^2 \quad (30)$$

$S_{x,n}{}^{(b)}$: set of transmit symbol vector candidates with $b_{x,n} = b$. .

To conclude this part, the ML-soft demodulation looks similar to the ML-hard detection problem but present more difficult in reality, such as the search space for the transmit symbol vector with the minimum EDs is limited to $S_{x,n}{}^{(b)}$, and the partitioning transmit symbol vector candidates into $S_{x,n}{}^{(1)}$ is all different depending on x and n.

III. Results and Analysis

In this section, we present our simulation results of the Bit Error Rate (BER) evaluation for different precoding techniques based on the system model shown in Fig. 2.

For these simulations, we consider the (2×2) and (4×4) MIMO precoding systems. The channel is disrupted by Rayleigh distribution. The transmission structure is a Bit Interleaved Coded Modulation (BICM) type, resulting from the concatenation of a channel encoder, a bit interleave and a bit-to-symbol conversion. The channel coding (Convontinal Code) is performed by an encoding rate R = 1/2, and a constraint length K = 7. The polynomials generator of the convolutional coder are [133, 177].

The encoded data frame is subsequently interleaved randomly and converted into complex symbols belonging to the constellation alphabet of the modulation, we used the 4QAM and 16QAM modulations. One of the quantified precoding techniques (Max-dmin, OSM and POSM) follows this BICM emission structure. Soft or hard decoding is performed for channel decoding, using the Max-log-MAP algorithm. For this simulation, 10 000 frames of 800 bits each were transmitted. The channel is quasi-statistical, so that the matrix H is assumed constant during the transmission of 800 bits. The number of transmitted streams is limited to $b = 2$. For simplicity, we consider perfect estimation of the CSI. We use the "Spatial Multiplexing (SM)" MIMO system as reference to compare all the precoders techniques.

In order to evaluate behavior of our proposed association (Max-dmin with soft ML detection), we simulated four different scenarios. The first one is a (2x2) MIMO transmission without any precoder; this is "Spatial Multiplexing MIMO (SM-MIMO)" system. The second one, consists in the same MIMO scheme with OSM precoder

(OSM-MIMO" system). The third one consists in a MIMO scheme with the POSM precoder (POSM-MIMO system). The last scenario is the MIMO scheme with Max-dmin precoder (Max-dmin –MIMO system).

Fig. 3, give performances comparison between two conventional detection algorithms, the MMSE and the maximum likelihood structures, for the different systems cited above. In this case, we use 2x2 MIMO system with 4QAM modulation.

As shown in Fig. 3, the maximum likelihood (ML) receiver is more efficient than MMSE for SNRs greater than 6dB. The ML receiver archive BER around 10^{-3}, were MMSE does not exceed 10^{-2} for SM-MIMO system. We can also see from this curves that the Max-dmin -MIMO and POSM-MIMO have closely performances and exceed the SM-MIMO and OSM-MIMO systems performances respectively. For example, POSM-MIMO and Max-dmin-MIMO provide 6dB SNR gain for 10^{-4} BER compared to SM-MIMO and OSM-MIMO.

Fig. 4 show the performances of the four simulated systems in terms of BER for 2x2 MIMO system with 4QAM modulation. The POSM–MIMO transmission with ML-soft decision has the best performance compared to Max-dmin–MIMO, SM-MIMO and OSM- MIMO transmissions with ML-Soft and ML-hard detectors respectively. The SNR gains of the POSM-MIMO-ML-soft are from 1dB to 2.5dB, compared to the other precoders.

In Fig. 5, we illustrate the simulation of BER for the P-OSM and max-dmin for 2x2 MIMO system with 4QAM and 16 QAM modulation. The result allows to verify that the modulation order has a big impact on the performance of the MIMO precoders systems. Since the ML detector estimates the Euclidean distance between the constellation points, the expansion of this constellation increases the number of errors on the transmit symbols. We achieve the best performance for 4-QAM modulation.

Fig. 6 shows the simulation of BER for n_t=n_r=4 using 4QAM modulation. The curve shows that the max-dmin-MIMO with ML Soft system remains the best BER precoder with a SNR gain of 1.5 dB and 2 dB at BER of 10^{-4} compared to the POSM-ML Soft and OSM-ML Soft systems respectively. For the same SNR the Max-dmin--MIMO with ML Soft system has a gain of 5dB on the SM-ML soft-MIMO system. These results show clearly that the SVD method is more efficient than antenna selection specifically when the receive antennas is equal or greater than four elements.

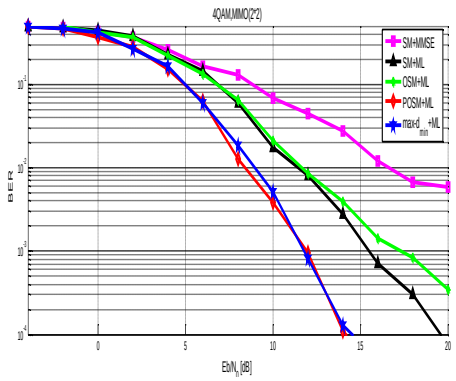


Fig. 3. BER for (2x2) MIMO systems with 4QAM modulation.

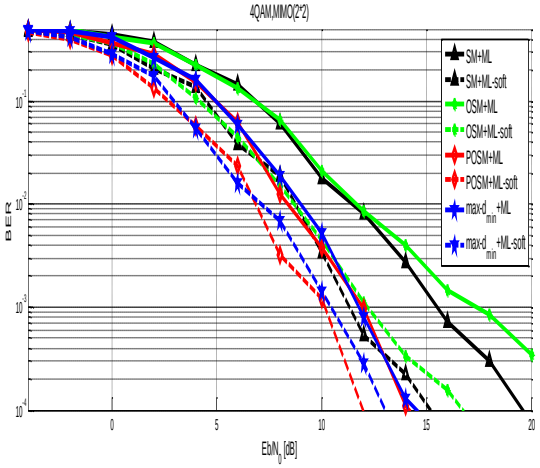


Fig. 4. BER for (2x2) MIMO systems with 4QAM modulation.

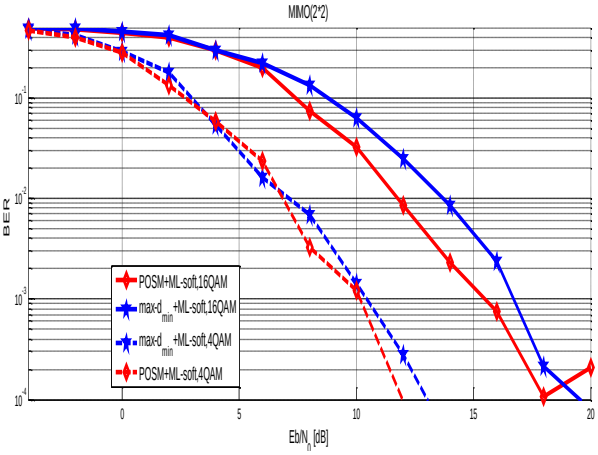


Fig. 5. BER for (2x2) MIMO systems with 4QAM and 16 QAM modulations.

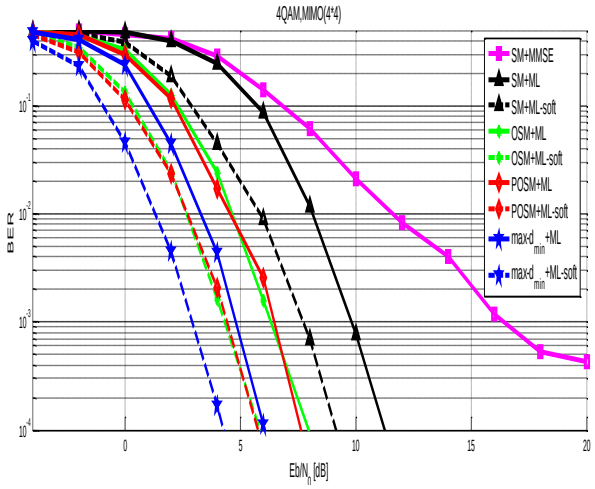


Fig. 6. BER for (4x4) MIMO systems with 4QAM modulation.

IV. Conclusion and Future Work

In this paper, we study the behavior of the concatenation of the Soft-ML detector with the precoder based on the minimum Euclidian Distance criterion. Given perfect CSI at the transmitter, we provided simulations results that demonstrate the efficiency of the proposed concatenation. The simulations show that the Soft ML detector concatenated with precoder based on the minimum Euclidian distance criterion using convolutive error correcting code provide better performance than the hard-ML and MMSE detectors in term of BER. In addition, we conclude from the results that Max-dmin and P-OSM precoders in MIMO systems have the best BER performance with Soft-ML detector compared to the other systems. When two transmitted antennas are used, the two precoders are equivalent in terms of BER performances. When n_t >2, the POSM is limited and should be associated with the antenna selection algorithm, in this case the Max-dmin based on the SVD method outperforms POSM with a significant SNR gain. Finally, our proposed system presents a real interest in the evolution of transmission techniques in general and those of LTE and LTE-A more precisely. The optimization of MIMO transmitters and receivers proposed in this work will have an impact on the energy consumption of these systems, a major challenge of our century.

As a future work, we want to implement non-binary LDPC codes (NB-LDPC) in precoded MIMO systems, The use of non-binary LDPC codes (NB-LDPC) is a promising solution because it allows to obtain excellent performance in error correction in the case of short frames and / or large size constellations. These properties are particularly well suited to use NB-LDPC codes with precoded MIMO systems.

References

[1] Rizwan Akhtar, Supeng Leng, Imran Memon,Mushtaq Ali, Liren Zhan, Architecture of Hybrid Mobile Social Networks for Efficient Content Delivery, Springer Science+Business Media New York 2014.

[2] XIONGWEN ZHAO, ADAM MOHAMED AHMED ABDO, CHEN XU, SUIYAN GENG, JIANHUA ZHANG, AND IMRAN MEMON, Dimension Reduction of Channel Correlation Matrix Using CUR-Decomposition Technique for 3-D Massive Antenna System, IEEE ACCES, February 14, 2018.

[3] Qasim Ali Arain, Imran Memon, Zhongliang Deng, Muhammad Hammad Memon, Farman Ali Mangi ,Asma Zubedi, Location monitoring approach: multiple mix-zones with location privacy protection based on traffic flow over road networks, Springer Science+Business Media New York 2017.

[4] Q.Li, G. Li, W. Lee, M.Lee, D. Mazzareszs, B. Clerckx, and Z.Li, MIMO techniques in WiMAX and LTE: a feature overview, IEEE Commun. Mag, vol.48, no.5, pp.88-92,2010.

[5] V. Tarokh, H. Jafarkhani, and A.R. Calderbank, Space-time block codes from orthogonal designs, IEEE Transactions on Information Theory, 45(5):1456–1467,1999, DOI :10.1109/49.753730.

[6] Qingfeng Jing, Jiajia Wu, Performance comparison of space-time block and trellis codes in the MIMO land mobile satellite channels, Radioelectronics and Communications System, Vol. 60, No. 1, pp. 3–17, 2017, DOI:10.3103/S0735272717010010.

[7] Nhat-Quang Nhan, Philippe Rostaing, Karine Amis, Ludovic Collin, Emanuel Radoi, Complexity Reduction for the Optimization of Linear Precoders over Random MIMO Channels, DOI 10.1109/TCOMM.2017.2716375, IEEE Transactions on Communications.

[8] Hyoungju Ji, Younsun Kim, Juho Lee, Eko Onggosanusi, Younghan Nam, Jianzhong Zhang, Byungju Lee, and Byonghyo Shim,Overview of Full-Dimension MIMO in LTE-Advanced Pro, EEE Communications Magazin, 2017.

[9] E. Telatar, “Capacity of multi-antenna Gaussian channels”. Eur. TransTelecommun. 10(6), 585–595 (1999), DOI :10.1002/ett.4460100604.

[10] P Rostaing, O Berder, G Burel, L Collin, Minimum BER diagonal precoder for MIMO digital transmissions. IEEE Signal Process. 82(10), 1477–1480 (2002).

[11] P Stoica, G Ganesan, Maximum-SNR spatial-temporal formatting designs for MIMO channels. IEEE Trans. Signal Process. 50(12), 3036–3042 (2002).

[12] H Sampath, P Stoica, A Paulraj, Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion. EEE Trans Commun. 49(12), 2198–2206 (2001).

[13] A Scaglione, P Stoica, S Barbarossa, G Giannakis, H Sampath, Optimal designs for space-time linear precoders and decoders. IEEE Trans. Signal Process. 50(5), 1051–1064 (2002).

[14] L. Collin, O. Berder, P. Rostaing, and G. Burel. Optimal minimum distance based precoder for mimo spatial multiplexing systems. IEEE Transactions on Signal Processing, 52(3):617–627, March 2004.

[15] Tarek Chehade, Ludovic Collin, Philippe Rostaing, Emanuel Radoi, and Oussama Bazzi , Power Allocation Optimization: Linear Precoding Adapted to NB-LDPC Coded MIMO Transmission, Hindawi Publishing Corporation International Journal of Antennas and Propagation, Article ID 975139, Volume 2015, doi.org/10.1155/2015/975139

[16] J.-M. Kwadjane , B. Vrigneau , C. Langlais , Y. Cocheril1 , and M. Berbineau, Performance of the max-dmin precoder in impulsivenoise for railway communications in tunnels, 2013 13th International Conference on ITS Telecommunications (ITST). MAX DMIN

[17] Nhat-Quang Nhan, Philippe Rostaing, Karine Amis, Ludovic Collin, and Emanuel Radoi, EXIT-based optimization of linear precoder for MIMO encoded systems assuming turbo detection, IEEE ICC 2017 Wireless Communications Symposium. RELATED WORK

[18] F Perez-Cruz, MRD Rodrigues, S Verdu, MIMO Gaussian channels with arbitrary inputs: optimal precoding and power allocation. IEEE Trans. Inf. Theory. 56(3), 1070–1084 (2010).

[19] Y. Kim, H. Lee, S. Park, and I. Lee, Optimal precoding for orthogonalized spatial multiplexing in closed-loop MIMO systems, IEEE Journal on Selected Areas in communication, vol. 26, no. 8, 2008.

[20] Yi Jiang, Member, Mahesh K. Varanasi, Fellow, ,Jian Li, Fellow, Performance Analysis of ZF and MMSE Equalizers for MIMO Systems: An In-Depth Study of the High SNR Regime, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 4, APRIL 2011.

[21] Ming-Xian Chang, Wang-Yueh Chang, Maximum-Likelihood Detection for MIMO Systems Based on Differential Metrics, OI 10.1109/TSP.2017.2698411,IEEE Transactions on Signal Processing.

[22] D.Chauhan and J.Bhalani, Performance of Nonlinear Detectors in Spatially Correlated Channels, of Radioelectronics and Communications Systems Journal.Vol.60,No.7, pp.297-302, 2017. DOI: 10.3103/S0735272717070020.

[23] V.L. Seletkov, "Statical characteristics of Soft Docoding Transformation", of Radioelectronics and Communications Systems Journal.Vol.54, No.1, pp.32-36, 2011. DOI:10.3103/S07352711010055.

[24] P. Rajani Kumari, K. Chenna Kesava Reddy, K.S. Ramesh, Hybrid Low Complex near Optimal Detector for Spatial Modulation, International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, No. 2, April 2017, pp. 818~822.

Comparative Analysis of Evolutionary Algorithms for Multi-Objective Travelling Salesman Problem

Nosheen Qamar[1]
Department of Computer Science and Information Technology, University of Lahore, Lahore, Pakistan

Nadeem Akhtar[2], Irfan Younas[3]
Department of Computer Science
National University of Computer and Emerging Sciences, Lahore, Pakistan

***Abstract*—The Evolutionary Computation has grown much in last few years. Inspired by biological evolution, this field is used to solve NP-hard optimization problems to come up with best solution. TSP is most popular and complex problem used to evaluate different algorithms. In this paper, we have conducted a comparative analysis between NSGA-II, NSGA-III, SPEA-2, MOEA/D and VEGA to find out which algorithm best suited for MOTSP problems. The results reveal that the MOEA/D performed better than other three algorithms in terms of more hypervolume, lower value of generational distance (GD), inverse generational distance (IGD) and adaptive epsilon. On the other hand, MOEA-D took more time than rest of the algorithms.**

Keywords*—Evolutionary computation; algorithms; NSGA-II; NSGA-III; MOEA-D; comparative analysis

I. Introduction

The optimization problems with a single objective are relatively easy to solve but in case of more than one objectives the optimization become harder and these kinds of problems are very common in the existing world. It is difficult to come up with unique solution for problems having more than one objective. The two or more objectives optimization problems are called Multi-Objective Optimization Problems (MOP). Most of the MOP are of NP-hard nature and require complex optimization algorithms to solve them. Evolutionary Algorithms (inspired by biological evolutionary theory) is a relatively new field which came into existence from the last few years and has widely been discussed in the last decade [1].

The aim of Traveling Salesman Problem (TSP) is to come across the possible trip with least length for salesman who had to complete his cycle of visiting all the cities with a constraint of visiting each city exactly one time. The nature of Traveling Salesman Problem (TSP) is NP-hard [1]. When there is not just one objective i.e. the minimum distance, but also time, cost and risk etc., then it will become a Multi Objective Traveling Salesman Problem (MOTSP).

In the case of Multi Objective Traveling Salesman Problem (MOTSP), it cannot be solved using deterministic methods, especially when there are large numbers of cities to visit. Heuristic Methods are based on approximations of Pareto Solutions (PS) and Pareto Front (PF) of multi objective traveling salesman problem (MOTSP). The Evolutionary Algorithms (EA) are most promising from other heuristic methods due to their ability to give approximate solutions in a single go. In most of the cases, the target of Multi Objective Evolutionary Algorithms (MOEA) is to come up with approximate PS/PF that would be as close and as diverse as possible to actual PS/PF. The convergence (close to actual/real PF) and diverse (fully spread on the PF) are two important challenges to take care while finding the PF [2], [3].

Two most famous multi objective optimization approaches are Vector Evaluated Genetic Algorithm (VEGA) and Multi Objective Genetic Algorithm (MOGA). The VEGA converts multiple objective functions into one composite function by assigning weights to given functions. But challenging part of this approach is careful assigning of weights to each solution function. This is a difficult task for the assigner to assign some weight to any objective function without deep knowledge of that specific domain [4]. The Second approach Multi Objective Genetic Algorithm (MOGA) aims to find a set of pareto optimal solutions (PS) and then choose a subset of solutions from PS which will then be called pareto optimal front (PF). As going forward from one solution to another, it needs some sacrifices to one objective while optimizing the other. The non-dominated sorting genetic algorithm (NSGA) based on MOGA was proposed in [5]. Later on, the NSGA-II [6] was proposed by avoiding the problems associated with NSGA to deal with Multi Objective Optimization Problems. To deal with more than three objectives problems, (Many-Objective) the NSGA-II did not prove to be very effective hence a new solution was proposed called NSGA-III [7] which was an extension of NSGA-II algorithm.

The Multi Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [8] is a recently developed algorithm inspired by evolutionary algorithms suggesting optimization of multi objectives by decomposing them. The MOEA/D performs better than Non-dominated Sorting Genetic Algorithm II (NSGA-II) and Multi Objective Genetic Local Search (MOGLS). To solve different complex Multi Objective Problems (MOPs), different extensions of Multi Objective Evolutionary Algorithm based on Decomposition (MOEA/D) have been practiced. Multiple initially developed MOEA/D and its multiple extensions are already being applied on MOTSP problem. A new extension named Multi Objective Evolutionary Algorithm derived from Decomposition with Ant Colony Optimization (MOEA/D-ACO) [9] which was proposed based on the idea that each ant will be responsible for one sub problem. The MOEA/D-ACO was compared with BicriterionAnt [10] algorithm by applying it on dual objectives traveling Salesman Problem (b-TSP) and improvement has been clearly observed.

The popularity of Traveling Salesman Problem, its NP-hard nature and it is well known and widely used problem has motivated us to use this problem to test our comparative analysis. In this study, we have applied NSGA-II, NSGA-III SPEA2, MOEA/D and VEGA. This study is a comparative analysis of the above mentioned five algorithms to find out that which algorithm proves to be the best for MOTSP problem.

This paper is structured as follows, Section II discusses the Literature Review, and Section III highlights the comparative analysis of evaluation that which algorithm works best for MOTSP. Finally, Section IV discusses the conclusion along with future work.

II. Literature Review

The Traveling Salesman Problem (TSP) is a combinational optimization problem [11] with an aim of finding shortest tour visiting all cities (from a given set) exactly at once. This could be the most popular NP-hard optimization problem and lots of studies could be made to get an optimized solution for this problem. There are different variants [12], [24] of Traveling Salesman Problem proposed including multiple-traveling salesman problem [13], [14], Multi-objective two-depot traveling salesman problem [15], probabilistic traveling salesman problem [16], Multi-objective Multiple Traveling Salesman Problem [17], Multi-objective Physical traveling [18], [19] and Multi-objective generalized Travelling Salesman Problem [20], etc.

The classic Travelling Salesman Problem (TSP) includes a number of variants, the Multi-Objective Traveling Salesman Problem (MOTSP) is the one which has been explored by a large number of researchers where multiple objectives i.e. time, cost, distance, etc. need to be optimized [21]. Due to its NP-hard nature, it is very difficult to get the optimal solution in the reasonable time. That is because multiple approximation techniques were proposed in three major categories i.e. classical heuristics, population based meta-heuristics and meta-heuristics based on single solution. The chapters of Johnson and McGeoch [22], [23] from the book of Gutin and Punnen [14] discuss the symmetric heuristics and asymmetric heuristics versions of Travelling Salesman Problem respectively. The [25] discussed the survey of local search (meta-heuristics for TSP), while the [26] describes genetic algorithms (GA) and [27] covers mimetic algorithms used for TSP.

In 1999 Preux and Talbi [28] describe the search algorithm's behavior with intent that the structure of the search space may improve the performance of the algorithm. In their study they reviewed the knowledge related to search spaces of combinational optimization problems and discussed the hybridization in detail. They also presented different techniques of hybridization based on their knowledge, on search space structure and the performance of an algorithm.

Borges and Hansen in 2000 [29] discussed the Multi-Objective TSP. The authors discussed the "global convexity" in Multi-Objective Combinational Optimization Problems generally and Multi-Objective TSP specifically. The paper focused on local optima landscape by using classical two-opt neighbors (without breaking the tour it will replace two edges with single possible solution, and two edges would get removed) with help of famous scalar functions i.e. Tchebycheff or weighted sum of multiple objectives.

The [30] in 2004 discussed the solution for TSP based on hybrid evolutionary algorithm, authors proposed an algorithm with strategy of distance preserving crossover (DPX) integrating memory as ant pheromone during the city selection process aiming to compliment the successful results of genetic algorithm (GA). The probability of distance and previous success for city selection along with combination of genetic algorithm (GA) and DPXwould be considered as additional information and would help in finding optimized quality solutions for TSP with reduced computational complexity.

The Pareto Converging Genetic Algorithm (PCGA) was proposed by Kumar and Singh [31] in 2007, doing hybridization of Pareto Rank Genetic Algorithm with Local Search. The evaluation criterion for each solution was its rank and total numbers of dominating individuals. The two individuals were selected based on raffle wheel and the distance preserving crossover (DPX) operation was performed to generate offspring. The produced offspring were again merged with population based on its rank. After doing mutation operation the converging criterion was defined depending on "rank-histograms" and within the population the rank of individuals is one plus the total number of individuals dominating it aiming to assign all non-dominated individuals to one. The union of new population with older one was ranked. As close as possible the pareto will be converged to rank histogram equal to a single value which is not equal to zero entry of 1/2for rank equal to 1 correspond to that no solution is better than the previous (older) population originated in evolving the new population.

Changdar et al. [32] in the year 2014 considered two objectives, cost as first and time as second to solve the multi-objective Static TSP in their suggested multi-objective genetic algorithm (MOGA). The nature of proposed algorithm was not clear. In the same year, Li [33] managed to propose an algorithm for multi-objective dynamic TSP with two and three objectives with a parallel search system. Moreover, Florios and Mavrotas's [34] proposed solution for Multi Objective Travelling Salesman Problem (MOTSP) and Set Covering Problem (SCP) which was based on Pareto front for dual objectives functions with help of AUGMECON2 method. Another contribution by Bouzoubia et al. [35] in the same year, made a difference by using couple of variations derived by Multi Objective Chemical Reaction Optimization (MOCRO) to get good solutions for multi-objective TSP by the use of non-dominated sorting technique which was already used in NSGA-II algorithm. He [36] also contributed to solve multi-objective TSP using membrane algorithms. Labadie et al. [37] were also one of those who put their part to get optimise solution for multi-objective TSP using two objectives with profits (BOMTSPP) in same year.

Bolano et al. [38] in 2015 proposed NSGA-II algorithm to solve multi-objective TSP using a NSGA II algorithm. Wang et al. [39] suggested hybrid NSGA-II algorithm to achieve optimal good solution for multi-objective TSP initially and

then he proposed a new hybrid algorithm [40] which combined an uncertain approach with Artificial Bee Colony (ABC) algorithm. Ariyasingha and Fernando [41] conducted a review of Colony Optimization Algorithms (COA) for MOTSP for bi-objective and tetra-objective functions.

The 2016 researches on multi objective travelling salesman problem contain a research by Cornu et al. [42] proposed a novel multi objective decomposition algorithm called perturbed Decomposition Algorithm (PDA). The newly proposed PDA algorithm suggests combination of decomposition methods, data perturbation and local search. Authors claimed that PDA performs better than existing algorithms available on multi objective travelling salesman problem (MOTSP).

Author in [43] suggest a new solution for Multi Objective Travelling Salesman Problem (MOTSP) with imprecise Multi Objective Genetic Algorithm (iMOGA) with fuzzy age selection. The proposed algorithm also used adaptive crossover and mutation which depends on generation. The fuzzy age was replaced by fuzzy extended age.

III. Algorithms Selected for Experiment

A. NSGA-II

The NSGA-II is a faster and better algorithm than the MOEA algorithms in terms of close coverage and correct pareto optimal front. The NSGA-II works as the initial population has been defined with some set of solutions, then λ solutions are generated with help of stochastic variation operators. The λ generated solutions evaluated and then ranked on pareto from as best solutions on first non-dominated front and so on. The main reason behind the selection of this algorithm was its low complexity, good coverage and better diversity.

B. NSGA-III

The many objective optimization problems are very challenging to optimize and are difficult to handle. The NSGA-III is the algorithm used to handle many objective problems. The reason behind selection of this algorithm was that during our experiments, we had up to 5 objectives and in that scenario this was an effective algorithm to measure results.

C. SPEA-2

The SPEA-2 is an improved version of SPEA algorithm and it starts its working with initial population and an empty archive. Then the fitness values of solutions are evaluated and then the solutions with best fitness are added to the archive (with a specific number) with non-dominated solutions and if there is still space, the good dominated solutions can also be added. After fulfilling the termination criteria, binary tournament is performed and the next generation is created after recombination and mutation operation and this process repeat with some specific set of generations. The reason behind selection of this algorithm is its focus on dominance count and rank and that good coverage can be achieved with the help of this algorithm.

D. MOEA/D

The main idea behind the MOEA/D algorithm is the decomposition of multi-objective optimization problem into a number of small scalar optimization problems and then optimizes those scalar problems simultaneously. Every sub problem was optimized with help of its multiple neighbors providing information. The motivation behind using this algorithm was its lower computational complexity due to breaking a larger and complex problem into multiple scalar problems and then their optimization based on their neighbors.

E. VEGA

The VEGA (Vector Evaluated Genetic Algorithm) is pioneer algorithm to find non-dominated solution for multi objective optimization problems. It is an extension of single objective genetic algorithm to optimize the multi objective problems. We used this algorithm due to its efficiency and higher speed.

IV. Experiments and Results

As mentioned above different experiments were conducted on TSP problem using five different (NSGA-II, NSGA-III, SEPA-2, MOEA/D and VEGA) algorithms and this section discusses the experimental setup and has the results of those experiments.

