International Conference on Data Science

Proceedings of the International Conference on Data Science

7-9 February, 2019
Karachi

Organized by

Mohammad Ali Jinnah University, Karachi

Proceedings of the

International Conference on Data Science 2019

7-9 February, 2019

Organized by:

Mohammad Ali Jinnah University, Karachi

Patron

- Dr. Zubair Ahmed Shaikh

Organizing Committee

- Dr. Asim Imdad Wagan (Chair)
- Dr. Shaukat Wasi
- Dr. Syed Imran Jami
- Dr. Tafseer Ahmed (Convener)

Technical Committee

- Dr. Abdul Samad, Habib University, Karachi
- Dr. Adnan Masood, UST Global, USA
- Dr. Arjumand Younus, Insight Centre for Data Analytics, Ireland
- Dr. Awais Athar, European Bioinformatics Institute, UK
- Dr. Chan Naseeb, KPMG, Netherlands
- Dr. Dost Muhammad Khan, Islamia University Bahawalpur, Bahawalpur
- Dr. Faisal Shifait, NUST, Islamabad
- Dr. Faraz Rasheed, Microsoft, Canada
- Dr. Faraz Zaidi, Region of Peel, Canada
- Dr. Hafeez Ur Rehman, FAST NU, Peshawar
- Dr. Jemal Abawajy, Deakin University, Australia
- Dr. Khurram Junejo, PAF-KIET, Karachi
- Dr. M. Abdul Rehman Soomrani, IBA, Sukkur
- Dr. M. Arshad Islam, Capital University of Science & Technology, Islamabad
- Dr. M. Atif Qureshi, CeADAR (the Centre for Applied Data Analytics Research), Ireland
- Dr. M. Imad Khan, Analytics & Artificial Intelligence, Australia
- Dr. M. Tanvir Afzal, Capital University of Science & Technology, Islamabad
- Dr. Mohammad Nauman, FAST NU, Peshawar
- Dr. Muazzam Siddiqui, King Abdulaziz University, KSA
- Dr. Muhammad Khurram, NED University, Karachi
- Dr. Muhammad Saeed, University of Karachi, Karachi
- Dr. Niladri Sett, Insight Centre for Data Analytics, Ireland
- Dr. Noman Islam, Iqra University, Karachi

- Dr. Quan Le, CeADAR (the Centre for Applied Data Analytics Research), Ireland
- Dr. Rauf Shams Malick, FAST NU, Karachi
- Dr. S. Saleha Raza, Habib University, Karachi
- Dr. S. Saqib Bukhari, German Research Center for Artificial Intelligence (DFKI), Germany
- Dr. S. Zafar Shazli, Juniper Networks, USA
- Dr. Sadaf Abdul Rauf, Fatima Jinnah Women University, Islamabad
- Dr. Saima Hassan, Kohat University of Science & Technology, Kohat
- Dr. Sajjad Haider, IBA, Karachi
- Dr. Shabbar Naqvi, Balochistan UET, Khuzdar
- Dr. Shahzad Cheema, IBM, Germany
- Dr. Shahzad Mumtaz, Islamia University Bahawalpur, Bahawalpur
- Dr. Shariyar Murtaza, TELUS Communications, Canada
- Dr. Sibt Ul Hussain, FAST NU, Islamabad
- Dr. Tahir Syed, FAST NU, Karachi
- Dr. Tahseen Jilani, University of Nottingham, UK
- Dr. Tariq Mahmood, IBA, Karachi
- Dr. Veronique Eglin, INSA de Lyon, France
- Dr. Yuichiro Kobayashi, Nihon University, Japan
- Dr. Zain Abbas, Scotiabank, Canada
- Dr. Zeeshan-ul-hassan Usmani, Data Scientist, USA

Table of Contents

Poster Papers

Short Term Authentication of ATM Customers through Characteristics of Body and Face

Muhammad Yasir Imam[1]
Department of Computer Sciences and IT
Alhamd Islamic University
Bhara Kahu, Islamabad, Pakistan
Yasirimam5110@gmail.com

Nabila Jannat[2]
Dept. of Computer Sciences and IT
Alhamd Islamic University
Bhara Kahu, Islamabad, Pakistan
Nabila.jannat@alhamd.pk

Farzana Bibi[3]
Dept. of Computer Sciences and IT
Alhamd Islamic University
Bhara Kahu, Islamabad, Pakistan
farzanasajawal12@gmail.com

***Abstract*—This research expresses the major framework for biometric, where the person is authenticated at an Auto Teller Machine (ATM), and has to be re-recognized from a camera within a very limited time, under very challenging clarification and pose situation, and utilizing data from a single session. The application is the automatic refund of forgotten card or money at an Auto Teller Machine, which occurs often, and reasons inconvenience public and reduction of financial gain for the banks. We present a multimodal verifying system that works under the constraints enforce by this application outline, and implement face identification and color based body characteristics identification to generate a system that make ATM behavior better in case of forgotten ATM card or money by re-recognize the customer from an implant ATM camera. We focus on the outline and the platform, and report tests with the present system under demanding terms, obtained from Auto Teller Machines putted in the field. To make the transmission more secure, we fix a camera in front of the ATM machine. Apart from biometric security, we can make a transaction by confirming the user's face through this camera. This will improve the security of ATM machine. That means we will also verify the face as well as biometric verification. And it will be very useful. This will not only save the account but will also satisfy the user and care user things.**

Keywords—Biometric and Face recognition, online banking, security, customer care, ATM security, Recover Forgotten things

I. INTRODUCTION

Banking systems have been increasing with the Development of information and communication technology. At present time, all banks want to decrease their basic structure amount by changing transactions of their customers to ATM and Internet services e.g. the websites. Investors mostly promote ATM for manual deals, like the money withdrawal or the money deposit. On these basis, Customer knowledge at the Automatic Teller Machine is a main concern for banks.

One of the most serious matters that ATMs machine from is (CMF) ATM card or/and Money forgetting, which is an unexpectedly usual case. In CMF, after completing the transaction process customer drop or forgets the ATM card or/and Money and left the system. After a short and limited period of time, these things will be devoured by the Automatic Teller Machine, and the customer has to go along a tedious and expensive operation to get back the ATM card/Money or have a new ATM card. Besides, Money is kept secure in a different box after CMF, and it requires a manually checking before that was back in circulation. It means that much of our time and money is being destroyed into handling with outcomes of CMF. Relevant statistics reachable from MasterCard, one ATM have 8-9 or may be more card returns and also 18-19 or more money returns per annum. In the regional or local city banking department, amount of ATM must be up to ten thousand or more, and the CMF extra cost for single country is US $ 1 million every year. The cost of CMF matters can be decreased if ATM allows to re-recognize users in short period of time which are coming back to Automatic Teller Machine to fetch those things which is forgotten by them. This needs the grip of things should delay, and next user are identified and checked whether the user is changing or no.

In this article, we present how ATM can pick this attitude using a camera-based system. In the current system, when a customer enters an ATM card in the ATM Machine and a session starts, the system begins identifying the face and the body appearance by operating the fix camera which is near the ATM and creates a non-permanent identity dataset for customer. If unfortunately user left ATM without picking her or his ATM card/Money, ATM Machine is waiting for user to come back, alternatively to get back the unremembered things. When detect that another user reached ATM, thus the items is retracted immediately. This outline differs to other biometric affirm outlines, where a person's photo is compared with a gallery photo obtained, may be, a long time ago comparing and under different situations. In this outline, the comparing photo and gallery photo both are divided in less than one minute at least. The basic problems in this outline the unrestrained and very low standard original world pictures, which is captured by the camera attached near ATM machine (mostly the face photos of the coming back ATM card holder/another reaching users), and also high lighting situations. Besides, the processing and decision formation should be done in very short interval of time-frame to be for real use.

The remnants of this article is arranged as following. The section 2 gives the summary of related work. The section 3 brings all entire system model design. The section 4 gives a detailed procedure, including face and body recognition and confirmation. The section 5 gives the experiments(test) and analyze the system working and performance. Lastly, the section 6 concludes this article.

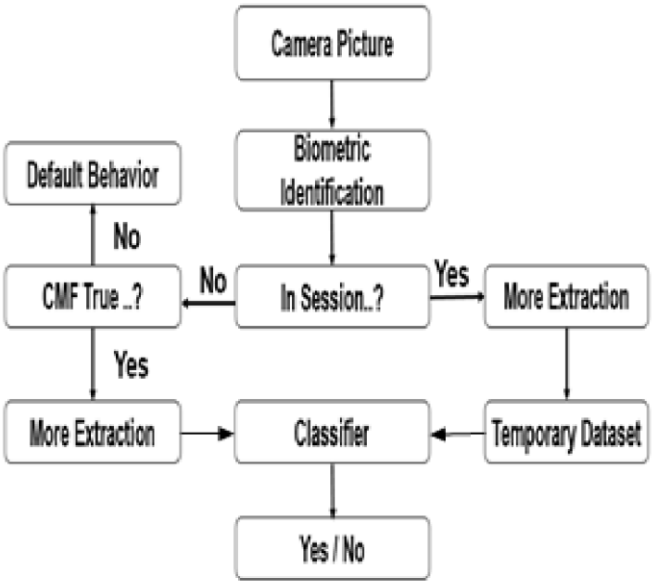


Fig.1 General Flow Chart of presented System

II. RELATED WORK

Since our attention is on the program allied demands of the recognition outline, we do not give a detail associated work in face recognition from video. While the face expo has been learned at large scale for biometric Objectives, it is relatively too less considered for ATM-based utilization. The basic problem is that ATM machines are as usually put outside, and work below at large scale alter lighting situations.

Babaei et al. proposed an idea for using face identification of ATM machine customers jointly with another biometric property similar to fingerprint, the iris identification and hand or finger geometry [1]. But, they failed in providing any kind of method for an authentic ATM machine customer outline. Peter et al. proposed face identification based way to upgrade ATM machine security [2]. In this presented system model, face identification executed on the motionless pictures of ATM machine customers, and compared with pictures which stored in gallery for making decision. Although, this proposal relies on an already created gallery, those which is usually includes pictures obtained much time earlier than the real apply, and under different lighting conditions.

In [3], Der man et al. presents idea for real CMF outline for face photos catch with a fix ATM machine camera. This effort was first original structure for focal point on CMF outline, but managed only for one biometric, a face and for identification. Because of hard high lighting situations in area, the face picture creates by its own self cheap outcomes. In this effort, we work on an alike appeal outline, but the use of profile face pictures rather front faces, and thus we also present merging the face and the body appearance identification outcomes. Our outline needs the beginning gain and recognition. To occur in very little time situations are usually steady, but not constant/manage.

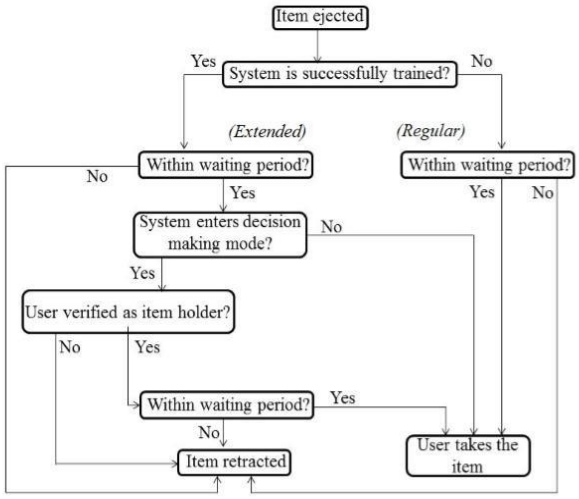


Fig. 2. All possible scenarios after an item is ejected by ATM.

III SYSTEM DESIGN

The system prototype relative to identified ATM & CMF outline presented in Diagram 1. In this work, therefore when a customer begins a links with ATM machine, 1st we gather pictures of customer from fixed ATM machine camera and next work on face recognition to cut face area out of the background. Following that step, an area of interest (AOI) is placed for exhibition of body appearance. We apply color based Interpretation work for this section. The AOI is respective relative to the face region. The system performs System monitoring whether the present condition "In-Session", that is, until the customer is still producing any kind of actions with the

ATM. While "In-Session", the system executes feature extraction to create a non-permanent face and body appearance information of the customer (i.e. owner of the ATM card) using property of recover face and body appearance pictures. On this short-term database, a classifier is trained. The classifier made decision mode just when it is completely instructed, that is, when it has minimum a sufficient unit of tests, in other case system model acts as a regular ATM machine. When "InSession" Condition is upgraded "No" that means customer has departed from ATM. If deal done usually, the non-permanent database is rejected. We pay attention here only those cases when customer has forgot to grab card/money after the deal/transaction. In this situation, the system model operation to the initial mortal reaching the system to compare them with the non-permanent database. Feature extraction are build on the latest "Out-of-Session" face and appearance picture. Which then Pull out features and pass it to the classifier. If the last outcome indicates that "Out-of-Session" the pictures belong to card owner, after that ATM machine wait rather of get back the card or money. Other hand, system determines that new customer is to be in services, and ATM machine instantaneously fetch the card/Money for the protection. The present ATM system, when experience with a CMF event, waits for an already decided time period (as usually 15-25 seconds) to get back the forgotten things. In our system model, we increase this time period to 25 seconds and after that give ATM to fetch the things for security. Although, if another mortal reaches the

ATM machine within this time limit, our system model moves into the decision building mode. Likely outline can be described as shown in the Figure 2 [11].

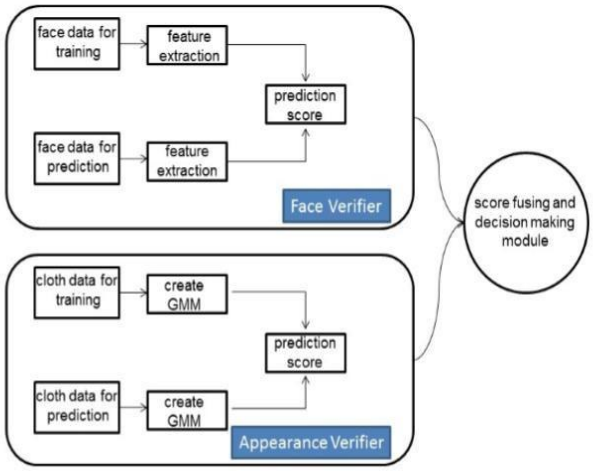


Fig. 3. Score level fusion of the two different modalities

IV. METHODOLOGY

4.1. CONFIRMATION SYSTEM

The confirmation system depends on training and forecast methods. It hires a face confirmation model, and also an appearance-based model, whose outcomes are fused at the comparing score level (as shown in Figure 3 [11]). We investigate our systems working with terms of the false/wrong acceptance rate (FAR), the genuine acceptance rate (GAR) and the receiver operating characteristic (ROC) curves, which is shown in graph.

4.2. FACE RECOGNITION

Face observation on the clicked photo is executed using OpenCV built-in Haar cascade profile photo face detector[1]. Since the profile photo face detector is only instructed for left side profile photo faces, the clicked frame should be flipped as according to the real condition. Considering the reality that the Automatic Teller Machine user's head location may differ (as shown in Figure 4), to at most the face observation rate, we execute a rotation on the clicked frame.

A) Definition of Variables

There Y is the number of faces observed, θ is angle of rotation, n is face area, Z_0, Z_1, Z_2,…., Z_{n1} sub areas, S is forecast and foreground score, α is the sigmoid factor, D the signed distance by SVM, A&B already defined parameters, N is the foreground pixel number, Ex, Ey the real coordinate (point E) as cloth area, d the set parameter and SFV, S_{CV}, W_{FC}, W_{CV} the fused final score, cloth and face identification score, weight of face and cloth identification.

The most rotation angle for at most the face identification rate is getting as:

$$Y_{max} = arg_{\theta bestmax} Y(\theta best) \ldots\ldots\ldots\ldots (1)$$

Where Y shows the number of faces observed under rotation angle at θ, and possible values are in 5 additions are judge for this kind of angle between 0−40. This stage is repeated for every frame in both tutoring and confirmation stages, such that using the at most rotation angle is regularly assured.

Fig.4 Customers with different kind of head position

Fig.5 Right: flip and rotate photo by utilizing by best angles Left: Real photo

4.3. PREPROCESSING

We operate various preprocessing operations. The clicked RGB photo first changed into grayscale. The profile photo face identifier works with some edge and incorporate background, which is rejected with a fixed mask. The remaining photo is more

[1] http://www.opencv.org

smoothed with a Gaussian filter and changed into 64x64 pixels. At last we execute a histogram equalization.

4.4. FACE IDENTIFICATION

1) More extraction: After preprocessing stage, LBP (Local Binary Pattern) mechanism is executed for as face descriptor [4, 5, and 6]. Here, we apply the uniform binary-based (LBP) with the eight examining objects within one-pixel region. After that, we present face area as n subareas Z0, Z1... Zn−1, and calculate histogram separately within one of these origins. At last, we merge following n histograms for use to produce horizontally better histogram vector.

2) Classification: We apply SVM (Support Vector Machines) as our classifier [8]. More vectors of the Automatic Teller Machine customers "In-Session" photos are operated as positive trails, while more vectors of face photos got previously are operated as negative trails for classifier instruction. The nonpositive samples can be kept on the system. The instructed sample size must meet the already defined least sample number to successfully instruct the classifier. "Out-of-Session" photos of the Automatic Teller Machine customer are operated for the forecast step. The measured signed distance value is moved into a score between 0 and also 1 using a sigmoid function [9]:

$\alpha = D * A + B$ …… (2)

$S = \frac{1}{1+\exp(-\alpha)}$ …….. (3)

There S is the forecast score in equation 3and 4. There α indicate the sigmoid factor, D indicates the got signed distance by SVM after forecast, A and B are already defined parameters. We get values of A and B as −5.2 and−0.5 by operating a nonlinear regression method [10].