A. Experiment Setup

In the multi-objective (K-objective) Traveling Salesman Problem, K objective functions need to be defined. These objectives can be cost of the tour, travel time or any other factor which need to be optimized. Table I demonstrates the experimental setup. The Cent OS, 8cores platform with Java8 and MOEA framework were used. The population size was decided as 50 and 100 with 50, 100, 1000 and 10000 generations. The experiment was repeated for 10,100 and 1000 iterations. The results were compared based on Hypervolume, Generational-Distance (GD), Inverted-Generational-Distance (IGD), Additive-ε and Time taken to conduct the experiment. Let's assume all contributing factors are on different graphs with same number of vertices but have different values for edges. In order to simulate multiple objectives for the TSP, different TSPLIB problem situations which have the same number of nodes were used. Each situation was considered to be a single objective which requires to be minimized. Multiple experiments were conducted for 5 objectives (5 cyclic tours for 5 libraries) of TSP problem by using TSPLIB standard dataset library [44]. System was implemented in Java language and the use MOEA framework [45] for the conditions of experiments.

B. Results and Analysis

Results have been compared by using four indicators as Hypervolume, Generational-Distance (GD), Inverted-Generational-Distance (IGD) and Adaptive-ε.

TABLE I. EXPERIMENTAL SETUP

Platform	Cent OS, 8 cores 8 GB Memory (6 GB user memory)
Framework	Java 8 MOEA framework
TSPLIB Libraries	kroA100, kroB100, kroC100, kroD100, kroE100
Algorithms	NSGA-II, NSGA-III, SPEA-2, MOEA/D, VEGA
Population size	50, 100
# Generations	50, 100, 1000, 10000
# Iterations	10, 100, 1000
Result Indicators	Hypervolume, Generational-Distance (GD), Inverted-Generational-Distance (IGD), Additive-ε and Time

1) Hypervolume: In Table II and Fig. 1 below, all the experimental results have been shown for the Hypervolume indicator. The results in Table II and Fig. 1 clearly show that the MOEA/D performed well for the population size 50 and 100, generations 10000 and the iterations 10,100 and 1000. The highest gain hypervolume produced by MOEA/D is between 0.172291 to 0.206567. Based on the given data we can say that the MOEA/D has performed better than the other algorithms for the TSP problem.

TABLE II. HYPERVOLUME COMPARISON FOR ALL ITERATIONS

Population-Size	Generations	Iterations	NSGA-II	NSGA-III	SPEA-2	MOEA/D	VEGA
50	50	10	0.002788	0.002561	0.002653	0.003207	0.003035
50	50	100	0.002701	0.002741	0.002682	0.002830	0.002638
50	50	1000	0.002712	0.002691	0.002744	0.002680	0.002730
50	100	10	0.003655	0.004549	0.004026	0.003426	0.002398
50	100	100	0.003721	0.003631	0.004269	0.003277	0.002720
50	100	1000	0.003792	0.003777	0.004420	0.003154	0.002697
50	1000	10	0.005910	0.007092	0.007180	0.012533	0.002285
50	1000	100	0.006470	0.007180	0.007078	0.010248	0.002643
50	1000	1000	0.006407	0.007465	0.007502	0.011357	0.002739
50	10000	10	0.008954	0.040115	0.013966	**0.206567**	0.016630
50	10000	100	0.009051	0.022306	0.013521	**0.172291**	0.015493
50	10000	1000	0.009039	0.027165	0.013355	**0.184121**	0.017165
100	50	10	0.003925	0.003964	0.003511	0.004261	0.003702
100	50	100	0.003828	0.003838	0.003736	0.004014	0.003931
100	50	1000	0.00379	0.003747	0.003797	0.003777	0.003827
100	100	10	0.003868	0.004339	0.003717	0.003901	0.004148
100	100	100	0.003757	0.003687	0.003882	0.003732	0.003876
100	100	1000	0.003749	0.003780	0.003763	0.003830	0.003752
100	1000	10	0.008978	0.008367	0.008371	0.009061	0.003326
100	1000	100	0.008202	0.008122	0.008506	0.008382	0.003980
100	1000	1000	0.008323	0.008314	0.008699	0.008442	0.003806
100	10000	10	0.011123	0.012378	0.013859	**0.179984**	0.004620
100	10000	100	0.011630	0.012774	0.015254	**0.136949**	0.003826
100	10000	1000	0.011542	0.012093	0.014958	**0.136986**	0.004074

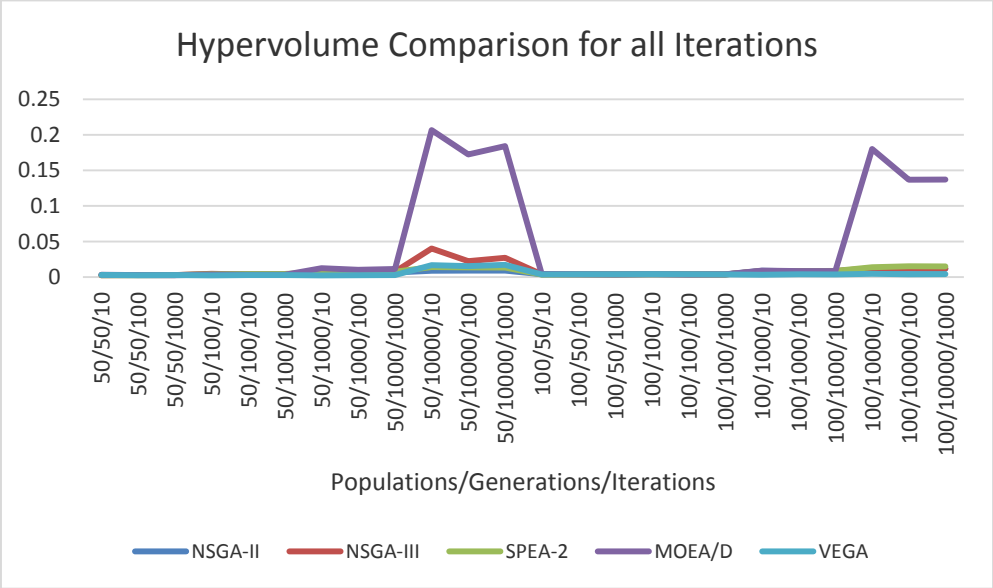


Fig. 1. Hypervolume comparison for all iterations.

2) Generational Distance (GD):

Table III contains the comparison data and based on that, Table II demonstrates that MOEA/D performed better than the rest of the three algorithms in terms of generational distance. The NSGA-III was less better with higher value of generational distance. The figure shows that while comparing base of GD, the results shows the MOEA/D performed least for the population size 50 and 100, the generations 10000 and the iterations 10,100 and 1000. NSGA-II, SPEA2 and VEGA performed almost equal for TSP problem with multiple objectives. Fig. 2 is a graphical representation of Table III.

TABLE III. GENERATIONAL DISTANCE (GD) COMPARISON FOR ALL ITERATIONS

Population-Size	Generations	Iterations	NSGA-II	NSGA-III	SPEA-2	MOEA/D	VEGA
50	50	10	0.139000	0.146823	0.140682	0.148243	0.163895
50	50	100	0.140185	0.142021	0.140469	0.143294	0.185806
50	50	1000	0.140536	0.140822	0.141191	0.140571	0.237233
50	100	10	0.115427	0.111052	0.102978	0.159342	0.170148
50	100	100	0.110141	0.108757	0.102570	0.148458	0.191301
50	100	1000	0.109818	0.109527	0.102568	0.147099	0.240964
50	1000	10	0.097494	0.094884	0.095641	0.096621	0.157229
50	1000	100	0.096075	0.094440	0.095249	0.101171	0.199204
50	1000	1000	0.096173	0.094236	0.095051	0.100675	0.234012
50	10000	10	0.090736	0.064280	0.086743	0.024569	0.150873
50	10000	100	0.090904	0.074443	0.088067	0.029978	0.169089
50	10000	1000	0.091203	0.070661	0.087542	0.027356	0.168105
100	50	10	0.115540	0.114182	0.107821	0.115765	0.117951
100	50	100	0.109221	0.108482	0.110741	0.110618	0.146849
100	50	1000	0.110426	0.109321	0.109581	0.109903	0.157706
100	100	10	0.108834	0.110043	0.107679	0.104872	0.121678
100	100	100	0.109281	0.108790	0.110245	0.110162	0.146884
100	100	1000	0.109025	0.109573	0.109145	0.109929	0.159126
100	1000	10	0.067435	0.067517	0.067733	0.089163	0.133054
100	1000	100	0.067604	0.067645	0.067243	0.088394	0.137151
100	1000	1000	0.067518	0.067437	0.067328	0.08828	0.279085
100	10000	10	0.064252	0.062686	0.062755	0.019408	0.127604
100	10000	100	0.063733	0.063687	0.062082	0.026588	0.165126
100	10000	1000	0.064086	0.063720	0.062128	0.027225	0.156897

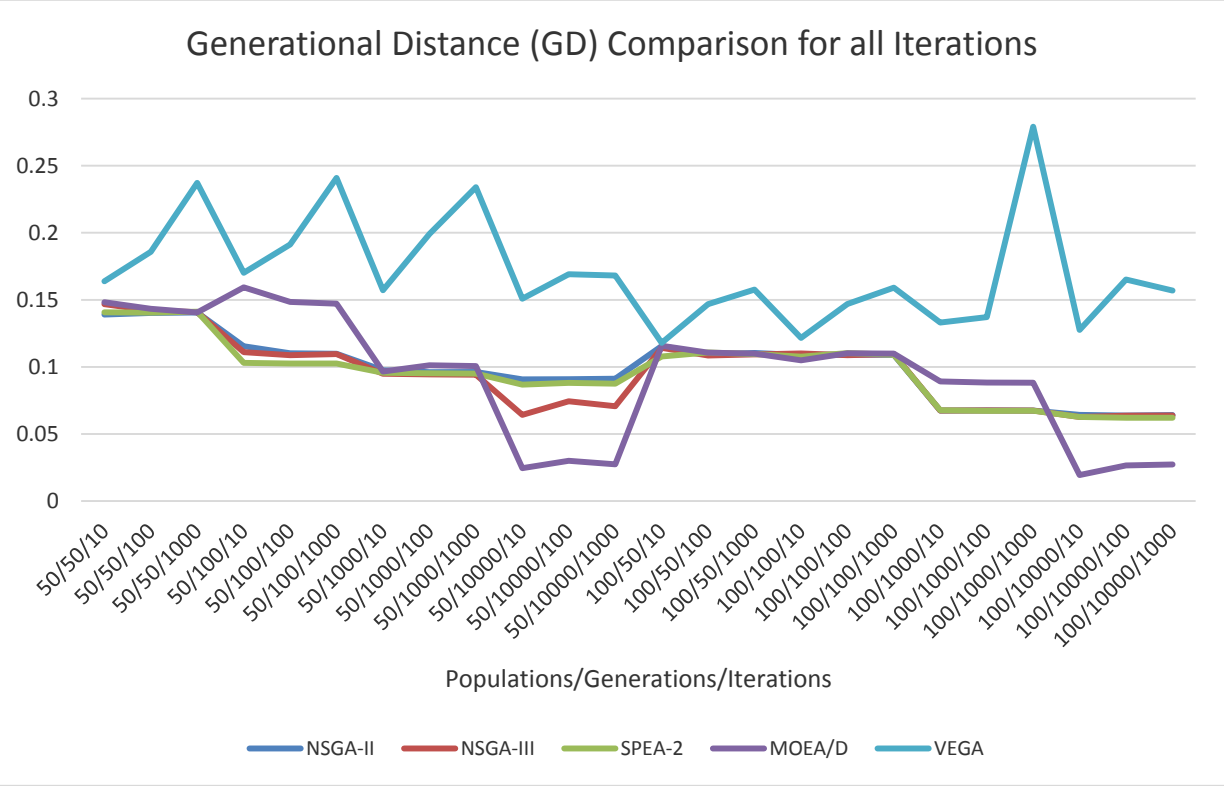


Fig. 2. Generational Distance (GD) comparison for all iterations.

3) Inverted-Generational Distance (IGD):

Table IV and so Fig. 3 (constructed from the data available in Table IV) represents that the MOEA/D performed better from rest of the four algorithms, specifically at the noticeable point of 50 and 100 population size, 10000 generations and 10,100, 1000 iterations. The rest of the three algorithms were with almost equal results.

TABLE IV. INVERTED-GENERATIONAL DISTANCE (IGD) COMPARISON FOR ALL ITERATIONS

Population-Size	Generations	Iterations	NSGA-II	NSGA-III	SPEA-2	MOEA/D	VEGA
50	50	10	0.876195	0.883478	0.875846	0.851304	0.841886
50	50	100	0.870580	0.871930	0.876809	0.867707	0.871765
50	50	1000	0.871787	0.872499	0.869325	0.873542	0.871335
50	100	10	0.840716	0.817917	0.838676	0.83858	0.890045
50	100	100	0.841584	0.850365	0.830244	0.849765	0.873776
50	100	1000	0.843890	0.844259	0.828859	0.854852	0.872269
50	1000	10	0.787684	0.776834	0.768491	0.698727	0.896334
50	1000	100	0.781947	0.774108	0.771899	0.730504	0.878822
50	1000	1000	0.782636	0.769031	0.766264	0.712183	0.873151
50	10000	10	0.730746	0.556786	0.690203	0.261163	0.651132
50	10000	100	0.740443	0.620768	0.685077	0.294764	0.656258
50	10000	1000	0.736205	0.592826	0.684586	0.283587	0.643210
100	50	10	0.838765	0.836466	0.849648	0.820317	0.845035
100	50	100	0.841628	0.844725	0.843842	0.838111	0.829266
100	50	1000	0.843522	0.845131	0.843501	0.845035	0.841485
100	100	10	0.847687	0.834806	0.853649	0.833991	0.837925
100	100	100	0.845193	0.845431	0.841668	0.840722	0.835674
100	100	1000	0.845553	0.845276	0.844949	0.842207	0.844789
100	1000	10	0.759297	0.770600	0.770172	0.745906	0.857111
100	1000	100	0.773435	0.771152	0.768211	0.755973	0.836586
100	1000	1000	0.767053	0.766924	0.762013	0.757285	0.844768
100	10000	10	0.734211	0.718311	0.702650	0.258231	0.810780
100	10000	100	0.721530	0.713585	0.685823	0.316330	0.841607
100	10000	1000	0.722457	0.718744	0.690895	0.314585	0.834649

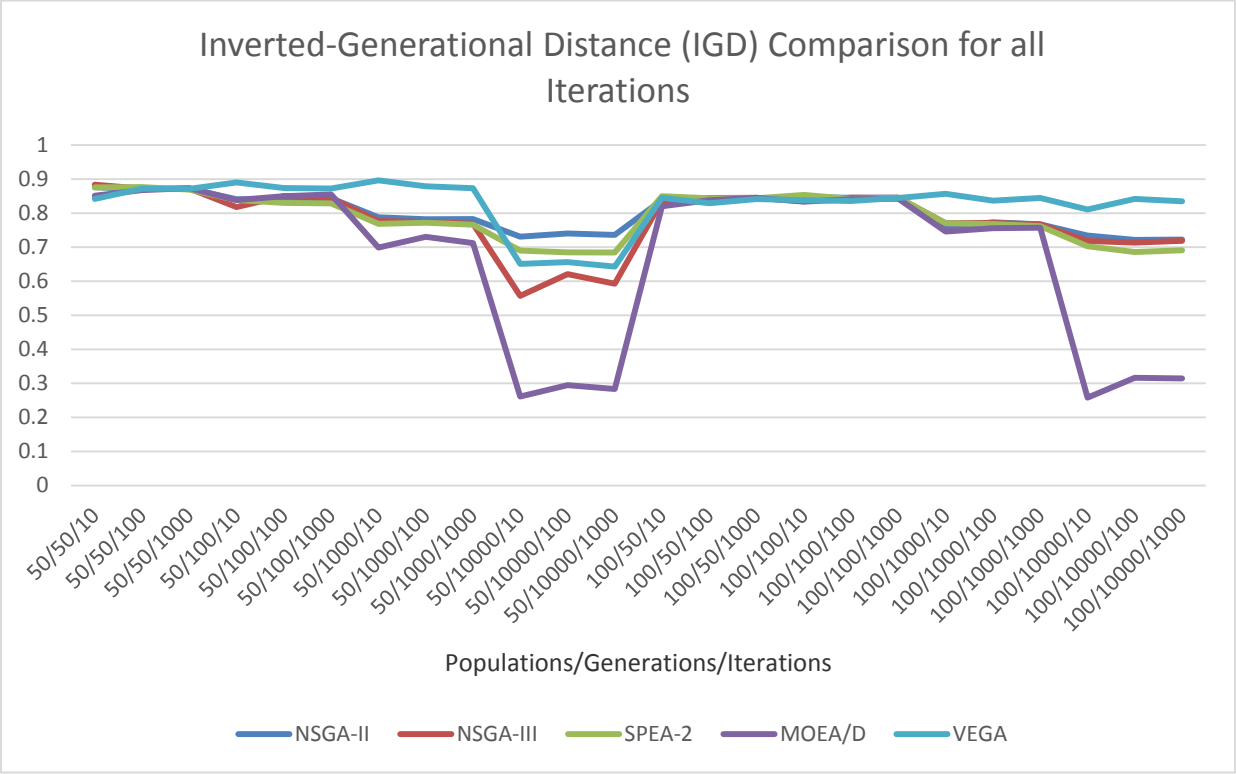


Fig. 3. Inverted-Generational Distance (IGD) comparison for all iterations.

4) Adaptive-ε:

The data in Table V was used to make the graph represented in Fig. 4 which illustrates that the MOEA/D is better when analysing the results based on Adaptive epsilon. As we know that lower epsilon value is better to achieve and when comparing with other four algorithms the MOEA/D is comaratively better.

TABLE V. ADAPTIVE EPSILON COMPARISON FOR ALL ITERATIONS

Population-Size	Generations	Iterations	NSGA-II	NSGA-III	SPEA-2	MOEA/D	VEGA
50	50	10	0.648598	0.652171	0.646459	0.654711	0.639249
50	50	100	0.653534	0.651258	0.653652	0.653722	0.649986
50	50	1000	0.651104	0.651957	0.652818	0.652084	0.652297
50	100	10	0.621849	0.614253	0.603638	0.645525	0.667193
50	100	100	0.623724	0.625773	0.609900	0.645131	0.654024
50	100	1000	0.622009	0.622433	0.608512	0.650904	0.652784
50	1000	10	0.589174	0.574628	0.568762	0.555046	0.653651
50	1000	100	0.581862	0.571760	0.569654	0.560187	0.652413
50	1000	1000	0.581643	0.571665	0.566568	0.557785	0.652627
50	10000	10	0.566406	0.503615	0.542692	0.308017	0.575066
50	10000	100	0.577152	0.529691	0.540753	0.327255	0.584646
50	10000	1000	0.575764	0.514632	0.539941	0.322224	0.574605
100	50	10	0.619609	0.619759	0.625357	0.617392	0.619747
100	50	100	0.625448	0.618305	0.621055	0.624260	0.623730
100	50	1000	0.621742	0.622209	0.621561	0.622711	0.621343
100	100	10	0.618887	0.624665	0.627339	0.619170	0.622793
100	100	100	0.621013	0.622225	0.621658	0.624664	0.617485
100	100	1000	0.621829	0.622342	0.621681	0.622295	0.622017
100	1000	10	0.550720	0.551214	0.550506	0.561869	0.631512
100	1000	100	0.553415	0.554696	0.550566	0.570871	0.622054
100	1000	1000	0.552924	0.553912	0.549637	0.568403	0.622550
100	10000	10	0.547026	0.539875	0.517950	0.314718	0.611036
100	10000	100	0.540999	0.532733	0.514461	0.355844	0.619680
100	10000	1000	0.543220	0.535499	0.513611	0.358324	0.619954

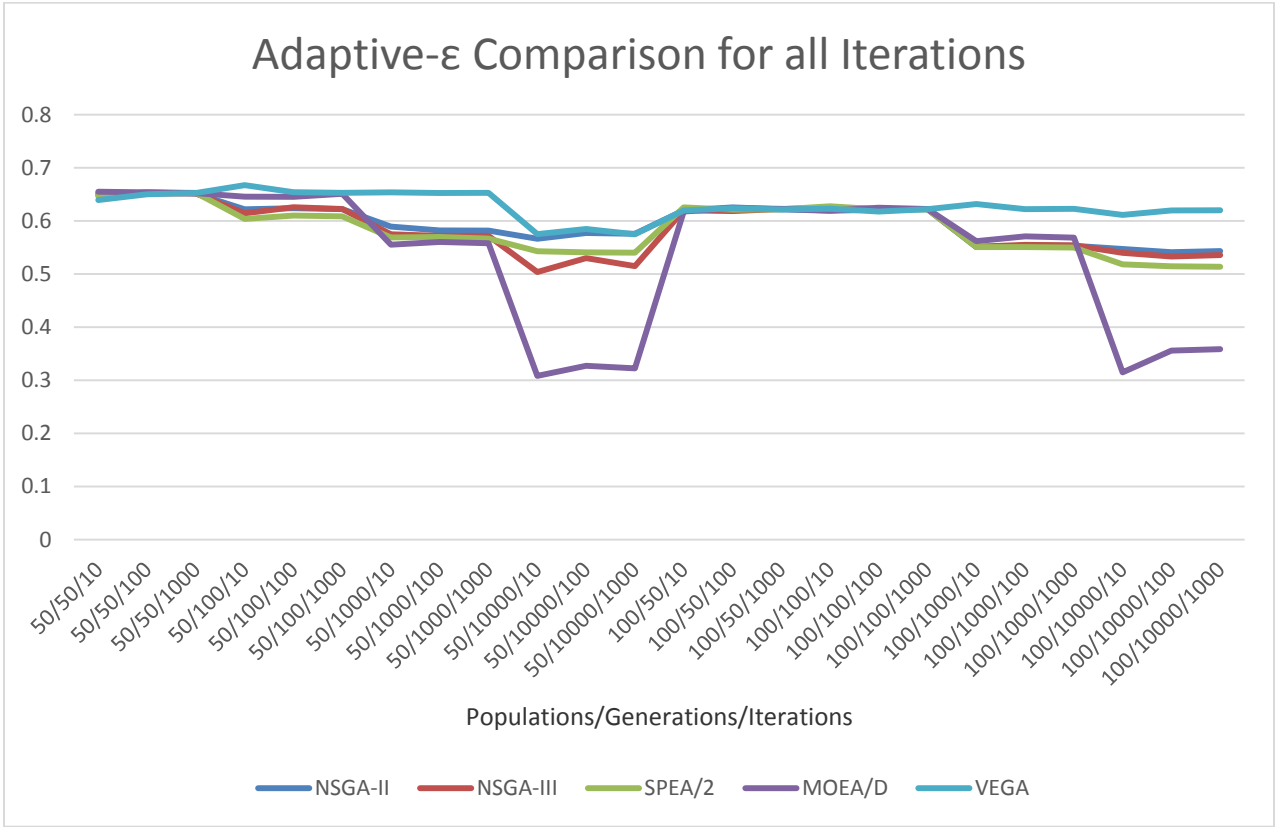


Fig. 4. Adaptive Epsilon comparison for all iteration.

5) Time Comparison:

Fig. 5 give us representation of the time comparison of our applied algorithms. Fig. 6 demonstrates the comparison of time for all the values including different populations, generations and iterations. The results clearly show that MOEA-D took more time than all other algorithms. The SPEA and NSGA-III took almost equal time and NSGA-II and VEGA took lowest time for the combination of 50 population size, 10000 generations and 1000 iterations. Almost same is the case for 100 population, with same number of generations and iterations.

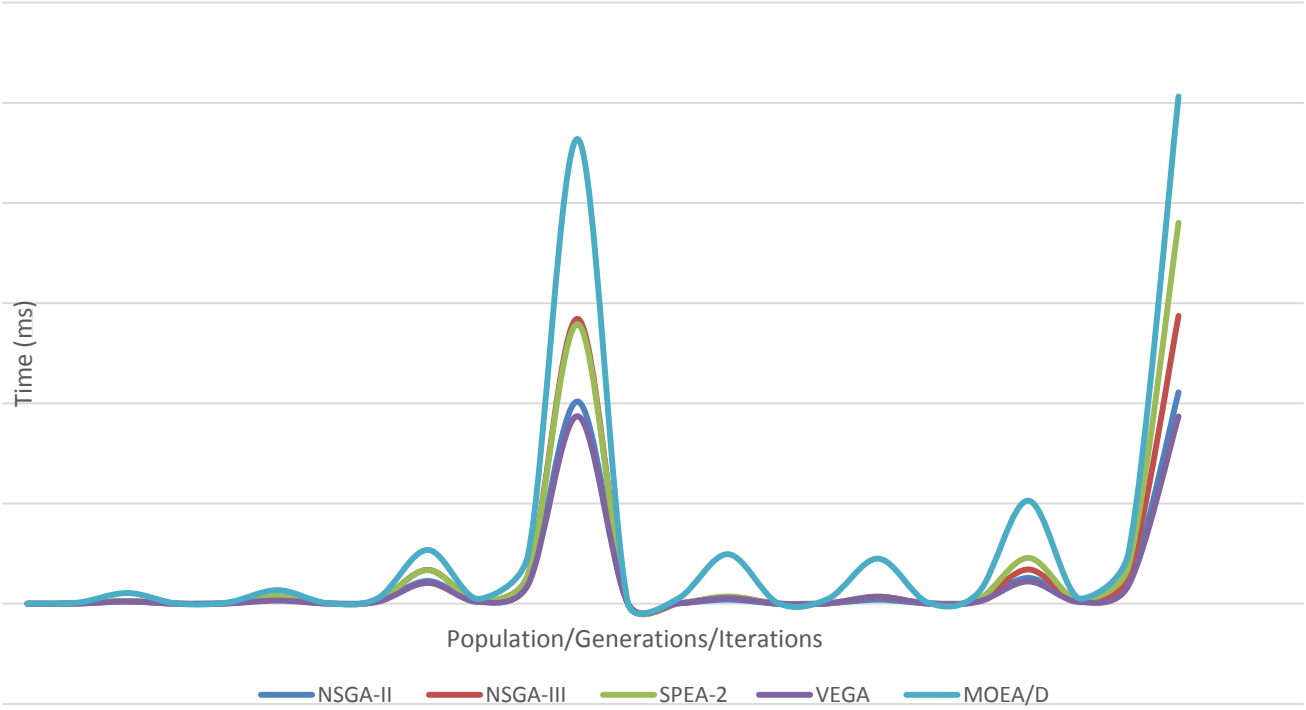


Fig. 5. Time comparison of all the algorithms.

V. Conclusions and Future Work

The TSP is a widely evaluated single objective problem. This problem can be expanded by converting it into multi objective or many objectives by considering different objectives like cost, time, speed etc. In our study we have applied five popular evolutionary algorithms (NSGA-II, NSGA-III, SPEA2, MOEA/D and VEGA) to solve the TSP problem and came up with the results that MOEA/D performed better on different (hypervolume, generational distance, inverted generational distance and adaptive epsilon) indicators more specifically with the population size 50 and 100 with 10000 no. of generations and 10, 100 &1000 iterations. The results further show that although MOEA-D performed better than other algorithms but it took more time in comparison to the rest of algorithms. With respect to time the NSGA-II and VEGA took the lowest time.

The future work can cover the implementation of the above algorithms on some other well-known problem like the problem of knapsack or different combination of algorithms can be used (other than ours) to find out what algorithm works best for kanpsack problem. It would also be interesting to know what would be the results when using different number of populations (greater than 100 as our maximum population size was 100) and different number of iterations and generations.

References

[1] R. Baraglia, J. Hidalgo, and R. Perego, "A hybrid heuristic for the traveling salesman problem," IEEE Transactions on Evolutionary Computation, vol. 5, no. 6, pp. 613 –622, 2001.

[2] K. C. Tan, E. F. Khor, and T. H. Lee, Multiobjective evolutionary algorithms and applications. Springer, 2005.

[3] C. A. Coello Coello, D. A. Van Veldhuizen, and G. B. Lamont,Evolutionary algorithms for solving multi-objective problem. Kluwer, 2002.

[4] J.D. Schaffer, "Multiple Objective Optimization with Vector Evaluated Genetic Algorithms," In Proc. 1st International Conference on Genetic Algorithms, pp.93-100, 1985.