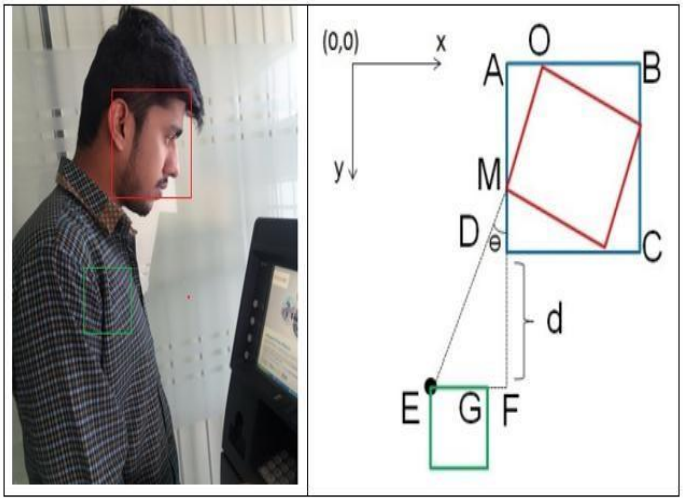


Fig.6 Choosing the appearance position by utilizing parameters of face identification model

4.5. APPEREANCE IDENTIFICATION

In a usual CMF circumstance, we imagine to see the customer come back the Automatic Teller Machine with the same body appearance as in the transaction stage, since mostly at least one minute is gone between the accession of the gallery photo and the test photo. In the preparing stage, the system memorizes the background data of the appearance area (of size 30x30 pixels) by operating a GMM (Gaussian Mixture Model). Through the forecast stage, the system judges whether the present body appearance area is related to the background or not. This outcome can more be changed into a forecast score between the zero (0) and one (1) by:

$S = 1.0 - N_{foreground}/N_{total}$ …… (4)

There S, N foreground and N total refer to foreground pixel number, forecast score, and total pixel number, respectively.

Choice of a correct body appearance area is (both during teaching and during working of the system) carried out as follows (Figure 6). We first get the customer's face area (red rectangle), its enclose box (blue rectangle) and the best rotation angle is θ. Then the points (Ex, Ey) of the real coordinate (point E) of the cloth area (green rectangle) are tally:

$E_x = A_x - [(D_y - M_y) + d]\tan(\theta)$ …….. (5)

$E_y = A_y + AD + d$ …… (6)

There, d is an already-set parameter. Operating validation set outcomes on original information; we empirically set d values as:

$$d = \begin{cases} 450 & if\ Ey \le 450 \\ 450 - Ay - AD & if\ Ey > 450 \end{cases} \ldots\ldots (7)$$

4.6. FUSION

The scores get by face and body appearance identification are mixed together by the Sum Rule [7]. Scores of non-identical methods are set a weight value between the 0 and the 1. This mixture form can be shown as follows:

$S_{fused} = S_{FV} * W_{FV} * S_{CV} - W_{CV}$ ……(8)

$W_{FV} + W_{CV} = 1$ …….. (9)

$W_{FV}, W_{CV} \in R^p, 0 \le W_{FV}, W_{CV} \le 1$ …(10)

$S_{FV}, S_{CV}, S_{fused} \in R_p, 0 \le S_{fv}, S_{CV}, S_{fused} \le 1$ ….(11)

in which, S fused, S_{FV} , S_{CV} , W_{FV} and W_{CV} shows the final fused score, the face identification score, the cloth identification score, the weight of face identification, and the weight of cloth verification, respectively. The final fused score, S fused, is matched with an already defined approach value, h. If it is more than h, classifier considers its related customer as comparing identity, or else and as fake.

V. EXPERIMENTAL DIAGNOSIS

5.1. DATABASE COLLECTION

We build our own database, which depends of original Automatic Teller Machine conversation sessions saved from 57 subjects.

These subjects were given an already-defined outline as follows: First of all, the customer begins conversation with the Automatic Teller Machine transaction cover cash returns. When customer withdraws money, she or he left the Automatic Teller Machine without taking the card or money. Then beside very little time, customer understand that card or money was lost or unremembered, and comes back to Automatic Teller Machine to get back card or money. The at most time for system model to Worth to Deposit "Out of-Session" the photos for 1customer at least 2 second, when customer comes back to collect the unremembered card or money Early first entering in the camera radius. And also, photo size and photo quality of

"Out-of-Session "photo are very cheap quality, form to distance in the middle of customer and Automatic Teller Machine camera changes. This movement blur generates largest dare for customer face affirm works in this case, and this is major cause why this outline is notably differed differentiate for a normal face affirm outline (as shown in Figure 6). Now, we only think about cases with a lone Automatic Teller Machine customer. New information is gathered for the case of more than two customers. In the number for "In-Session" photos are 57 subjects of our dataset differ between 1 and 1000, consisting on the time period they spending for customer transactions and customer head pose, while that of "Out-of-Session" photos differ between

1 and 29, consisting on their time come back to the Automatic Teller Machine, and time which take for collecting unremembered money or card etc. Some customers may be instantly collect their cards or money, while other customers may take more time. The both genders are involved in this observation. The highness of this test objects is in the middle of 159 cm and 190 cm, as that different customers with differ heights are observed. It's compulsory, because camera is fixed, high and short subjects can give unusual cropped face photos.

5.2. EVALUATION STEPS

Authentic and fake scores are generated for both (face and body appearance identification) modules. Later, these scores are operated to make final decision. We determine non-identical limited values for obtaining FAR and GAR values to get ROC curves on graph form, which point out the system's working limitation, to judge each module's independent and fused working. The account speed of our presented system on this Automatic Teller Machine is 15 fps (frames per second), which meets the processing needs for timely reactions.

5.3. EXPERIMENTS AND TEST SCENES

1) Two class classification: We distribute the "Out-ofSession" photos into two segments, such as negatives (impostors) and positives (matching identities). We tested our presented model on our database of 57 objects/subjects. Module weight $W_{F\,V}$ is selected as

   0.7, GMM to detect background uses 10 clusters, the cloth region size is set to 50x50, minimum number of photos for the positive and the negative classes is 100 for training, and maximum number of photos are 300 for positive class (gain from camera during session), and 600 for negative class (saved for offline). During supplying negative patrons for each customer, we select a total of 26 other persons "In Session" the face photos for training, and remaining 26 other persons "Out-of-Session" face photos are utilized for (negative) impostor accesses. The outcome ROC curves of GAR and FAR values are as shown in Figure
   8.
2) "In-session" vs. "Out-of-Session": The standard of the "Out-of-Session" photos are very cheap compared to those of "In-Session" photos. We tested the algorithm for "In-Session" photos which are not utilized in the training process phase to observe the effect of photo standard on algorithm working. The related ROC curves are shown in figure 9.

   The fused score threshold h is determined by experimental outcomes and utilized to accept or reject the predicted new photo. If model only utilizes face identification and operates at a FAR value of 5%, we obtain a GAR value of 81.5%. If the model utilizes both face and body identification with a hopeful FAR value of 5%, we obtain a GAR value of 91%. The pleasing aspect of the CFF outline is that there can hardly be any deliberately attacks on this model, as the frequency of CFF is low. It is not possible for a negative (impostor) to wait until someone forgets card/Money at an ATM. Therefore, a much higher FAR than a simple bio-metric outline is acceptable in this model. From these outcomes, we can observe that fused model gives good performance than any single system of the face or body identification. For our ultimate use, rely on our supposition of FAR and GAR, we can choose the real threshold value from these outcomes. In our case, the operation range allows a FAR around 5% and for a GAR around 90%. During this, by comparing outcomes of Figure 8 with that of Figure 9, we see that the growth in photo standard outcomes in a growth in the system performance. The performance outcomes for good standard "In-Session" photos point out that the presented model gives acceptable outcomes when tested photos have enough high standard. To evaluate our planning requirements during our experiments, in practice, we used one Traditional system for ATMs operating in the field. Especially, the ATM used for our test is an Intel Core 2 Duo 3.00GHz CPU and 2GB RAM, using the Windows XP Professional Version 2002 Service Pack 3 or higher as the operating system. Our proposed system counting speed on ATM is 15 frames per second (FPS), which meets processing need for timely reply.

5.4. ANALSIS OF FAILURE CASES

Face identification is essential for our presented system model, and the system fails when we cannot successfully get face pictures for associated sessions. In such cases some potential causes are encountered in which the user is covered by hair or clothes (such as shown in Figure 7). In these conditions, our presented system is ignored in case of any CCF events and restart a retreat outline of normal time-out-based Remove, which is how the unlimited Automatic Teller Machine Security. This means that if the proposed system fails, the default fake cannot be believed to mimic the Original customers in CCF. Any kind of up gradation in true positive rate straight convert and money rescue for banking organization, and poorly case (0 right positive rate) relevant to what

present Automatic Teller Machine systems have. For future work, we have idea to upgrade the photo the quality and resolution to reduce the (-) effects of movement blur during putting within the computational boundaries imposed by the Automatic Teller Machine system. Besides that, other facing challenges are multiple customer's faces in the scene are possible research instructions. Fusing data from the Automatic Teller Machine camera own self is possibly helpful, but in practice, Automatic Teller Machine have various camera places.

Fig7. Failure Situation: Face Position hide by garment s or other physical appearance

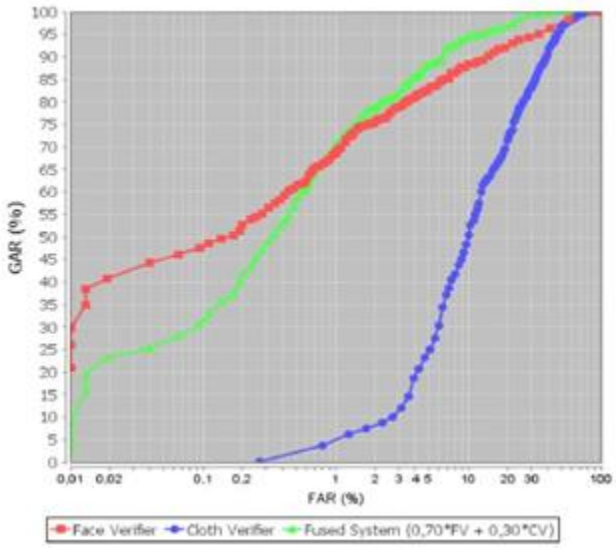


Fig. 8 ROC curves of presented system

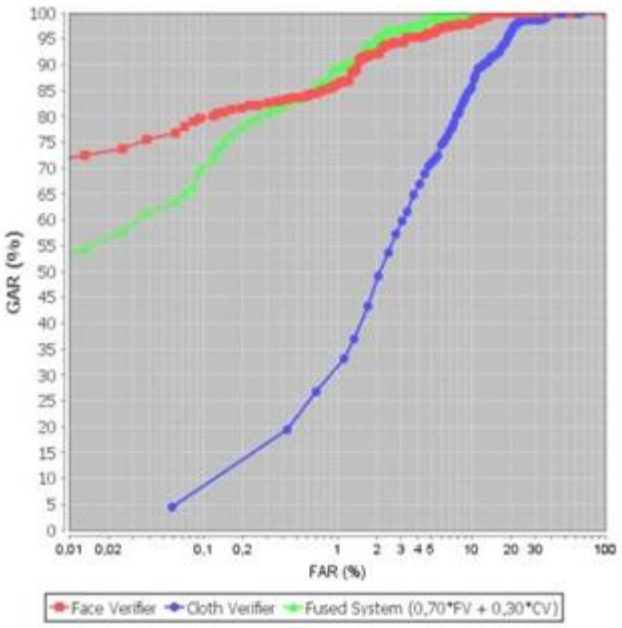


Fig. 9 ROC curves utilizing "In Session" photo for forecast

VI. CONCLUSION

In this work, we present a computer vision based Automatic Teller Machine customer verification framework using face and body appearance identification to decrease card and money retraction. Our proposed system also increases the ATM security. We Overview the proposed system under different situations, and with our own dataset, based on an original outline. The experimental outcomes obviously our presented model promise that to reduce Money/card unremembered problem.

Our "Out-of-Session" observation situation is 1 that is nearest to imagine original world program. This is further important for this program to hold higher true positive rate (Facility) as the behavior is displayed. Any successful delivery of the card or money to the customer (original or fake) can be logged.

REFERENCES

[1] H. R. Babaei, O. Molalapata and A. A. Pandor, Face Recognition Application for Automatic Teller Machines (ATM), in ICIKM, 3rd ed. vol.45, pp.211-216, 2012.

[2] K. J. Peter, G. Nagarajan, G. G. S. Glory, V. V. S. Devi, S. Arguman and K. S. Kannan, Improving ATM SecurityviaFaceRecognition,inICECT,Kanyakumari,2011, vol.6, pp.373-376.

[3] E. Derman, Y. K. Gecici and A. A. Salah, Short Term Face Recognition for Automatic Teller Machine (ATM) Users, in ICECCO 2013, Istanbul, Turkey, pp.111-114

[4] T. Ahonen, B. Hadid and M. Pietikainen, Face Description with Local Binary Patterns: Application to Face Recognition, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28,pp.2037-2041,Dec. 2006.

[5] T. Ojala, M. Pietikainen and D. Harwood, A Comparative Study of Texture Measures with Classification Based on Featured Distributions, in Pattern Recognition, vol. 29, pp. 51-59, Jan. 1996.

[6] T. Ojala, M. Pietikainen and T. Maenpaa, Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 971987, Aug. 2002.

[7] A. Ross and A. Jain, Information Fusion in Biometrics, in Pattern Recognition Letters, vol.24, pp.2115-2125, 2003.

[8] C. C. Chang and C. J. Lin, LIBSVM: A Library for Support Vector Machines, in ACM Transactions on Intelligent Systems, 2:27:1-27:27, 2011 Available: http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf

[9] B. Zadrozny and C. Elkan, Transforming Classifier Scores Into Accurate Multiclass Probability Estimates, in ACM SIGKDD 2002, pp. 694-699.

[10] G. A. F. Seber and C. J. Wild, Nonlinear Regression, Hoboken, NJ: WileyInterscience, 2003.

[11] G. Kayım, E. Derman and A A. Salah Re- configuration of automatic teller machine with Body & Face 978-1-4673-9448-2- 2016 IEEE

Corpus Construction and Structure Study of Urdu Language using Empirical Laws

Nuzhat Khan
Department of Information Technology
Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS)
Quetta, Pakistan
salafshara1416@gmail.com

Muhammad Paend Bakht
Department of Telecom Engineering
Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS)
Quetta, Pakistan
paend.bakht@gmail.com

Raja Asif Wagan
Department of Information Technology
Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS)
Quetta, Pakistan
raja.asif@buitms.edu.pk

***Abstract*—This paper presents its findings about statistical analysis of Urdu language structure. The proposed method is useful to uncover the composition pattern of Urdu language stored in human brain. Two empirical linguistic laws Zipf's law and Heap's law are tested for Urdu with help of natural language text written in its correct genuine script ''Nastalik'. The frequency table is constructed from Urdu corpus, specially collected for this research. Application of both laws on Urdu language is tested with the help of frequency table and logarithmic plots using Python. Statistical facts are formed by analyzing each word, word's frequency, ranks of words, rank-frequency relationship and high frequency words ratio in the language. It is concluded that Urdu language obeys zip's law and Heap's law like other previously studied languages.**

Keywords—Zipf's law, Heap's law, empirical laws for Urdu, Urdu Structure, Urdu statistical analysis, NLP

I. INTRODUCTION

Languages are composed of random objects with some hidden connection rules/patterns that make comprehensible structure. The hidden composition of language reveals its paradigm statistically. In machine learning and artificial intelligence languages are modeled using these mathematical facts to design reliable systems equipped with linguistic skills. Progressive tools for linguistic tasks in English language are result of deep structural study. Statistical study of human languages helps in finding the similarities and differences amongst languages to develop linguistic featured tools [1]. Human language is not simply a natural way of communication when it comes to learning language structure and patterns instead of using language for simple communication. It converts to an object full of wonders when stepping out from the human language for looking into its deep structure. Language can be explored on many levels from phonemes to phrases. Most significant structural analysis is held to study the hidden patterns of language in the form of general composition of language and grammatical rules of combining and contrasting its components [2]. Natural language processing (NLP) is one of the high-tech methods to explore natural language text with blend of concepts and technology. There are multiple definitions, levels and fields of NLP. In the domain of NLP, language structure detection is directly linked with Artificial Intelligence (AI) for modeling the natural language in machine understandable form. NLP is the immature form of Natural Language Understanding (NLU). Scientists' efforts are in the way to proceed the NLP to NLU for future intelligent systems. These systems are supposed to be capable of performing multiple linguistic tasks with multilingual features [3]. Reliable systems are likely to overcome the language barriers in this global village of numerous thousand living languages [4]. Zipf's law and Heap's law are two most popular linguistic laws for language structure analysis [5] [6] [7]. Zipf's law is also known as power law [8]. These laws explain two scale dependency among language components [9]. According to Zipf's law, rank of a word starts decreasing when its frequency tends to increase in a large set of natural language text. If we denote the rank of a word with 'R' and the frequency of word with 'f' in text, then the rank and frequency relationship is presented by Equation (1).

$$R(f) \sim f^{-\beta} \tag{1}$$

where $\beta>0$ [10]. While Heap's law is about corpus size and unique terms relationship. Heap's law simply explains the ratio of unique terms according to its vocabulary size. Arrival of new unique terms starts decreasing in the vocabulary on adding more text in corpus. Size of vocabulary and number of unique terms depend on each other in opposite manner [7, 11, 12]. If we denote number of unique terms (types) with 'N_u' and text length (tokens) with 'N_t' then according to Heap's law 'N_u starts decreasing on increasing N_t'. Type token relationship is given by Equation (2) [13].

$$N_t \propto k\,(N_u)^{-\gamma} \tag{2}$$

Heap's law and Zipf's law have been examined for various languages like English, Italian, Spanish, French and many more [14-17]. For English and Russian these laws coexist with different exponents but Chinese, Korean and Japanese don't confirm these laws [8]. Urdu is a widely spoken language with a treasure of awesome literature in the form of text books. This is national language of Pakistan and official language of many Indian states. Urdu is ranked as 20th largest language of world with almost 64 million

people speak it as their first language and more than 104 million people speak Urdu as second language [18, 19]. Urdu writing script is unique in the way that all characters change their shape according to position in the word. Urdu characters are different from standard ASCII characters as they have no upper or lower cases. Urdu is written from right to left and follows Unicode UTF-8 standard. Due to its unique script, uncommon standard and lack of electronic text samples, Urdu remained unexplored and is unable to meet current requirement by AI and NLP tools and applications [20]. Urdu language has multiple forms for one word and its difficulties in stemming/lemmatization make it a challenge for researchers [21, 22]. To overcome the lack of corpus, many efforts have been made [23, 24] but still no standard corpus is publicly available. For Urdu language structure analysis, large amount of Urdu text was collected in electronic form. The text is combination of various books including Quran and Bible translated in Urdu [25]. Corpus also includes large set of Urdu news prints. Therefore, effort is carried out to clean the text by removing non-Urdu words, characters, symbols, numbers, misspelled words, alphabets and text of other languages written in Urdu script etc. The most important cleaning step was removing incorrect spacing between Urdu characters, as there is no standard system of spelling and grammar correction currently offered for Urdu language. After pre-processing, an open and unannotated corpus is constructed which contain 3084039 words in Nastalik script and it is named as Urdu Language Corpus (ULC). Roman Urdu [26] text was avoided to manage a well-organized, standard and authentic corpus based research. Unlike some previous researchers utilized roman Urdu text samples to overcome unavailability of machine readable Urdu text in Nastalik script [27]. Main objective of this research is to analyze Urdu structure with help of universal laws. This work will contribute in Urdu structure discovery which has been done for other languages [27, 28]. We tried to provide basic information about words, frequencies, ranks and composition pattern of Urdu language. Furthermore, this work has performed fundamental quantitative analysis of most common words in Urdu language.

II. CORPUS TEXT COLLECTION

For this purpose, ULC was collected from multiple sources which include two religious books Holy Quran and Holy Bible translated to Urdu from Arabic and Hebrew respectively, various other books on different topics written in Urdu language and Urdu newspapers. To add diversity in the corpus, text was collected from various fields such as religion, sports, health, crimes and poetry. Variation in sources of text facilitated to capture maximum Urdu words. The assortment of text is foundation of open corpus [29] as well. An Open and unannotated corpus [30] is composed of numerous individual raw text files as shown in Figure 1.

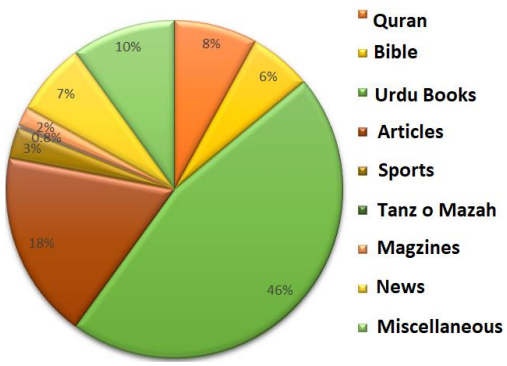


Figure 1. Text resources

Parts of corpus along with token size (number of words) and percentage of each part in total text are shown in Table 1.

Table 1. Text Collection

Text	Tokens	Types	Percentage
Quran	269991	8027	8
Bible	203927	8995	6
Urdu books	1528659	96862	46
Articles	588978	44396	18
Sports	105578	9924	3
Tanz o Mazah	6835	2223	0.2
Magazine	48157	8164	1.8
News	256000	45750	7
Miscellaneous	305696	15643	10

III. CORPUS CLEANING PROCESS

Text cleaning steps have been carried out on raw text. To keep only correct Urdu words, all non-Urdu characters were removed. Because raw text was comprised of uppercase and lowercase alphabets, numbers, mathematical signs, currency symbols, dates, page numbers, punctuation marks and a lot of text in Arabic and local languages. Non-Urdu characters of all kinds were removed step by step and only Urdu alphabets (huroof e tahajji) were captured. Afterwards, misspelled words and incorrectly spaced words were removed. Words with incorrect spacing took two steps to clean. Words were sliced according to the given space. Every single character was removed automatically, and remaining part was included to frequency table. Then it became easy to capture incorrect words due to their least frequency in the table. Example of cleaning process for incorrect spacing is explained in Table 2. It describes detailed cleaning process of correct words containing inappropriate spaces with help of five columns. First Column contains wrongly spaced words, second column holds single character after slicing and third column showing big part of fragmented word. Column four and five describe process of removal for both fragments respectively.

TABLE 2. TEXT CLEANING PROCESS INCORRECT SPACE

Word	Part A	Part B	Auto Remove	Manual Remove
ا ردو	‘ا’	’ردو‘	‘ا’	’ردو‘
احتسا ب	‘احتسا’	’ب‘	‘ب’	’احتسا‘
ا ستعمال	'ا'	'ستعمال'	'ا'	'ستعمال'
استعما ل	'استعما'	'ل'	'ل'	'استعما'
خا ک	'خا'	'ک'	'ک'	'خا'

Urdu Language Corpus (ULC) contains 3084039 tokens and more than 78750 types (unique words) after cleaning process.

IV. EXCEPTIONS IN URDU LANGUAGE

Urdu language is apparently different and hard to explore because of some exceptional features. It is observed that there are various acknowledged and highly adopted words of other languages that are included in Urdu vocabulary. These terms are accepted as part of vocabulary for being in form of proper nouns or due to common approval. The Urdu natural language vocabulary tends to continuously get rich because of these ‘immigrant’ words. These words are included in the corpus. Multiple identical words also exist in Urdu language. Unlike pair of words, these words don’t vary in terms of spelling or character combination. The identical words are called homographs [31, 32].

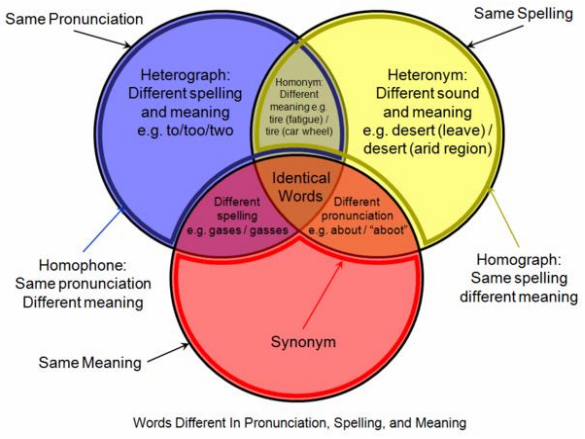


Figure 2. Homographs.

Homographs can be identical or unidentical in terms of pronunciation. Identical words with identical pronunciation are called homonyms and identical words with unidentical pronunciation are named as heteronyms [33]. Most of Urdu words become identical when written without diacritics and some are identical even with diacritics. Some of these identical words are not distinguishable even through pronunciation. So, the homographs are included as single body in the corpus. Some identical words are mentioned in mentioned in Table 3.

TABLE 3. HOMOGRAPHS IN URDU

Word1	Meanings	Word2	Meanings	Type
اس (oos)	that	اس (ees)	this	heteronym
پر (par)	on	پر (pur)	fill	heteronym
تو (to)	then	تو (tuoo)	you	heteronym
کیا (kiya)	what	کیا (kiya)	done	homonym
بس (bus)	authority	بس (bus)	stop	homonym
بھیڑ (bheedh)	crowd	بھیڑ (bhair)	sheep	heteronym
میں (mai’n)	me	میں (me’n)	in	heteronym
کان (kaan)	ear	کان (kaan)	Mine/s (salt or gold)	homonym
سو (so)	sleep	سو (so)	then	homonym
سو (sou)	hundred	سو (soo)	direction	heteronym

V. RESULTS AND DISCUSSIONS

To explore Zipf’s law and Heap’s law, frequency of each term is calculated individually. Rank is assigned to each word according to how frequent a word appeared in the corpus**.** Word with highest frequency was assigned lowest rank starting from 1. While second most frequent word received rank 2, third most frequent word established rank 3 and so on.

Table 4. WORDS, FREQUENCY, RANK

Word	Frequency	Rank
کے	110119	1
کی	86193	2
اور	79911	3
میں	75856	4
ہے	73044	5
سے	61307	6
کا	47081	7
اس	46510	8
کو	45059	9
نے	39488	10

Table 4 shows that frequency and rank are inversely proportional, as Zipf’s law states. According to Table 4, most frequent words do not appear with equal frequency in the text. Some words are observed with high frequency and most are with low frequency. Large part of corpus consists of high frequency words where low frequency words are small portion of the language. Some words appeared less than 20 times, most of words are part of the corpus but with low frequency, equal to or less than hundred. First most frequent term کے is more than 3% of the corpus with highest frequency 110119 times appeared in text. Table 5 shows 20 most frequent words with their frequency and percentage in total text.

Table 5 MOST FREQUENT WORDS WITH PERCENTAGE

Word	Frequency	Rank	Percentage in Text
کے	110119	1	3.57060984
کی	86193	2	2.794809015
اور	79911	3	2.591115093
میں	75856	4	2.459631671
ہے	73044	5	2.368452539
سے	61307	6	1.987880179
کا	47081	7	1.526601966
اس	46510	8	1.508087284
کو	45059	9	1.461038593
نے	39488	10	1.280398854
کہ	36587	11	1.186333895
ہیں	30734	12	0.996550303
پر	30699	13	0.995415428
بھی	27626	14	0.895773367
کر	23661	15	0.767208197
ان	21387	16	0.693473721
وہ	20954	17	0.679433691
یہ	20774	18	0.673597189
تو	20731	19	0.672202913
نہیں	19320	20	0.626451222

First 200 most frequent words are 52.62096232 % of total text in the corpus and 500 most frequent words are 62.83393303% of the corpus. Table 6 describes that small number of high frequency words are large part of the corpus.

Table 6. WORDS AND THEIR PART IN ULC

Number of High Frequency Words	Percentage in Corpus
200	52%
500	62%
1000	71%
1500	76%
2000	79%
3000	84%
5000	88%
6000	90%
7000	91%

Urdu also confirms 80/20 rule of Pareto distribution which directs that ’20% most frequent words are 80% of the language’[34]. In Urdu language less than 10% of most frequent words are more than 90% of total text in the corpus. This feature makes Urdu language learning easy for new speakers. With memorizing only 10% of vocabulary, and some loose grammatical order, Urdu learning will become faster. The similar approach can be applied on machine learning and auto text generation. For machine learning, high frequency words can be stored according to their frequencies. Zipf’s law is also verified with log plot curve against frequency and ranks of first 3000 high frequency words in ULC. Log plot of Urdu words and ranks confirms Zipf’s law. In Figure 3, the curve clarifies frequency rank relationship in Urdu language.

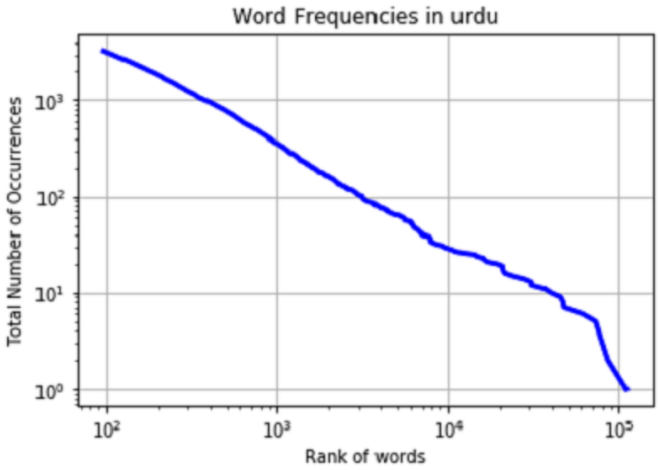


Figure 3. Log plot of frequency and rank

In logarithmic plot of frequency and rank horizontal axis shows ranks and vertical axis displays frequencies or number of occurrences of words in the text. In this log plot of more than 3000 high frequency words, Slope of the curve is -1.00049926628. It is nearly equal to ideal Zipf’s slope of curve which is -1[35]. The resulting curve of Urdu language is comparable to Arabic log plot of words [36] and frequencies with curve shape in Figure 4.

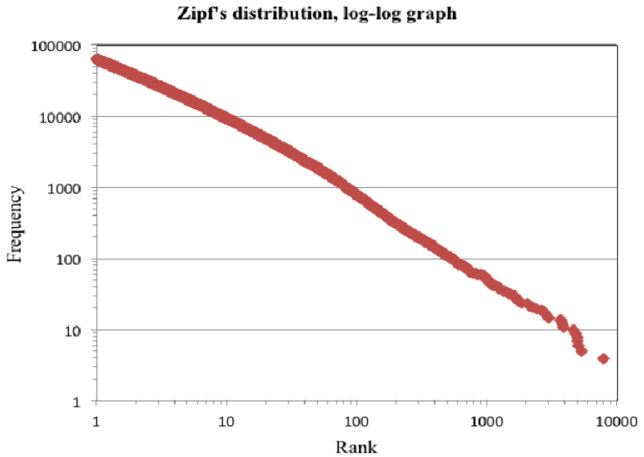


Figure 4. Log plot of Arabic language

On analyzing the curve of frequency-rank it is confirmed that Urdu language is a systematically evolved language

with some pre ordered structure. ULC has been tested for Heap's law for type token relation. The dependency of new terms (words) coming on size of vocabulary is verified by calculating number of words and number of unique words(types) in specific vocabulary proportions. For small set of tokens, existence of types is high in the corpus. The resulting graph of type token relation shows ratio of new terms encounter on increasing vocabulary size in Figure 5. Vocabulary size is on horizontal axis with label 'Number of Tokens' and the number of unique words depending on vocabulary size are located vertically with label 'unique Tokens'.

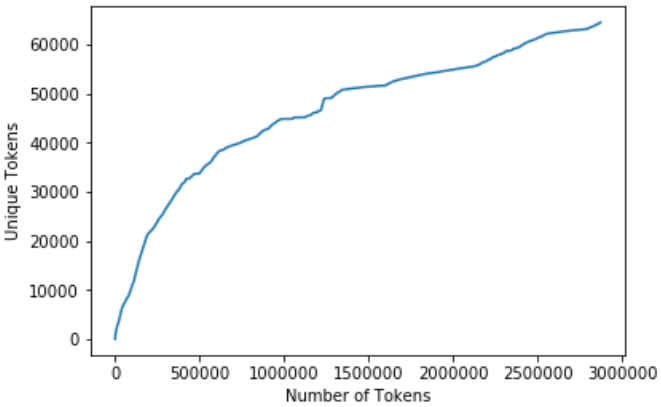


Figure 5. Type token dependency

It is clear from the curve on plot that ratio of new words growth is getting slower on increasing vocabulary size. The types are outwardly growing more in the beginning and started decreasing in the middle. The curve pattern showing constant decline of types on replication of tokens in the vocabulary. When size of vocabulary is less than 500k, the number of unique words is more than 10k and rapidly growing with vocabulary as on 500k the number of unique words is more than 30k but on next phase vocabulary is doubled, tokens are 1 million and types raised by 10k only. The vocabulary has been improved to 3 million words and number of unique words are only 60k. This relation confirms implementation of Heap's law of inverse type token relation on Urdu language. From the curve pattern it can be concluded that on some point, when number of words cover full vocabulary, unique term arrival is approximately extinct.

VI. CONCLUSION

From frequency table, word frequency-rank dependency curve, logarithmic plot of frequency-ranks and logarithmic plot of types-tokens, it is concluded that Urdu follows Zipf's law and Heap's law. Urdu is observed a systematically evolved language with some hidden patterns. These patterns can give deep understanding of Urdu structure on statistically revealing the composition of Urdu language. Various tasks in fields of linguistics and Artificial Intelligence such as artificial language generation, Urdu learning for human/machine, preferential text storage/ retrieval etc. can be achieved after getting enough statistical information of language elements, its composition patterns and probability of a word occurrence in the language. It is significance of Urdu language structure that it is composed of numerous elements with some specific predictable and statistically measurable order. It provides deep insight into language with help of universal laws.

VII. FUTURE WORK

Construction of annotated Urdu corpus with part of speech tagging will contribute more to model Urdu language which can intelligently differentiate words with/without diacritics. In depth statistical analysis of Urdu is vital to find exponential values of type token and frequency rank dependency ratios for comparing Urdu with other languages.

REFERENCES

[1] Lin, R., Q.D. Ma, and C. Bian, Scaling laws in human speech, decreasing emergence of new words and a generalized model. arXiv preprint arXiv:1412.4846, 2014.

[2] 2. Akmajian, A., et al., Linguistics: An introduction to language and communication. 2017: MIT press.

[3] 3. Liddy, E.D., Natural language processing. 2001.

[4] 4. Katzner, K. and K. Miller, The languages of the world. 2002: Routledge.

[5] 5. Zipf, G.K., Selected studies of the principle of relative frequency in language. 1932.