[5] N. Srinivas and K. Deb, "Multiobjective function optimization usingnondominated sorting genetic algorithms," *Evol. Comput.*, vol. 2, no.3, pp. 221–248, Fall 1995.

[6] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, Vol. 6, APRIL 2002.

[7] K. Deb, H. Jain, "An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints," IEEE Transactions of Evolutionary Computation, Vol. 18, Aug 2014.

[8] Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition." IEEE Trans. Evolutionary Computation, vol. 11, no. 6, pp. 712–731, 2007.

[9] L. Ke, Q. Zhang, and R. Battiti, "Moea/d-aco: A multiobjective evolutionary algorithm using decomposition and ant colony." IEEE Trans.Cybern., vol. 43, no. 6, pp. 1845–1859, 2013.

[10] S. Iredi, D. Merkle, and M. Middendorf, "Bi-criterion optimization with multi colony ant algorithms." in EMO, ser. Lec. Notes in Comp. Science, E. Zitzler, K. Deb, L. Thiele, C. A. C. Coello, and D. Corne, Eds., vol.1993. Springer, 2001, pp. 359–372.

[11] G. Gutin, ,A. Punnen, "The traveling salesman problem and its variations,"Kluwer Academic Publishers, Dordrecht, 2002.

[12] G. Reinelt, "TSPLIB A traveling salesman problem library," ORSA J.Comput., 1991.

[13] H. Zhou, M.Song, "An improvement of partheno-genetic algorithm to solve multiple travelling salesmen problem" in Proc. 15th International Conference on Computer and Information Science (ICIS), IEEE/ACIS, 2016. DOI. 10.1109/ICIS.2016.7550780

[14] J. Li, M. Zhou, Q. Sun, X. Dai, X. Yu, "Colored Traveling Salesman Problem" in IEEE Transactions on Cybernetics, vol. 45, pp. 2390 – 2401, Nov. 2015. DOI. 10.1109/TCYB.2014.2371918

[15] H. Zhong, N. Zhuang, Z. Wu, W.Cai, "Multi-objective two-depot traveling salesman problem with uncertain fleet size for hazardous materials" in Proc. 8th International Conference on Supply Chain Management and Information Systems (SCMIS), 2011.

[16] A. Henchiri, M. Bellalouna, W. Khasnaji, "Probabilistic traveling salesman problem: a survey" in Proc. Federated Conference on Computer Science and Information Systems, pp. 55-60, 2014. DOI. 10.15439/2014F381

[17] V. A. Shim, K. C. Tan, and C. Y. Cheong, "A Hybrid Estimation of Distribution Algorithm with Decomposition for Solving the Multiobjective Multiple Traveling Salesman Problem" IEEE Transactions on systems, man and cybernetics, vol. 42, pp. 682-691, Oct. 2012. DOI. 10.1109/TSMCC.2012.2188285

[18] D. Perez, E. Powley, D.l Whitehouse, S. Samothrakis, S. Lucas, P.I. Cowling, "The 2013 Multi-Objective Physical Travelling Salesman Problem Competition," in Proc. IEEE Congress on Evolutionary Computation, 2014,pp. 2314-2321.

[19] E.J. Powley, D. Whitehouse, P.I. Cowling, "Monte Carlo Tree Search with macro-actions and heuristic route planning for the Multiobjective Physical Travelling Salesman Problem," in Proc. IEEE Conference on Computational Intelligence in Games (CIG), 2013. DOI. 10.1109/CIG.2013.6633658

[20] J.V.Pinxten, M. Geilen, T. Basten, U. Waqas, L. Somers, "Online Heuristic for the Multi-objective Generalized Traveling Salesman Problem" Design, Automation & Test in Europe Conference & Exhibition, 2016.

[21] F. Samanlioglu, W.G. Ferrell,M.E. Kurz, "A memetic random-key genetic algorithm for a symmetric multi-objective traveling salesman problem,"Computers and Industrial Engineering, vol. 55, pp. 439-449, 2008.

[22] D.S. Johnson, L.A. McGeoch, "Experimental analysis of heuristics for the ATSP," in The Traveling Salesman Problem and its Variations, G. Gutin, A. Punnen, eds. Kluwer, Dordrecht, 2002.

[23] D.S. Johnson, L.A. McGeoch, "Experimental analysis of heuristics for the STSP," in The Traveling Salesman Problem and its Variations, G. Gutin, A. Punnen, Kluwer, Dordrecht, 2002.

[24] G. Gutin, A. Punnen, "The Traveling Salesman Problem and its Variations," Kluwer, Dordrecht. 2002.

[25] D.S. Johnson, L.A. McGeoch, "The Traveling Salesman Problem: A Case Study in Local Optimization," in Local Search in Combinatorial Optimization, E.H.L. Aarts, J.K. Lenstra, Eds. John Wiley and Sons Ltd., Chichester, 1997, pp. 215–310.

[26] P. Larranaga, C.M.H. Kuijpers, R.H. Murga, I. Inza, S. Dizdarevic, "Genetic algorithms for the travelling salesman problem: A review of representations and operators," Artificial Intelligence Review, vol. 13, pp. 129–170, 1999.

[27] P. Merz, B. Freisleben, " Memetic algorithms for the traveling salesman problem. Complex Systems," vol. 13, pp. 297–345, 2001.

[28] P. Preux, E.G. Talbi, "Towards hybrid evolutionary algorithms," International Transactions in Operational Research, Wiley, vol. 6, Nov. 1999, pp. 557-570

[29] P.C. Borges, M.P. Hansen, "A study of global convexity for a multiple objective travelling salesman problem," in Essays and surveys in metaheuristics, C.C. Ribeiro, P. Hansen, Eds. Kluwer, Dordrecht, pp. 129–150, 2000.

[30] C.M. White, G.G. Yen, "A hybrid evolutionary algorithm for traveling salesman problem" in Proc. Congress on Evolutionary Computation, IEEE, 2004. DOI. 10.1109/CEC.2004.1331070

[31] R. Kumar,P.K. Singh, " Pareto evolutionary algorithm hybridized with local search for biobjective TSP" in Hybrid Evolutionary Algorithms, C. Grosan,A. Abraham, H. Ishibuchi, Eds. Springer, Heidelberg, 2007.Ff

[32] C. Changdar,G.S. Mahapatra, R.K. Pal, " An efficient genetic algorithm for multi-objective solid travelling salesman problem under fuzziness," Swarm and Evolutionary Computation, vol. 15, pp. 27-37,2014.

[33] W. Li, "A parallel search system for dynamic multi-objective traveling salesman problem," Journal of Mathematics and System Science, vol. 4, pp. 295-314, 2014.

[34] K. Florios, G. Mavrotas, "Generation of the exact pareto set in multi-objective traveling salesman and set covering problems," Applied Mathematics and Computation, vol. 237, pp. 1-19, 2014.

[35] S. Bouzoubia, A. Layeb, S. Chikhi, "A multi-objective chemical reaction optimization algorithm for multi-objective travelling salesman problem," International Journal of Innovative Computing and Applications, vol. 6, pp. 87-101, 2014.

[36] J. He, "Solving the multiobjective multiple traveling salesmen problem using membrane algorithm, Bio-Inspired Computing, Theories and Applications," Communications in Computer and Information Science, vol. 472, pp. 171-175, 2014.

[37] N. Labadie, J. Melechovsky, C. Prins, "A parallel search system for dynamic multi-objective traveling salesman problem, Applications of Multi-Criteria and Game Theory Approaches," Springer-Verlag, London, 2014.

[38] R.I. Bolanos, M.G. Echeverry, J.W. Escobar, " A multiobjective nondominated sorting genetic algorithm (NSGA-II) for the multiple traveling salesman problem," Decision Science Letters, vol. 4, pp. 559-568, 2015.

[39] P. Wang, C. Sanin, E. Szczerbicki, "Evolutionary algorithm and decisional DNA for multiple traveling salesman problem," Neurocomputing, vol. 150, pp. 50-57, 2015.

[40] Z. Wang, J. Guo, M. Zheng, Y., Wang, "Uncertain multiobjective traveling salesman problem," European Journal of Operational Research, vol. 241, pp. 478-489, 2015.

[41] I.D. Ariyasingha, T.G.I. Fernando, "Performance analysis of the multiobjective ant colony optimization algorithms for the traveling salesman problem," Swarm and Evolutionary Computation, in Press, 2015.

[42] M. Cornu, T. Cazenave, D. Vanderpooten, "Perturbed Decomposition Algorithm applied to the multi-objective Traveling Salesman Problem," Computers & Operations Research elsevier, Apr 2016.

[43] S. Maity, A. Roy, M. Maitic, "An imprecise Multi-Objective Genetic Algorithm for uncertain Constrained Multi-Objective Solid Travelling Salesman Problem," Expert Systems with Applications elsevier, vol. 46, pp. 196-223, Mar 2016.

[44] S http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/.

[45] http://moeaframework.org/.

Teen's Social Media Adoption: An Empirical Investigation in Indonesia

Ari Kusyanti, Harin Puspa Ayu Catherina, Dita Rahma Puspitasari, Yustiyana April Lia Sari
Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

***Abstract*—Social media has reached their popularity in the past decade. Indonesia has more than 63 million social media users who are accessing their account through mobile phone and therefore Indonesia is the third largest users in the world after United States and India. This study is attempted to determine the factors affecting user behaviour intention of social media usage. TAM (Technology Acceptance Model) for Social Media by Rauniar et al. is adopted to provide empirical evidence of teens in Indonesia. Data were collected through questionnaire survey and hypotheses are analyzed with SEM (Structural Equation Modeling). Result shows that factor affecting Indonesian teens in using social media is perceived usefullness (PP), while Trustworthiness (TW) has no significant influence towards thier intention to use social media.**

Keywords*—Social media; Technology Acceptance Model (TAM); user behavior; perceived usefulness; trust; intention; actual use; Structural Equation Modeling (SEM)

I. Introduction

In the past decade, social media has reached their popularity rivaling search engines as the most visited sites [1]. Among the many social media sites that have emerged, Facebook is the most popular social media site in the world, although many new social media sites arise [2]. Social media nowadays has over 70 million active users who log in daily to check their account [3]. Because of its popularity, some people even said if someone hasn't connected with social media, then he doesn't exist [4].

In Indonesia, there are around 63 million social media users who access their account via mobile phone, and therefore Indonesia is the third largest social media user's population in the world after United States and India [5]. Emarketer also stated that 80,9% of internet users on mobile phone aged range from 16 to19 years old in Indonesia always access their social media account every week [6].

This study using TAM (Technology Acceptance Model) Social Media developed by [7] investigate user's intentions to use social media and their behavior. TAM introduced by [8] and only assumed that the information system was used only as an organization's arrangement to improve the efficiency of its employees. TAM excludes the fact that system information can be used outside of organizational settings that can also be used by individual users. If TAM used by an individual user then an "entertainment" factors can be added, and known as Perceived Playfulnes (PP) factor [7]. Furthermore, TAM does not discuss the role of other users in influencing individual attitudes toward social media. It is considered problematic because many psychological researches proves that individual behavior can be influenced by the behavior of others arround them. Therefore, the mass number of users in social media can be an important factorknown as Critical Mass (CM) [7].

In a previous study conducted by [9] found that young people or teens in Greece use social media because it is influenced by 'social surfing' and 'wasting time'. The results of this study found that social networking applications are successful in helping them to communicatee with others even though the person is not located near them.

This study focuses on Indonesian teens as the research object to determine what factors influencing teens' behaviour intentions toward social media usage. Data used in this research are collected through a questionnaire survey and hypotheses are analyzed with Structural Equation Modeling.

II. Model Structure and Hypothesis

A. Technology Acceptance Model (TAM)

Technology Acceptance Model (TAM) was introduced by Davis based on Theory Reasoned Action (TRA). TAM exploring the main concept of TRA that someone's intention and behavior is determined by trust. TAM explains that user acceptance on new system is affected by their behavior towards the new system and this theory is acceptable to the diversity of technology, gender and groups [10]. The model of TAM is shown in Fig. 1.

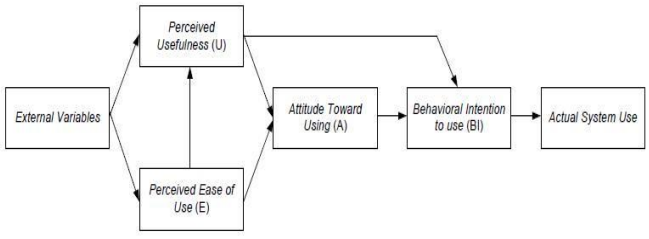


Fig. 1. Technology Acceptance Model (TAM).

According to [8], the constructs which build TAM are explained and presented in Table I.

TABLE I. DEFINITION OF EACH CONSTRUCT ON TECHNOLOGY ACCEPTANCE MODEL

Constructs	Definition
Perceived Usefulness	The degree of an individual believes that by using a specific system will improve his or her job performance.
Perceived Ease of Use	The degree of an individual expects the specific system needs less effort.
Attitude Toward Using	The degree of attitude of an individual towards a specific system.
Behavioural Intention	The intention of an individual in using a specific system.
Actual System Use	The degree of an individual's performance of a specified behaviour towards using a system.

B. Technology Acceptance Model (TAM) Social Media

Rauniar et al. [10] developed TAM Social Media from the original TAM which is proposed by Davis. The TAM Social Media is used by Rauniar et al. [7] on their research to understand user behaviour on using social media sites [7]. The model of TAM Social Media is shown in Fig. 2.

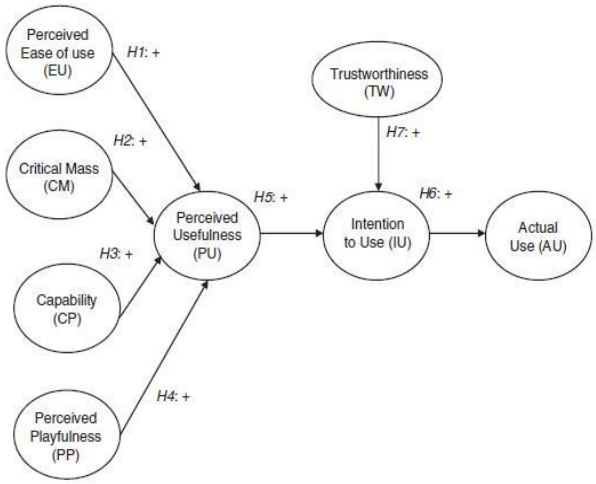


Fig. 2. Technology Acceptance Model (TAM) social media.

The construct which build TAM are explained and presented in Table II based on [7].

C. Research Model

This paper uses a model that adopts by [7] TAM Social Media to explore factors that affecting Indonesian teens' behavior intention on using social media. The difference between this study with prior study lies on research object which this study focusing on teens' behavior on using social media site in Indonesia.

D. Research Model

This paper uses a model that adopts by [7] TAM Social Media to explore factors that affecting Indonesian teens' behavior intention on using social media. The difference between this study with prior study lies on research object which this study focusing on teens' behavior on using social media site in Indonesia.

TABLE II. DEFINITION OF EACH CONSTRUCT ON TECHNOLOGY ACCEPTANCE MODEL SOCIAL MEDIA

Constructs	Definition
Perceived Ease of Use	The degree of an individual in assesing the social media based on how easy it is to use and how effective it is in helping them accomplish their social-media-related activities.
Critical Mass	The degree of the membership of society that matters most in a user's social media network.
Capability	The social media site's capability including features, applications, and other tools that help users accomplish their social-media-related activities.
Perceived Playfulness	The degree of an individual perceived that the social-media features and applications are fun and enjoyable.
Perceived Usefulness	The extent to which the social media user believes that using a particular social media site helps them accomplish their goal.
Trustworthiness	The degree of an individual feel safe and secure about their data and activities in social media
Intention to Use	The degree of a decision that an individual has to take whether to perform behaviour or not.
Actual Use	Frequency of social media used by the user.

E. Hypotheses for the construct

On the research conducted by [8], they explained that perceived ease of use have a direct effect on perceived usefulness. Study conducted by [7] also indicates that easier a website to use therefore it can enhance users' experience on doing activities on it and helping them to accomplish their social-media-related activities.

H1: "Perceived ease of use" will have positive effect on "perceived usefulness".

TAM Social Media indicates that social media sites is used by user to mainly communicate with people that they already know in real life and being part of their social life [7].

H2: "Critical mass" will have positive effect on "perceived usefulness".

TAM Social Media explain that social media sites are providing tools and features to enhance their users' experience and fulfill their social media activities so that they can get benefit of the site [7].

H3: "Capability" will have positive effect on "perceived usefulness".

A study conducted by [7] states that mixing work and play can enhance people's productivity and performance. When a person enjoys using a website, then the frequency of visiting that website will increase [7].

H4: "Perceived playfulness" will have a positive effect on "perceived usefulness".

TAM states that user intention formed by believes that using a system will increase their productivity [8]. A person's intention to use a social media site is determined by the benefits they get if they use it [7]. And intention itself is a reflection of people's decision whether they will use the particular system or not [7].

H5: "Perceived usefulness" will have a positive effect on "intention to use".

H6: "Intention to use" will have a positive effect on "actual use".

Malhotra, Kim, & Agarwal [11] stated that trust has a significant effect on intention to use. And when a person decided to be a user of a social media site, then he must trust the social media related to their personal information safety [7].

H7: "Trustworthiness" will have a positive effect on "intention to use".

III. Data Analysis

Data used in this research were collected through paper-based questionnaire survey with respondents are teens with age ranged from 15 to 18 years old. The survey took place in a public senior high-school in Indonesia. A critical sample size needed in this research according to SEM is 200 samples [12].

A. Descriptive Analysis

Pilot study was conducted before the full-scale study to measure the reliability to ensure the consistency of respondents in answering the questionnaire [13]. Reliability of each construct was measured using Cronbach's alpha coefficient. The reliability of each construct is shown in Table III.

TABLE III. Reliability of Each Construct

Factor	Cronbach Alpha
Limit Value	**>0.6**
Perceived Ease of Use (EU)	0.739
Critical Mass (CM)	0.889
Capability (CP)	0.672
Perceived Playfulness (PP)	0.894
Perceived Usefulness (PU)	0.881
Perceived Ease of Use (EU)	0.739
Intention to Use (IU)	0.798
Actual Use (AU)	0.725
Trustworthiness (TW)	0.854

TABLE IV. Characteristic of Respondent

Age	Total	%	Gender	%
15	72	24.49	Female	53
			Male	19
16	158	53.74	Female	115
			Male	43
17	57	19.39	Female	35
			Male	22
18	7	2.38	Female	4
			Male	3
Total	294	100		100

Questionnaires obtained from teens with age ranged from 15 to 18 years old who are actively using Facebook. The characteristic of respondents is shown in Table IV.

B. Missing Data and Outlier

Outlier data is data which have very different value with the average data and should be removed to avoid problems in further analysis [14]. Outlier data is examined by using Mahalanobis Distance. Data that exceeds the value of Mahalanobis Distance should be deleted. The collected data has Mahalanobis Distance value of 49.58. There are 30 data which have Mahalanobis Distance value of more than 49.58, therefore they should be eliminated.

C. Sample Adequacy Test

Kaiser-Meyer-Olkin Sample Adequacy Test was conducted to see the value of KMO to decide whether the data is appropriate for factor analysis or not [14]. The result of KMO test is 0.854 which means the data are appropriate for factor analysis [14].

D. Normality Test

Kolmogorov-Smirnov test is conducted to determine the normality of the data. The result of this test is 0.078 which means that the data sample is normally distributed.

E. Structural Equation Modeling

To analyse the data, Structural Equation Modelling (SEM) is used. There are two steps used in SEM: 1) measurement model fit to determine goodness of fit indices, and 2) structural model fit to evaluate the relationship of each construct. [15].

The result of the model goodness-of-fit indices is shown in Table V. Based on this result can be concluded that the model is fitted to the data.

TABLE V. Goodness of Fit Indices of The Research Model

Fit Index	Value	Recommended Value [16]
x^2/df	2.764	$1< x^2/df <3$
GFI	0.881	>0.8
AGFI	0.840	>0.8
RMSEA	0.078	<0.08
CFI	0.920	>0.9

TABLE VI. Hypotheses Testing Result

Hypothesis	Relationship	P	Supported
		<0.05	
H1	PU← EU	***	Yes
H2	PU← CM	0.024	Yes
H3	PU← CP	0.500	No
H4	PU← PP	***	Yes
H5	IU← PU	***	Yes
H6	AU← IU	0.035	No
H7	IU← TW	0.403	No

The next step is evaluating the relationship between construct, in this case is the hypotheses, by using structural mode fit. The hypotheses is supported if the significant value is p-value<0.05.The results of structural model fit can be seen in Table VI.

IV. Research And Discussion

This paper evaluates teens' attitude and behaviour on using social media in Indonesia. An extended TAM model is used to understand the relationship between constructs. The research data sample was collected through paper-based questionnaire survey from a public senior high-school in Indonesia. The respondents are teens with age range from 15 to18 years old.

A. H1 Result Discussion

Based on the test results from Hypothesis 1, it can be concluded that respondents assumes that social media is very easy to use and does not require much effort to use it so it

makes respondents can feel the benefits of using this social media. It shows that in this research the factor of ease of use (EU) has a significant influence to the factor of perceived usefulness (PU). Therefore, in this study Hypothesis 1 is accepted.

The results of this study is similar to the results of research conducted by [7] which suggests that when a person feels that in using a social media does not require much effort and easy to use then the user will tend to feel the benefits of usingsocial media.

B. H2 Result Discussion

Based on the test results from Hypothesis 2, it can be concluded that the respondents are using social media because of the influence of people around them, i.e their classmates. It shows that in this study critical mass factor (CM) has a significant influence on the factor of benefit in the use (PU). Therefore, in this study Hypothesis 2 is accepted.

The results of this study are similar to the results of research conducted by [7] which suggests that when a group of people use a new technology then the news about the features and benefits in the technology will spread to the people around them so it will attract people to use the technology.

C. H3 Result Discussion

Based on the test results from Hypothesis 3, it can be concluded that respondents though that social media is usefull though it does not provide clear instructions or explanation about its features, i.e how to post, how to upload/download pictures or videos. It shows that in this research, factor capability (CP) has no significant influence on perceived usefulness (PU) factor. Therefore, in this study Hypothesis 9 is rejected.

The results of this study are similar to the results of research conducted by [7] which suggests that the capability is one of the factors that influence users to use social media.

D. H4 Result Discussion

Based on the test results from Hypothesis 4, it can be concluded that respondents thought that features and applications offered by social media are delightful, exciting thrilling, and fun so that respondent thought it benefits them. It shows that in this research perceived playfulness (PP) have significant influence to perceived usefulness (PU). Therefore, in this study Hypothesis 4 is accepted.

The results of this study is similar to the results of research conducted by [7] which suggests that when a user feels that the application used provide fun for them then users can feel the benefits of the application.

E. H5 Result Discussion

Based on the test results from Hypothesis 5, it can be concluded that respondents can feel the benefits obtained from the use of social media so that respondents have an interest to continue using the social media. It shows that in this research; perceive usefulness (PU) has a significant influence towards intention to use (IU). Therefore, in this study Hypothesis 5 is accepted.

The results of this study are similar to the results of research conducted by [7] which suggests that when a user can feel the benefits of using an application then the user will tend to continue using the application.

F. H6 Result Discussion

Based on the test results from Hypothesis 6, it can be concluded that respondents think that although they intend to use social media this does not affect the frequency of using social media.In this research most of respondents' access thier social media couple times in a week and spending 2 hours on average. It shows that in this research intention to use (IU) has no significant influence towards actual use (AU). Therefore, in this study Hypothesis 6 is rejected.

The results of this study are similar to the results of a study conducted by [17] which suggests that measuring the intentions of the use of an application by a user of an application is relative and can't accurately represent the actual use.

G. H7 Result Discussion

Based on the test results from Hypothesis 7, it can be concluded that respondents thought that the social media does not guarantee the security of their profile and their post in social media therefore it affects their intention to use the social media.It shows that in this research, respondents' trust (TW) has no significant influence towards intention to use (IU). Therefore, in this study Hypothesis 7 is rejected.

The results of this study are similar to the results of a study conducted by [18] who suggested that users have no intention to disclose their personal information to an application unless they trust the application and know the risks that may occur.

V. CONCLUSION

Based on the results, this study provides empirical evidence of Indonesian teens' behaviour intention in using social media. Factors that influence them in using social media is Perceived Usefullness (PU) since they thought the features of social media benefits them in accomplishing social-media-related activities. This is due to the features of social media are easy to use (EU) and enjoyable (PP), eventhough there are no clear instructions or explanation about its features (CP). Moreover, the respondents use the social media since their fellows are also using social media therefore it can benefits them in comunicating with their friend (CM). However, the respondents thought that social media can not be trusted (TW) since it does not guarantee the security of their profile and post. Therefore, this factor does not affect the intention to use social media.

REFERENCES

[1] Steinfield, C., Ellison, N., Lampe, C. and Vitak, J., 2012. Online Social Network Sites and the Concept of Social Capital. Frontiers in new media reasearch, 15, p.115-131

[2] Emarketer. 2016a. Facebook Remains the Largest Social Network in Most Major Markets. Emarketer

[3] Lewis, K., Kaufman, J. and Christakis, N. 2008. The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. Journal of Computer-Mediated Communication, 14(1), p.79-100

[4] Svensson, A.E.C. 2011. Facebook – the Social Newspaper That Never Sleeps-A study of Facebook eWOM's persuasiveness on the receivers,

[5] Emarketer. 2015. Facebook Users in Indonesia Have Highest Mobile Usage Rate Worldwide. Emarketer.

[6] Emarketer. 2016b. In Indonesia, Facebook Remains the Most Popular Site. Emarketer.

[7] Rauniar, R., Rawski, G., Yang, J. And Johnson, B. 2014. Technology acceptance model (TAM) and social media usage: an empirical study onFacebook. Journal of Enterprise Information Management, 27(1), p.6-30

[8] Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. 1989. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models.Management Science, 35(8), p.982-1003. U.S.A: The Institute of Management Sciences.

[9] Giannakoss, M. N., Pateli, A.G., and Chorianopoulos, K. 2013 Investigating Facebook's acceptance and satisfaction: a study in the Greek university community. International Journal Social and Humanistic Computing, Vol. 2, Nos. 1/2, 2013

[10] Wu, Mei-Ying., Chou, Han-Ping., Weng, Yung-Chien. dan Huang, Yen-Han. 2011. TAM2-based Study of Website User Behavior—Using Web 2.0 Websites as an Example. WSEAS Transactions on Business and Economics, 4(8), p.133-151,

[11] Malhotra, N.K., Kim, S.S. and Agarwal, J. 2004. Internet users's information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information systems research, 15(4), p.336-355

[12] Hoe, S.L. 2008. Issues and procedures in adopting structural equation modeling technique.Journal of applied quantitative methods, 3(1), p.76-83

[13] Kusyanti, A., Haq,. 2016. "How do I look": Self-disclosure of Instagram users in Indonesia. Journal of Education and Social Sciences, Vol. 5, issue 2, (October) ISSN 2289-1552

[14] Field, A., 2009. Discovering statistics using spss.3rd ed. [e-book].Sage Publications.

[15] Chandio, F.H., 2011. Studying acceotance of online bankin information system: a structural equation model. S3. Brunel Bussiness School, Brunel University London.