[6] 6. Gelbukh, A. and G. Sidorov. Zipf and Heaps Laws' coefficients depend on language. in International Conference on Intelligent Text Processing and Computational Linguistics. 2001. Springer.

[7] 7. Heaps, H.S., Information retrieval, computational and theoretical aspects. 1978: Academic Press.

[8] 8. Lü, L., Z.-K. Zhang, and T. Zhou, Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. Scientific reports, 2013. 3: p. 1082.

[9] 9. Fernholz, R.T. and R. Fernholz, The Universality of Zipf's Law for Time-Dependent Rank-Based Random Systems. arXiv preprint arXiv:1707.04285, 2017.

[10] 10. i Cancho, R.F., The variation of Zipf's law in human language. The European Physical Journal B-Condensed Matter and Complex Systems, 2005. 44(2): p. 249-257.

[11] 11. Takahashi, S. and K. Tanaka-Ishii, Do neural nets learn statistical laws behind natural language? PloS one, 2017. 12(12): p. e0189326.

[12] 12. Serrano, M.Á., A. Flammini, and F. Menczer, Modeling statistical properties of written text. PloS one, 2009. 4(4): p. e5372.

[13] 13. Bernhardsson, S., L.E.C. Da Rocha, and P. Minnhagen, Size-dependent word frequencies and translational invariance of books. Physica A: Statistical Mechanics and its Applications, 2010. 389(2): p. 330-341.

[14] 14. Mehri, A. and M. Jamaati, Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. Physics Letters A, 2017. 381(31): p. 2470-2477.

[15] 15. CALUDE, A., The use of heaps as quantifier and intensifier in New Zealand English 1. English Language & Linguistics, 2017: p. 1-26.

[16] 16. Kanter, I. and D. Kessler, Markov processes: linguistics and Zipf's law. Physical review letters, 1995. 74(22): p. 4559.

[17] 17. i Cancho, R.F. and R.V. Solé, Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences, 2003. 100(3): p. 788-791.

[18] 18. JONES, P.A. The World's Top 20 Languages—And The Words English Has Borrowed From Them. AUGUST 25, 2015 [cited 2018 July 05]; http://mentalfloss.com/article/67766/worlds-top-20-languages-and-words-english-has-borrowed-them].

[19] 19. Ali, A., A. Hussain, and M.K. Malik, Model for english-urdu statistical machine translation. World Applied Sciences, 2013. 24: p. 1362-1367.

[20] 20. Bradby, H., Translating culture and language: a research note on multilingual settings. Sociology of Health & Illness, 2002. 24(6): p. 842-855.

[21] 21. Riaz, K. Challenges in Urdu Stemming (A Progress Report). in Proceedings of the 1st BCS IRSG conference on Future Directions in Information Access. 2007. BCS Learning & Development Ltd.

[22] 22. Hussain, S. Complexity of Asian writing systems: a case study of Nafees Nasta'leeq for urdu. in Proceedings of the 12th AMIC Annual Conference on e-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore. 2003.

[23] 23. Becker, D. and K. Riaz. A study in urdu corpus construction. in Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12. 2002. Association for Computational Linguistics.

[24] 24. Sharjeel, M., R.M.A. Nawab, and P. Rayson, COUNTER: corpus of Urdu news text reuse. Language Resources and Evaluation, 2017. 51(3): p. 777-803.

[25] 25. Al-Ma'adeed, S., D. Elliman, and C.A. Higgins. A data base for Arabic handwritten text recognition research. in Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on. 2002. IEEE.

[26] 26. Sharf, Z. and S.U. Rahman, Lexical normalization of roman Urdu text. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 2017. 17(12): p. 213-221.

[27] 27. Martynyuk, S., Statistical approach to the debate on Urdu and Hindi. The Annual of Urdu studies, 2003. 8.

[28] 28. Németh, G. and C. Zainkó, Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. Acta Linguistica Hungarica, 2002. 49(3-4): p. 385-405.

[29] 29. McEnery, A.M. and A. Wilson, Corpus linguistics: an introduction. 2001: Edinburgh University Press.

[30] 30. Wiebe, J. and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. in International Conference on Intelligent Text Processing and Computational Linguistics. 2005. Springer.

[31] 31. Hofler, D.B., Vocabulary development - classifying homonyms, homophones and other word terms. 1982.

[32] 32. Wikipedia. Homograph. 2018; Available from: https://en.wikipedia.org/wiki/Homograph#References.

[33] 33. Hobbs, J.B., Homophones and homographs: An American dictionary. 2006: McFarland.

[34] 34. Dunford, R., Q. Su, and E. Tamang, The Pareto Principle. The Plymouth Student Scientist, 2014. 7(1): p. 140-148.

[35] 35. Osgood, C., The nature of meaning. Psychological bulletin, 1952. 49: p. 197-237.

[36] 36. Masrai, A. and J. Milton, How Different Is Arabic from Other Languages? The Relationship between Word Frequency and Lexical Coverage. Vol. 3. 2016. 15-35.

ECG Signal Analysis for the Recognition and Classification of Premature Ventricular Contractions Arrhythmia

Qurat-ul-ain Mastoi[a], Hira Farman[b], Dr.Teh Ying Wah[a], Dr. Ram Gopal Raj[a] , Sanaullah Mastoi[c]

[a]Faculty of Computer Science and Information Technology, University of Malaya, KL, 50603, Malaysia.
[b]Faculty of Computer Science , Mohammad Ali Jinnah University, Karachi Pakistan
[c]Faculty of Science ,Institute of Mathematics, University of Malaya, KL, 50603, Malaysia

Abstract—**This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. Premature ventricular contraction (PVC) is an early sign of potential cardiac morbidity, which is not lead to mortality but it leads to severe heart disease. Therefore, it is necessary to recognize PVC from the electrocardiogram (ECG). We proposed the methodology based on several steps including data modification, windowing function, feature extraction, and classification. In data modification part our methods separate limb leads for signal processing while windowing function set the width of rectangular space for signal analysis. However, feature extraction and PVC detection algorithm identify the pattern of PVC and normal beats.In the end, classification part playing a significant role in training and testing the PVC beats for automatic recognition. Thus, our proposed solution obtained an excellent result using four features of PVC. The performance of our system is measured with sensitivity, positive predictability, accuracy and precision which achieved 100%, 71.54%,99.54% and 62.48% respectively. The accurate and timely detection of PVC may reduce the rate of sudden cardiac death. Our proposed classifier the bagged decision tree proves that it plays a vital role in the detection of PVC beats and it can be valuable in the medical field for classifying such type of irregularities in the bioelectrical signal.**

Keywords—PVC, classification, pattern recognition, data modification and feature extraction

I.INTRODUCTION

Premature ventricular contraction is the type of arrhythmia which arises commonly from the ventricles of the heart. Premature ventricular contractions may act as early predictors of underlying cardiac morbidity. While the existence of isolated PVCs may not be life-threatening, the presence of frequent and multiple PVC may indicate an increased risk of sudden cardiac death. According to a recent study from the American Heart Association[1], approximately 2,200 Americans die each day from cardiovascular diseases. The potential of PVCs in quantifying the risk of cardiac death and guiding life-sustaining treatment has yet to be established. In the recent years, researchers have conducted large studies in the automatic detection of heart diseases based on different classification methods and feature extraction methods including replacing strategy[2], learning vector quantization neural network[3], Support vector machine ,K-nearest neighbor[4], cardioid based technique[5], fractional linear prediction[6], morphological feature-finding through wavelet transform[7], pattern matching with fuzzy neural network system[8] and neuro-fuzzy classifier[9].

Although all studies have successfully identified premature ventricular contractions, the accuracy and sensitivity levels of detection could still be improved. Furthermore, most studies have utilized the MIT-BIH dataset which is already well modified and has been used repeatedly by many studies. In addition, these all studies considered the R-R interval as a primary criterion for the identification of any premature beat. Our paper proposes a solution based on the alternative approach of identifying PVC beats through the recognition of broadened QRS complexes and calculating the shortening and previous intervals of the beat. Aforementioned problems are addressed by our system to recognize accurately the patterns of PVCs by segmenting the individual beat. Our proposed algorithm has been developed from an unmodified secure database, obtained from a Malaysian local longitudinal study of aging. The objective of this study was to modify bio-signals and accurate pattern recognition for classification of PVC beats by using the bagged decision tree

2. MATERIALS AND METHOD

A. The overall Architecture of the proposed work.

This proposed system includes five main stages: data modification, signal processing within the fixed window, feature extraction and segmentation of individual heartbeats by using a windowing algorithm, limiting the main complex structure duration greater than 3mm and then extracting those special features which are relevant to PVC, for instance non-uniformity beat of R-R interval with respect to the previous and shortening of the R-R interval and then classifying the dataset into two classes. The workflow diagram is shown in figure 1. Data modification is an important step of our method due to the usage of a naive dataset. The signal was preprocessed by using a low-pass filter and a high-pass filter in order to smooth the signal. Segmentation and feature extraction broke up the signal into individuals beats in order to extract valid information from each heartbeat. The PVC checker then verified the presence of a ventricular beat by individual beat features. The classified part of the dataset is used to train and test the dataset.

B. ECG data Acquisition

This study employed the ECG dataset acquired as part of the Malaysian Elders Longitudinal Research study. The sampling and recruitment methods of MELoR has been published in [10]. The conventional limb leads signals I, II, III, aVR, aVL, and aVF were obtained through high resolution 6 channel recordings (Taskforce, CNSystems, Austria). 10-minutes of ECG recording were available for all 6 channels with the participant rested in the supine position at a constant sampling frequency. Only signals from leads II were utilized in this study.

C. Data Modification

This first stage of the study protocol involved modifying data for further signal processing. We used ead I, II and III as the combination of signals from MELoR it was needed to be modified automatically to remove signal noises and artifacts. Separate directories were created through Matlab for individual leads' numerical data for further preprocessing.

$$\sum_{i=1}^{n} n_i x_{lead} \quad (1)$$

Where n represents the length of a number of recordings in the directory where all recordings were stored and x represents the ECG lead signal. In the modification process, all signals were decomposed on the basis of separate standard leads.

D. Signal Preprocessing

Contaminated signals are major problems in bioelectrical signal processing. Variations in signal in terms of noise may hide many important features. Therefore, it is important to clean the signal before extracting the main features of particular abnormalities. In the first part of our proposed algorithm, signals were normalized to reduce DC offset. After normalization, we used a FIR low pass filter to remove the distorted edges of the signal, whereas the notch filter was used for removing 60Hz power line interference.

E. Windowing Algorithm

The windowing algorithm is a mathematical function which is used to set windows for intervals which have zero value outside of the particular intervals. This function helps to visualize data in a rectangular shape with graphics visualization. We used this formula to show correct features of ECG signals within rectangular window width size set 0.4 seconds by using this equation.

$$w = 0.4 * constant\ frequency \quad (2)$$

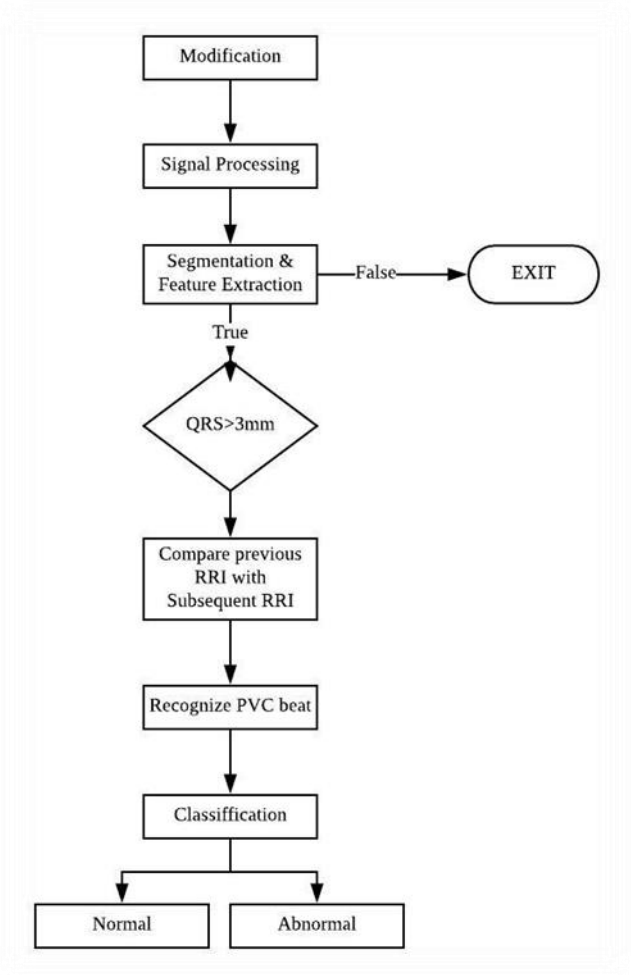


Fig: 1 Work flow of premature Ventricular Contraction recognition system

F.Feature Extraction

Feature extraction is the major step of any classification. In this work, we extracted the timing and morphological features of PVC beats. Our main features were R peak, RR intervals, and QRS complexes.

a. R-peak detection and segmentation

The R peak is prominent periodical beat with the ECG signal. The R peak is detected in each window with the threshold value set to the maximum value of the window area to detect the R peak index.

$$R(x) = \max(x)\{i, i + w - 1\} = \begin{cases} peak, R < 0.55 \\ i: end \end{cases} \quad (3)$$

Where w denotes the width of the window and x represents the filtered ECG signal. All prominent peaks are stored in the R variable matrix. In addition, we also extracted local timing features of R waves to calculate the distance between two consecutive or more than two consecutive beats.

$$RR = \sum_{m=1}^{R_l}[R(m-1)] - R(m) < w \quad (4)$$

Where R_l denote the length of peaks in signal. The RR variable creates a separate matrix for values of intervals of peaks.

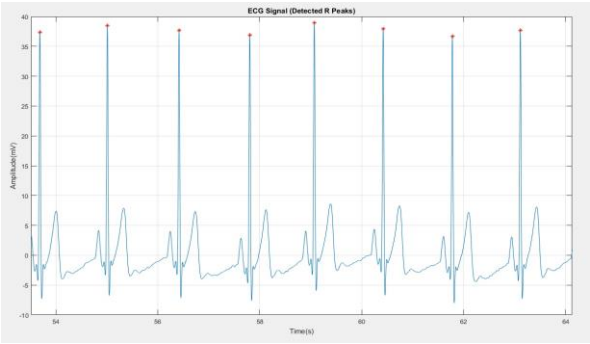


Fig 2 identifies prominent signal peaks

After finding the prominent signals peaks, our method involved dividing the beat into the individual segment and calculating other features on the basis of the individual segments.

b. Q and S Fiducial mark

In order to obtain full QRS complexes, we applied a separate formula for each because all other markers relying on the length of the R peak. The Q and S clinical markers were detected by using this given equation.

$$Q_{rl} = \sum_{q=2}^{R_l}(x[R_l(q)(-1) - fix(0.3 * constant\ frequency)]) = \begin{cases} Qrl, Qrl \neq 0 \\ false, Qrl = 0 \end{cases} \quad (5)$$

$$S_{rl} = \sum_{s=2}^{R_l}(x[R_l(s)(-1) + fix(0.3 * constant\ frequency)]) = \begin{cases} Srl, Srl \neq 0 \\ false, Srl = 0 \end{cases} \quad (6)$$

Q_{rl} was selected by identifying the minimum value in the window starting from 20ms before the occurrence of the maximum peak value, whereas S_{rl} is calculated in the same way but after the maximum peak. After finding individual values of Q_{rl} and S_{rl} the complex duration was measured by subtracting those values with respect to the sampling frequency to obtain the QRS complex value.

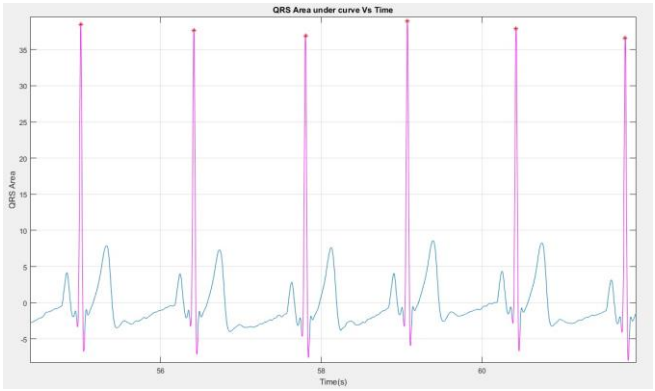


Fig 3: Area of QRS complex

G. *PVC Detection Algorithm*

Premature Ventricular Contractions were defined as premature beats initiating from the ventricles of the heart as advised by our medical expert. We designed our methods on the basis that the QRS complex of a PVC is wider than 3mm. A further algorithm then processed the selected potentially abnormal segment for the identification of PVC unless the system was unable to proceed. After identification of a candidate segment based on QRS duration, the RR intervals of adjacent beats were then measured. If the RR interval between the previous narrow QRS complex and the broadened QRS beat was at least 80% less than the RR interval of the preceding beat, and the RR interval of the abnormal QRS complex and the subsequent beat was at least 10% greater the last 'normal' RR interval, representative of the compensatory pause which regularly follows each PVC. Fig 4 displays the pattern of PVC beats.

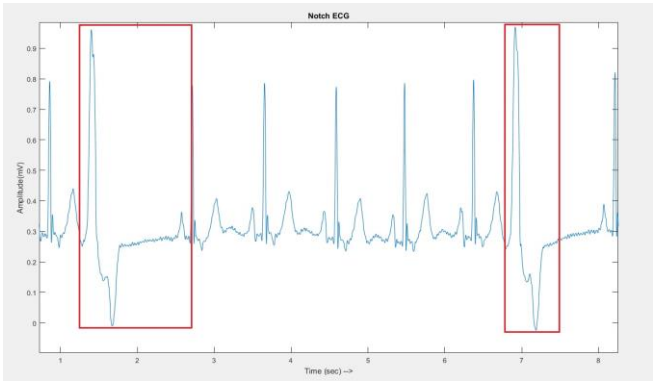


Fig 4 pattern of PVC arrhythmia

H. Classification

We used the bagged decision tree classifier. In ensemble algorithms, bagging methods form a class of algorithms which build several instances of a black-box estimator on

random subsets of the original training set and then aggregate their individual predictions to form a final prediction. It is generally the most powerful tool to make decisions and identify patterns within large data sets[11].The bagged decision tree classifier demonstrated greater accuracy and better capacity to evaluate different classes. The bagged decision trees combine the results of many decision trees, which reduces the effects of overfitting and improves generalization process.

III. Experimental Results

In this study, we used real-time ECG signals from the MELoR study. Using only lead II for PVC detection. We used data modification method for the takeout most necessary lead for identification of PVC beat from limb lead cluster. Windowing function used to reserve a particular area for signal processing. The feature extraction method used to calculate intervals and complexes for identifying pattern. Consequently, our proposed algorithm for PVC checker used to evaluate the PVC beat. Whereas, classification is the main important step for automatic recognition of PVC beat. The tenfold cross-validation technique was used for training phase and testing phase. We used five transformed features for training phase while testing we used four features.

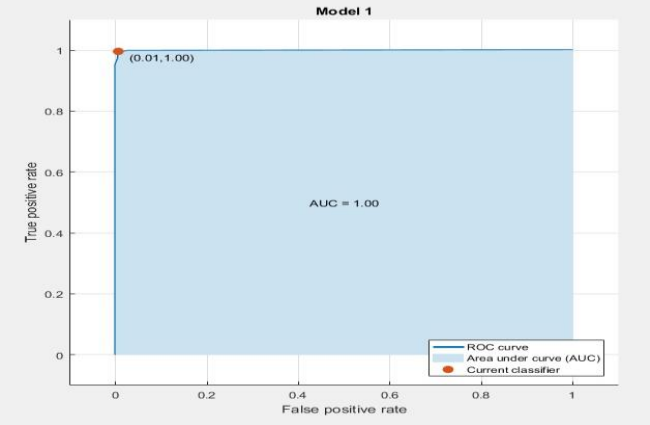


Fig 5 Training performance of the classifier.

In this study, we randomly selected 7 sample ECG recordings from the MELoR dataset with corresponding files no S007, S0011, S0019, S0015, S0020, S0017, S0014 and 1 sample recording from MIT-BIH with file no 106.We collected 1000 samples for training, 500 of PVC beats and 500 of normal beats using the 5 features include RRprevious, RRinterval, subsequent RRinterval, QRS, and class. In this experiment, we tested the beats individually with all features and overall results were measured by taking the mean of each beat. The training performance measured 99.5% accuracy using bagged decision tree classifier which is shown in fig 5. After training the classifier we tested the classifier by using unknown samples according to the training.

A. *Performance evaluation parameters*

To evaluate the performance of parameters we used four performance parameters calculated by the readings from the test dataset which include sensitivity (Se), positive predictive value (Pp), error rate and accuracy, where the true positive (TP) rate is the number of correctly detected PVC beats and the false positive (FP) rate is the number of wrongly identified beats selected by the classifier, whereas Er describes the total number of classification errors of PVC and accuracy defines the overall performance of the classifier. These performance measures recorded individual beats and overall performance computed using average readings.

$$S_e = \frac{TP}{TP+FN} * 100 \quad (7)$$

$$P_p = \frac{TP}{TP+FP} * 100 \quad (8)$$

$$E_r = \frac{FN+FP}{N} * 100 \quad (9)$$

$$A_{cc} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$\mathbf{Prec} = \frac{TP}{TP+FP} \quad (11)$$

B. *Discussion*

In this paper, we demonstrated that our proposed algorithm for pattern recognition and classification of PVC is able to extract and identify patterns of PVC successfully. To evaluate the performance of the proposed algorithm 37 sorted signals were taken from the MELoR dataset. It can be seen from our experimental results in table 1 ECG samples S0024, S0027, S0030, and S0035 had no PVC beats. The Se, Pp, and accuracy were all 100% in signal samples S001, S003, S004, sm0011, S0012, sm0014, sm0015, S0018, S0025 and S0036. Consequently, the remaining signal samples had accuracies exceeding 95%. As we can see, the sensitivity was 100% for all beats because we had 0% FN beats in the classification result. The overall sensitivity and accuracy approximated to 100%. The positive predictive value was 72%. In comparison to the findings of recent studies, the results of our classification were marginally better in terms of sensitivity and accuracy.

The authors in [2] proposed replacing the strategy to evaluate the effects of a heartbeat on the variation in principal directions and the authors successfully achieved the sensitivity and accuracy of 96.12% and 98.77% respectively.

Table 1- Experimental result of classification performance									
File	Total beats	PVC beats	TP	FP	Se(%)	Pp(%)	Er(%)	Precision(%)	Accuracy (%)
S001	666	1	1	0	100	100	0	100	100
S002	843	39	39	17	100	69	2.01	69.6	97.9
S003	698	1	1	0	100	100	0	100	100
S004	687	1	1	0	100	100	0	100	100
S005	782	5	5	1	100	83	0.12	83.3	99.8
S0024	779	-	-	-	-	-	-	-	-
S0027	859	-	-	-	-	-	-	-	-
S0030	788	-	-	-	-	-	-	-	-
S0035	717	-	-	-	-	-	-	-	-
S006	690	1	1	1	100	50	0.14	50	99.8
Sm007	296	1	1	1	100	50	0.33	50	99.6
S008	723	1	1	2	100	33	0.27	33.33	99.7
S009	835	24	24	8	100	75	0.95	87.5	99.04
S0010	722	1	1	1	100	50	0.13	50	99.86
sm0011	650	1	1	0	100	100	0	100	100
S0012	617	2	2	0	100	100	0	100	100
S0013	489	1	1	4	100	20	0.81	20	99.1
sm0014	543	2	2	0	100	100	0	100	100
sm0015	639	1	1	0	100	100	0	100	100
S0016	650	3	3	8	100	27	1.23	27	98.7
sm0017	632	9	9	1	100	90	0.15	90	99.8
S0018	723	3	3	0	100	100	0	100	100
sm0019	674	5	5	4	100	55	0.59	55.55	99.4
sm0020	556	13	13	3	100	81	0.53	81.25	99.4
S0021	660	12	12	1	100	92	0.15	92.3	99.8
S0022	806	14	14	4	100	77	0.49	77.77	99.5
S0023	721	1	1	7	100	12.5	0.97	12.5	99
S0025	761	3	3	0	100	100	0	100	100
S0026	639	2	2	1	100	66	0.15	66.66	99.8
S0028	650	1	1	10	100	16	1.36	9.09	98.6
S0029	736	1	1	3	100	25	0.40	25	99.59
S0031	732	45	45	27	100	62.5	3.6	62.5	96.3
S0032	645	30	30	9	100	76.9	1.39	76.9	98.60
S0033	612	3	3	1	100	75	0.16	75	99.8
S0034	914	2	2	1	100	66	0.10	66.66	99.8
S0036	564	1	1	0	100	100	0	100	100
S0037	772	1	1	1	100	50	0.12	50	99
Total(Avg)	25470	231	231	116	100	71.54	0.7	62.48	99.45

We also compare our method with [3] which proposed a novel method for automatic detection of PVC by using a learning vector quantization neural network and achieved 90.26% and 98.90% for sensitivity and accuracy respectively. The limitations of previous studies included the use of the most prominent peak as a primary source of detection of PVC whereas in our proposed method for feature extraction we used the QRS complex as a primary strategy for detecting PVC and we also collect features of adjacent RR intervals to the PVC for accurate identification of beats. The bagged decision tree method helped to obtain accuracy in the testing phase while making the right decision for PVCs.

Table 2- Comparison analysis

Study	Technique	Se%	Acc%
[2]	PCA	96.12%	98.77%
[3]	LVQ neural network	90.26%	98.90%
Our proposed method	bagged decision tree	100%	99.45%

C. Conclusion

In this paper, we proposed a novel algorithm for feature extraction for the detection of PVC using the windowing function and segmentation of individual beats. Accurate pattern recognition of PVCs was achieved using the modification of raw data and newer feature extraction methods. We considered a QRS complex exceeding 3mm as the primary criteria for identification of PVCs. The proposed algorithm was validated using the MELoR dataset. The experimental statistics disclose that our method was robust, achieving good classification performance using bagged decision tree classifier in terms of sensitivity and accuracy with only four relevant features of PVC. In our future work, we will analyze the role of abnormalities in T-wave repolarization in the identification of potentially life-threatening cardiac conditions.

Conflict of Interest

The authors declare no conflict of interest

Acknowledgment

We are grateful to the Malaysian Elders Longitudinal Research (MELoR) team for their role in data acquisition. Special thanks to Mr. Goh Choon Hian who personally recorded nearly all the signals, and Professor Chee Kok Han and Dr. Ang Choon Chin from the Department of Cardiology, University of Malaya Medical Centre for their role in ECG interpretation.

References

[1] Heart Disease and Stroke Statistics 2017 At-a-Glance American Herat Association 2017, pp. 1-5.
[2] R. Zarei, J. He, G. Huang, Y. Zhang, Effective and efficient detection of premature ventricular contractions based on variation of principal directions, Digital Signal Processing, 50 (2016) 93-102.
[3] X. Liu, H. Du, G. Wang, S. Zhou, H. Zhang, Automatic diagnosis of premature ventricular contraction based on Lyapunov exponents and LVQ neural network, Computer methods and programs in biomedicine, 122 (2015) 47-55.
[4] A. Orozco-Duque, F. Martinez-Tabares, J. Gallego, C. Rodriguez, I. Mora, G. Castellanos-Dominguez, J. Bustamante, Classification of premature ventricular contraction based on discrete wavelet transform for real time applications, Health Care Exchanges (PAHCE), 2013 Pan American, IEEE, 2013, pp. 1-5.
[5] V. Mai, I. Khalil, A cardioid based technique to identify Premature Ventricular Contractions, Computing in Cardiology, 2011, IEEE, 2011, pp. 673-676.
[6] A. Ebrahimzadeh, A. Khazaee, Detection of premature ventricular contractions using MLP neural networks: A comparative study, Measurement, 43 (2010) 103-112.
[7] R.C.-H. Chang, C.-H. Lin, M.-F. Wei, K.-H. Lin, S.-R. Chen, High-Precision Real-Time Premature Ventricular Contraction (PVC) Detection System Based on Wavelet Transform, Journal of Signal Processing Systems, 77 (2014) 289-296.
[8] P. Li, C. Liu, X. Wang, D. Zheng, Y. Li, C. Liu, A low-complexity data-adaptive approach for premature ventricular contraction recognition, Signal, Image and Video Processing, 8 (2014) 111-120.
[9] M.A. Chikh, M. Ammar, R. Marouf, A Neuro-Fuzzy Identification of ECG Beats, Journal of Medical Systems, 36 (2012) 903-914.
[10] M. McStea, K. McGeechan, S.B. Kamaruzzaman, R. Rajasuriar, M.P. Tan, Defining metabolic syndrome and factors associated with metabolic syndrome in a poly-pharmaceutical population, Postgraduate Medicine, 128 (2016) 797-804.
[11] K. Machová, F. Barcak, P. Bednár, A bagging method using decision trees in the role of base classifiers, Acta Polytechnica Hungarica, 3 (2006) 121-132.

Mobile Crowdsensing Application of Road Condition Detection

Nauman Mazhar[1], Afzal Ahmad[2], Abdul Razzaq[3], Khizzar Abbas[4], Muhammad Adnan[5], Hamid Abdullah[6], Nabeel Rasheed[7]

[1]*Lecturer, Department of Information Technology, University of Gujrat*

[2]*Department of Information Technology, University of Gujrat*

[3]*Department of Information Technology, University of Gujrat*

[4] *Department of Information Technology, University of Gujrat*

[1]*nauman.mazhar@uog.edu.pk*

[2]ahmadafzaal702@gmail.com

Abstract - This application has been developed for android phones to determine road conditions. It occasionally detects the hurdles and jumps whenever found on the road in order to determine the road condition. It is free from user input, which works in the background to perform operations. It automatically sends road condition information to the server whenever it detects the hurdles on the road. This information is stored in the form of a map on the server which further can be used by the municipal authorities to repair the broken road. This application is not perfectly accurate because there is some factor which can affect the working of the application. This application succeeds approximately 90% in sensing the potholes and 95% of the speed bumps, whereas producing incorrect sensing around 10% of the potholes and 5% of the speed bumps.

I. INTRODUCTION

A. Background and motivation

Almost all the roads of Pakistan are damaged. This condition can cause accidents and extensive damage to vehicles. Many accidents occur daily due to the poor condition of roads. Funds given by the government to repair these roads are spent without giving any visibility to the general public. In monsoon, the rain damages the roads to intolerable levels in Pakistan. Due to the poor drainage system, the road condition in Pakistan becomes worse. The government of Pakistan is improving the road conditions by making new flyovers and roads. Unfortunately, this can be implemented in large cities but is difficult to implement in the small cities and rural areas of Pakistan. Sometimes, citizens report the location and poor condition of roads to help the authorities in its repairment. In addition, due to disrepair, citizens make illegal speed bumps on the road to slow down the traffic near the residential areas. They do not report these illegal speed bumps. Such illegal speed require years to actually remove them. The Municipality and authority have the data of speed bumps illegally made by the citizens and they can specify all these illegal speed bumps. Municipalities can allocate more funds in these areas to repair speed bumps and roads.

At the same time, the population of Pakistan is increasing rapidly. With the increasing population, the number of cars and vehicles are also increasing. But unfortunately, the network of roads has not grown at the same rate of cars and other vehicles. These roads need to be repaired since damaged roads cannot be easily mapped. Measuring the road conditions and mapping the location with the road is not an easy task. It requires extensive resources i.e. time, expertise, transportation, and manpower. One approach which is not feasible is the government hiring people to identify the conditions of roads. The conditions of the road may have already changed by the time teams survey an area.

B. Problem Statement

Road roughness and potholes (road condition) cannot be easily measured and mapped. Although the roads cannot be repaired immediately, the authorities should be informed about the conditions of roads after a specific time. Making government departments for handling the road conditions is an expensive and time taking process. This process is not suitable for mapping and measuring road conditions. Therefore, the objective is to create a reliable system which is based on mobile crowd sensing to measure the state of roads regularly without any interaction with a user.

C. Inability of previous work

Mednis and his fellows proposed a system to monitor road conditions using Android OS smart phones [6] with Accelerometers. They use an automated approach to detect potholes and ensure more accurate data with fewer errors. The limitation of this approach of monitoring road and collecting data was they used standalone smart phones and did

not use centralized database i.e. it cannot process data taken from different mobile phones present at one location.

Jakob Eriksson and his fellows designed a pothole detection system called Pothole Patrol which is used to collect data from sensor vibrations and GPS sensors, and this system analyzes the data collected from the sensors fixed in a collection of vehicles to measure the road conditions. [2] This application is also using a centralized database like we have used in our system. But this system is different from our system because they have used GPS system and accelerometer fixed in cars to identify the road conditions, speed bumps, and potholes.

Mohan et al, designed a system called Nericell, [1] used to calculate and measure road conditions and also the conditions of traffic. They are also using additional GPS sensors and accelerometer sensor of cell phones which are not smart cell phones. By using these external sensors, speed bumps and stop and go events of traffic are measured. In this system, almost all functions of detection were implemented but not in a smart phone. The drawback of this system is that when the data was collected from the smart phone, it was suspended by the system.

D. Proposed Approach

Almost all the people in urban and rural areas of Pakistan are using smart phones. Additionally, 3G and 4G services of the internet are also present in Pakistan and most of the citizens are using these internet services. This usage of internet in Pakistan is the solution of our problem statement i.e. making an android smart phone application which has the ability to sense speed bumps and road conditions. This paper is going to discuss the application using crowd sensing to sense potholes and speed bumps. The application is designed for Android smart phone that has no interaction with the user and takes no input from user i.e. zero input application of smart phone. This application uses a centralized approach of a database which collects data from multiple users i.e. the benefit of crowd sensing, and updates this data as the user travels through different routes every single day. Another benefit of this application is that as several users travel through the same place, this application filters this data and separate the false collections of data from correct data.

This smart phone application does not use any external GPS and accelerometer sensors other than a smart phone because it is able to run this application without using these external sensors. There is no need to buy any hardware i.e. accelerometer sensor, and there is no need to place accelerometer sensor at the fixed position in a car because this application uses smart phone's accelerometer during the journey and it will increase the application usability.

This application runs in the background as a service so that whenever the user is on the road, it can gather data from the vibration of a car when it moves over a speed bump or pothole. This information is processed in the application and the processed data [4] is kept in the database created on the server. The meta-classifier on the server side decides, if the same users detect the presence of pothole or speed bump from the same location, this information is skipped otherwise it is stored in the database. The information stored in the database is then available on the website so that public can read the road condition information

II. RELATED WORK

Mednis et al. described accelerometer data based pothole detection algorithm with hardware/software resources using Android OS based smart phones. The system senses events of vehicles like cars and buses but motorcycles and bikes [7] are not supposed. However, accelerometer sensors combining different other sensors were used to collect data. Tmote Mini hardware, which is a sensor node using Texas instrument and 3-axis accelerometer analog devices were used with accelerometer sensor to collect data about road conditions. Also, a specific location was considered to place the Android smart phone. We want that while using our application, the application detects the user activity if the user is using mobile, the application stops collecting data from the sensor if not it collects data.