[16] Oruç, Ö.E. and Tatar, Ç. 2017. An investigation of factors that affect internet banking usage based on structural equation modelling. Computers

[17] Tao, D. 2009. Intention to use and actual use of electronic information resources: further exploring technology acceptance model (TAM). AMIAAnnual Symposium Proceedings, 9, p.629

[18] Wu, J. and Liu, D. 2007. The effects of trust and enjoyment on intention to play online games.Journal of electronic commerce research, 8(2), p.128

Securely Eradicating Cellular Dependency for E-Banking Applications

Bisma Rasool Pampori, Tehseen Mehraj
Department of Info. Tech.
Central University of Kashmir
Srinagar, Kashmir, India

Asifa Mehraj Baba
Department of ECE
School of Technology
IUST, Awantipora, Kashmir

Burhan Ul Islam Khan
Department of CSE
School of Technology
IUST, Awantipora, Kashmir

Zahoor Ahmad Najar
Department of Info. Tech.
Central University of Kashmir
Srinagar, Kashmir, India

***Abstract*—Numerous applications are available on the Internet for the exchange of personal information and money. All these applications need to authenticate the users to confirm their legitimacy. Currently, the most commonly employed credentials include static passwords. But people tend to behave carelessly in choosing their passwords to avoid the burden of memorizing complex passwords. Such frail password habits are a severe threat to the various services available online especially electronic banking or e-banking. For eradicating the necessity of creating and managing passwords, a variety of solutions are prevalent, the traditional ones being the usage of One-Time-Password (OTP) that refers to a single session/transaction password. However, the majority of the OTP-based security solutions fail to satisfy the usability or scalability requirements and are quite vulnerable owing to their reliance on multiple communication channels. In this paper, the most reliable and adoptable solution which provides better security in online banking transactions is proposed. This is an initiative to eradicate the dependency on Global System for Mobile communication (GSM) that is the most popular means of sending the One-Time-Passwords to the users availing e-banking facilities.**

Keywords*—*E-banking; one time password (OTP); global system for mobile communication (GSM); authentication

I. Introduction

The Internet has proved to be the fastest emerging medium in the present day for delivering services in both retail and corporate banking sectors [1]. Electronic banking (E-Banking) is one of the significant developments that have revolutionised the banking sector. Electronic banking is defined as "the technology which allows customers to access the banking services electronically like to pay bills, transfer funds, and view the accounts details and advices" [2]. It provides automatic delivery of conventional and new banking services/products directly to users via interactive and electronic communication channels. Electronic or online banking comprises of the systems that facilitate the customers of financial institutions and individuals for accessing their accounts, performing business transactions, or acquiring information on commercial products/services by using a private or public network. E-banking is a standard term that encompasses mobile banking, internet banking, telephone banking, etc. Furthermore, the evolvement of E-banking services using numerous electronic communication channels such as cell phone, telephone, internet, etc. has presented a convenient and feasible way of providing banking services [3].

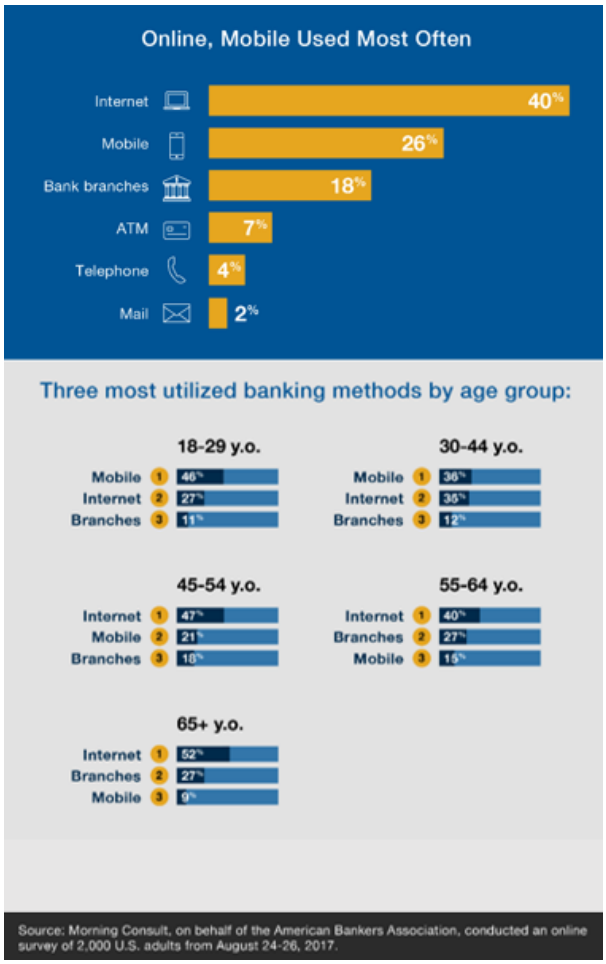


Fig. 1. Preferred banking method (*Adopted from source*[1]).

The emergence of e-banking is not just coerced by the banks' need to minimise costs, but it is an actualisation of customers' demand who crave for online access to their bank services anytime and anywhere [1]. Numerous reasons make the importance of E-banking evident. First and foremost, it provides unparalleled convenience to its customers by accrediting 24×7 access to a wide variety of services. Moreover, it presents a cost-efficient alternative to branch and telephone banking owing to the comparatively low maintenance and capital cost, along with its ability to offer entirely automated processing of transactions [1].

With the analysis, it has been observed that most of the users all over the world prefer to use net-banking. Fig. 1 illustrates the use of net-banking which gives the fair idea of the popularity of Internet banking among various age groups in the United States.

As per survey [1], 40% of nationals in America prefer managing their bank accounts online as compared to other methods. About 26% customers perform their banking transactions using mobile phones whereas only 18% visit the bank branches on their own.

Even in Canada, more than 68% of the citizens prefer mobile and online banking. Although online banking is popular, mobile banking is gaining popularity. About 51 percent of Canadian nationals make use of the Internet as the primary banking method which is 4 percent less than the margin in 2014. This is due to the rise in mobile banking among them with 17 percent of the individuals utilizing mobile phones as their principal banking mode which is 9 percent higher than that in 2014[2].

In India, online banking has brought about a significant change in the banking industry with the introduction of National Electronic Funds Transfer (NEFT), Real Time Gross Settlement (RTGS), Immediate Payment Service (IMPS), mobile banking, credit cards, prepaid cards, debit cards, etc. [4]. There has been a drastic jump in the values of transactions conducted on mobile wallets and PPI (Prepaid Payment Instruments) cards from Rs. 82 billion and Rs. 105 billion respectively in 2014-2015 to Rs. 532 billion and Rs. 277 billion respectively in 2016-2017[3].

It is thus evident that banking industry has made unparalleled success in delivering its services ranging from telephone banking to computer-based banking applications progressing to Automated Teller Machine (ATM) and Internet banking. There has been a dramatic development in the field of mobile communication technologies with the introduction of WAP, 3G technology, SMS, etc. which has led to the broad expansion of mobile telecommunication. The latest developments in technology in conjunction with the increased rate of acceptance of mobile phones have persuaded the business enterprises to launch and develop a variety of services via mobile phones [5]. On the same lines, the banking sector resolved to build banking applications on mobile devices in order to provide mobility in addition to the inherent features of mobile technology to their customers which was absent in the conventional electronic banking methods. Mobile banking – a subset of E-banking – provides personalisation, dissemination, flexibility and ubiquity which promises unmatched productivity, profitability and market potential to businesses. Moreover, the users find mobile banking more convenient and feasible on the ground of security and adoptability as compared to the traditionally employed approaches [6]. Furthermore, in India, the Governor of Reserve Bank of India (RBI) is sturdily in favour of making use of mobile phones as devices for carrying out bank transactions. Keeping in view the drastic fall in tariffs associated with mobile phones coupled with the flawless connectivity of mobile and fixed lines, mobile banking has evolved as a cost-efficient banking channel today [2].

Despite the numerous facilities offered by the internet banking, it is not devoid of shortcomings. The existing internet banking system has been found to be open to the danger of hacking and a range of other attacks [7], [8]. Although the various attack methods are diverse, the primary objective of the intruders is to acquire confidential information of users, e.g., social security numbers, usernames and passwords, credit card numbers, etc. All these are static credentials, i.e. these do not change. Primarily, the users tend to select passwords which are unsafe but easy to memorise. As a result, these are vulnerable to a variety of attacks such as dictionary attacks. With the current availability of numerous powerful and sophisticated hacker tools as well as keylogger software, it has become an easy task for the intruder to retrieve the password. In the existing system, neither private nor public network is fully secure. There have been numerous events in the past where reputable enterprise agencies such as Walmart, eBay, World Bank, ICICI Bank, etc. have been hacked which resulted in massive loss of both confidential information and property [9]. Thus, there is an urge to adopt a strategic and long-term approach which can meet the much-desired security concerns of the users.

A. OTP Utilization in Banking

With the rapid expansion in computerisation of small and large businesses, authentication is the need of the hour to provide security against the emerging threats. Authentication assures secure online transactions and facilitates the development of trust among the trading partners dealing online [9]. Authentication is a procedure by which an entity establishes a claimed property to another entity. In other words, it deals with testing or verifying who an information resource or person claims to be, which adequately convinces the authenticator that the claimed identity is valid. It is followed by authorisation which evaluates whether the authenticated entity has the privilege to access the resources [10]. In general, authorisation can be referred to as the permission to access and the determination of privileges that an entity has on a system and what the entity is allowed to do with the resources [11].

[1] "Survey: Online, Mobile are Most Popular Banking Channels", Aba.com, 2017. [Online]. Available: https://www.aba.com/Press/Pages/092117ConsumerSurveyBankingPreferences.aspx. [Accessed: 15- Jan- 2018].

[2] "Issue Brief: How Canadians Bank", *Cba.ca*, 2017. [Online]. Available: https://www.cba.ca/technology-and-banking. [Accessed: 15- Jan- 2018].

[3] "Digital revolution in the Indian banking sector | Forbes India", *Forbes India*, 2017. [Online]. Available: http://www.forbesindia.com/article/weschool/digital-revolution-in-the-indian-banking-sector/47811/1. [Accessed: 15- Jan- 2018].

Authentication can be achieved in one of the three ways [12], [13]:

a) Something the user only knows, e.g., a private key, a password, a Personal Identification Number (PIN) or a secret key.

b) Something possessed by the user or a physical item exclusively owned by the user, e.g. a driver's license, a credit card, a passport, a smart card or an identification card.

c) Something the user is, e.g. facial characteristics, retinal pattern, or fingerprints.

d) Something the user produces, e.g. voice pattern, handwriting characteristics and current signature.

It has been observed that the authentication based only on the first type, i.e. something a user knows (static password) is vulnerable to phishing attacks. Online banking applications require rigorous user authentication. Most often, user authentication is realized by employing two-factor authentication (2FA) technique, which is a two-level security approach -- based on something the user only knows, viz. a static password, and something the user possesses, viz. a One Time Password (OTP) [14]. Biometrics is not used for authentication purposes in banking for reasons like cost, reliability, privacy, and complexity. It has been found that the most commonly used authentication technique is based on the usage of passwords as they are easy to implement, convenient, inexpensive and highly adopted by masses but they are relatively simple to be stolen or broken [11], [9]. Consequently, a reliable and hard-to-crack authentication is needed which can be provided by one-time-password system [15]. One-time-passwords, also referred to as single-use passwords, are dynamic, i.e. these are changed every time they are used. OTPs are considered to be the most reliable variant of passwords and provide an effective solution for security [10]. In internet banking, one-time passwords play a prominent role in providing authentication to enhance the security. OTPs impart an additional layer of security above the conventional static passwords which are vulnerable to replay attacks. OTPs offer immunity against replay attacks as the unique password generated once can never be repeated which implies that even if the attacker procures the OTP, it will be futile [16]. In banks, OTPs are used in conjugation with static passwords which offer a strong defence against numerous online attacks [9]. Regardless of this win-win proposal, the distribution of OTPs to the concerned user is a significant issue. There are certain pitfalls in the existing method of delivery of OTPs particularly in regions like Kashmir where most of the time SMS service is banned. Our proposed work is an initiative to eradicate those issues by putting forward an alternate solution for distribution of OTP to the user.

This paper is comprised of five sections. Section 2 discusses the variety of authentication/authorization schemes currently employed followed by the architecture of the proposed system in Section 3. Section 4 covers the illustration of the implementation details of the authentication mechanism proposed. Section 5 provides the analysis of results, and finally, the concluding remarks along with future scope have been elucidated in Section 6.

II. Evaluation of Existing Security Solutions

There has been a considerable change in the security means of online banking since the short span it has been in use. Particularly, the authentications schemes utilized in previous systems have been found to be susceptible to numerous attacks like man-in-the-browser, phishing, password sniffing, malware, etc. This section investigates the prevailing authentication and authorization schemes concerning the usability and security of online transactions.

Authors in [17] have set forth a novel two-factor authentication mechanism based on OTP using mobile phones. Two nested hash functions, i.e., SHA-1 and MD-5 have been employed in this system to provide online authentication. In this system, SMS based OTPs are eradicated as OTP is generated on a user mobile phone thus eliminating the use of separate tokens providing usability. However, the proposed system fails to give efficient storage and user ergonomic authentication as the generated OTP is 128-bit data which is difficult to be entered manually by the user.

A scheme has been put forward in [18] referred to as 'Infinite Length Hash Chains' to enhance the extensibility and efficiency of the traditional idea of hash chaining by employing public key techniques. In this scheme, the hash chain length has been extended infinitely thus avoiding the need to reboot the system. A one-way hash function based on a public key algorithm provides an infinite source that forms the core of OTP production.

In [9], authors have performed a detailed study of various authorisation and authentication systems proposed beforehand, with the goal to highlight the numerous issues still prevailing in the existing field. The paper embroils schemes offered to alleviate the number of security threats present in access management both in private as well as public networks. The review has been carried out by categorising numerous authentication techniques and approaches into non-OTP and OTP based schemes. Although OTP based methodologies were found to be more robust in providing the required security than non-OTP-based schemes, all the techniques surveyed in this paper have been found to have some loopholes ranging from the type of communication medium used, to the computational complexity of the technique, to their technical adoptability. Some methods have been found to be costly for the service provider along with the associated customer, hence proving deficient in user ergonomics and obstructing their technical adaptability in general. Many issues have been highlighted resulting from techniques which involve dependency on outside parties, for instance, GSM or certain authorised entities either for OTP distribution or communication with the user, all of them are found to face obstructions. Further, some techniques discussed in this paper have been found to make use of multiple communication channels which burdens the customers with substantial service charges and at the same time have proven to be infeasible. The paper concludes that a major portion of the considered authentication mechanisms was found to be weak in providing security owing to various password generation schemes like MD-5, SHA-1, and AES, etc. employed by the techniques, which have been found to be unsuccessful both in providing strong security and in

maintaining their continual existence as there exists an exponential growth in the network security issues.

A system has been presented in [19] that makes use of OTP as the One Time Key to perform AES encryption along with Quadratic Residue Cipher (QRC). This leads to spoofing of the source IP addresses multiple times, therefore, increasing the complexity required to find the sequence of packets and IP addresses. The OTP and the random numbers generated are transmitted to the user via SMS which poses several restrictions to the client in terms of cost, security and transmission.

An authentication technique based on mobile/web has been put forward in [20] to enhance multi-factor authentication which is used for verification of the user as well as the current transaction. The proposed system makes use of PingPong 128 stream cipher for the generation of OTP keys which are encrypted using Advanced Encryption Scheme (AES). The dissemination of ICs to the end user's device is done by an authorised person via the web browser, or Bluetooth enabled device.

Authors in [21] have presented a novel authentication scheme which generates OTP by using time as well as location information of mobile device. The proposed scheme enhances security by restricting the validity of OTP generated to a confined geometrical location and a definite time-period. The user is facilitated with transparent authentication provided the user moves consistently, hence enabling the user to avoid entering credentials manually every time. This system provides user authentication for accessing crucial online services for instance e-banking transactions and e-commerce; resistance against numerous attacks, e.g., replay, man-in-the-middle (MIM), brute force, eavesdropping, user impersonation and dictionary attacks. However, clock synchronization between server and mobile device is difficult to acquire as mobile phones are likely to go out of network coverage resulting in failure in synchronization.

In [22], an OTP based authentication mechanism was introduced that eliminates the issue of counter de-synchronisation by implementing the one-way hash function and symmetric encryption algorithm. Further, the proposed system has successfully been able to minimise several attacks such as replay, Denial of Service (DoS) and guessing attacks. Symmetric encryption technique, i.e. Advanced Encryption Standard (AES) in conjunction with one-way hash function (MD5) has been employed by the system. The scheme offers easy implementation in current enterprise applications, and server-side security requirements are also minimised.

An authentication mechanism designed for home networks has been introduced in [23]. The framework employs OTP-based smart cards which offer secure authentication. Mutual authentication is achieved using three-way challenge-response handshake. The system makes home networks resilient against various passive attacks as well as phishing attacks. System security is centred on the one-way characteristic of the hash function and a nonce that has been used to thwart the problem of time synchronisation. However, this system fails to protect against numerous active attacks.

A technique was proposed in [24] that extend the password generator to improve OTP-manageability. The proposed system presents better performance in terms of computation cost related to the generation and verification of password in addition to the bandwidth associated with transmission. A Manageable One Time Password (M-OTP) module has been used in this scheme that may be some firmware module or any software program on the user device. Advanced Encryption Standard (AES) algorithm used in this system provides the necessary encryption along with one-way functionality.

A One-Time-Password (OTP) MIDlet has been put forward by authors in [25] that work on the user's mobile phone to provide integrated authentication for various critical internet services. The proposed system uses two different channels, i.e., GSM and internet to send and receive authentication messages. A challenge-response approach is employed by the system for OTP generation. This model uses Java MIDlet embedded on the Java-enabled mobile phone on the user side and an applet on the user's computer to transmit OTP to servlet acting as an authentication server. The fundamental characteristic of the scheme is that it involves closed loop formation among system components. HTTPS connection is used by the authentication server to connect applet and SMS facility provided by GSM to connect the mobile phone. Moreover, the mobile phone needs to possess Bluetooth facility to enter the credentials else the user has to enter those manually.

In [26], authors have put together the advantages of software as well as hardware tokens by embedding both on mobile phones with hardware Mobile Trusted Module (MTM) enabled in them. This technique offers better usability along with security and scalability options for OTP generation based on mobile phones embedding trusted computing technology. The system presented uses SHA-1 for OTP generation. Two different channels, i.e. GSM network and the internet, used in this system, result in a considerably complex system making it difficult for the intruder to carry out a man-in-the-middle attack.

The limitations as well as the main contributions of the researches conducted by various authors so far in the field of authentication and authorization have been presented in Table I.

As can be observed from the findings in Table I, the prevailing authentication solutions face potential threats and may not be able to support in the long run with the advancements in attacks conducted by intruders and introduction of quantum computing. Other significant issue includes the distribution of OTP via GSM networks which face issues like delay and cost associated with SMS, roaming constraints, weak security, government regulatory restrictions, etc. Also, the existing authentication solutions fail in terms of providing scalability, flexibility, cost-effectiveness, reliability and technical adoptability. To tackle these challenges, the alternative solution proposed has been given in the following section:

TABLE I. SUMMARY OF FINDINGS

AUTHOR	CONTRIBUTION	RESULT OBTAINED	LIMITATIONS
(Eldefrawy et al., 2011) [17]	Designed a novel two-factor authentication mechanism using mobile phones.	• Less computation time offered since no dependence on public key techniques.	• 128-bit OTP generated is neither storage efficient nor user-friendly. • Reported attacks on SHA-1 make it impractical to be used for security purposes.
(Bicakci and Baykal, 2002) [18]	Presented a flexible Infinite Length Hash Chain based on a public-key algorithm for OTP generation.	• The user is granted the freedom to move in any direction in the chain further offering no restriction on the length of the chain • Complexity because of restarting the system and communication overhead is averted.	• Low computation devices (e.g. mobiles phones) fail to implement this technique due to higher complexity offered by system owing to use of public key operations.
(Long and Blumenthal, 2007) [24]	Proposed an efficient Manageable One Time Password (M-OTP) via the extension of the password generator.	• Successfully thwarts offline dictionary attacks. • Enhances the manageability of OTP for consumer applications thereby resulting in better user convenience.	• Usage of MD-5 and AES-128 is not apt which have eventually proven to be compromised.
(Hallsteinsen and Jorstad, 2007) [25]	Designed a mobile phone-based OTP-MIDlet for unified authentication.	• Lowers management cost of hardware tokens than previous OTP-based approaches. • Provides resilience against a range of attacks like hacking, eavesdropping, man-in-the-middle, replay, guessing and sniffing attacks.	• Dependence of the client terminal on Bluetooth facility hampers user-friendliness. • Exchange of keys and user communication is done by authentication server making use of GSM network.
(Davaanaym et al., 2009) [20]	Designed an OTP authentication mechanism based on PingPong128 stream cipher	• Execution can be done with no extra expense charged from the user, thus easily implementable. • Provides immunity against attacks based on key-stream properties.	• Usage of AES as the encryption algorithm for the OTP generated. • Employment of two channels of communication viz. GSM and TCP/IP
(Jeong et al., 2008) [23]	Presented smart card-based OTP authentication scheme for the home network.	• Eliminates serious time synchronisation problem by making use of timestamp. • Offers immunity against phishing attack and a range of passive attacks like Man-In-The-Middle, passive eavesdropping, stolen verifier, Denial of Service, and replay attacks. • A decrease in computational overhead along with communication cost. • The home user has free will to choose password thus improving convenience.	• The system fails to provide security against active attacks. • Smart card – a hardware token, is used for authentication which hampers user convenience.
(Liao et al., 2009) [22]	Proposed an authentication scheme that eradicates counter de-synchronisation.	• Successfully eliminates counter de-synchronization problem. • Effectively thwarts replay as well as guessing attacks. • Easily implementable in enterprise applications and offers reduced server security requirements.	• AES-128 in conjunction with MD-5 has been employed which have proven to be compromised previously.
(Alzomai and Josang, 2010) [26]	Presented usage of mobile phone as scalable, trusted computing-based OTP device	• OTP generation providing usability and scalability. • Reduced man in the middle attacks.	• SHA-1 acts as an OTP generator that has been compromised. • An attacker can masquerade the server resulting in the generation of useless OTPs at the user side. • User terminal to connect service provider and mobile phone used to generate OTP are separate which restricts usability. • Not widely adoptable.
(Srivastava et al., 2011) [19]	Proposed an improved knock sequence algorithm using AES	• Avoids attacks like denial of service (DoS), man-in-the-middle (MIM), etc. • Detection and interpretation of consecutive knock sequences is difficult • Disclosure of data is prevented via multi-packet authentication mechanism • Out of sequence delivery of packets is eradicated.	• Transmission of OTP via GSM network
(Hsieh and Leu, 2011) [21]	Presented an authentication framework based on two parameters, i.e. location and time of the mobile phone.	• Provided secure authentication of user to access crucial internet services • Resilience against a range of attacks, e.g. brute force, eavesdropping, user impersonation, and replay attacks. • Offers transparent user authentication.	• The requirement of mobile phones with GPS facility. • Mobile phone and server should be clock synchronised.

(Moon et al., 2012) [27]	Three solutions for fuzzy fingerprint vault were proposed to enhance the security of biometric information.	• Biometric information that is highly reliable can't be lost, duplicated, changed or speculated. • Since it works with unordered sets, it is apt for crypto-biometric frameworks. • Since polynomial reconstruction is quite infeasible, thus guaranteeing its security. • Immune to correlation assault. • Improved execution of GAR is brought about without influencing FAR.	• Fuzzy vault cannot be repudiated, once it has been compromised. • Some modern assaults can compromise biometric data. • Since this framework can't scale well to a huge service pool, so it is more relevant to restricted applications.
(Ma et al., 2013) [28]	An authentication mechanism based on identity was designed which puts the speech features into use.	• Better client ergonomics is provided.	• It can be rendered useless if an attacker gets access to authentic user's pre-recorded voice.
(Castiglione et al., 2014) [29]	An effective end-to-end OTP validation mechanism which involves AKE convention and the keyed HMAC.	• Since it is quite simple and involves less computational overhead, it can work independently along these lines. • In addition to efficiency, it also provides transparency. • Immune to extensive variety of assaults, e.g., offline dictionary attack, password guessing attack, replay attack, brute force attack, stolen verifier attack, denial of service attack and eavesdropping • Its adoptability is quite suitable.	• It can't be utilized when the number of iterations surpasses the length of the commonly settled upon Master Key. • Security of the mechanism depends entirely upon storage and secure handling of the Master Key.
(Avhad and Satyanarayana, 2014) [30]	A password/user ID, single biometric and OTP based authentication scheme is proposed.	• Improved assurance information at reduced cost. • Preserves privacy of client in distributed systems. • Easily implementable configuration.	• Dependence on GSM network for transmission of OTP. • Vulnerable to imitation and MIM attacks.
(Oruh, 2014) [31]	Presented a system for ATMs that integrates biometric authentication with user PIN and smartcard.	• Relatively more secure, reliable and accurate user authentication technique for ATMs.	• No complete exclusion of false matches and non-matches in the biometric feature (i.e., fingerprint).
(Boonkrong, 2017) [32]	Designed and developed a multi-factor authentication protocol, starting from a registration system, which generates authentication factors, to an actual authentication mechanism.	• Requires only three messages between the bank's server and the client to complete the authentication process. • Utilizes a username, a password, number of iterations, a public key, a private key, a symmetric key, a digital signature and an IP address as factors of authentication, all being unique to each user.	• Vulnerable to the threat of password reuse since password is not changed every time the user logs in. • Usage of SHA1 and MD5 which have been reported to be vulnerable.
(Akinyede and Esese, 2017) [33]	Proposed a model that performs hashing using Salted SHA 512, authentication using OTPs, encryption and decryption using AES.	• More advantageous and comfortable model to help conquer challenges to online banking. • Validity and dependability of the model was ensured using a password recovery tool.	• SHA-2 has been employed which makes the system susceptible to length extension attacks. • Usage of AES for encryption/decryption.

III. PROPOSED SECURITY ARCHITECTURE

A. Problem Formulation

Short Message Service (SMS) is the most common approach used by the traditional systems to distribute the OTPs to the users for performing online transactions [11]. The 2FA technique based on SMS is the most widely used approach in India, particularly in banks because of reasons like end-device liability and zero logistics charges [34]. However, there are various limitations associated with SMS-based 2FA which are discussed as follows:

1) Delay in SMS delivery:

Though SMS transmission occurs in no time in normal circumstances, there may be a delay associated with the transmission of SMS in cases of network congestion. Since SMSs are not transmitted point to point, instead these are queued and sent to the desired network cell where they are again queued before being sent to the end user. It is this queuing which accounts for SMS delays. This delay may lead to session time-out which may last for few minutes thus impeding the transaction/authentication to happen [35].

2) Low security:

Various possible security concerns are associated with SMS-based OTP. The SMS encryption employed in India is generally basic in nature. At the outset, in between user and service provider, there are various mobile phone operators who happen to be part of the trust chain and as such, have to be trusted. A major security breach is possible if and when a gateway is attacked in the instance of roaming. Moreover, SMS encryption is easily decrypted by an intruder. SIM swap attack is another emerging danger faced by SMS-based OTPs [11].

3) Service unavailability/Coverage areas:

SMS-based 2FA OTPs result in inconvenience to the users in the reception of incoming SMSs when they move outside the network coverage [35].

4) Roaming Restrictions:

It is one of the significant shortcomings of SMS-based systems that a user travelling abroad faces restrictions and may

be deprived of availing SMS facility. Even if the services are available, users have to pay high roaming costs resulting in constraints on SMS services. When the roaming services are turned off, then banks will fail in SMS-OTP distribution to users, which leads to the conclusion that users are stopped from continuing any further processes. Roaming facility for global GSM mobile phones is not permitted in Kashmir. Moreover, pre-paid GSM connections in the valley can't benefit roaming facility in the rest of India, restricting customers stationed outside the state from availing net-banking services offered by banks [17], [35].

5) Government Regulatory Regulations

In critical emergency and sensitive conditions, the government of India has to follow some set of standard rules which involves blocking bulk SMS services thus affecting facilities provided by SMS-based authentication mechanisms [16], [36]. In Kashmir, a similar situation was observed in the year 2010, when the government announced to block SMS service in Jammu & Kashmir. On June 30, 2010, SMS facility on mobile phones was snapped by the government, however, only after a short duration of 6 months, the SMS services were resumed on post-paid networks, but pre-paid network continued to face the ban. The ban on SMS services for pre-paid subscribers which formed 70% customers to telecom companies were deprived of using the SMS service, and lastly, the ban was lifted after four years on May 20, 2014[4].

Though user authentication is ensured by employing OTP when OTP generation is based on GSM, it is vulnerable to compromise [9]. Thus, a serious cost-effective solution and a secure authorisation mechanism which caters the much-needed convenience of customers without compromising on security aspect are required. To overcome all the short falls, device-details based two-way authentication mechanism needs to be developed; further improvements need to be achieved by considering SHA3, truncated SHA1 and RIPEMD-128 in place of SHA1 and MD5 and overcoming human ease of short-data entries as opposed to large length data.