Jakob Eriksson and his fellows designed a Pothole detection system called Pothole Patrol which is used to collect data from sensor vibrations and GPS sensors, and this system analyzes the data collected from the sensors fixed in vehicles to measure the road conditions [2]. Using this system, the potholes and other road conditions can be recognized using accelerometer sensor. The system uses embedded GPS and accelerometer of 3-axis in the car. But in our application, there is no need to buy any other hardware because the smartphone is enough to perform task i.e. estimated condition of roads.

"Potholes hunters" is basically is an app that map potholes, and potholes, checking existing potholes and rating the worst potholes in the country [6]. This app requires user help to map that pothole to fill a form and take pictures of the pothole and record that data. Our app has the advantage that does not require a user to take a picture and map that pothole. Our app automatically detects potholes and send that data to cloud for public and administrative authority to repair that pothole.

III. SYSTEM DESIGN

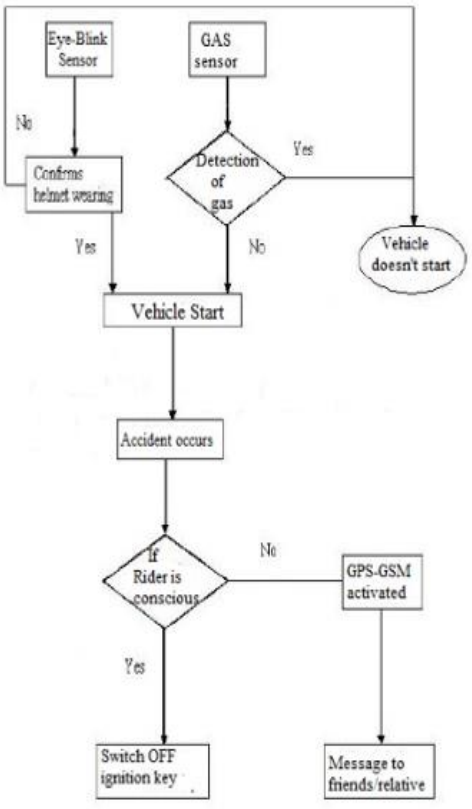
This mobile app according to the android also have two background process that receives and send data from one point to the other point through sensors. One process gets data from the users through sensors and the second process upload data for the user requirements. Simply some sensor is used in this process to get data from the user and provide the view of the roads conditions to the users according to the data that are received by the user [8]. In the process, the database also uses a procedure that saves the data of all users collected by the user's mobile and also mobile orientation. App engine also used in this process receives the request from the database, processes the data and again sends it to the database. The database has many tables of user data in which every table is separate but linked to each other.

A website is also used in this process which sends and receives queries according to the user requirement. It also incorporates the map location of the user. The main aim of this website is to show the current data of the conditions of the roads. Some labeled data also use in this process, which is described at the first stage, helps the user if the current map location is same as the map location of the this labeled data. This data is show to the user without any processing or delay. The data is collected or saved by just visiting the main area. The android app picks the map location and gets this data by the movement of the mobile that has the android app in on mode.

A. Sensor Types

All data collected from the user mobile or the current location of the user is collected by different sensors. These sensors sense the main event and then collect the data according to that event. Then this data send to the database for the more process for the user. The current location of the user is traced through an API. After tracing, the application collects the raw data about the user's location through sensors and sends back the information about the roads to the user.

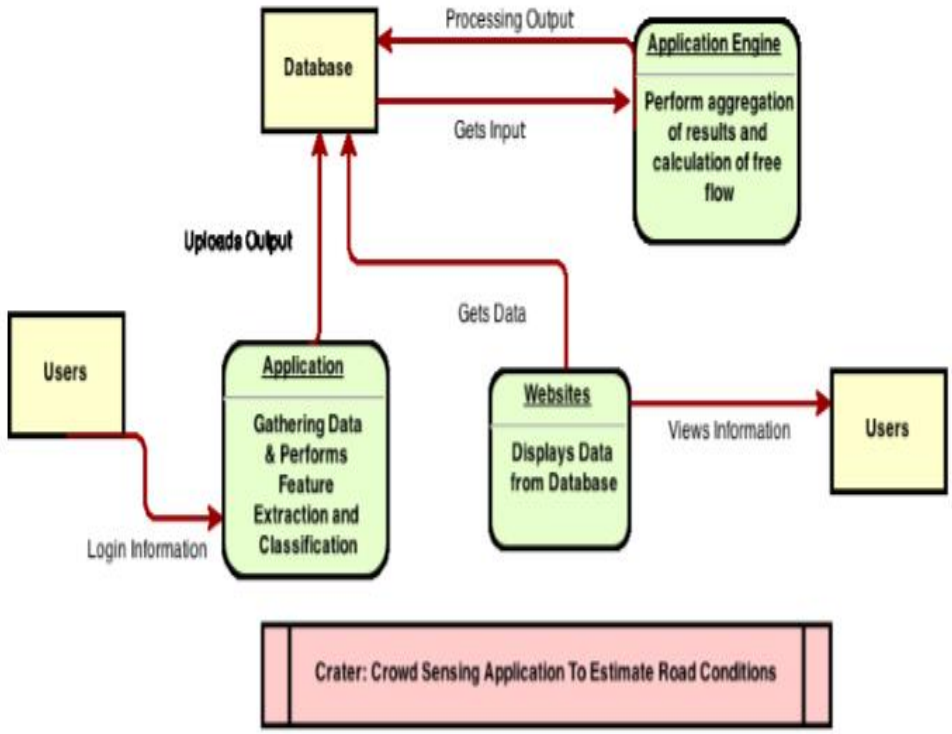
Sensor Delay Fastest: Whether road is clean or damaged is reflected by the speed of the car of the user because sensors differentiate between the speed of the car at the different times. Sensor Delay Fastest is the sensor that senses the changes in speed of the user's car. These sensor show the recordings to the user about the condition of the road in the application.



Sensor Manager: One sensor manages the orientation of the mobile. The x-axis shows the position of mobile in the horizontal form; Y-axis shows the position in the vertical form and Z-axis show the position of mobile in downward. Now, this data is recorded in the application for the user.

B. Process of Android APP

Login: At the first stage the user login the app for getting the information about the road conditions. When the user login then the database receive the data through the sensors and show the results according to the data or location of the user. All data of each user is recorded separately in different tables. This data is stored until the current user logout or delete the data.



Sensor: The sensor gets the data from the user mobile and its position which is sent to the data storage. Important data is selected for further processing. The sensor shown in the above paragraph senses all data of the user.

Database: The database has a lot of data about all users in the form of tables. Each table has different and useful data for every user. There are two types of table:

1. User contributed Data

2. Voted Views

Contribution of user in data collection: Data collected by users at different times in the raw form is stored in the tables of the database. This table has the data about user location and the mobile orientation. The data is sent in selective form where it is selected according to the requirements of the users.

Voted Views: Data that is retrieved by the processing of application is saved in the table of voted view. This data informs the user about the condition of the road i.e. whether the road is clear or damaged and also tells the location of the clear or damaged road. This table is very important according to the information that is provided to the user.

The database then sends data to the engine app for further processing according to the requirement of the user.

App Engine: The database has data about the users in the form of tables. It sends data according to the user requirement to the app engine which acts as a server in this process. App engine processes the data and sends back the resulting one for the user. Restful APIs [3] provide access to show data on the website and mobile app. App Engine is a server at this stage that deletes old data from the app and uploads new information. The website just informs the user about the condition of roads. This information also stored in the android app.

Website: The website is also a significant component of this process. This website sends query to the database that hold all data about every user that has an account on this application. Database answers to these queries and retrieves results to send back to the user. The website is just used to display data for the user. It has both the internal and the external interface. Mobile APIs are not able to give user's speed directly but speed is calculated by a specific formula which uses the difference between the distances of two points.

IV. Detector Design

We used a supervised learning approach for app classifier which uses labeled training sets for creating a distinct Android app that is stated as a data collection app. Android app - stated as a data collection app, allows the user to start collecting data and stop by pressing the button that is given in the app. Data from GPS and three accelerometers is stored in data collection in a file. Data collection app also offers two buttons that allow a user to spot instantly as the vehicle goes above a pothole or over a speed bump. By pressing the button, the location and timestamp are stored in the file.

Without preprocessing the data that is collected from data collection app, this data cannot be used directly on behalf of feature abstraction and

arrangement. Furthermore, we do not need to fix or place the smart phone in a specific position when collecting data. This necessitated rotating the accelerometer axes to readjust axes in a typical direction. The Axes reorientation then aligns the path of maximum variation in accelerometer analyses through 'z-axis', that point down to the earth/ground. The second maximum variation is orthogonal to 'z-axis', and is aligned to 'y-axis', and the path of this variation points to onward motion of a vehicle.

By determining directions of both y, z axes, an 'x-axis' is aligned by residual orthogonal direction, and direction of this axis is typically sideward to the direction vehicle. As per successive exploration, this is usually the Type equation here. property of the reoriented 'z-axis', to a lesser degree 'y-axis', that offers valuable characteristics to sense speed bumps and potholes. It includes defining the accelerometer covariance matrix and each accelerometer reading is switched to align through the x, y, and z-axis by performing a matrix vector multiplication.

Moreover, the vehicle's speed is not shown directly by Android API. By using the Haversine formula, changes in geographical coordinates to distances is converted and then the speed is calculated from consecutive changes in longitude and latitude. Once the distance between two successive locations is recorded, it is divided by time. When a speed bump or pothole is met with vehicle, through likewise reviewing of the accelerometer and by taking the timestamp of this encountered hurdle, we can find features that can be worth further discovering for possible usage in ordering speed bumps and potholes.

Figure shows a photograph of a cemented road, currently on our university grounds. It was used at the development stage to collect data.

A. *FEATURE EXTRACTION:*

Data collection app executes at one-second interval for the speed bump and pothole detectors. Each time a speed bump and pothole is detected, the accelerometer sensor readings are composed during the previous 12sec. The figure shows the accelerometer readings with all three axes above time for a vehicle that is moving over the pothole and speed bump. But without additional processing, it is possible the coming signals are noisy if high-frequency components are involved so we need to process the signals to make them noiseless. During the data collection through high frequencies components, we discover noisy signal through two different bases.

1) The vehicle regular frequency component vibrate when an engine of the vehicle is running which is a cause of noise signal, even when the vehicle is not moving.

2) The irregular high-frequency component can be a source of noise due to the bumpiness of the road surface and the vehicle's speed.

As we know high-frequency component is not beneficial for our objective since it produces noise for signals so we can remove it by passing signals through the low-pass (referred as an LP) filters. We use LP filtered signals for pothole and speed bump detectors. The best important features involved are finding,

•Standard deviation for accelerometer sensor readings through three axes(x, y, and z)

• Mean deviation for accelerometer sensor readings through three axes (x, y, and z)

•Standard deviation of low-pass filtered through three axes (x, y, and z)

•Mean deviation of low pass filtered through three axes (x, y, and z)

•Minimum/maximum value on x/y/z-axis

Through supervised learning approach we measured five different sorts of classifiers, decision tables, naive Bayes, and SVM.

V. RESULTS

A. *IN-APP SENSORS:*

The app collects data, switches the mobile phone sensors axes to reproduce in the way of motion of the vehicle. It extracts characteristics and use them to detect speed bumps and potholes on the road segment. Design selections and judgments have been labeled for these sensors. The results of the classifiers helps speed bump/pothole sensors and the timestamp of the recording and the location is saved in the database. App automatically uploads this data to the database in the cloud when it gets connected to a Wi-Fi network. Data is the reflection of the achievement of the speed bump and pothole sensors. These outcomes were made using the considered exercise set defined.

Grouping	True positives rate	False positives rate	Accuracy	Recall	Final-measure	Infected area
Standard VM	50.5%	4.6%	8.2%	60.4%	98.8%	93.5%
Native Bayes	9.0%	91.9%	23.5%	30.8%	90.4%	18.2%
C5	6.0%	84.6%	68.8%	64.5%	68%	79.5%
Choice Table	6.9%	50.3%	32.7%	44.7%	61.3%	47.9%
Sub-clustering	18.9%	45.6%	4.5%	37.9%	96.9%	60.8%

Table 1: In-app speed bump groupings

1) GROUPINGS FOR SPEEDBUMP SENSOR:

Speed bumps are comparatively infrequent incidents on the road. We tried classification through great false positives rates from more attention. Built on the false positives in Table 1, this disqualified native Bayes. This leaves us with standard VM, C5choice tree and conclusion table that has better performance compared to true positives and false positives rates. Out of these three, standard VM was the better alternative to true positives forms, C5 conclusion tree and conclusion table in terms of accuracy.

Grouping	True positives rate	False positives rate	Accuracy	Recall	Final-measure	Infected area
standard VM	60%	5%	82%	60%	70%	90%
Native Bayes	90%	91%	23%	10%	90%	19%
C5	60%	0.8%	68.8%	62.5%	65%	75.5%
Choice Table	65.9%	5.3%	32.8%	40.7%	65.3%	90%
Sub-clustering	18%	45.6%	34.5%	87.9%	56.9%	55.8%

Table 2: In-app pothole groupings

2) GROUPINGS FOR POTHOLE SENSOR:

A related study of the similar five classifiers for pothole sensor rules out native Bayes and overseen clustering due to exact extra ordinary false positives rates. Out of the three remaining methods, decision tables have a significantly lower true positives rate than standard VM and C5choice tree, therefore we discard that as well. Among standard VM and C5 decision tree, standard VM is better than C5 in terms of recollection, final-measurements and infected are a by significant boundaries. We selected two standard VM classifiers for a speed bump and pothole sensors.

B. IN-CLOUD SENSOR:

App machine obtains results of the app through speed bump and pothole sensors noticeable with time stamps and Global Positioning System directs. Asimple edge based in-cloud sensor is being consumed, if more users report sensing a speed bump at a network location, a position is noticeable as overwhelming on the speed bump. Thus, the final determination of whether speed bumps observed mistakenly or not is done by the in-app sensor. The consensus based results of the in-cloud sensor are shown on the website and is also made available on the app itself.

VI. CONCLUSION

In this research paper, we have discussed the smartphone application to sense the potholes, speed bumps and other condition of roads. The working of the application based on the systems already developed. The application is different from the other similar applications because there is no need to buy any external hardware for sensing the location and potholes except the android phone which has its own GPS and accelerometer sensor. Our application does not require any input from the user and also does not require a particular location to be placed. It just has to be installed in the android phone and it will run in the background. According to our findings, the application is not 100% accurate. In some situations, it does not measure the hurdles and speed bumps accurately.

VII. LIMITATIONS OF OUR WORK

Some problems have been noticed that can affect the data collected by the application. It is possible highly inaccurate information can be collected about the road condition using an accelerometer to determine whether the road is paved, unpaved or for detecting the dirt on roads.

Another problem on which further research can be conducted is the detection of the traffic quantity on

the road. The traffic quantity on the road can be measures by using the mobility traces of each vehicle on the road.

VIII. REFERENCES

[1] Chen, K., Lu, M., Fan, X., Wei, M., & Wu, J. (2011). *Road condition monitoring using on-board three-axis accelerometer and GPS sensor*. Paper presented at the Communications and Networking in China (CHINACOM), 2011 6th International ICST Conference on.

[2] Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., & Balakrishnan, H. (2008). *The pothole patrol: using a mobile sensor network for road surface monitoring.* Paper presented at the Proceedings of the 6th international conference on Mobile systems, applications, and services.

[3] Ganti, R. K., Ye, F., & Lei, H. (2011). Mobile crowdsensing: current state and future challenges. *IEEE Communications Magazine, 49*(11).

[4] Ghose, A., Biswas, P., Bhaumik, C., Sharma, M., Pal, A., &Jha, A. (2012). *Road condition monitoring and alert application: Using in-vehicle smartphone as internet-connected sensor*. Paper presented at the Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on.

[5] Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N. Y., Huang, R., & Zhou, X. (2015). Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys (CSUR), 48*(1), 7.

[6] Mohan, P., Padmanabhan, V. N., &Ramjee, R. (2008). *Nericell: rich monitoring of road and traffic conditions using mobile smartphones.* Paper presented at the Proceedings of the 6th ACM conference on Embedded network sensor systems.

[7] Pan, B., Zheng, Y., Wilkie, D., &Shahabi, C. (2013). *Crowd sensing of traffic anomalies based on human mobility and social media.* Paper presented at the Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

[8] Perttunen, M., Mazhelis, O., Cong, F., Kauppila, M., Leppänen, T., Kantola, J., . . .Ristaniemi, T. (2011). *Distributed road surface condition monitoring using mobile phones.* Paper presented at the International Conference on Ubiquitous Intelligence and Computing.

[9] Wan, J., Liu, J., Shao, Z., Vasilakos, A. V., Imran, M., & Zhou, K. (2016). Mobile crowd sensing for traffic prediction in internet of vehicles. *Sensors, 16*(1), 88.