B. Design Methodology

The proposed system is much more secure than the before-mentioned approaches since it provides security and supports the performance with continued existence. The proposed system is cost-effective and ensures privacy, confidentiality, non-repudiation as it employs secure hash algorithms (SHA3 and truncated SHA1) which can produce One Time Passwords (OTPs) on Android environment thereby providing password security besides improved protection.

The proposed work makes use of SHA3, truncated SHA1 and RIPEMD128 as standard algorithms for generation of One-Time-Passwords (OTPs) from an initial germ with no dependence on GSM network. The hardware profiles (IMEI, IMSI and timestamp) and software profile (index number) form the initial seed. Two random numbers 'N' and 'M' are generated which specify the number of SHA3 and RIPEMD128 iterations. A third coordinate, i.e. |N - M%N| determines the number of iterations of truncated SHA1. RIPEMD128 results in 128-bit data which is later collapsed into the 64-bit result. The 64-bit key may be converted into six words in a user readable format using FRC 1751. This system, however, does not employ any traditional encryption/decryption method. The proposed work is implemented on an Android platform using Java as a programming tool. The proposed work has been built and tested on Android which is gaining popularity among the users globally as compared to tablet PCs and other smart-phones. The Android market share is 87.7%, which is the highest than the market share of any mobile OS [37]. Thus, the proposed work is highly adoptable technically.

The principal objective of the proposed work is to design a model for OTP authentication. The whole process can efficiently be visualized with the help of modules which have been described as follows:

1) Registration Phase

The primary module in the proposed system comprises of the registration phase which is given in Fig. 2. In this phase, the user registers or enrols for the proposed authentication system. The generation of a user interface on Android is followed by the formation of unique initial seed from the hardware profiles of the user (comprising of Timestamp, IMSI and IMEI) and the software profile (consisting of the index number).

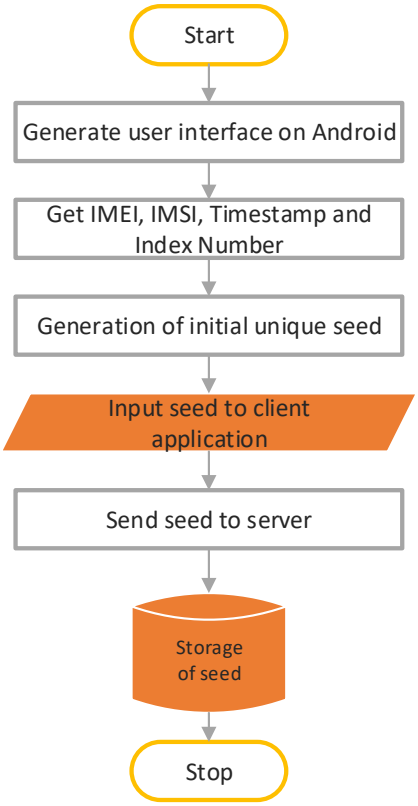


Fig. 2. Registration of user.

[4] "Jammu and Kashmir government lifts ban on SMS on pre-paid mobiles | Latest News & Updates at Daily News & Analysis", *dna*, 2014. [Online]. Available: http://www.dnaindia.com/india/report-jammu-and-kashmir-government-lifts-ban-on-sms-on-pre-paid-mobiles-1990107. [Accessed: 17-Jan- 2018].

2) OTP Generation Phase

The next phase involves the generation of OTP for authentication purposes as illustrated in Fig. 3. In this phase, a one-time-password that can be easily read by a human is generated on the user's mobile device. It begins with the user logging-in to the website of the service provider using a distinct username-password combination. Then, the server is allowed to compute the seed from the status as submitted by the user. This is followed by the transmission of a random challenge to the client by the server for computing OTP at his/her side.

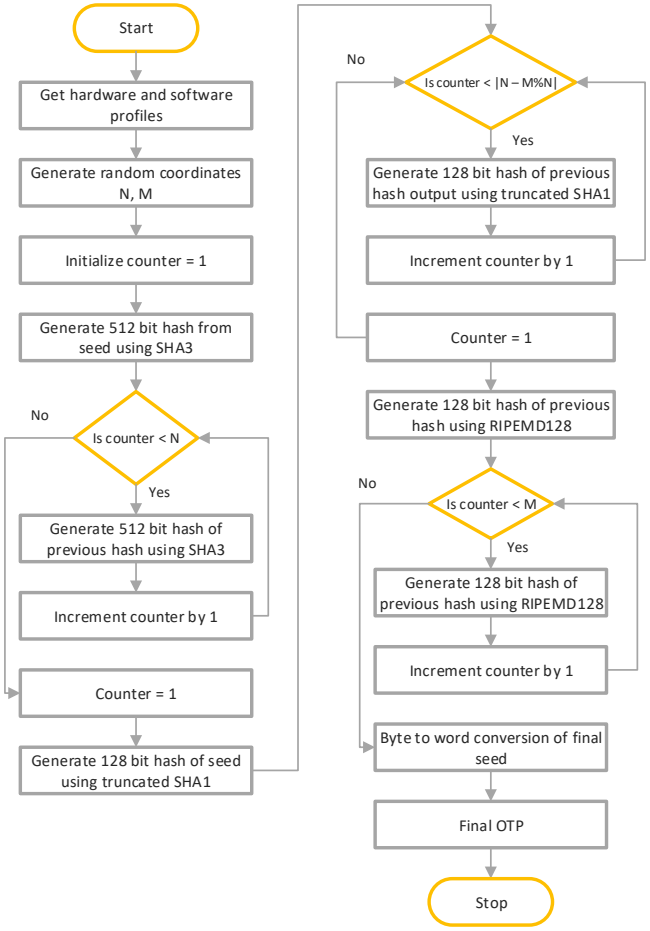


Fig. 3. OTP generation on the client side.

At the client side, this challenge is read and used for generating the final hash with the help of SHA-3. The output value of SHA-3 is input to truncated SHA1 whose output is in turn fed to RIPEMD128. Byte to word conversion is performed on the resulting output yielding a human-readable OTP.

3) OTP Authentication Phase

In this phase, the final authentication of the OTP generated is performed as shown in Fig. 4. This is done by comparing the OTP generated at the client side and that at the server side. Only when the two match, the user is provided access to the online services by the service provider.

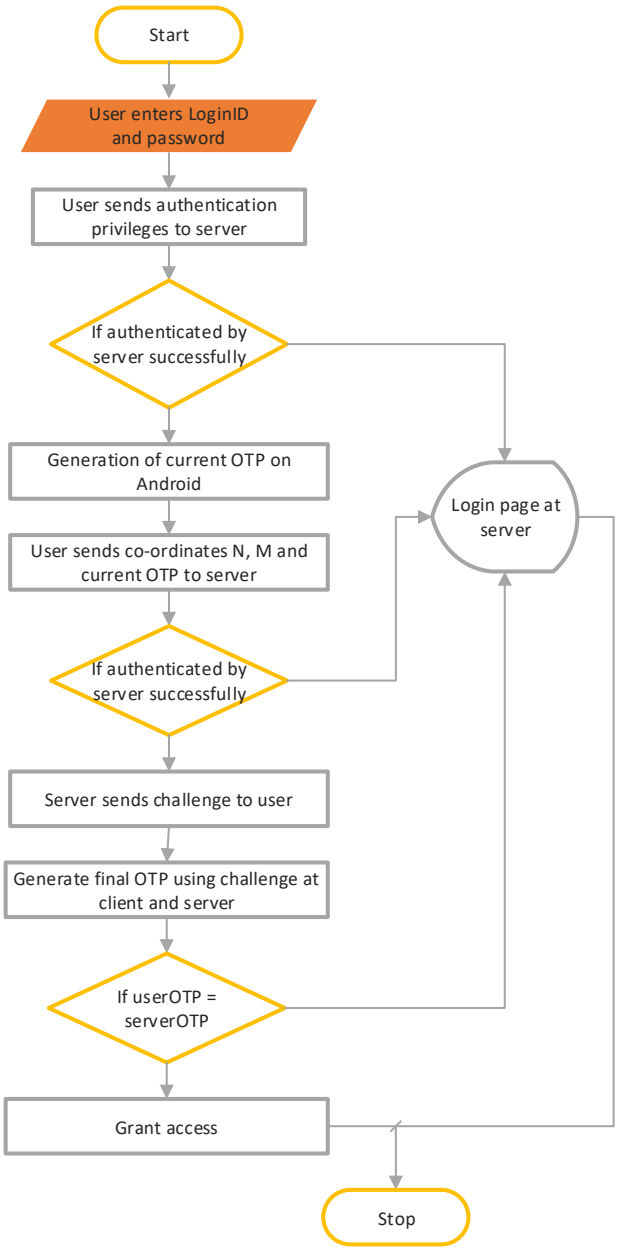


Fig. 4. Final authentication of generated OTP.

IV. Design Implementation

The implementation details have been highlighted in this section including the installations required. For writing the Android application, the requirements comprise of Java Development Kit (JDK), the Android SDK, an Editor or Eclipse (an Integrated Development Environment for Java) for making the development process easier. Besides, a plug-in is also required for Eclipse, i.e. Android Development Tools (ADT) for providing added support. The list of One-Time-Passwords has been computed before-hand to enhance the efficiency and performance. This is done using SQLyog as the database management tool owing to the ease of use and automaticity it provides.

A. Illustration of the Proposed System

The screenshots that follow illustrate the process of authentication of the user in a stepwise manner. Each of the steps that are involved has been described with the help of corresponding figures. These have been given as under:

a) Bank Employee Login:

At the server (bank) site, the employee who is privileged can log in to the system. The employees and their details, i.e. employee id and passkey are stored on the server. It is only the registered employee who can manage the customers/users who want to avail the online banking services. The login page of the employee appears as shown in Fig. 5. After logging in successfully, the employee can open an account for the user or can credit amount.

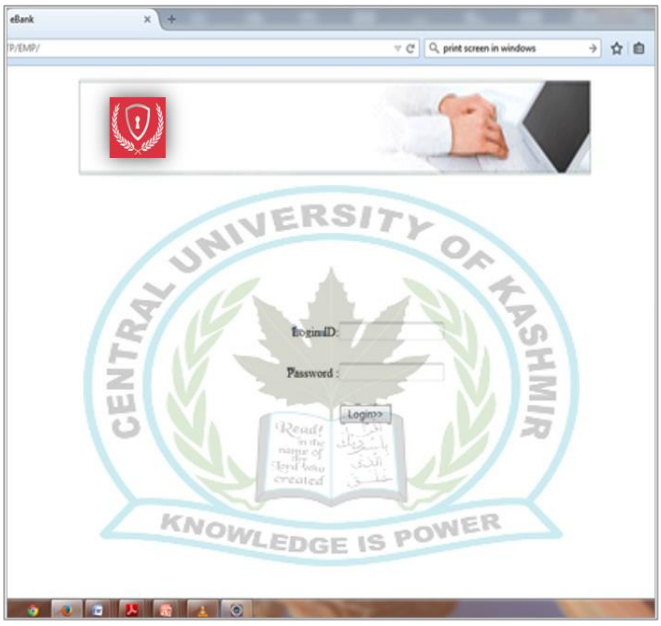


Fig. 5. Employee login page.

b) Open new account for user:

Only the registered employees can add the user's account after obtaining the hardware/software profiles and other details of the user as shown in Fig. 6. The authentication details of the user must be submitted manually by the users to the server (bank) which is considered as the most secure form of passing on the details. The hardware profiles constitute the IMSI (International Mobile Subscriber Identity), the IMEI (International Mobile Equipment Identity), the Timestamp and the software profile comprises of the Index number that uniquely identifies your application. These details are taken from the user's device and are placed under System Info as shown in Fig. 7. Even if the application is uninstalled on the device, the index number remains there with the server. It shall be flushed only if the user removes the application file (.apk) from his/her device. In that case, the user will have to request for the application from the server (bank) again and re-register his/her details manually. After registering, the user is provided with an account number to be used for accessing the online services. The details of all the users registered are stored at the server side.

Fig. 6. Screenshot showing how to open a new user account.

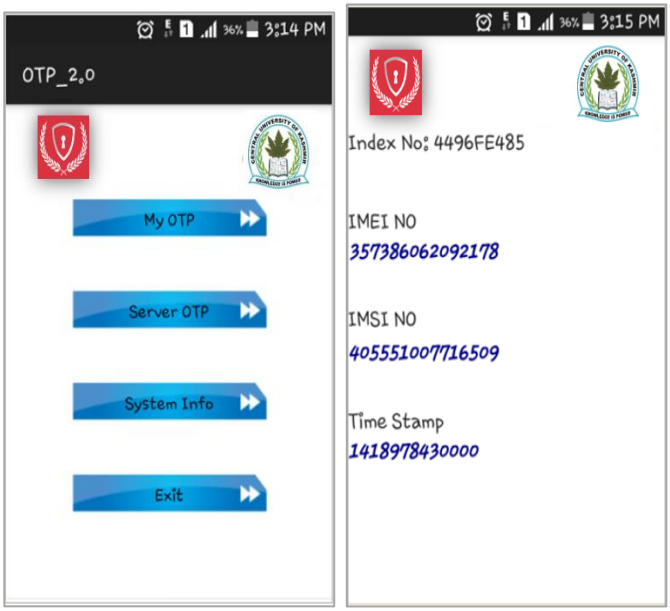


Fig. 7. Screenshots showing system info on the user device.

c) Credit amount:

Again, it is the employee who has the privilege to credit amount to any account which is shown in Fig. 8 below.

d) User login page:

When provided with an account by the server (bank), a new user submits his/her login id and password to account for the first layer of authentication. The user must activate his/her account by logging in to his/her account at first. The login page for user appears as in Fig. 9.

Fig. 8. Screenshot for crediting amount.

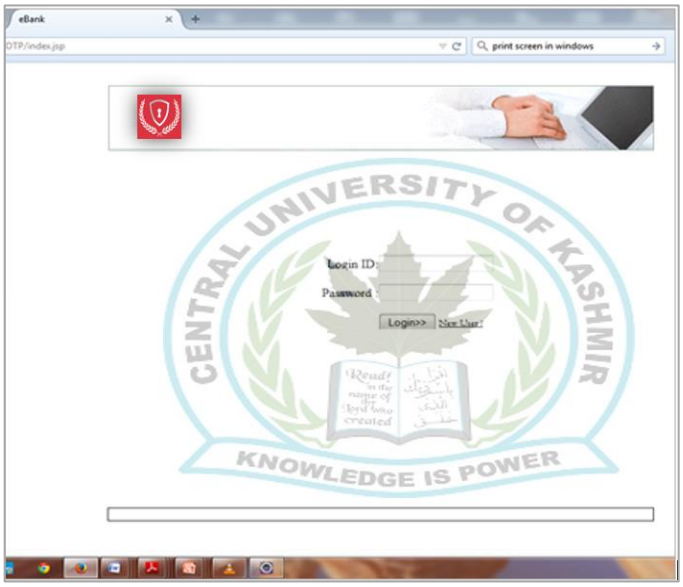


Fig. 9. User login page.

The server (bank) stores the login details of the users which comprises of the user id and the password. It is worth mentioning that it is the hash value of the password (shown in Fig. 10) that is stored rather than the password itself. This protects the system from various attacks such as stolen-verifier attack where it may be possible that the intruder gets access to the password file stored on the server.

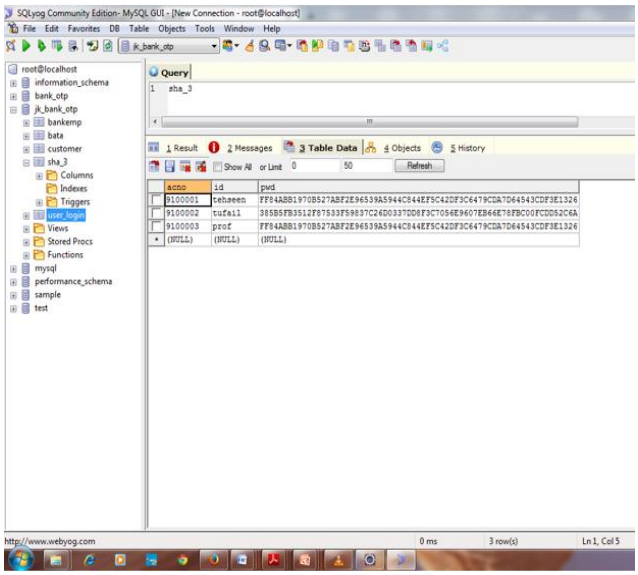


Fig. 10. Screenshot showing users' login details.

e) *Sending current OTP to server:*

As soon as the user logs in, the server prompts the user (shown in Fig. 11) for the current OTP, the coordinate N (number of SHA3 iterations) and the coordinate M (number of RIPEMD128 iterations). The current OTP and the values of M, N are generated at the user device in 'My OTP' as depicted in Fig. 12. Every time the user taps 'My OTP', a new set of OTP, M and N is produced. The OTP generated comprises of 6 human-readable words which provide increased user convenience.

Fig. 11. Screenshot showing how current OTP is sent to the server.

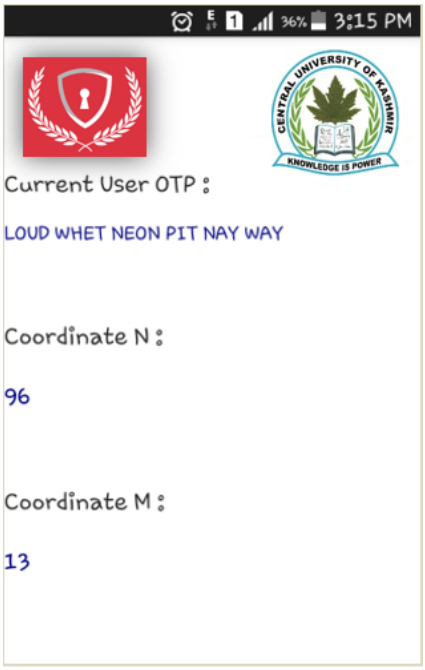


Fig. 12. Screenshot showing OTP generation on the client side.

f) Server challenge:

The next step in the authentication process consists of a challenge sent by the server to the user to validate the user by his/her response. In the challenge, the server sends the coordinates M and N to the user and asks it to generate the final OTP on his/her handheld device. The server challenge has been shown in Fig. 13, and the generation of new/final OTP at client side has been illustrated in Fig. 14. When the server receives the OTP from the client, it matches this OTP with the one generated at the server side. Only when the two matches, the server grants access privileges to the user, i.e. only when the user shall be allowed to transact from his account.

Fig. 13. Screenshot showing challenge sent by the server.

Fig. 14. Screenshots of final OTP generation on the client side.

The user when authenticated can access his banking services online. A user may transfer funds, view his/her account summary or change his/her password very conveniently. The account summary includes the date and time of transactions made from and to the user account. This can be seen from Fig. 15.

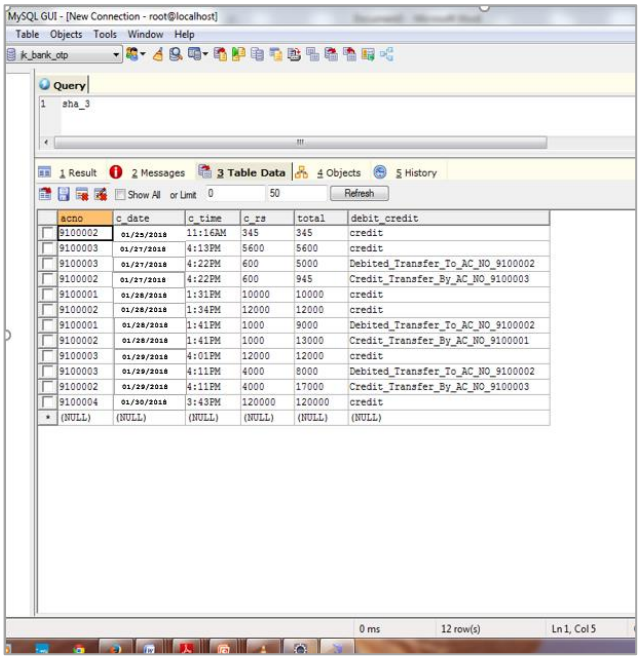


Fig. 15. Screenshot showing users' transaction details.

The server keeps a record of the transaction details of all the customers in its database.

g) List of One-time passwords:

The initial seed is formed by the concatenation of IMEI, IMSI, timestamp and index number. The values of SHA3 from 0-99 are pre-calculated to enhance efficiency. In this way, reliable and efficient One-Time-Passwords are generated which have been shown in Fig. 16.

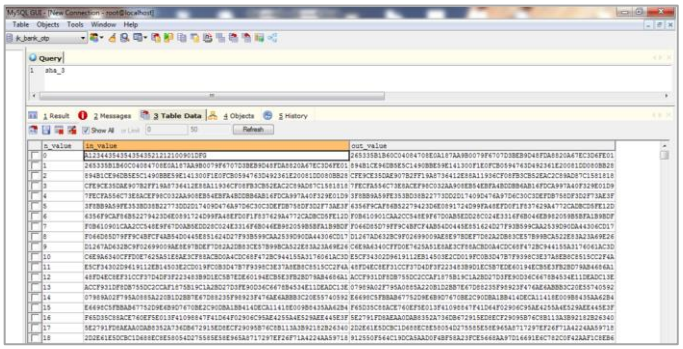


Fig. 16. Screenshot showing OTP list.

V. Results and Analysis

The primary usage of the proposed system is that it can be utilized for an efficient identification of a legitimate member for having access rights to their privilege account. As the entire framework is based on the enhanced operation of One-Time Password (OTP) as well as updated cryptographic hash functions (SHA3, truncated SHA1) and RIPEMD128, so the system excels some of its security measures. It should be also known that although there are abundant micro stages of implementation of the system, the fundamental concept of the framework basically uses the trusted handheld devices as a medium to perform authentication procedure of the system. Underlying concepts of SHA3, truncated SHA1 and RIPEMD128 makes all the differences in the security incorporations that make the authentication more robust as compared to the standard techniques adopted so far. Therefore, this section discusses the performance analysis of the system using processing time as a parameter for performance evaluation.

A. Processing Time

An effective time required to process the complete OTP generation for performing the user authentication is termed as 'processing time'. The effectiveness of the system against mitigating common types of attack/intrusion from an illegitimate member and an effective processing time required to perform the procedure is the attribute of performance analysis.

The OTP generator is installed on 100 Android mobile phones from various vendors with varying hardware specifications for the evaluation of the processing time of the application which is defined with regard to our application as the time it takes for an OTP generator to create an OTP from the initial seed. The time it takes in OTP generation is dependent on the values of random numbers N and M. This is because these values specify how many times the hash functions SHA3, truncated SHA1 and RIPEMD128 are to be used. These hash functions are used for the calculation in which the values for N and M are randomly being taken by the OTP generator. This random feature in selection of coordinates lends credibility to our results. In Fig. 17, it can be seen that irrespective of the hardware specification, it does not take more than 49 ms for the application to generate OTP. The results have been formulated using a timer. Furthermore, the average time it takes for a mobile to generate an OTP is 26.72 ms. Both the time calculations conform to the tolerable time thus proving the efficiency of the system. It is important to mention here that the correctness of the timer depends on the internal threads of the CPU; thus, to eradicate that factor influencing the final application throughput, the running application were stopped on all the test beds. Fig. 17 shows the final results accomplished while evaluating the system processing time.

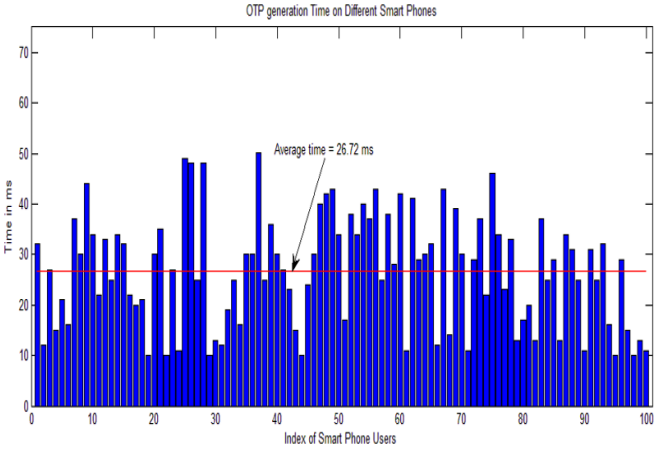


Fig. 17. Total processing time.

To emphasize on the user friendliness of the same by inclusion of Byte to word conversion, a survey was conducted where 40 users were asked randomly to key in the original 128 generated OTP to be entered as 32 digits. The analysis of the results shows that it takes on an average 30.185 seconds for the user to enter the same into his/her computer via keyboard. That is not all, 17.5% of the times an error occurred with it. Errors in keying the mobile generated OTP are shown by red rectangles in Fig. 18.

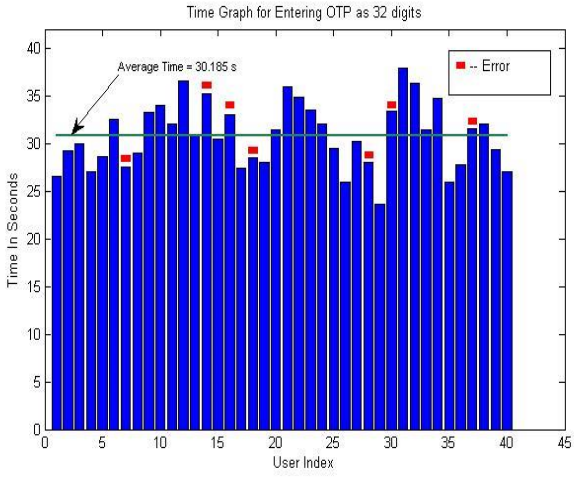


Fig. 18. Time graph for entering OTP as 32 digits.

However, when the same users were asked to enter a OTP as a six-dictionary word, the average time taken by the user decreased to 11.87 seconds (as shown in Fig. 19). Also, the occurrence of errors also diminished to 2.5 % which is extremely low as compared to the previous occurrence and thus shows the user-friendly attribute of the system. Therefore, the inclusion of Byte to Word mechanism is proven to be important while providing applicability from user perspective.

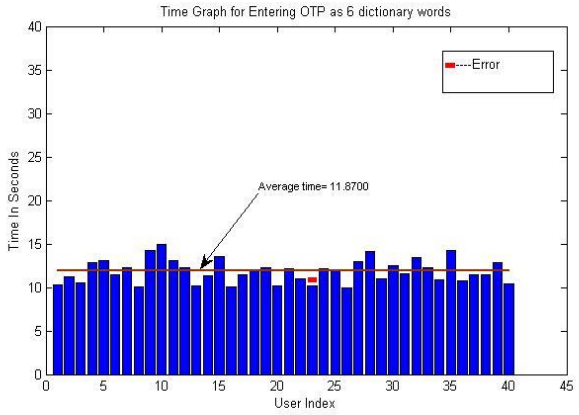


Fig. 19. Time graph for entering OTP as 6 dictionary words.

B. Security Analysis

The evaluation of security analysis critique lies in the efficient operation of the security functionality for performing secure transaction, for the considered application on financial institutions. The OTP authentication implemented makes use of OTP which is valid only for a single session and is generated with the help of strong mechanism performing concatenated cryptographic functions (SHA3, truncated SHA1 and RIPEMD128), which results in a system impossible to break until the quantum computing comes in picture. The proposed scheme can resist an off-line guessing attack because it uses strong passwords generated from strong hash functions. Also, it can protect the online transactions from replay attacks, key-loggers, shoulder surfing attacks, etc. Moreover, replaying reusable passwords are restricted by encoding passwords to be used one time.

While reviewing the base work, it was found that SHA-1 can be a potential security algorithm for mechanizing robust authentication system. However, it is not true as SHA-1 is also explored with multiple security incapability and reported attack evidences found in history. Therefore, the implementation has been carried out using SHA3 and truncated SHA1. SHA3 is designed to mitigate the security loopholes that SHA-0 or SHA-1 couldn't afford to furnish in a secure and efficient authentication mechanism. Henceforth, SHA3 is adopted in our application taking into consideration the effectiveness in security for longer time (software lifetime). Moreover, the proposed security policy ensures better privacy and confidentiality by not permitting the server and the user to secretly create and exchange the authentication token in external communication environment. Therefore, the system considers the hardware profiles and associates the exclusive generation of one-time password with it, so that it cannot be replicated for other mobile devices for particular session usage. It is almost impossible for an attacker to have the possession of actual mobile device of a legitimate user along with the static password thus rendering the system secured. Furthermore, even if the attacker has the possession of the above credentials he/she again needs to switch ON the stolen mobile phone while keeping the SIM of legitimate user intact before he/she can start using the OTP generator. This will mean that his/her location is known, and he/she can be caught. The system can thus be seen as fool proof because of the employment of the multilayer security factor as illustrated.

The results show that the client can access the privilege account in fast track proving its compatibility in mobile devices too. The performance of the system is tested by installing the application in Android mobile device, which shows no significant changes in the service delivery. Hence, it can be said that a performance which is user-friendly and secure is recorded in this evaluation process. The performance of the system can be stated as better and secure as it has two inherent characteristics: 1) The OTP processing speed is acceptable; 2) using SHA-3, it is almost impossible to explore the message mapping with the pre-determined hash function.

VI. Conclusion and Future Work

Authentication, as well as authorisation, plays an important role in securing transactions conducted on any communication network particularly over the cellular network. As observed from this study, a one-time-password authentication scheme needs to be developed on the mobile platform that provides security as well as performance in the long run. Various authentication schemes have been explored in this study which points out the vulnerabilities in the prevalent GSM-based systems. Thus, a unique password authentication mechanism is presented where the user generates OTPs on his/her mobile phone using its hardware and software profiles. For OTP generation, three hash functions, i.e. SHA-3, truncated SHA-1 and RIPE-MD128 are used which are concatenated with each other. However, the major limitation associated with the proposed scheme is that the mobile phone becomes the only point of failure in case it is stolen or lost or malfunctions. However, one method to overcome the limitation will be to design a backup procedure to restore its status after any untoward incident.

Further plans for the research undertaken include the extension of the authentication mechanism to other mobile phone platforms like Windows, Symbian, iOS, etc. There is a vast scope for researchers to make further enhancements in the solution proposed to have more security in their online transactions. This system has a large expanse in providing secure authentication and authorisation means in banking as well as other finance-based applications calling for higher security. The presented work makes use of two factors of authentication – something you know and something you have. In future, the other factor, i.e. something you are (biometrics), can be made use of to enhance the security further. Moreover, the performance analysis of the authentication system may be conducted based on user ergonomics as compared to prior systems.

References

[1] *Two-Factor Authentication for Banking - Cryptomathic*, 1st ed. Jægergårdsgade 118, DK-8000 Aarhus C, Denmark: Two-Factor Authentication for Banking – Building the Business Case, 2012, pp. 4-16.

[2] P. Vashishta and S. Kapoor, "E-Banking: Perspective for Survival in Current Market", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 1, no. 1, pp. 42-46, 2012.

[3] R. Jatana and R.K. Uppal, *E-banking in India: challenges and opportunities*. New Century Publications, 2007.

[4] B. Khan, R. Olanrewaju, A. Mehraj, A. Ahmad and S. Assad, "A Compendious Study of Online Payment Systems: Past Developments, Present Impact, and Future Considerations", *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 256-271, 2017.

[5] M. Masihuddin, B.U.I. Khan, M.M.U.I. Mattoo and R.F. Olanrewaju, () "A Survey on E-Payment Systems: Elements, Adoption, Architecture, Challenges and Security Concepts", *Indian Journal of Science and Technology*, vol. 10, no. 20, pp. 1-19, 2017.

[6] T. Laukkanen and J. Lauronen, "Consumer Value Creation in Mobile Banking Services", *International Journal of Mobile Communications*, vol. 3, no. 4, pp. 325-338, 2005.

[7] S. Choudhary, R. Temkar and N. Bhatta, "QR Code Based Secure OTP Distribution Scheme for Authentication in Net-Banking", *International Journal of Information Science and Intelligent System*, vol. 2, no. 4, pp. 115-121, 2013.

[8] R.F. Olanrewaju, B.U.I. Khan, M.M.U.I. Matto, F. Anwar, R.N. Mir and A.N.B. Nordin, "Securing Electronic Transactions via Payment Gateways – A Systematic Review", *International Journal of Internet Technology and Secured Transactions*, vol. 7, no. 3, pp. 245-269, 2017. (Accepted to be published)

[9] T. Mehraj, B. Rasool, B.U.I. Khan, A. Baba and A.G. Lone, "Contemplation of Effective Security Measures in Access Management from Adoptability Perspective", *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 6, no. 8, pp. 188-200, 2015.

[10] J.M. Stewart, E. Tittel and M. Chapple, *CISSP: Certified Information Systems Security Professional Study Guide*. Sybex, 2005.

[11] J. F. Kizza, *Computer network security*. New York: Springer, 2005.

[12] A.F. Behrouz, *Cryptography and network security*. Tata McGraw Hill Education Private Limited, 2010.

[13] D. Salomon, *Elements of computer security*. Springer Science & Business Media, 2010.

[14] C. Yan-ping, L. Dong-liang and G. Rui, "Security and precaution on the computer network", in *Future Information Technology and Management Engineering (FITME), 2010 International Conference on*, Changzhou, (Volume: 1), 2010, pp. 5-7.

[15] R.E. Smith, *Internet cryptography*. Addison-Wesley Longman Publishing Co., Inc., 1997.

[16] H. Ma, S. Yan, X. Bai and Y. Zhu, "The Research and Design of Identity Authentication Based On Speech Feature", in *Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013 International Conference on*, Nangang, 2013, pp. 166 - 169.

[17] M. Eldefrawy, K. Alghathbar and M. Khan, "OTP-Based Two-Factor Authentication Using Mobile Phones", in *Eighth International Conference on Information Technology: New Generations (ITNG)*, Las Vegas, NV, 2011, pp. 327-331.

[18] K. Bicakci and N. Baykal, "Infinite Length Hash Chains and Their Applications", in *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2002. WET ICE 2002. Proceedings. Eleventh IEEE International Workshops on*, Ankara, Turkey, 2002, pp. 57 - 61.

[19] V. Srivastava, A. Keshri, A. Roy, V. Chaurasiya and R. Gupta, "Advanced Port Knocking Authentication Scheme with QRC Using AES", in *Emerging Trends in Networks and Computer Communications (ETNCC), 2011 International Conference on*, Udaipur, 2011, pp. 159 - 163.

[20] B. Davaanaym, Y. Lee, H. Lee and S. Lee, "A Ping-Pong Based One-Time-Passwords Authentication System", in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, Seoul, 2009, pp. 574 - 579.

[21] W. Hsieh and J. Leu, "Design of a Time and Location Based One-Time Password Authentication Scheme", in *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*, Istanbul, 2011, pp. 201 - 206.

[22] S. Liao, Q. Zhang, C. Chen and Y. Dai, "A unidirectional one-time password authentication scheme without counter desynchronization", in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on (Volume: 4)*, Sanya, 2009, pp. 361 - 364.

[23] J. Jeong, M. Young Chung and H. Choo, "Integrated OTP-Based User Authentication and Access Control Scheme in Home Networks", in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, Waikoloa, HI, 2008, pp. 294.

[24] M. Long and U. Blumentthal, "Manageable One-Time Password for Consumer Applications", in *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, Las Vegas, NV, 2007, pp. 1 – 2.

[25] S. Hallsteinsen, I. Jørstad and D. Van Thanh, "Using the mobile phone as a security token for unified authentication", in *Systems and Networks Communications, 2007. ICSNC 2007. Second International Conference on*, Cap Esterel, 2007, p. 68.

[26] M. Alzomai and A. Josang, "The Mobile Phone as a Multi OTP Device Using Trusted Computing", in *Network and System Security (NSS), 2010 4th International Conference on*, Melbourne, VIC, 2010, pp. 75 – 82.

[27] K. Moon, D. Moon, J. Yoo and H. Cho, "Biometrics Information Protection Using Fuzzy Vault Scheme", in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, Naples, 2012, pp. 124-128.

[28] H. Ma, S. Yan, X. Bai and Y. Zhu, "The research and design of identity authentication based on speech feature", in *Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013 International Conference on*, Nangang, 2013, pp. 166 - 169.

[29] A. Castiglione, A. De Santis, A. Castiglione and F. Palmieri, "An Efficient and Transparent One-Time Authentication Protocol with Non-Interactive Key Scheduling and Update", in *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*, Victoria, BC, 2014, pp. 351-358.

[30] P. R. Avhad and R. Satyanarayana, "A Three-Factor Authentication Scheme in ATM", *International Journal of Science and Research (IJSR)*, vol. 3, no. 4, pp. 656-659, 2014.

[31] J. N. Oruh, "Three-Factor Authentication for Automated Teller Machine System", *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS)*, vol. 4, no. 6, pp. 160-166, 2014.

[32] S. Boonkrong, "Internet Banking Login with Multi-Factor Authentication", *KSII Transactions on Internet & Information Systems*, vol. 11, no. 1, pp. 511-535, 2017.

[33] R.O. Akinyede and O.A. Esese, "Development of a Secure Mobile E-Banking System", *International Journal of Computer (IJC)*, vol. 26, no. 1, pp. 23-42, 2017.

[34] A. Arara, E.B.E. Fgee and H.A. Jaber, "Securing e-Government Web Portal Access Using Enhanced Two Factor Authentication", in *Advances in Engineering Sciences & Applied Mathematics (ICAESAM'2015), 4th International Conference on*, 2015, pp. 65-69.

[35] M.H. Eldefrawy, M.K. Khan, K. Alghathbar, T.H. Kim and H. Elkamchouchi, "Mobile One-Time Passwords: Two-Factor Authentication Using Mobile Phones", *Security and Communication Networks*, vol. 5, no. 5, pp. 508-516, 2012.

[36] W. Stallings, *Cryptography and network security*, 5th ed. India: Pearson Education, 2011.

[37] "Mobile OS market share 2017 | Statista", *Statista*, 2017. [Online]. Available: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/. [Accessed: 18- Jan- 2018].

A Review and Classification of Widely used Offline Brain Datasets

Muhammad Wasim, Muhammad Sajjad, Farheen Ramzan, Usman Ghani Khan, Waqar Mahmood
Al-khawarizmi Institute of Computer Science,
UET Lahore, Pakistan

***Abstract*—Brain Computer Interfaces (BCI) are a natural extension to Human Computer Interaction (HCI) technologies. BCI is especially useful for people suffering from diseases, such as Amyotrophic Lateral Sclerosis (ALS) which cause motor disabilities in patients. To evaluate the effectiveness of BCI in different paradigms, the need of benchmark BCI datasets is increasing rapidly. Although, such datasets do exist, a comparative study of such datasets is not available to the best of our knowledge. In this paper, we provided a comprehensive overview of various BCI datasets. We briefly describe the characteristics of these datasets and devise a classification scheme for them. The comparative study provides feature extractors and classifiers used for each dataset. Moreover, potential use-cases for each dataset are also provided.**

***Keywords*—BCI; dataset; brain-computer interface; amyotrophic lateral sclerosis; classification**

I. Introduction

Human brain controls all internal and external body functions. It is responsible for activities such as learning, creativity, memories and many others [1]. The functions and structure of the human brain have always fascinated researchers [2], [3]. A clear understanding of human brain helps in disease diagnostics and a similar mechanism can be used to develop intelligent machines [4], [5]. To reverse engineer the magnificent brain, many sensory computational models have been developed [6], [7]. These models cover various aspects of human brain and the data collected from these models is available for further exploration. The data collected from these models have various applications ranging from medical diagnostics to autonomous robot navigation [8], [9].

The sensory models used for brain recordings are broadly divided into two categories: Invasive and Non-Invasive [10]. The invasive method is used for medical diagnoses of diseases such as seizure, sclerosis, tumors, epilepsy, and spinal cord trauma. The treatment of these diseases requires surgery for the placement of electrodes in brain gray matter. Invasive methods record brain activity from the cortical surface and include techniques such as Electrocorticography (ECOG) and Intracranial Electroencephalography (IEEG). Non-invasive methods do not need surgery or insertion of an instrument in the patient's body. These methods include models such as Electroencephalography (EEG), Magnetoencephalography (MEG), Functional Magnetic Resonance Imaging (FMRI), Positron Emission Tomography (PET), Infrared (IR) Imaging, Near Infrared Spectroscopy (NIRS), and Fetal Magnetoencephalography (FMEG).

EEG recordings are measured in Hertz (cycles per second) and are divided into different frequency bands as slow, moderate, and fast waves [11] as shown in Table I. To record brain activity, a complete paradigm needs to be designed. The paradigm involves the presentation of cues to the subject. Electromagnetic activity as a result of these cues is recorded using a headset. The cues might include audio or visual information which facilitates us in the collection of brain activity for particular classes which is later used for testing and evaluation.

The brain datasets developed based on various computational models using both invasive and non-invasive techniques. In this paper, we present short description of each datasets and describe different characteristics of the datasets in tabular format. We further classify the datasets according to our proposed classification scheme.

Moreover, these datasets are saved using different formats. These datasets can be used for conducting new experiments and hypothesis validation. Using existing datasets saves the time and energy of researches. These existing datasets and availability of the work already done on them accelerate further development in the area of brain-computer interfacing. The datasets will be help full for fresh researcher in this domain has limited resources and can take full advantage of these datasets.

Distribution of paper is given as: In Section 2 we present the review of datasets. The datasets classification scheme is illustrated in Section 3. In Section 4 we summarize classification of datasets and discuss their importance features.

TABLE. I. Type of EEG Brain Waves Ramadan Et Al. (2015)

Wave	Frequency	Activated
Delta Waves	0.5Hz-3.5Hz	During sleep and normally found in young children
Theta Waves	4Hz-7Hz	During sleep
Alpha Waves	8Hz-12Hz	During walking or when mentally active
Mu Waves	8Hz-12Hz	During movement or intent of movement
Beta Waves	12Hz-38Hz	alert, attentive, engaged in problem-solving or decision making
Gamma Waves	38Hz-42Hz	State of universal love, altruism and high virtues

II. Description of Datasets

In this section, we describe different characteristics of each dataset. Many of the datasets are collected using different devices, sampling rate, filter, and classes. Most of the data set are collected by using more than 64 electrodes device for experimentation. Furthermore, we also described how much participants (healthy or unhealthy) takes part in each of the experiment. Next, each of the datasets is concisely explained.

1) Motor Imagery, Uncued Classifier Application:

The dataset on motor imagery uncued classifier [12] was recorded from seven healthy subjects. Data was collected using 64 EEG channels (0.05-200Hz) with a sampling rate of 1000Hz. The task for collecting data was motor imagery and without subject feedback. Motor Imagery dataset from Institute for Knowledge Discovery contains data from 9 healthy subjects[13]. 22 Ag/AgCl electrodes EEG channels and 3 EOG channels were used and data was recorded with a sampling frequency of 250Hz. Data was notch filtered between 0.5Hz and 100Hz.

2) Hand movement direction in MEG:

Dataset Hand Movement Direction in MEG was recorded from 2 right-handed healthy subjects [14]. It was recorded using 10 MEG channels and a band pass filtered between 0.5 Hz to 100 Hz with a sampling frequency of 400 Hz. The subjects performed wrist movements in four different directions using MEG with 625 Hz sampling rate.

3) Finger movements in ECoG:

Finger Movements in ECoG dataset was recorded from epileptic patients at Harborview Hospital in Seattle[15]. The data was recorded using 48 to 64 ECoG channels band pass filtered between 0.15Hz to 200Hz having 1000Hz sampling rate. The dataset contains ECoG data during individual flexions of the five fingers; movements acquired with a data glove.

4) Error-related potentials (ERPs) during continuous feedback:

Error-Related Potentials (ERP) During Continuous Feedback dataset [16] was recorded from 10 healthy subjects having age between 24 years to 25 years. 28 EEG electrodes were used to record EEG and 3 EOG electrodes were used to measure EOG. The data has a sampling rate of 512 Hz, notch filterer at 50 Hz and band-pass filtered between 0.5 Hz and 60 Hz.

5) Two-finger game-play with deliberately failing controller:

Another dataset on Two-Finger Gameplay with Deliberately Failing Controller comprises of data on 12 subject performing a paceman game [17]. For recording purpose the EEG and Biomedical signals, BioSemi ActiveTwo EEG system was used having a sampling rate of 512 Hz. 32 Ag/AgCl active electrodes were used to record the EEG signals. 4 EOG was used to measure ocular and muscle artifacts and 4 EMG signals over the muscles used to press with the index finger.

6) Covert and overt ERP-based BCI:

Covert and Overt ERP based BCI Dataset [18] contains recordings of P300 evoked potentials. It was recorded with BCI2000 using two different paradigms like overt attention and covert attention based on P300 Speller [19] and GeoSpell interface [20] respectively. The EEG signals digitized at 256 Hz with frequency range in between of 0.1 Hz and 20 Hz were recorded with 16 Ag/AgCl electrodes using g.USBamp amplifier. It was recorded from 10 healthy female subjects having a mean age of 26.8+- 5.6.

7) Neuroprosthetic control of an EEG/EOG BNCI:

The dataset on Neuroprosthetic Control of an EEG/EOG BNCI used 26 5 EEG channels for EEG recordings using an active electrode EEG system. EEG signals had a sampling rate of 200 Hz and band-pass filtered between 0.4 Hz to 70 Hz. 1 EOG channel was also used with a sampling rate of 200 Hz. The dataset was collected from a highly defected spinal cord patient with upper limb paralyzed.

8) Individual imagery:

Another dataset on Individual Imagery [21], in which EEG data was recorded from 9 patients instructed to relax and avoid eye moments, suffering from spinal cord injury and brain stroke. EEG signals were recorded using 30 channels. The g.tec GAMMAsys a system with g.LADYbird active electrodes and 2 g.USBamp bio signal amplifiers were used for recording. The EEG signals were band-pass filtered between 0.5-100 Hz with a notch filter at 50 Hz having a sampling rate of 256 Hz.

9) ECOG-based BCI based on cognitive control:

ECOG-Based BCI Based On Cognitive Control dataset [22] is about cognitive control network for BCI purposes. They used FMRI for non-invasive localization of the cognitive control network and recorded data from an epilepsy patient, who was implanted with subdural grid electrodes over the left and right frontal cortex temporarily. The subject performed two target tasks in several runs using the high-frequency power of 55-95 Hz.

10) Emergency braking during simulated driving:

Emergency Braking During Simulated Driving dataset [23] was collected from 18 subjects by using 59 EEG electrodes and 2 bi-polar EOG with Ag/AgCl electrodes mounted on a cap with a sampling rate of 200 Hz. BrainAmp hardware was used to amplify and digitize the EEG and EMG signals. TORCS software was used to provide information about technical and behavioral markers.

11) Mental arithmetic:

Mental arithmetic dataset [24] was recorded from 8 subjects (3 male and 5 female) having a mean age of 26 years with a standard deviation of 2.8 years. A multi-channel system was used which contain 16 photo detectors and 17 light emitters, and it was a 3 x 11 grid having a total of 52 FNIRS with a sampling rate of 10 Hz. To record brain oxygenation continues wave system was used. An aggressive hemodynamic response [24] was shown during the tasks of mental arithmetic.

12) Auditory oddball during hypnosis:

Auditory oddball during hypnosis dataset was recorded from 2 healthy subjects, one male and one female of right-handed by using 27 EEG active electrodes and 4 EOG channels were used to record eye movement by using a sampling frequency of 512 Hz. Data was band-pass filtered between 0.01-100 Hz and notch filtered at 50 Hz.

13) SCP Training in stroke:

The dataset on SCP Training in Stroke was recorded from 2 chronic stroke patients by using a single electrode Cz with a Nexus-10 MKII DC amplifier having a sampling frequency of 256 Hz. For a record of eyes movement, 2 bipolar EOG electrodes were used.

14) Mental imagery, multi-class:

Mental Imagery Multi-Class dataset [25] has 32 integrated electrodes (DC-256Hz) having a sampling rate of 512 Hz were used to collect data from 3 normal subjects during 4 non-feedback sessions. The dataset is presented in two ways, raw EEG signals, and precomputed features. Raw EEG signals have a sampling rate of 512 Hz. On the other hand, in precomputed features, Surface Laplacian was used to spatially filter raw EEG signals. Imaginary repetitive self-paced right-hand movements and generating words start with same random letter.

15) Motor imagery, small training sets:

Motor Imagery dataset [26] is focused on applying machine learning approach to BCI. The EEG data were recorded from 5 healthy subjects by using BrainAmp amplifiers, in which 118 out of 128 channel Ag/AgCl electrode were used and data were band-pass filtered between 0.05-200 Hz, digitized at 1000 Hz. For analysis purpose another version of data with a down-sampled rate of 1000Hz.

16) Monitoring error-related potentials:

Monitoring error-related potentials dataset [27] is about Error Related Potential (ERP) recorded via EEG. The subject had to monitor the performance of an external device that was not controlled by subject. The EEG data were recorded from 6 subjects having a mean age of 27.83 +- 2.23 by using 64 electrodes of Biosemi active two systems at full DC with a sampling rate of 512 Hz.

17) Emotion recognition using EEG signals:

Another dataset is Emotion Recognition Using EEG Signals [28], collected from 15, 7 males and 8 females had a mean age of 23.27 years and standard deviation of 2.37 years. ESI NeuroScan System is used, in which a total of 62 Ag/AgCl electrodes channels were used with a sampling frequency of 1000 Hz. A band-pass filter with a frequency range between 0-75 Hz was used.

18) Visual search within natural images:

Visual Search within Natural Images [29] dataset is a short demo of 5 experimental trials. Brain Products amplifiers having 25 recording channels were used to record EEG signals and this data was band-pass filtered between 1-40 Hz offline. For eye movements binocularly was used and SMI IView X tracker was used and data was recorded with a sampling rate of 500 Hz.

19) EEG Eye State (Planning &\ Relax):

EEG Eye State (Planning and Relax) dataset [30] contains EEG recordings that are used for classification for two mental stages namely planning of motor imagery actions and relaxed state. The dataset was recorded from a 25 years old healthy right-handed subject. A Medelec Profile Digital EEG machine was used for recordings which contain 8 EEG Ag/AgCI electrodes and has a sampling rate of 256 Hz This data was filtered with a high frequency of 50 Hz and low frequency of 1.6 Hz, notch filtered at 50 Hz.

20) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity:

The dataset for Indications of Nonlinear Deterministic and Finite Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State[31] was provided to study the dynamic properties of brain EEG signals recorded from different brain regions and different physiological and pathological brain states. The datasets were recorded from 5 healthy subjects by using 128 electrodes with a sampling rate of 173.61Hz. The datasets were low-pass filtered at 40 Hz and band-pass filtered at 0.53 Hz to 40 Hz.

21) Self-regulation of SCPs:

The dataset for Self-Regulation of SCPS[32] was presented to study cortical positivity and cortical negativity. The dataset was recorded from an artificially respired Amyotrophic Lateral Sclerosis (ALS) patient using 6 channels using PsyLab EEG Amplifier with a range of +/-1000 microV with a sampling rate of 256 Hz.

22) Self-paced:

Self-Paced dataset [33] was presented to predict the upcoming finger movement for both hands, 130 milliseconds before the key press. The data was recorded from a healthy subject with no feedback provided by using 28 EEG channels with a sampling rate of 1000Hz. NeuroScan amplifier and Ag/AgCl electrode cap from ECI was used for recording EEG signals. The dataset is band pass filtered between 0.05-200Hz and down sampled with 100Hz.

23) EEG Motor Movement/Imagery Dataset:

Motor Movement and Imagery Dataset[15] was provided by the developers of the BCI2000 instrumentation system[34]. The data were collected from 109 subjects by using 64 EEG channels and each signal sampled at 160 samples per second. The data is provided in European Data Format (EDF+). DREAMS Subject dataset [35] was recorded during DREAMS project to analyze, test and train classification algorithms for automatic sleep stages.

24) The DREAMS Subject database:

The datasets were collected from 20 subjects 14 females, 4 males having age between 20 years to 65 years by using polysomnographic (PSG) for the whole night with a sampling frequency of 200Hz. The PSG recording was annotated in different sleep stages according to criteria of Rechtschaffen and Kales (R and K) [36] standards introduced by American Academy of Sleep Medicine (AASM). The data was acquired

in sleep lab of Belgium Hospital. For each recording, at least 1 EOG channel, 3 EEG channels and 1 EMG channel were used.

25) The DREAMS REMs Database:

DREAMS REMs database [35] is about Rapid Eye Movement (REM) [37] sleep, which is a period of sleep during which one experiences clear dreams. This dataset was recorded for DREAMS project to analyze, train and test the classification algorithms for automatic detection. There were 9 excerpts, each of which was 30 minutes long. The dataset was acquired in a sleep lab of Belgium Hospital and recorded from 5 healthy subjects of both males and female having age between 20 years to 46 years by using 32 channel polygraph of BrainnetTM System of MEDATEC, Brussels, Belgium. The recordings involved 2 EOG channels, 3 EEG channels and 1 EMG channel with a sampling frequency of 200Hz.

26) 4 Class EEG Data:

The dataset for Multi-Class Motor Imagery EEG[38] is about multi-class cued motor imagery having four classes as left hand, right hand, foot, and tongue. The data were recorded from three subjects with 60 EEG channels amplifier from Neuroscan, where left mastoid was used as a reference and right mastoid was used as ground. The EEG data were sampled at 250 Hz and filtered between 1-50 Hz with notch filter on. There were 60 trials per class and data across all trials were concatenated. This dataset was stored as Geographic Data Files (GDF) [39].

27) Motor imagery in ECoG recordings, session-to-session transfer:

Dataset for Motor Imagery in ECOG Recordings, Session-to-Session Transfer[40] was provided with the goal that classifiers for BCI systems usually do not perform better for the data acquired on different days and sessions from the same subject without retraining. The dataset was recorded by using 8x8, 64 channel ECoG platinum electrode grid, placed on the right motor cortex also partly covering surrounding cortex areas and was sampled at 1000 Hz. The data was also filtered at a frequency range of 0.016-300 Hz. The data was about the cued motor imagery of left pinky and tongue from 1 subject.

28) P300 speller paradigm:

The dataset for P300 Speller Paradigm [19] was presented with the goal to estimate the probability of subject paying attention to the letters in a 6 x 6 matrix by intensifying the rows and columns respectively. This dataset was collected from 2 healthy subjects by using BCI2000 system with 64 EEG channels, which is digitalized at a sampling rate of 240Hz and data were band-pass filtered between 0.1 Hz to 60 Hz.

29) ERP-based Brain-Computer Interface recordings:

The dataset for ERP-based Brain-Computer Interface[41], [42] was provided with the goal to identify the factor affecting the performance of BCI based on event-related potentials and to improve the usability and transfer rate of these interfaces. The data were recorded from 12 subjects by using a BioSemi ActiveTwo EEG system. A total of 64 EEG electrodes were used with a sampling frequency of 2048 Hz.

30) EEG brain wave for confusion:

The data was stored in European Data Format (EDF+) with signals and annotations. EEG dataset for confusion [43], [44] contains EEG signals recorded from 10 college students who watched 10 Massive Open Online Course (MOOC) video clips. There are two types of videos confusing and non-confusing for each student. For recording signals, single-channel wireless-Mindset over frontal lobe of subjects are used. There are 100 data points and sampling rate was 0.5 seconds and high-frequency signals were reported during this 0.5 seconds.

31) Motion VEP Speller:

Motion VEP Speller dataset [45] was provided with the goal to estimate the usability of gaze-independent communication. The data were recorded from 16 healthy subjects, 10 males, and 6 females, having age between 21 years to 30 years with a mean age of 23.8 years. To record EEG signals, Brain Products (Munich, Germany) actiCap active electrode system with 64 electrodes and a BrainAmp EEG amplifier was used. The data were sampled at 1000 Hz and band-pass filtered at a frequency range of 0.016-250 Hz. An Intelligence IG-30 (Alea Technologies) eye-tracker was used to control eye movements with a sampling rate of 50 Hz.