Fintech based Agricultural Supply Chain Framework for South Asian Countries

Haleema Sadia
Computer System Department
Mehran University of Engineering and Technology
Jamshoro, Pakistan
haleema.sadia@unifiedcrest.com

Adnan Ashraf
Faculty, CSE
Mehran University of Engineering and Technology
Jamshoro, Pakistan
adnan.arain@unifiedcrest.com

Manzoor Hashmani
Faculty,
University Technology Petronas
Malaysia
mhashmani@yahoo.com

Syed Adila Afghan
Faulty of informatics
University of Debrecen
Debrecen, Hungary
adila@eng.unideb.hu

Sanam Narejo
Faculty, CSE
Mehran University of Engineering and Technology
Jamshoro, Pakistan
sanam.narejo@faculty.muet.edu.pk

Rizwan Iqbal
Faculty, CSE
Bahria University
Karachi, Pakistan
mail.rizwaniqbal@yahoo.com

Abstract— Economic growth of an agrarian country is considered as risk-free due to stable performance of its agriculture sector. Asia / South Asia produces 75% of crops, live-stock but farmers get meager facilities. This is due to poor access to field-research (seed development, cultivation methods, crop protection and production). The misuse of water and fertilizers is found very common in South Asian countries, especially India, Pakistan, Iran, and Bangladesh. These challenges made Agricultural sector less attractive for investment. Consequently, the economies of few countries are on the verge of collapse whereas, others are not risk free.

This paper gives a way-out to address these challenges by designing a Fintech based supply chain framework. In this paper comparative analysis of both traditional and proposed system is performed. After study of the incumbent system, a blockchain based agricultural supply chain system-Agri-BC- is proposed. Elements of the framework (Agri-BC) effectively address the challenge of intermediaries' commission that results in the increased profit share of farmers. The survey, about framework (Agri-BC) conducted from farmers, reveals that issues can be resolved, and investment can be effectively utilized in this sector. Ultimately, this gives an opportunity for new investors to become major stakeholders.

Keywords— Agriculture, Fintech, Blockchain, supply chain.

I. INTRODUCTION

Agriculture sector is a major contributor in the economy of an agrarian country. It provides input to industries and food to mankind for survival. Other sectors directly or indirectly rely on this sector for input as raw material and collectively contributes towards the country economy. Its importance can't be denied, as this sector is cause of human survival [1][2]. Effective performance of agriculture sector is key to satisfy the needs of ballooning population. If agricultural sector fails to meet the demands of the population, it will have a negative effect on the health of mankind and economy of country [3]. Even economic depression does not affect the demand curve of agriculture sector. A country which has exponential population growth and an agricultural state, there exist inequality between urban and rural areas of the country. To bridge this gap, it is need of time to develop and flourish rural agriculture sector of the country.

Agricultural sector contributes a substantial portion of capital formation for a country in form of: export of agricultural products, employment, and agricultural taxation. The Agriculture supply chain is one of the areas in agriculture sector, improvement in it can have a significant change in profitability. Traditionally farming product was only traded in domestic market. Now, with the advent of science and modern technology, product sales are not only confined to domestic market but, are exported to other countries results in revenue generation and improves country GDP [4]. Growth in the country economy can be achieved by exports of farming product, and it also proves to be in the better interest of its producer [5] [6]. Along with profit generation, export transactions also includes taxation complexity, intermediaries, discrepancies in information, ineffective (time-consuming) transaction and lack of transparency.

Lack of agricultural research, resistance to adoption of modern agricultural techniques and processes, indirect link between farmer and consumer results in large number of intermediaries ultimately reduces profit, and tax burden is superimposed on the farming community are few of agriculture related factors contribute toward depreciated GDP [4].

Analyzing current situation it is need of time to the unleash potential of agriculture sector by integrating it with IT and the digital world. By proposing a system to resolve all issues which are faced by stakeholders to bring agricultural output from farm to market for trade. With the rapid development of other sectors in this digital era, it is mandatory to keep agriculture sector abreast with other economy contributing sector [7].

In the traditional supply chain, proper coordination among harvest, operations, logistics, and optimization of operation is complex and unwieldy errand [4]. Risk of imbalance in production capacity and improper metric for quality evaluation of the crop are significant causes of the decline in the supply chain system. Blockchain technology

adoption in agricultural supply chain sector promises to resolve above mentioned problems.

II. TRADITIONAL AGRICULTURE SUPPLYCHAIN

In the most of developing countries of South Asia, its native farmers are deprived of prosperous life. Their income is not on same scale as farm yield. Multiple factors contribute towards low income of farmers.

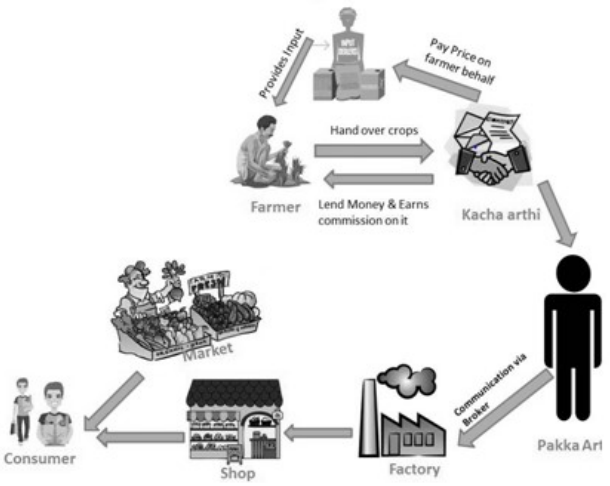


Fig. 1.Traditional Agriculture System based on intermediaries

Fig.1 shows the prevailing Agricultural System [8]. The process starts from crops collection from farm to be supplied to market and made available for consumption, involves changing of multiple hands. Chain of intermediaries charges specific commission percentage for every errand. i) *Kacha arthi [8]* (a middleman, between the pakka arthi and the farmer, charges commission from farmer but doesn't takes crop title, crop ownership is still with farmer). ii) *Pakka arthi [8]* (is wholesaler who purchases agriculture output in the open market, either via bidding process or directly from the kacha arthi).

Farmer when in need approaches kacha arthi for input of next crop, farming goods and loan. Crops once harvested are handed over to kacha arthi and after deduction of loan and agricultural input cost or any other liability along with some specific percentage commission fee, a small share of profit is given to the farmer once crops are sold at market.

Kacha arthi transfers the goods to the pakka arthi, another agent, who either directly sell it to in market or through brokers to the factories. Further processes take place at factories and are made available for the consumption to customers in different forms. A large number of intermediaries are added to the chain and commission is charged on every stage, result in the low share of profit for farmer.

Most of the time farmers sell crops at an offered rate in the nearest market as they are unaware of government decided rates and other market rates of crops. No proper metrics for crop quality evaluation are defined, pakka arthi decide it along with rates paid for crop at the time of dealing. Moreover, the demand of crop is not properly communicated to farmer due to lopsided relation between farmer and market, results in excessive production than demand. To avoid crop wastage farmers accept rate offered in the nearest market, most of the time it not even covers actual cost invested from sowing to bringing it to market.

Tax system for agriculture is neither properly formulated nor implemented and revised from long time. Tax is being charged in two portions, a fixed amount on agricultural land. Second is Variable tax with fixed rates on agricultural income [9]. Tax evasion and embezzling occur due to false record keeping of crop production and its sell in market.

Demand forecasting of agricultural product is a difficult task to predict due to absence of proper record keeping related to product yield, indirect communication between farmer and customer, and lack of transparency in supplychain. Imbalance in supplychain of agriculture product occur due to unpredictable demand forecasting,it sometimes results in lack of crop availability and mostly excessive production. Resulting in large amount of crop wastage and reduced profit.

In the traditional supply chain, the farmer gets a small portion of overall profit, which deviates towards zero due to crop wastage. Profit maximization and wastage deduction can be achieved by adoption of an openly accessible and transparent system. Bridging communication gap between farmer and customer, make it feasible for farmer to acquire information about market trend and estimate production for next season crop based on market demand.

Farmers are reluctant to go for global export due to hectic procedure of product delivery tracking and the multiple numbers of stages and agents. All these can be made easy by introducing a new system that provides transparent and open access. Here, transparent and visible or open access means information accessiblility to any party (node) participating in supplychain. It resultantly reduces operating cost and improves performance, product status tracking, encourage global export and attracts foreign investors.

III. PROPOSED FINTECH BASED SOLUTION (AGRI-BC)

To satisfy all requirements, a low cost and access controllable database system is considered. It is a distributed database system which emerged a few years ago. A cryptocurrency based system that works on blockchain technology. It allows users to transfer currency securely without involvement of centralized authority using public distributed ledger known as blockchain [10]. Agri-BC based framework is proposed to overcome issues spotted in the traditional agriculture system. Core stakeholders of the agriculture system: Farmer and Buyer (Market) are focused entities of the proposed framework, along with participation of regulatory authorities and food insurer. Framers can utilize this medium to remove intermediaries and to introduce transparency in the network.

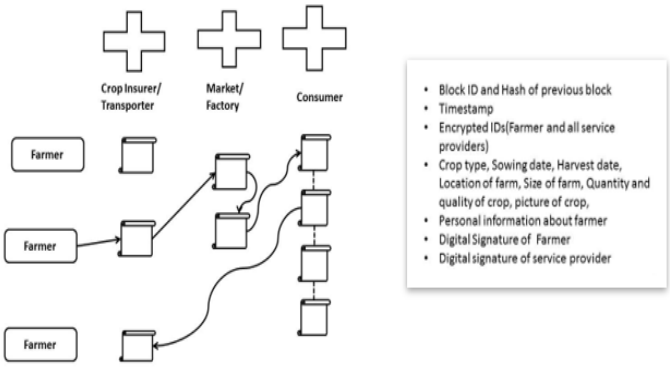


Fig. 2. 2-dimensioal view of proposed framework of Agri-BC (blockchain)

Fig.2 depicts the 2-Dimensional view of the framework based on blockchain technology. Regulatory authority registers the farmer along with its land ownership. Farmer populates the ledger, provide crop details, quantity, sow date, and harvest date. Quality of crop is sensed via sensors. Transaction is verified, if majority of the parties agree, termed as Consensus in blockchain. Transaction block is added to the previous blockchain. All ledgers on nodes participating in blockchain are updated.

Transporter or crop insurer check harvest date and accordingly contact farmer and bind in contract with him. This contract is programmed contract or programming logic termed as SMART Contract in blockchain terminology [6]. Once smart contract binding occurs, transporters collect crop from the farm, thus the authority of crop is shifted to the transporter.

Transporter handovers crops to Factory or Market from where the currency is transferred in the account of farmer and transporter. After processing food, it is made available for consumption. Once it is made available, consumer based on product id or QR code can track all the details of food consumed from sowing till processing.

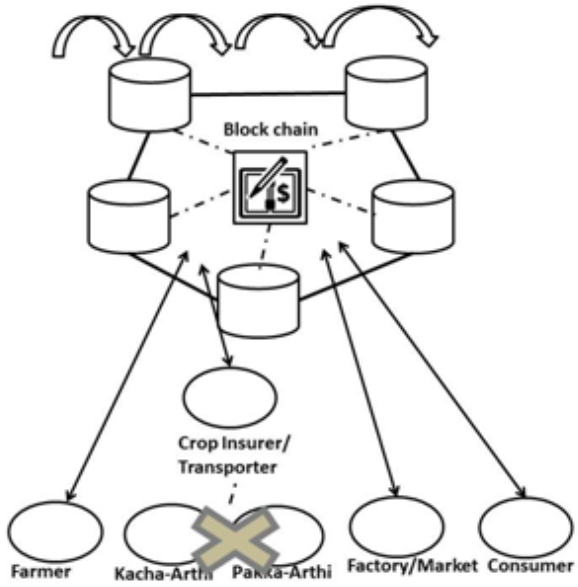


Fig. 3. Entities of Agri-BC(blockchain)

In Fig.3 [11], participating entities: farmer, transporter, factory, consumer and regulatory authority in the proposed Agri-BC system. It demonstrates how roles of kacha and pakka arthi are replaced by a single entity. Bitcoin mechanism, a cryptographic proof is used instead of trust in the third party (kacha arthi) for transaction execution online between farmer and market [12].Transactions are protected using digital signature [13], sent to the market trader (receiver) digitally signed by private key of sender(farmer). For transporting material, owner (farmer) has to prove ownership of private key. Receiving entity (market/trader) will verify the digital signature (ownership of private key) on the transaction using sender's public key.

IV. LAYERED REPRESENTATION OF PROPOSED FRAMEWORK

Fig.4 depicts the architectural representation of the proposed Agri-BC framework. Inspirationally, at initial stage 3 layers are considered: *application layer, block chain layer and storage layer* as shown in CrowdBC framework [14]. Application layer is dedicated for interaction with the user. Second layer is designated with the task related to block chain activities. Third layer works as storage, as vast amount of data on blockchain can have impact on its execution.

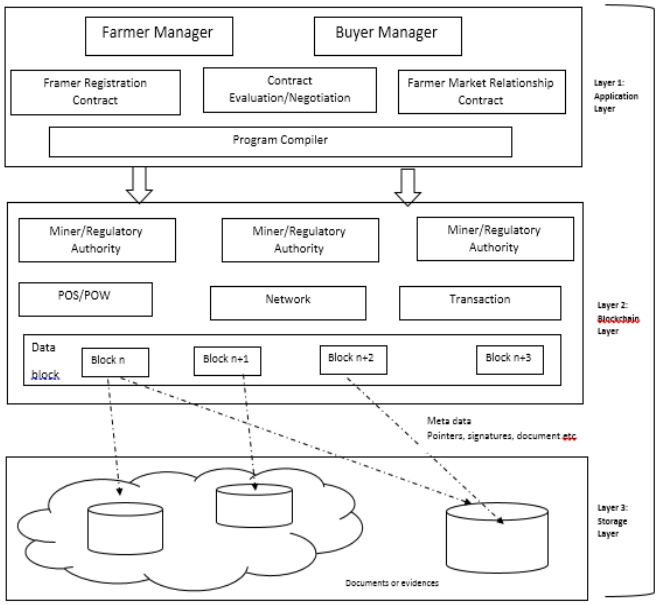


Fig. 4. Layered representation of Agri-BC framework

Farmer uploads data related to his crop before harvest; it is posted on application layer from where buyer access this information. Users are entitled to provide their true identity for registration with key pair for (public and private key). Blockchain layer contains programming compiler for smart contract execution. A contract processing state machine is considered which changes states of transaction on the trigger of the event. For every transaction, a new state machine is produced and the status of the transaction is updated on blockchain.

To overcome limited storage data capability on the blockchain, logical and storage layers are separated. Transaction and entities basic information, which are frequently accessed are stored on blockchain layer whereas details: evidence and documentation are stored in storage layer.

A. Application Layer

This layer contains two core modules in Agri-BC application layer: Farmer Management (FM) and Buyer Management (BM). FM and BM are bounded by smart contract. It is mandatory to register Farmer and buyers at the initial stage on Agri-BC. Then framer populates detail information regarding farm, crop sowing, and harvesting date, yield and others.

At second stage farmer decides terms for smart contract and demands for negotiable rates. On the other side, Buyer Management uploads details demand based on consumption, quality and possible rates. BM decides the transaction fees to be paid to miners in smart contract.

An instance for reputation is also created by default at inception for both farmer and market for evaluation purpose [15]. Reputation of both entities will be updated by each other based on the transaction performance and feedback. Buyer Management Module will traverse uploaded details and undergo the contract state.

B. Blockchain Layer

This middle layer serves two purposes: provides consensus on the order of smart contract framed and runs the state machine which changes transaction states on the trigger. The Smart contract is written on the blockchain, once confirmed by the miner. State machine defines task status and it is updated with transactional status on arrival of input on the blockchain layer [16]. Users are provided access to state machine status on blockchain layer. Information uploaded by farmer and buyer is sent and updated on blockchain layer by the smart contract. Excessive information is routed to the storage layer for the efficient performance of the discussed layer. Hash pointer serves as a connecting point between blockchain layer data and storage layer data.

C. Storage Layer

It is the third layer in Agri-BC architecture, it stores the actual data value, evidence, and details. Data values are signed by the owner's public key. Final Contract conditions and terms are encrypted using the buyer's public key and can decrypt using its private key. This layer is hidden from users: farmers and buyer are unaware of this layer existence. They access data, validate its authenticity and integration using data hash values and digital signature in the blockchain layer.

V. STEPS

Steps followed on Agri-BC framework are discussed below.

1. The first step involves registration of new farmer and buyers. Each user is provided with public and private key pairs.
2. Second step comprises of transaction conformation and status transformation.
3. Third step is for posting of finalized contract terms after negotiation. Buyer pays for the crop in advance and payment is deposited on blockchain for security purpose. An evaluation function is required, when farmer post the details of crops and field, it is evaluated by the miner/food insurer on blockchain instead of the buyer.
4. Fourth step involve evaluation of the contract deal. Each buyer deposit some coins or posses reputation status or past deals record to ensure the trust. Higher reputation indicates a good buyer with secure business transaction. Low reputation shows the unsatisfactory past performance. A minimum threshold is set for the evaluation of the reputation for finalization of the dealings.
5. Last step is reserved for exchange of goods, contract evaluation and assignment of the reward.

VI. CONCLUSION

This research presents a Fintech based supply chain framework to help countries overcome issues faced while practicing the traditional agricultural model. The Blockchain based agricultural supply chain adoption is directly linked to the prosperity of farmers. New directions are also discussed for creating direct and transparent recordkeeping, and communication among all stakeholders.

In this paper comparative analysis of both traditional and proposed system is performed. After study of the incumbent system, blockchain based agricultural supply chain system-Agri-BC- is proposed. Numbers of intermediaries are also reduced in the Agri-BC framework. As market legal bindings with transporter and farmer are based on smart contract, thus the role of kacha arthi and legislative representative is eliminated. No doubt, availability of data on every node, overcomes the issue of data tempering and misrepresentation. As blockchain indelible feature record crop production and sales, ultimately tax authority can accurately charge variable tax on farm yield.

VII. FUTURE WORK

The team is currently working on a dynamic algorithm for implementation of Agri-BC framework from end-user's perspective. Algorithm is expected to facilitate timely transportation of crop and assist farmers in marketing and making smart contract regardless of literacy levels. Minimizing crop wastage will be one of the primary goals.