32) Center Speller:

Center Speller dataset [46] was provided with the goal to develop a visual speller named Center Speller that does not require eye movements. The data was recorded online from 13 healthy subjects, 8 males 5 females aged 16 years to 45 years with mean age of 27 years. The EEG data was recorded with Brain Products (Munich, Germany) actiCAP using 64 electrodes with a sampling rate of 1000 Hz. A band pass filter was used at a frequency range of 0.016-250 Hz.

33) SSVEP-based BCI with LED:

SSVEP-Based BCI with LED dataset [39] was recorded from 5 male and female subjects having age between 22 years to 30 years by using 8 EEG channels with g.Mobilab+ device with a sampling frequency of 256 Hz.

34) Global datasets for autism disorder:

Global Datasets for Autism Disorder [47] was collected from 18 boys, 10 normal and 8 abnormal having age between 10 years to 16 years. To record the EEG signals, a recording system from BCI2000 with active electrodes and the Active digital EEG amplifier was used. The recording system contained 16 Ag/AgCl, g.tec EEG cap, electrodes, g.tec GAMMAbox, g.tec USBamp and BCI2000 with a sampling frequency of 256 Hz. The data were band-pass filtered at a frequency between 0.1 Hz to 60 Hz and notch filtered at 60 Hz.

35) Event-related potential datasets based on a three-stimulus paradigm:

The dataset for Event-Related Potential Datasets Based On a Three-Stimulus Paradigm[48] was provided with the goal to introduce three-stimulus paradigm for the P300 component and provide datasets for three-stimulus paradigm EEG/ERP

freely available to the researcher. This dataset was collected from 25 healthy subjects, 9 males and 11 females having age between 20 to 26 years and 19 of them were right-handed. The EEG data was recorded of event-related potentials (ERP) of 20 subjects by using BrainVision Recorder 1.2 and data was stored in BrainVision format. The data of other 5 subjects were discarded due to excessive eye blinking. The data was sampled at 100 Hz and low-pass filtered with a cut-off frequency of 250 Hz.

III. Classification

In this section, we present the classification background and the proposed classification scheme under which various datasets can be classified.

A. Classification Background

The datasets can be classified based on the cognitive behavior of human or functional atlases of the brain. The classification helps us understand which part of the brain is being activated and which brain processes are generated in response to a particular cognitive action. We present a classification scheme for datasets that cover various aspects of human behavior, cognitive states, and abilities.

Behavioral neuroscience [49] is the study of physiology, genetics, behavioral evolution, and development mechanism in animals and humans under the principles of biological sciences. It mainly deals with brain functions and components, neural activity, neurotransmitters, hormonal changes, behavioral evolution and their effects on behavioral changes. It is also termed as biological psychology or psychobiology [50].

Studies in the field of behavioral neuroscience are mainly directed towards animals and humans to better understand the human pathology and mental processes. Due to the technological advancements and development of non-invasive methods, behavioral neuroscience now also deals with linguistics, philosophy, and psychology [51]. The datasets can be broadly divided into following categories:

- Sensation and Perception
- Motivated Behavior
- Control of movement
- Learning and memory
- Sleep and biological rhythms
- Emotion
- Language
- Reasoning and decision making
- Consciousness

1) Sensation and Perception:

Human is considered to have five basic senses as proposed by Aristotle [52]. The sensation is the body's way of detecting some external or internal stimulation. Particular brain regions generate, receive and interpret specific signals based on sensation. The various senses are as follows [53]:

- **Sight:** This sense to see something. There are two distinct receptors present related to sight, one for color (cones) and one for brightness (rods) [54].
- **Taste:** sweet, salty, sour, bitter, and umami (umami receptors detect the amino acid glutamate, which is a taste of meat and some artificial flavoring) [55].
- **Touch:** This has been found to be distinct from pressure, temperature, pain, and even itch sensors [56].
- **Pressure:** It is a type of skin pressure which results from persisting pressure on the skin.
- **Itch:** This is a distinct sensor system from other touch-related senses.
- **Thermoception:** This is the ability to sense heat and cold. There are different types of thermoreceptors for detecting heat or cold in the brain. These thermoreceptors in the brain are used for monitoring internal body temperature.
- **Sound:** Detecting vibrations along some medium, such as air or water that is in contact with your ear drums [57].
- Smell: This sense is due to sensors that work off of a chemical reaction. This sense combines with taste to produce flavors.
- **Proprioception:** This sense gives you the ability to tell where your body parts are, relative to other body parts. This sense is one of the things police officers test when they pull over someone who they think is driving drunk. The close your eyes and touch your nose test is testing this sense. This sense is used all the time in little ways, such as when you scratch an itch on your foot, but never once look at your foot to see where your hand is relative to your foot [56].
- **Tension Sensors:** These are found in such places as your muscles and allow the brain the ability to monitor muscle tension.
- **Nociception:** In a word, pain. This was once thought to simply be the result of overloading other senses, such as touch, but this has been found not to be the case and instead, it is its own unique sensory system. There are three distinct types of pain receptors: cutaneous (skin), somatic (bones and joints), and visceral (body organs) [58].
- **Equilibrioception:** The sense that allows you to keep your balance and sense body movement in terms of acceleration and directional changes. This sense also allows for perceiving gravity. The sensory system for this is found in your inner ears and is called the vestibular labyrinthine system.

- **Stretch Receptors:** These are found in such places as the lungs, bladder, stomach, and the gastrointestinal tract. A type of stretch receptor, that senses dilation of blood vessels, is also often involved in headaches.
- **Chemoreceptors:** These trigger an area of the medulla in the brain that is involved in detecting blood born hormones and drugs. It also is involved in the vomiting reflex.
- **Thirst:** This system more or less allows your body to monitor its hydration level and so your body knows when it should tell you to drink.
- **Hunger:** This system allows your body to detect when you need to eat something [59].
- **Magnetoception:** This is the ability to detect magnetic fields, which is principally useful in providing a sense of direction when detecting the Earth's magnetic field. Humans do not have a strong magnetoception, however, experiments have demonstrated that we do tend to have some sense of magnetic fields. It is theorized that this has something to do with deposits of ferric iron in our noses [60].
- **Chronoception:** This refers to how the passage of time is perceived and experienced. Humans have a startling accurate sense of time, particularly when younger. Long term time keeping seems to be monitored by the superchiasmatic nuclei (responsible for the circadian rhythm). Short term time keeping is handled by other cell systems [61].
- **Electroreception**: Electroreception (or electroception) is the ability to detect electric fields.
- **Hygroreception:** This is the ability to detect changes in the moisture content of the environment.
- **Equilibrioception**: Balance, equilibrioception, or vestibular sense is the sense that allows an organism to sense body movement, direction, and acceleration, and to attain and maintain postural equilibrium and balance.

B. Proposed Datasets Classification Scheme

By analyzing different components of the brain and their associated functions they perform, we can classify the datasets on the basis of our classification scheme as shown in Table I, which shows that there is a wide range of mental tasks that need to be considered for BCI research. But there are some potential problems in recording many brain activities. Dataset column in Table II, show the category of each of the dataset.

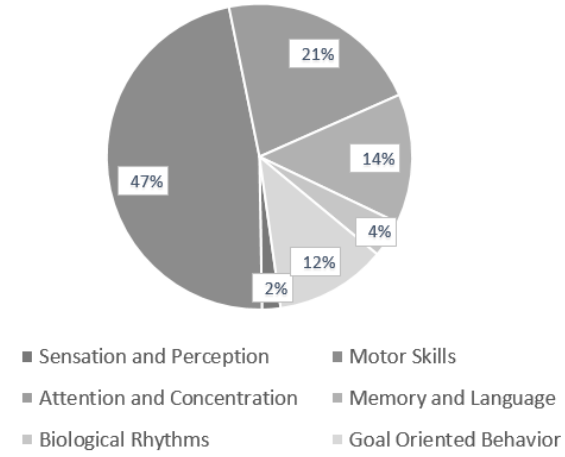


Fig. 1. Comparison chart of datasets.

TABLE. II. CLASSIFICATION OF DATASETS BASED ON DIFFERENT CATEGORIES

Category	Subcategory	Dataset
Sensation and Perception	Sense, Sight, Taste, Touch, Pressure, Itch, Thermoception, Sound, Smell, Proprioception, Tension Sensors, Nociception, Equilibrioception, Stretch Receptors, Chemoreceptors (vomiting reflex), Thirst, Hunger, Magnetoception, Chronoception (Time), Electroreception, Hygroreception	7), 12)
Motor Skills	Imaginary, Physical	1), 2),3), 4), 5),6), 7), 8), 14), 15), 16), 18),19), 20), 21), 22), 23), 26), 28), 29), 31), 32), 33) ,34)
Attention and Concentration	Attention, Concentration	6), 14), 35)
Memory and Language	Short-term memory, Long-term memory, Working memory, Communication	6), 8), 9), 11), 13), 17), 20), 28), 29), 30) ,35)
Biological Rhythms	Breathing, Heartbeat, Blood Pressure, Sleep, Dreams	25, 26
Goal Oriented Behaviours	Anticipation, Problem-solving, Decision making, Emotions, Sequencing, Inhibition	5, 9, 17, 20, 21, 23

IV. DISCUSSION

The brain contains about 100 billion neurons and each neuron is constantly sending and receiving signals through a complex mechanism. During certain activity performed by the human brain, neurons make thousands of connections through these processes which are difficult for EEG electrodes so the signals need to be disentangled. Our thoughts, movements, actions, learning, and decision are the result of complex electro-chemical processes in the brain. The BCI datasets target a limited number of mental tasks as it is very complex

to map brain signals into human actual intentions. Brain signals corresponding to certain activities such as sneezing is quite difficult to capture.

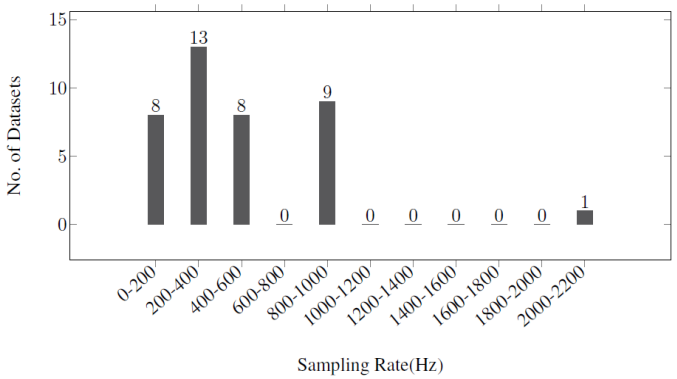


Fig. 2. Comparison of different datasets based on sampling.

It requires specific environment parameters and settings to get better results. Also, the cost of devices used for recording signals is reasonably high and complexity of brain structure is a hindrance in the way of recording, analyzing and mapping brain signals to certain human activities. These may be the reasons why such datasets available online are limited to simple activities. Comparing the sampling frequency of these datasets shows that most of the datasets lie in the range of 200 Hz to 1000 Hz but we found only one dataset with a high sampling rate of 2048 Hz. The bar chart in Fig. 2 shows the datasets with a specific frequency.

By looking at the dataset from another perspective, the Fig. 1 shows that 47% of the datasets are related to motor skills - either imagery or physical. 21% datasets are related to attention and concentration, 14% related to memory and language, 12% related to goal-oriented behavior, 4% related to biological rhythms and only 2% related to perception and concentration. This comparison reveals that brain activities corresponding to human senses and biological rhythms such as dreams and sleep are difficult to capture. Special equipment, environmental and experimental setup is required which is either costly or hard to achieve. Datasets are collected by many institutes, research groups who are continuously working and struggling to achieve accurate results. Many researcher groups have worked to collect different datasets.

Table III illustrates which paper make use of the datasets and which feature extractor and classifier used in the research. While a datasets references are also presented which includes references to research papers in which these datasets have been described and elaborated. The references to the papers that used or cited these datasets for their research work is also shown in Table III.

As the datasets contain records of the subject's specific activity for a limited period of time which contain noise ratio that creates a problem when getting output. Some of the researcher groups and institutions used a hybrid system for EEG signals acquisition where they used EEG, EOG, ECOG, MEG, and FNRI to get the precise outcome. It is also well known that EEG signals capturing devices are non-invasive, low-cost, and modest. It was the reason that most datasets were collected using EEG method as shown in Fig. 3.

The datasets are recorded for brain activities using different devices, mapped to some mathematical form and stored using different formats. Some formats in which datasets are stored are .edf (European Data Format), .dgf (General Data Format), .mat, .txt, .bdf (Glyph Bit Distribution Format), .dat, .vhdr, .vmrk, .set, .avg, .eeg, .cnt etc. As there is a wide range of data formats, the processing of the datasets is rather a complex task. Also, the software available for brain signal processing, support a limited number of data formats. There is a lack of standard formats and structures in which datasets are recorded.

The datasets presented here have been used by many researchers over time. Understanding and recognizing human intentions via brain signals is an important step and needs complex data analysis and processing. Various softwares are available for analyzing brain activity with limited techniques. Therefore, some standards and tools are required to make research easy in the field of brain-computer interface. Software tools can be helpful in determining a comparison between different methods of data processing, determining hyper-parameters required for particular algorithms and defining compatibility of certain concepts. BCI datasets are mostly goal oriented. Researchers working on specific BCI application prefer to generate their own datasets that are mostly not available publicly.

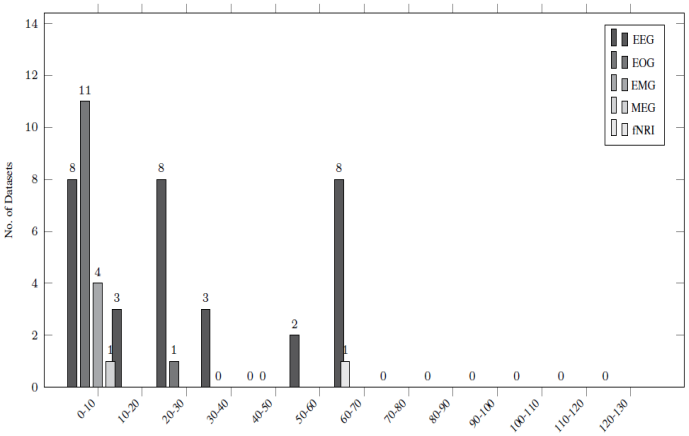


Fig. 1.Comparison of different datasets based on various numbers of channels used in EEG, EMG, ECOG, MEG, EOG and FNRI.

V. Conclusion

We have explored and discussed different datasets of the BCI research in which most of them are based on EEG. We presented a comparison of datasets with respect to the frequency of different datasets which shows that most of the datasets are collected in a frequency range from about 200-1000 Hz. A classification scheme – with six categories - was proposed for the datasets categorization. A comparison of datasets with respect to their respective categories shows that most of the datasets are related to brain activity during motor skills. Mostly, SVM and LDA classifiers were used to process and classify these datasets.

REFERENCES

[1] I. Wood, "Neuroscience: Exploring the brain," 1996.

[2] J. Kaufman and D. Charney, "Effects of early stress on brain structure and function: implications for understanding the relationship between child maltreatment and depression," Dev. Psychopathol., 2001.

[3] S. MacDonald, L. Nyberg, and L. Bäckman, "Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity," Trends Neurosci., 2006.

[4] C. Olsson and L. Nyberg, "Brain simulation of action may be grounded in physical experience," Neurocase, 2011.

[5] M. M. Waldrop, "Brain in a box," Nature, vol. 482, no. 7386, p. 456, 2012.

[6] R. Palaniappan, "Electroencephalogram-based Brain–Computer Interface: An Introduction," Guid. to Brain-Computer Music Interfacing, 2014.

[7] A. Gevins and A. Rémond, "Methods of analysis of brain electrical and magnetic signals," 1987.

[8] P. Jahankhani and K. Revett, "Data mining an EEG dataset with an emphasis on dimensionality reduction," Intell. Data …, 2007.

[9] A. M. Baldini et al., "Search for the lepton flavor violating decay$\$ mu^+$\$ rightarrow$\$ mathrm ${$e$}$^+$\$ gamma with the full dataset of the MEG experiment," Eur. Phys. J. C, vol. 76, no. 8, p. 434, 2016.

[10] H. Anupama and N. Cauvery, "Brain computer interface and its types-a study," Int. J., 2012.

[11] R. Ramadan, S. Refat, M. Elshahed, and R. Ali, "Basics of Brain Computer Interface," Brain-Computer Interfaces, 2015.

[12] B. Blankertz, G. Dornhege, M. Krauledat, and K. Müller, "The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects," Neuroimage, 2007.

[13] M. Fatourechi, A. Bashashati, R. Ward, and G. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey," Clin. Neurophysiol., 2007.

[14] S. Waldert, H. Preissl, E. Demandt, and C. Braun, "Hand movement direction decoded from MEG and EEG," J., 2008.

[15] G. Schalk, D. McFarland, and T. Hinterberger, "BCI2000: a general-purpose brain-computer interface (BCI) system," IEEE Trans., 2004.

[16] M. Spüler and C. Niethammer, "Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity," Front. Hum. Neurosci., vol. 9, p. 155, 2015.

[17] M. Witkowski and M. Cortese, "Enhancing brain-machine interface (BMI) control of a hand exoskeleton using electrooculography (EOG)," J., 2014.

[18] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain--computer interface using electrocorticographic signals in humansThe authors declare that they have no competing financial interests.," J. Neural Eng., vol. 1, no. 2, p. 63, 2004.

[19] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," Electroencephalogr. Clin., 1988.

[20] F. Aloise, P. Aricò, F. Schettini, A. Riccio, and S. Salinari, "A covert attention P300-based brain–computer interface: Geospell," Ergonomics, 2012.

[21] R. Scherer et al., "Individually adapted imagery improves brain-computer interface performance in end-users with disability," PLoS One, vol. 10, no. 5, p. e0123727, 2015.

[22] M. Vansteensel, D. Hermes, and E. Aarnoutse, "Brain–computer interfacing based on cognitive control," Ann., 2010.

[23] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components—a tutorial," Neuroimage, vol. 56, no. 2, pp. 814–825, 2011.

[24] G. Pfurtscheller and G. Bauernfeind, "Focal frontal (de) oxyhemoglobin responses during simple arithmetic," Int. J., 2010.

[25] J. R. Millan, "On the need for on-line learning in brain-computer interfaces," in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, vol. 4, pp. 2877–2882.

[26] G. Dornhege, B. Blankertz, and G. Curio, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," IEEE Trans., 2004.

[27] R. Chavarriaga and J. Millán, "Learning from EEG error-related potentials in noninvasive brain-computer interfaces," IEEE Trans. neural, 2010.

[28] W. Zheng and B. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Trans. Auton. Ment., 2015.

[29] N. Li, F. Niefind, S. Wang, W. Sommer, and O. Dimigen, "Parafoveal processing in reading Chinese sentences: Evidence from event-related brain potentials," Psychophysiology, vol. 52, no. 10, pp. 1361–1374, 2015.

[30] O. Dimigen, R. Kliegl, and W. Sommer, "Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials," Neuroimage, 2012.

[31] R. Andrzejak, K. Lehnertz, F. Mormann, and C. Rieke, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," Phys. Rev. E, 2001.

[32] N. Birbaumer, N. Ghanayim, T. Hinterberger, and I. Iversen, "A spelling device for the paralysed," Nature, 1999.

[33] B. Blankertz, G. Curio, and K. Müller, "Classifying single trial EEG: Towards brain computer interfacing," Adv. neural Inf., 2002.

[34] P. PhysioBank, "PhysioNet: components of a new research resource for complex physiologic signals," Circ. v101 i23. e215-e220.

[35] S. Devuyst, T. Dutoit, P. Stenuit, M. Kerkhofs, and E. Stanus, "Cancelling ECG artifacts in EEG using a modified independent component analysis approach," EURASIP J. Adv. Signal Process., vol. 2008, no. 1, pp. 1–13, 2008.

[36] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," 1968.

[37] E. Olson, B. Boeve, and M. Silber, "Rapid eye movement sleep behaviour disorder: demographic, clinical and laboratory findings in 93 cases," Brain, 2000.

[38] C. Vidaurre, T. Sander, and A. Schlögl, "BioSig: the free and open source software library for biomedical signal processing," Comput. Intell., 2011.

[39] A. Schlögl, O. Filz, H. Ramoser, and G. Pfurtscheller, "GDF-a general dataformat for biosignals version 1.25," 2005.

[40] T. N. Lal et al., "Methods towards invasive human brain computer interfaces," in Advances in neural information processing systems, 2004, pp. 737–744.

[41] L. Citi, R. Poli, and C. Cinel, "Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in Donchin's speller," J. Neural Eng., 2010.

[42] A. L. Goldberger et al., "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215--e220, 2000.

[43] I. Allen and J. Seaman, Going the distance: Online education in the United States, 2011. 2011.

[44] H. Wang, Y. Li, X. Hu, Y. Yang, Z. Meng, and K. Chang, "Using EEG to Improve Massive Open Online Courses Feedback Interaction.," AIED Work., 2013.

[45] J. Wolpaw, "Brain–computer interfaces as new brain output pathways," J. Physiol., 2007.

[46] M. Treder and N. Schmidt, "Gaze-independent brain–computer interfaces based on covert attention and feature attention," J. neural, 2011.

[47] E. Alsaggaf, "DISCOVERING AUTISM DISORDER BY ANALYSIS EEG SIGNALS USING DIFFERENT CLASSIFICATION ALGORITHMS," 2013.

[48] J. Polich, "Updating P300: an integrative theory of P3a and P3b," Clin. Neurophysiol., 2007.

[49] M. Mesulam, Principles of behavioral and cognitive neurology. 2000.

[50] [50] undefined Merriam-Webster, Merriam-Webster's collegiate dictionary. 2004.

[51] M. Breedlove, M. Rosenzweig, and V. Neil, "An Introduction to Behavioral and Cognitive Neuroscience," Biol. Psychol. Watson, 2007.

[52] D. Howes, Empire of the Senses. 2005.

[53] D. Hiskey, "HUMANS HAVE A LOT MORE THAN FIVE SENSES." 2010.

[54] S. Hecht, "Rods, cones, and the chemical basis of vision," Physiol. Rev., 1937.

[55] R. Mattes, "Is there a fatty acid taste?," Annu. Rev. Nutr., 2009.

[56] G. Robles-De-La-Torre, "The importance of the sense of touch in virtual and real environments," Ieee Multimed., 2006.

[57] A. Davis, C. McMahon, and K. Pichora-Fuller, "Aging and hearing health: the life-course approach," Gerontologist, 2016.

[58] R. Fulbright, C. Troche, and P. Skudlarski, "Functional MR imaging of regional brain activation associated with the affective experience of pain," Am. J., 2001.

[59] O. Farr, R. Chiang-shan, and C. Mantzoros, "Central nervous system regulation of eating: Insights from human brain imaging," Metabolism, 2016.

[60] A. Voustianiouk and H. Kaufmann, "Magnetic fields and the central nervous system," 2000.

[61] A. Hirsch, "Method of Altering Perception of Time," US Pat. App. 13/734,106, 2013.

[62] P. Kindermans, P. Buteneers, … D. V.-… M. L. for, and undefined 2010, "An uncued brain-computer interface using reservoir computing," biblio.ugent.be.

[63] G. Dornhege, B. Blankertz, … M. K.-I. transactions on, and undefined 2006, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," ieeexplore.ieee.org.

[64] S. Waldert et al., "Hand movement direction decoded from MEG and EEG," J. Neurosci., vol. 28, no. 4, pp. 1000–1008, 2008.

[65] S. Hajipour Sardouie and M. B. Shamsollahi, "Selection of Efficient Features for Discrimination of Hand Movements from MEG Using a BCI Competition IV Data Set," Front. Neurosci., vol. 6, 2012.

[66] R. Flamary and A. Rakotomamonjy, "Decoding Finger Movements from ECoG Signals Using Switching Linear Models," Front. Neurosci., vol. 6, 2012.

[67] N. Liang and L. Bougrain, "Decoding Finger Flexion from Band-Specific ECoG Signals in Humans," Front. Neurosci., vol. 6, 2012.

[68] A. Nijholt, D. P.-O. Bos, and B. Reuderink, "Turning shortcomings into challenges: Brain--computer interfaces for games," Entertain. Comput., vol. 1, no. 2, pp. 85–94, 2009.

[69] F. Aloise et al., "A covert attention P300-based brain--computer interface: Geospell," Ergonomics, vol. 55, no. 5, pp. 538–551, 2012.

[70] [70] M. Witkowski, M. Cortese, M. Cempini, J. Mellinger, N. Vitiello, and S. R. Soekadar, "Enhancing brain-machine interface (BMI) control of a hand exoskeleton using electrooculography (EOG)," J. Neuroeng. Rehabil., vol. 11, no. 1, p. 1, 2014.

[71] M. J. Vansteensel et al., "Brain--computer interfacing based on cognitive control," Ann. Neurol., vol. 67, no. 6, pp. 809–816, 2010.

[72] Y. Wang, Z. Zhang, Y. Li, X. Gao, … S. G.-I. T. on, and undefined 2004, "BCI competition 2003-data set IV: an algorithm based on CSSD and FDA for classifying single-trial EEG," ieeexplore.ieee.org.

[73] H. Yoon, K. Yang, C. S.-I. transactions on knowledge, and undefined 2005, "Feature subset selection and feature ranking for multivariate time series," ieeexplore.ieee.org.

[74] R. Chavarriaga and J. del R. Millán, "Learning from EEG error-related potentials in noninvasive brain-computer interfaces," IEEE Trans. neural Syst. Rehabil. Eng., vol. 18, no. 4, pp. 381–388, 2010.

[75] D. Nie, X. Wang, L. Shi, B. L.-N. E. (NER), and undefined 2011, "EEG-based emotion recognition during watching movies," ieeexplore.ieee.org.

[76] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," IEEE Trans. Auton. Ment. Dev., vol. 7, no. 3, pp. 162–175, 2015.

[77] R. B. Bhatt and M. Gopal, "FRCT: fuzzy-rough classification trees," Pattern Anal. Appl., vol. 11, no. 1, pp. 73–88, Jan. 2008.

[78] Y. Kerimbekov, H. Ş. Bilge, and H. H. Uğurlu, "The use of Lorentzian distance metric in classification problems," Pattern Recognit. Lett., vol. 84, pp. 170–176, Dec. 2016.

[79] Z. Yang, Y. Wang, and G. Ouyang, "Adaptive neuro-fuzzy inference system for classification of background EEG signals from ESES patients and controls.," ScientificWorldJournal., vol. 2014, p. 140863, Mar. 2014.

[80] G. Chen, W. Xie, @bullet Tien, D. Bui, and A. Krzy_ Zak, "Automatic Epileptic Seizure Detection in EEG Using Nonsubsampled Wavelet–Fourier Features," J. Med. Biol. Eng., vol. 37.

[81] G. Dornhege, B. Blankertz, … G. C.-… in N. I., and undefined 2003, "Combining features for BCI," papers.nips.cc.

[82] B. Blankertz, G. Curio, and K.-R. Muller, "Classifying single trial EEG: Towards brain computer interfacing," Adv. Neural Inf. Process. Syst., vol. 1, pp. 157–164, 2002.

[83] J. Sleight, P. Pillai, S. M.-A. A. U. of Michigan, and undefined 2009, "Classification of executed and imagined motor movement EEG signals," shiwali.me.

[84] M. Tolić, F. J.-K. I. journal of fundamental and, and undefined 2013, "Classification of wavelet transformed EEG signals with neural network for imagined mental and motor tasks," hrcak.srce.hr.

[85] E. J. Olson, B. F. Boeve, and M. H. Silber, "Rapid eye movement sleep behaviour disorder: demographic, clinical and laboratory findings in 93 cases," Brain, vol. 123, no. 2, pp. 331–339, 2000.

[86] S. Ge, R. Wang, D. Y.-P. one, and undefined 2014, "Classification of four-class motor imagery employing single-channel electroencephalography," journals.plos.org.

[87] Z. Chin, K. Ang, C. Wang, … C. G.-… in M. and, and undefined 2009, "Multi-class filter bank common spatial pattern for four-class motor imagery BCI," ieeexplore.ieee.org.

[88] W. Jakuczun, "Constructing discriminative biorthogonal bases for classification," 2004.

[89] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), vol. 2, pp. 985–990.

[90] M. Kaper, P. Meinicke, … U. G.-I. T., and undefined 2004, "BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm," ieeexplore.ieee.org.

[91] R. Fazel-Rezai, "Human Error in P300 Speller Paradigm for Brain-Computer Interface," in 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007, pp. 2516–2519.

[92] N. J. Hill et al., "Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects," IEEE Trans. neural Syst. Rehabil. Eng., vol. 14, no. 2, pp. 183–186, 2006.

[93] H. Wang, Y. Li, X. Hu, Y. Yang, Z. Meng, and K. Chang, "Using EEG to Improve Massive Open Online Courses Feedback Interaction.," in AIED Workshops, 2013.

[94] S. Schaeff, M. S. Treder, B. Venthur, and B. Blankertz, "Exploring motion VEPs for gaze-independent communication," J. Neural Eng., vol. 9, no. 4, p. 45006, Aug. 2012.

[95] C. Guger et al., "How many people are able to control a P300-based brain–computer interface (BCI)?," Neurosci. Lett., vol. 462, no. 1, pp. 94–98, Sep. 2009.

[96] M. S. Treder, N. M. Schmidt, and B. Blankertz, "Gaze-independent brain--computer interfaces based on covert attention and feature attention," J. Neural Eng., vol. 8, no. 6, p. 66003, 2011.

[97] H. Hwang, J. Lim, Y. Jung, H. Choi, … S. L.-J. of neuroscience, and undefined 2012, "Development of an SSVEP-based BCI spelling system adopted a QWERTY-style LED keyboard," Elsevier.

[98] E. Kalunga, K. Djouani, Y. Hamam, S. Chevallier, and E. Monacelli, "SSVEP enhancement based on Canonical Correlation Analysis to improve BCI performances," in 2013 Africon, 2013, pp. 1–5.

[99] R. Djemal, K. AlSharabi, … S. I.-B. research, and undefined 2017, "EEG-Based Computer Aided Diagnosis of Autism Spectrum Disorder Using Wavelet, Entropy, and ANN," hindawi.com.

[100] L. Vareka, P. Bruha, and R. Moucek, "Event-related potential datasets based on a three-stimulus paradigm," Gigascience, vol. 3, no. 1, p. 1, 2014.

[101] F. Barceló, J. A. Periáñez, and E. Nyhus, "An information theoretical approach to task-switching: evidence from cognitive brain potentials in humans," Front. Hum. Neurosci., vol. 1, 2007.

TABLE. III. DATASETS USED IN DIFFERENT RESEARCH

Dataset Ref	Feature Extractor	Classifier
1	FBCSP [62]	Reservoir Computing
	Common Spatial Patterns (CSP) & CSSP [63]	LDA
2	RLDA [64]	
	Time domain & frequency domain [65]	Support Vector Machine (SVM) and LDA
3	Smoothed Auto-Regressive (AR) [66]	time sample classification scheme
	Ridge Regression and Sparse Linear Regression [67]	
4	Time domain & frequency domain[16]	Support Vector Machine (SVM)
5	ERD and LRP[68]	Support Vector Machine (SVM)
6	Stepwise Linear Discriminant Analysis (SWLDA) [69]	Linear Support Vector Machine (SVM)
7	Autoregression Model [70]	Discriminability based on true and false positive classifications
8	Common Spatial Patterns [21]	Fisher's LDA
9	Autoregressive [71]	---
14	--- [25]	Gaussian classifier
15	CSP & Fisher Discriminant Analysis [72]	Neural network
	Feature Subset Selection (FSS) [73]	Support Vector Machine (SVM)
16	[74]	Gaussian classifier
17	Principal Component Analysis [75]	Linear Support Vector Machine (SVM)
	Differential Entropy [76]	KNN, LR, Support Vector Machine (SVM), DBN
19	--- [77]	Fuzzy-rough classification tree (FRCT)
	Two-Dimensional Lorentzian Space [78]	Classification via Lorentzian Metric (CLM)
20	wavelet coefficients [79]	Adaptive Neuro-Fuzzy Inference System (ANFIS)
	Fourier Transform [80]	Nearest Neighbour
21	Autoregressive models(AR) & Common Spatial Patterns (CSP) [81]	linear & Non-Linear Classifiers
22	Common Spatial Subspace decomposition and FD Fisher Discrimination [82]	Fisher Discriminant Analysis and SVMs
23	Principal Component Analysis [83]	Support Vector Machine (SVM)
	Discrete Wavelet Transform [84]	Bayes Quadratic Classifier
24	--- [35]	EAS, AF-EA, and ICA-EA
25	--- [85]	---
26	Common Spatial Pattern (CSP) [86]	Support Vector Machine (SVM)
	Filter Bank Common Spatial Pattern (FBCSP) [87]	
27	Decision-Tree Induction (Local-Feature Extractor) [88]	Support Vector Machine (SVM)
	Auto-Regression Model [89]	Single Hidden Layer Feedforward Neural Networks (SLFN)
28	[90] Continuous Wavelet Transform (CWT) & Student's t-Statistic	Support Vector Machine (SVM)
	[91] Averaged Mexican Hat Wavelet Coefficients	Custom classification method
29	Automatic Feature Selection Technique [92]	Event related desynchronization (ERD) based classification
30	Custom Feature Extract [93]	Different classifier used
31	Heuristic Search Temporal [94]	linear discriminant analysis (LDA)
	Custom features [95]	linear discriminant analysis (LDA)
32	Custom Spatial Features[96]	Linear Discriminant Analysis (LDA
33	--- [97]	Custom classification algorithm
	Canonical Correlation Analysis [98]	Support Vector Machine (SVM)
35	Entropy [99]	Artificial Neural Network (ANN)
36	Spatial, Temporal, and Spatio-Temporal [100]	Linear Discriminant Analysis (LDA
	--- [101]	Rules Based Classification

A Novel Design of Pilot Aided Channel Estimation for MIMO-CDMA System

Khalid Mahmood
University of Technology Nowshera, Pakistan

***Abstract*—In order to estimate a fading channel characteristics, a pilot signal is propogated with traffic channel. Fading channel parameter estimation is of paramount importance as it may be utilized to design different equalization techniques. It may also be utilized to allocate weights of rake receiver to sturdiest multipaths as well as coherent reception and weighted combination of multipath constituents of wireless communication systems. In this paper, a pilot aided channel estimating technique for MIMO-CDMA systems is presented. This technique utilizes minimum mean squared error estimation of corrupted information in a flat fading channel along with noise. Simulation results predicts theoretical predictions are strongly validated for different values of SNR and users.**

***Keywords*—Channel estimation; MIMO-CDMA; channel estimator; MMSE; Rayleigh fading channel; SNR**

I. Introduction

Design of an efficient receiver in code division multiple access (CDMA) system is perplexing as output (received) information carrying signal is compromised by Gaussian noise corrupted, interference and random nature of a flat fading channel.

Fading channel parameter estimation is of paramount importance as it may be utilized to design linear and decision feedback equalizers. Moreover fading channel estimating techniques may also be utilized in employing well-known maximum ratio combining (MRC) method [1]. Detecting multipath flat fading channel parameters could be utilized to assign the weights of rake receiver to strong multipath(s) for a DS-CDMA system [1]. Pilot channel based estimating technique was utilized in IS-95 forward link [2]. It is used as for 3rd generation (3G) system on forward as well as converse link, respectively [3].

A flat fading channel estimating technique is the one, where we utilize a fixed low pass filter whose cutoff frequency is made equal to maximum Doppler frequency (MDF) [4]. Though this technique is easier to design, but its performance degenerates when its cutoff frequency strays from MDF. Optimal filters such as Wiener and Kalman were employed in [5], [6], [7], but designing them is extremely difficult due to channel parameter statistics requirement. Pilot-aided Rayleigh fading channel estimating technique utilizing minimum mean squared error (MMSE) estimator for single-input single-output (SISO) CDMA systems is developed in [8]. In our paper, we have designed a channel estimation technique for MIMO-CDMA wireless system which utilizes MMSE estimation of the impaired signal in the presence of Rayleigh fading channel and Gaussian noise (AWGN). To the best of our knowledge, multiple access interference (MAI) is simply treated as additive noise [9]. But in this research work, MMSE estimation of channel parameters is formed in presence MAI and noise without considering MAI as Gaussian. A unified approach of evaluating channel estimation for Rayleigh fading is also performed in this paper.

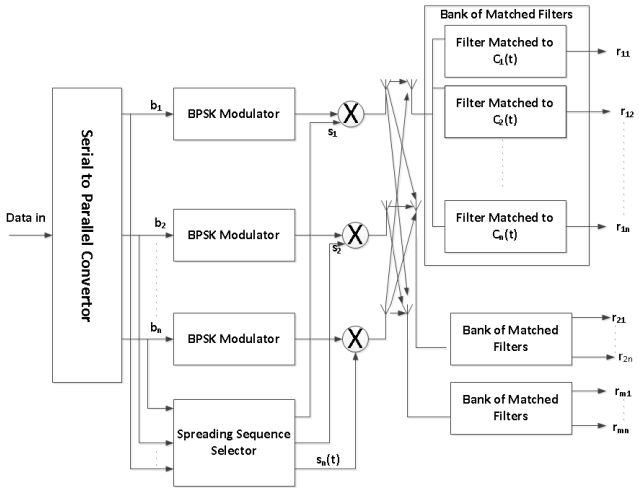


Fig. 1. System Model.

II. System Model

CDMA transmitter model of a wireless network having M transmit and N receivers is shown in Fig. 1.

We are considering a Rayleigh fading channel having complex impulse response between the n^{th} transmitter and m^{th} receiver for the l^{th} symbol is

$$H_{mn}^{l}(t) = h_{mn}^{l} e^{j\phi_l} \delta(t) \tag{1}$$

Where,

h_{mn}^{l} is the impulse response

ϕ_n is phase of Rayleigh fading channel for l^{th} symbol.

The pdf of h_{mn}^{l} for the Rayleigh fading channel is given as

$$f_{h_{mn}^{l}}(x) = \frac{ux}{\sigma_h^2} \exp\left(-\frac{u^2}{2\sigma_h^2}\right), \quad \text{for} \quad u > 0, \tag{2}$$

Where,

σ_h^2 represents Rayleigh fading channel variance

m^{th} receiver observes the following:

$$r_m(t) = \sum_{n=1}^{N}\sum_{l=-\infty}^{\infty}\sum_{p=1}^{P} A^p b_n^{l,p} s_n^{l,p}(t) h_{mn}^l + \nu_m(t), \quad m = 1,2,3\ldots M \tag{3}$$

Where,

P represents number of users,

$s_n^{l,p}(t)$ represents rectangular signature waveform with random signature sequence of the p^{th} user,

T_b and T_c represents bit period and chip interval,

$\left\{b_n^{l,p}\right\}$ represents input bit sequence of p^{th} user,

A^p represents transmitted amplitude of p^{th} user,

ν_m represents Gaussian noise with zero mean and variance σ_ν^2

Cross correlation (CC) between signature sequences of j^{th} and p^{th}users is

$$\begin{aligned}\rho_l^{p,j} &= \int_{(l-1)T_b}^{lT_b} s_n^p(t)\, s_n^j(t)\, dt \\ &= \sum_{i=1}^{N_c} c_{l,i}^p c_{l,i}^j\end{aligned} \tag{4}$$

,

$\left\{c_{l,i}^p\right\}$ is assumed to be normalized spreading sequence of k^{th} user.

Receiving side is comprised of matched filter and is matched to a signature waveform (signal) of desired user. We are assuming that user 1 is the desired one. So matched filter's output may be set up as

$$\begin{aligned}y_m^l &= \int_{(l-1)T_b}^{lT_b} r_m(t)\, s_m^{l,1}(t)\, dt \\ &= \sum_{n=1}^{N} A^1 b_n^{l,1} h_{mn}^l + \kappa + \nu_m, \quad m = 1,2,3\ldots M,\end{aligned} \tag{5}$$

Where,

κ is MAI at the m^{th} receiver in presence of flat fading Rayleigh channel. It may be set up as

$$\kappa = \sum_{n=1}^{N}\sum_{k=2}^{P} A^p b_n^{l,p} \rho_n^{k,1} h_{mn}^l, \quad m = 1,2,3\ldots M \tag{6}$$

It may also be written as

$$\kappa = \sum_{n=1}^{N} Q_{mn}^{l,p} h_{mn}^l, \quad m = 1,2,3\ldots M \tag{7}$$

The term, $Q^{l,p} = \sum_{p=2}^{P} A^k b^{l,p} \rho^{p,1}$ exhibits the Gaussian behavior [10], $Q^{l,p} \sim \mathcal{N}(0, \sigma_Q^2)$, and $\sum_{n=1}^{N} A^1 b_n^{l,1} h_{mn}^l$ represent required signal. It may be expressed as

$$\sum_{n\,=1}^{N} A^1 b_n^{l,1} h_{mn}^l = A^1 b^{l,1} \sum_{n=1}^{N} h_{mn}^l \tag{8}$$

$\alpha^l = \sum_{n=1}^{N} h_{mn}^l$ is being used as sum of Rayleigh fading channel and considering the fact that equal power is transmitted by all users, we are defining

$$x \triangleq b^1 \alpha \tag{9}$$

We are dropping the time index. Equation (5) can be written as

$$y_m = x + \kappa \tag{10}$$

Pdf of MAI plus noise as derived in [11] is

$$f_\kappa(\kappa) = \frac{1}{2\sqrt{\pi}} \sum_{n=1}^{N} \mathcal{A}\Gamma\left(\frac{1}{2}, \mathcal{B}; \frac{\kappa^2}{\mathcal{C}}\right) \tag{11}$$

Where,

$\mathcal{A} = \frac{C_n}{\sigma_{U_n}} \exp(\mathcal{B})$,

$\mathcal{B} = \frac{\sigma_\eta^2}{2\sigma_{U_n}^2}$,

$\mathcal{C} = 4\sigma_{U_n}^2$

$\sigma_{U_n}^2 = \sigma_I^2 \sigma_\alpha^2$ is an MAI variance in MIMO-CDMA systems [12].

III. MMSE Estimation of MIMO-CDMA Flat Fading Channel

MMSE estimating method technique is well known in literature for estimation in the presence of AWGN. In this research work, this technique is extended to include MAI variance in estimating Rayleigh channel for MIMO-CDMA systems. MSE cost function of MMSE estimation is given as

$$J = E\left[(u - \hat{u})^2\right] \tag{12}$$

Where,

$\hat{u}$ is the estimate of u

A cost function is defined as

$$\begin{aligned}\hat{u}(v) &= E[u \mid v] \\ &= \int_{-\infty}^{\infty} up(u \mid v)\, du\end{aligned} \tag{13}$$

Estimation of $\hat{\alpha}$ is obtained from (9)

$$\hat{\alpha} = \frac{\hat{u}}{b^1} \tag{14}$$

In order to determine $\hat{u}$, conditional pdf $p_{u|v}(u \mid v)$ is needed, which is given as

$$p_{u\,|v}(u \mid v) = \frac{p_{v|u}(v \mid u)\, p_u(u)}{p(v)} \tag{15}$$

in order to utilize (15), $p_u(u)$ should be evaluated first. As u is product of two random variables c^1 and α, so its pdf may be calculated by using random variable transformation

$$p_u(u) = \int_0^\infty \frac{1}{\lambda} p_\alpha(\lambda)\, p_c\left(\frac{u}{\lambda}\right) d\lambda \tag{16}$$

Where, $p_c(c) = 0.5\,[\delta(c+1) + \delta(c-1)]$and $p_\alpha(\alpha)$ is the pdf of sum of Rayleigh fading random variables. pdf of the sum of Rayleigh fading random variables is given as [13]

$$\begin{aligned} p_{SAA}(t) &= \frac{t^{2N-1} e^{-\frac{t^2}{2b}}}{2^{N-1} b^N (N-1)!}, \\ b &= \frac{\sigma^2}{N}\left[(2N-1)!!\right]^{1/N} \end{aligned} \tag{17}$$

Where,

$(2N-1)!! = (2N-1)(2N-3)\cdots 3.1$ and $t = u/\sqrt{N}$ are assumed to be normalized argument.

By applying transformation technique to (17), the pdf is written as

$$\begin{aligned} p_\alpha(\alpha) &= \frac{1}{\sqrt{N}} \frac{\left(\frac{\alpha}{\sqrt{N}}\right)^{2N-1} \exp\left(-\frac{\alpha^2}{2Nc}\right)}{2^{N-1} c^N (N-1)!}, \\ c &= \frac{\sigma^2}{N}\left[(2N-1)!!\right]^{1/N} \end{aligned} \tag{18}$$

Pdfs in (16) may be expressed as

$$p_c\left(\frac{u}{\lambda}\right) = 0.5\left[\delta\left(\frac{u}{\lambda}+1\right) + \delta\left(\frac{u}{\lambda}-1\right)\right] \tag{19}$$

and

$$\begin{aligned} p_\alpha(\lambda) &= \frac{1}{\sqrt{N}} \frac{\left(\frac{\lambda}{\sqrt{N}}\right)^{2N-1} \exp\left(-\frac{\lambda^2}{2Nc}\right)}{2^{N-1} c^N (N-1)!}, \\ c &= \frac{\sigma^2}{N}\left[(2N-1)!!\right]^{1/N} \end{aligned} \tag{20}$$

Now we can express (16) as

$$\begin{aligned} p_u(u) &= \frac{\mathcal{G}}{c^N} \int_0^\infty \lambda^{2N-2} \exp\left(-\frac{\lambda^2}{2Nc}\right) \\ &\quad \times \left[\delta\left(\frac{u}{\lambda}+1\right) + \delta\left(\frac{u}{\lambda}-1\right)\right] d\lambda \end{aligned} \tag{21}$$

Where,

$\mathcal{G} = \left(\frac{1}{\sqrt{N}}\right)^{2N} \left(\frac{1}{2}\right)^N \frac{1}{(N-1)!} x^{2N-1}$.

Integral in (21) is calculated as

$$p_x(x) = \frac{\mathcal{G}}{b^N} u^{2N-1} \exp\left(-\frac{u^2}{2Nb}\right) \tag{22}$$

Pdf of $p_{y|x}(v \mid u)$ is calculated by (11) to obtain

$$p_{v\,|u}(v \mid u) = \frac{1}{2\sqrt{\pi}} \sum_{n=1}^{N} \mathcal{A}\Gamma\left(\frac{1}{2}, \mathcal{B}; \frac{(v-u)^2}{\mathcal{C}}\right) \tag{23}$$

Since $\int_{-\infty}^{\infty} p_{u|v}(u \mid v)\, du = 1$, $p(v)$ is set up as

$$\begin{aligned} p(v) = \int_{-\infty}^{\infty} p_{v|u}(v \\ \mid u)\, p_u(u)\, du \end{aligned} \tag{24}$$

Putting in the terms from (22) and (23), we found

$$p(v) = \frac{1}{2\sqrt{\pi}} \frac{\mathcal{G}}{b^N} \sum_{n=1}^{N} \mathcal{A} T_{n(1)} \tag{25}$$

where $T_{n(1)}$ is

$$T_{n(1)} = \int_{-\infty}^{\infty} \Gamma\left(\frac{1}{2}, \mathcal{B}; \frac{(v-u)^2}{\mathcal{C}}\right) \exp\left(-\frac{u^2}{2Nb}\right) du \tag{26}$$

By Substituting pdfs of $p(v)$ and $p_{v|u}(v \mid u)$ in (15) and utilizing (13), $\hat{u}$ estimate is provided as

$$\hat{u} = \sum_{n=1}^{N} \frac{T_{n(2)}}{T_{n(1)}} \tag{27}$$

Where, $T_{n(2)}$ is given by

$$T_{n(2)} = \int_{-\infty}^{\infty} u^{2N} \Gamma\left(\frac{1}{2}, \mathcal{B}; \frac{(v-u)^2}{\mathcal{C}}\right) \exp\left(-\frac{u^2}{2Nc}\right) du \tag{28}$$

Since we could not find any closed form solution for the integrals in (26) and (28), we have solved them numerically.

IV. Simulation Results and Discussion

We are using following simulation setup:

- Synchronous MIMO-CDMA system with random signature sequence having length 30.
- Flat Rayleigh fading channel.
- Received signal having noise which is AWGN.
- 5 and 10 users, respectively.
- 2×2 MIMO system.

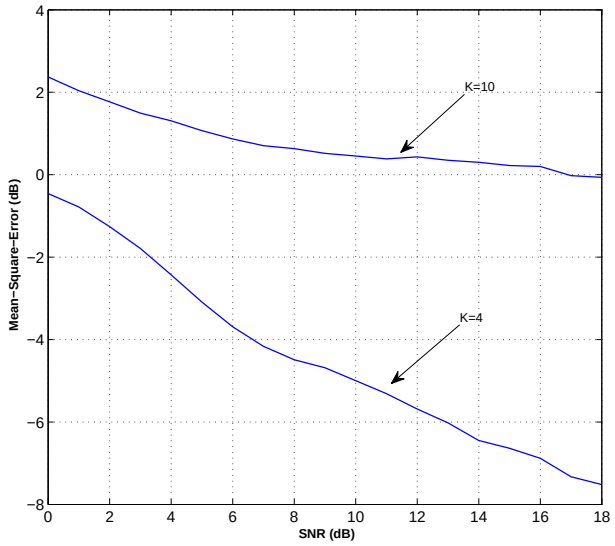


Fig. 2. Mean Squared Error (MSE).

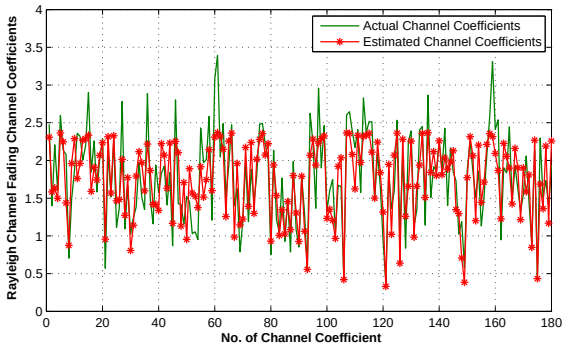


Fig. 3. Channel Coefficients (Rayleigh).

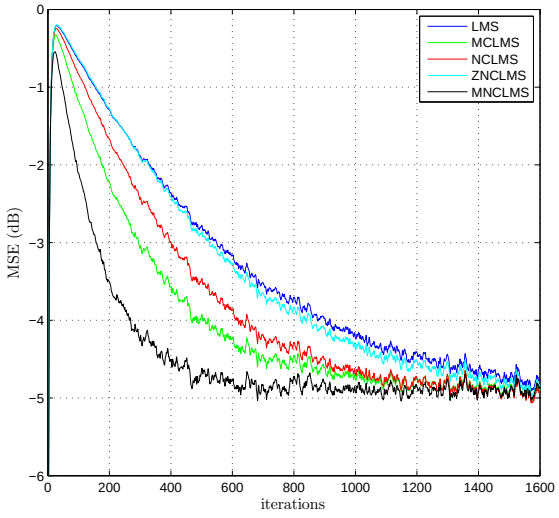


Fig. 4. MSE for competing algorithms for 6 users at SNR value of 10 dB.

Fig. 2 shows MSE of channel estimation for 6 and 12 users for different values of SNR. As shown, performance of channel estimator worsens if more users are added (from 6 to 12). This worsening is the result of MAI at the receiving end. At an SNR value of 17.5 dB, our designed channel estimator attains MSE of -7.7 dB for 6 users and MSE of 1.1 dB for 12 users. Channel estimates of Rayleigh fading channel are compared to that of actual random channel coefficients at 20 dB SNR in Fig. 3. As observed, our designed estimator tracks random channel coefficients effectively and efficiently.

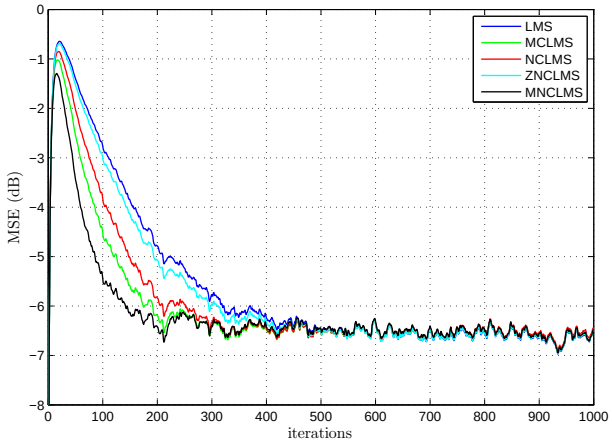


Fig. 5. MSE for competing algorithms for 12 at an SNR of 18 dB.

The proposed estimator is also tested with different constrained algorithms for the linear equalizer [12]. Simulation results in Fig. 4 and 5 show that proposed estimator is working efficiently as MAI and noise constrained algorithm is converging much faster than other constrained algorithms for linear equalization case when used with the proposed estimator.

V. Conclusion

In this paper, a pilot aided channel estimating technique for MIMO - CDMA systems is presented. This technique is utilizes minimum mean squared error estimation of corrupted information in a flat fading channel along with noise. Simulation results predicts theoretical predictions are strongly validated for different values of SNR and users.

References

[1] M. Salehi and J. Proakis, "Digital communications," *McGraw–Hill, New York*, 2008.

[2] J. S. Lee and L. E. Miller, *CDMA systems engineering handbook*. Artech House, Inc., 1998.

[3] E. Dahlman, B. Gudmundson, M. Nilsson, and A. Skold, "Umts/imt-2000 based on wideband cdma," *Communications Magazine, IEEE*, vol. 36, no. 9, pp. 70–80, 1998.

[4] S. Sampei and T. Sunaga, "Rayleigh fading compensation method for 16qam in digital land mobile radio channels," in *Vehicular Technology Conference, 1989, IEEE 39th*. IEEE, 1989, pp. 640–646.

[5] J. K. Cavers, "An analysis of pilot symbol assisted modulation for rayleigh fading channels [mobile radio]," *Vehicular Technology, IEEE Transactions on*, vol. 40, no. 4, pp. 686–693, 1991.

[6] A. Aghamohammadi, H. Meyr, and G. Ascheid, "Adaptive synchronization and channel parameter estimation using an extended kalman filter," *Communications, IEEE Transactions on*, vol. 37, no. 11, pp. 1212–1219, 1989.

[7] A. D'Andrea, A. Diglio, and U. Mengali, "Symbol-aided channel estimation with nonselective rayleigh fading channels," *Vehicular Technology, IEEE Transactions on*, vol. 44, no. 1, pp. 41–49, 1995.

[8] M. N. Iqbal and M. Moinuddin, "Pilot-aided rayleigh fading channel estimation using mmse estimator for ds-cdma system," in *Multitopic Conference (INMIC), 2011 IEEE 14th International*. IEEE, 2011, pp. 347–350.

[9] R. K. Morrow Jr and J. S. Lehnert, "Bit-to-bit error dependence in slotted ds/ssma packet systems with random signature sequences," *Communications, IEEE Transactions on*, vol. 37, no. 10, pp. 1052–1061, Oct 1989.

[10] M. Moinuddin, a. U. H. Sheikh, A. Zerguine, and M. Deriche, "A Unified Approach to BER Analysis of Synchronous Downlink CDMA Systems with Random Signature Sequences in Fading Channels with Known Channel Phase," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–13, 2008. [Online]. Available: http://www.hindawi.com/journals/asp/2008/346465.html

[11] K. Mahmood, S. Asad, M. Moinuddin, A. Zerguine, and S. Paul, "Statistical analysis of multiple access interference in rayleigh fading environment for mimo cdma systems," in *Statistical Signal Processing (SSP), 2014 IEEE Workshop on*, June 2014, pp. 412–415.

[12] K. Mahmood, S. M. Asad, M. Moinuddin, and S. Paul, "Design of MAI constrained decision feedback equalizer for MIMO CDMA system," *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–5, Nov. 2011.

[13] J. Hu, S. Member, and N. C. Beaulieu, "Accurate Simple Closed-Form Approximations to Rayleigh Sum Distributions and Densities," vol. 9, no. 2, pp. 109–111, Feb 2005.