REFERENCES

[1] J.Ricker-Gilbert, C. Jumbe, and J.Chamberlin, How does population density influence agricultural intensification and productivity? Evidence from Malawi. *Food Policy*, vol. 48, pp.114-128,2015.

[2] M. Usman, Contribution of Agriculture Sector in the GDP Growth Rate of Pakistan, (2016).

[3] U.S. Congress, Office of Technology Assessment, Agricultural Commodities as Industrial Raw Materials.Washington, DC: U.S. Government Printing Office, May, 1991.

[4] Food and Agriculture Section, Planning Commission Islamabad, *Final report of working group on Agricultural Marketing Infrastructure and post harvest management for the 10th Five Year People's Plan 2010-2015*. Islamabad, Pakistan: Dr. Iqrar Ahmed Khan, 2009.

[5] Cultivating Trade: The economic Impact of Indiana's Agricultural Exports

[6] C.P. Timer, "Agriculture and Economic Development," in *Handbook of agriculture economics*, B.L. Gardener and G.C. Rausser, Eds. Amsterdam, North Holland,2002,vol. 2A, pp. 487-546

[7] M. Iqbal, and M. Ahmad, Science & technology based agriculture vision of Pakistan and prospects of growth, 2005.

[8] A.Haq, A. Aslam, A. A. Chaudry, et al, Who is the "arthi": Understanding the commission agent's role in the agriculture supply chain, 2013.

[9] T. M. Bucha, Agriculture Income Tax – "The Reality".

[10] M.Crosby, P. Nachiappan, S. Pattanayak, V. Verma , and Kalyanaraman, Blockchain Technology: Beyond Bitcoin, 2015.

[11] M. Nakasumi, Information Sharing for Supply Chain Management based on block chain Technology, 2017.

[12] S.Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, 2008.

[13] O. Mazonka," Blockchain: Simple Explanation" Working Paper, 2016.

[14] M. Li, J. Weng, A. Yang, et al, CrowdBC: A Blockchain-based Decentralized Framework for Crowdsourcing, 2017.

[15] R. D. Pietro, X.Salleras, M.Signorini, and E.Waisbard, "A blockchain-based rust System for the Internet of Things," In Proc. ACM SACMAT 2018,April 2018.

[16] J. Debus, Consensus Methods in Blockchain Systems FSBC Working Paper.Frankfurt School Blockchain Center, 2017

A Rewriter Model for Urdu Document Concision with Neural Word Embeddings

Maida Shahid[1], Summra Saleem[1], Aniqa Dilawari[1], Usman Ghani Khan[1,2]

maaida.shahid@gmail.com, summra.saleem@kics.edu.pk, aniqa.dilawari@kics.edu.pk, usman.ghani@kics.edu.pk

1. Al-Khawarizmi Institute of Computer Science

2. Department of Computer Science and Engineering UET
Lahore, Pakistan

***Abstract*—Exponentially increased text data on the internet, led to plethora of work on English text concision. Native Languages are lacking in this regard for text concision. Our research work proposes a network to generate comprehensive document summary. Proposed Rewriter Model for the abstractive summarization of Urdu Dawn news generates compact and concise summary; closely related to the human generated summary. Comparison of the results with CNN/ Daily Mail dataset shows that proposed model performs better than the previously proposed baseline models. The calculated rouge score for proposed model are 42.65% using Urdu news dataset.**

Keywords— Urdu, Abstractive Concision, Word Embedding, Attention Distribution

I. INTRODUCTION

Text summarization is a method of compressing the text in such a way that precise and accurate summary is obtained, containing all the important information of the original text also called text concision. Available data on the internet is very vast and unstructured. Moreover, data is increasing day by day. Therefore, automatic text summarization is the need of the hour to address the challenges of growing data. Automated text summarization makes the selection process easier, lessens the reading time and improves the efficiency of indexing. It comes under the title of machine learning, deep learning and data mining.

There are basic two general approaches of summarization. 1) Extractive approach 2) Abstractive approach. In Extractive approach of concision, selected sentences contain most relevant information of the article. In comparison to this, abstractive approach of concision paraphrasing is used to describe the key idea of the article. Therefore, abstractive concision tries to imitate the summary generated by humans.

Urdu is an Indo-Aryan language that has more than 100 million native speakers in Pakistan and India. In Pakistan, many newspapers like Millat, Nawa-e-waqt, Daily Jang and Dawn are published in Urdu. Our research paper takes the advantage of abstractive summarization to generate a compact and more sophisticated summary of Urdu news articles that is closely related to the human generated summary. Word level attention is used for the abstractive summarization in our proposed Rewriter Model. The novel Seq2seq model is used in our Rewriter model that can elect words for summary from text by pointing to the exact location as well as it can generate new words from the vocabulary. The basic advantage of this pointing and generating idea is the ordering of words of the flexible length input sequence. Furthermore, by using bidirectional LSTM, Rewriter can look forward and backward over the input word embeddings.

Training, testing and validation of this model is done on the Urdu Dawn news dataset. In order to test the output of our Rewriter model ROUGE metric is used. Rouge scores are calculated and compared with the scores of recently proposed models on extractive and abstractive summarization. Results show that this model outperforms the state-of-the-art models.

Details of the rest of the sections of the paper are given here. Section II describes the related work that has been done in this field up till now. Section III describes the problem statement and the solution to this problem is given in section IV. Methodology of the Rewriter model is explained along with equations in section V. Section VI and section VII explain implementation details of our Rewriter and qualitative analysis of the results with CNN/Daily Mail dataset respectively. In the end, section VIII concludes the paper along with giving the future directions.

II. LITERATURE SURVEY

Most of the work on text concision is related to extractive summarization. Approaches like greedy [1], graph based [2] and constraint optimization [3] were the traditional ones for extractive summarization. In the past few years, the trend has been shifted towards the approaches based on neural networks. Nallapati has proposed a selector architecture that has the advantage of picking one sentence at a time arbitrarily to generate the summary [4]. Yasunaga proposed a model in which graph based convolutional network is embedded in a recurrent neural network to compute the sentence level attention [5].

Due to the difficulty of generating abstractive summary, a little significant work has been done from recent years. Initiation was done by Rush, who proposed an attention-based encoder to generate the abstractive summary [6]. Nallapati proposed a sequence-to-sequence model based on the works of Rush [7]. Advanced Sequence-to-sequence model was proposed by See that not only copy words via pointing, but also generates novel words. Coverage mechanism was introduced to keep track of the words that

had already been included in the generated summary to discourage repetition [8].

Literature survey related to Urdu text summarization was done and many papers were consulted. Burney presented an add-in for MS Word named "Auto Summarizer for Urdu Language" that can summarize enlightening articles such as economical, scientific and sports as well as news [9]. For the evaluation and development of single documented summary of Urdu, Humayoun proposed a benchmark corpus. Evaluations and experiments can be done by the researchers using open source software tools provided by the author [10].

III. Problem Statement

In this era of the internet, there is a huge amount of data in the form of text available on the internet that cannot be processed by humans manually. As there are a number of documents available, searching for the relevant document and then extracting the useful and related information from them is very tiresome. Therefore, the concept of Automatic text summarization has been introduced. But, all the work has been done in English as it is the international language spoken all over the world. Almost none or very little work has been done in native languages. Urdu is being spoken and understood by more than 100 million natives of Pakistan and India. In Pakistan, many newspapers like Millat, Nawa-e-waqt, Daily Jang and Dawn are published in Urdu and no method of Urdu Text Summarization has been introduced yet.

IV. Solution

To cope with the issues of Urdu Text Summarization, a Rewriter model is introduced that takes the articles and their human generated summaries in Urdu and provides the abstractive summary that closely relates to the human generated summary. Results comparison shows that our model outperforms the baseline model for English text summarization.

V. Methodology

Our Rewriter Model is similar to the sequence-to-sequence attentional model proposed in [8], as it can generate words from a fixed vocabulary as well as it can copy words from the source text at the same time to form a compact understandable summary. Proposed model is extension of sequence to sequence model in addition to Urdu word embedding generation for urdu dataset. Baseline model obtains current LSTM cell state for word prediction at timestep t. Our model employs concept of word dependency and utilize LSTM cell state preserved at time t-1 for word prediction.

. In order to encode the input text in the form of words, Rewriter model uses a single layer of bidirectional LSTM. To decode the encoded words of the text into human readable summary single layer of unidirectional LSTM is used. The Urdu tokenized words are fed into the encoder one after the other, thus producing the encoder hidden states h_{enc}. At every time step t, word embedding of the previous Urdu word is fed into the decoder to produce the decoder hidden states h_{dec}.

$$Score_i = v^t \tanh(W_{enc}h_{enc} + W_{dec}h_{dec} + offset) \quad (1)$$

$$a^t = \frac{\exp(Score_i)}{\sum_{j=0}^{k} exp(Score_j)} \; (i = 0,1,2 \dots j) \quad (2)$$

In the equations (1) and (2) W_{enc} , W_{dec} , v^t and $offset$ are trainable parameters. To understand the complete derivation of the attention distribution a^t, the research paper of Bahdanau [11] can be consulted.

Attention distribution helps the decoder to locate the words that will be used in summary, as it is the probability distribution over the article words. Weighted sum of encoder's hidden states corresponds to context vector and it can be calculated by the formula given in (3).

$$C_{enc}(a^t) = \sum_i a_i^t \, h_{enc\,(i)} \quad (3)$$

Two linear layers are used to produce the probability of vocabulary distribution by normalizing the concatenated hidden states of decoder h_{dec} and context vector C_{enc}.

$$P_{vocab}(C_{enc}) = \frac{\exp(W'(W[h_{dec},C_{enc}]+offset)+offset')}{\sum_{j=0}^{k} \exp(W'(W[h_{dec},C_{enc}]+offset)+offset')} \quad (4)$$

Where W', W, $offset'$, $offset$ are the trainable parameters. P_{vocab} is the probability of the decoded words. Generating probability can be represented by $p_{gen}(C_{enc})$ and copying probability by $1 - p_{gen}(C_{enc})$. The range of the probabilities is [0,1]. Derivation Formula of the generating probability can be seen in the paper by Abigail See [8]. Final vocabulary distribution can be calculated by adding all the words of the vocabulary as well as the words appearing in the articles.

$$P = p_{gen} P_{vocab} + (1 - p_{gen}) \sum_i a_i^t \quad (5)$$

From the above equation (5), it can be seen that Rewriter model can deal with both in vocabulary as well as out of vocabulary (OOV) words, whereas the previous baseline model was limited to the pre-set vocabulary.

A. Loss Minimization

The main objective of research is to minimize the negative log-likelihood of each decoded word on every time step t during training. Equation of the loss function can be written as:

$$loss = -\frac{1}{T} \sum_{i=1}^{T} \log(P) \quad (6)$$

Where T shows the total number of words in the generated summary.

B. Coverage Mechanism

In this Rewriter model, there was a problem of repetition of words in generated summary same as with all the sequence-to-sequence models. Therefore, we also introduce the coverage mechanism so that our model does not give attention repetitively to the same words. Coverage vector C_{cov} is calculated by the given equation (7) at every decoder step t, that helps to calculate the amount of attention given to every word of the article.

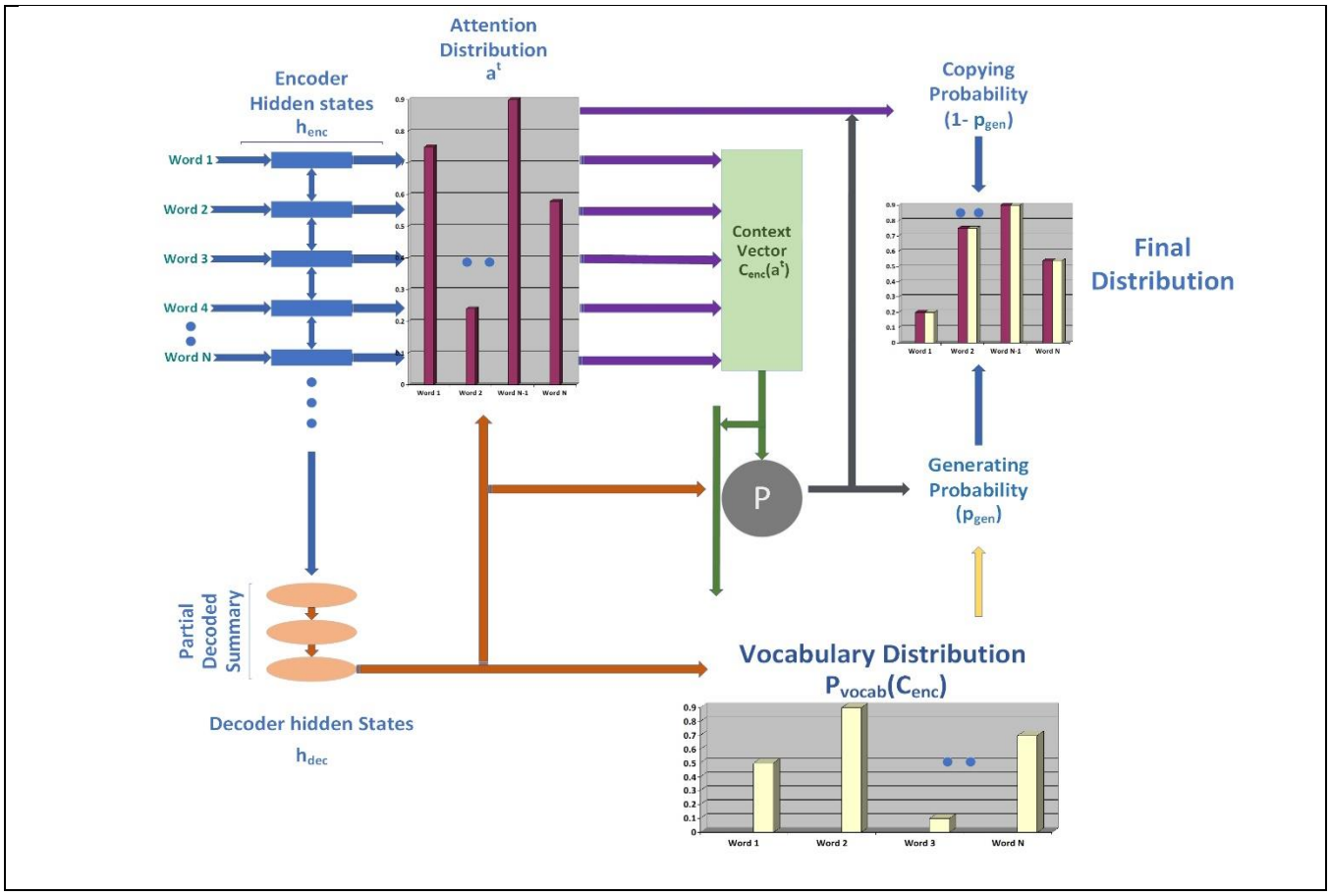


Fig. 1. System Architecture of Rewriter Model. Generating probability in the range of [0,1] is calculated for every timestep of the decoder which tells the probability of the words that are generated corresponding to the words that are directly copied from the article. The attention distribution and the vocabulary distribution are calculated and then added together to get the final distribution.

$$C_{cov} = \sum_{t'=0}^{t-1} a^{t'} \quad (7)$$

To understand the equation of Coverage Vector in depth, Abigail See's paper can be consulted [8].

C. Coverage Loss Minimization

The main objective during training is to minimize the coverage loss as well, so that, the attention is not given to the same words repeatedly. The equation is given as

$$loss_{cov} = \frac{1}{T} \sum_{t=1}^{T} \sum_{w=1}^{W} \min(a_w^t, C_{cov\,(w)}^t) \quad (8)$$

VI. Implementation

Pre-processing steps for Urdu articles and both the datasets of Urdu and English are explained in the next subsections.

A. Pre-Processing Steps

- News articles were saved in text files.
- Tokenization was done by breaking the sentences of the articles into words and extra punctuation marks were removed.
- Unique words were separated from the whole dataset to generate the vocabulary file.
- As the summary and the article were concatenated and stored in a new file having @highlight tag to identify each summary sentence.
- Using the tokenized text, bins files were generated as the model accepts the bin files.
- Bin files were chunked to three splits; train, test and validation split.
- The division was done in such a way that out of total articles 80% were stored in training file, 10% in testing and the remaining 10% in validation file.

B. Dataset

Most of the experiments for English text concision using various models have been done on the news articles of CNN and Daily Mail datasets. The news stories of the CNN/Daily Mail contain the reference summaries as well. Out of the total dataset, 287,226 articles are used in the training set, 11,490 articles in the testing set and 13,368 articles in the validation set.

Our dataset of Urdu contains 5,030 articles of Dawn News, out of which there are 2,382 articles of Dunya News and 2,648 articles of Pakistan News. From the total dataset, 4026 articles are used in the training set, 502 articles in the testing set and remaining 502 articles in the validation set.

C. Training

128-dimensional word embeddings for Urdu were used in all experiments. Vocabulary size was 23,919 unique Urdu words. Articles were truncated to 400 words during the training phase and summaries were limited to 100 words. Batch size was kept to 16 and training was done on the GPU. Training took 1.5 days and 50,597 iterations were done.

Word embeddings were not pre-trained. Rather, they were trained from the initial stage during the training of rewriter model for Urdu articles summarization. Ada-grad optimizer was used with the accumulator's initial value of 0.1 and learning rate was 0.15. Firstly, the trained model was obtained without coverage, then the model was converted into coverage model by introducing the coverage mechanism. Beam search was used for the testing of the model with the beam size of 4. Figure 2 shows the loss graph in which the loss keeps on decreasing as the training progresses.

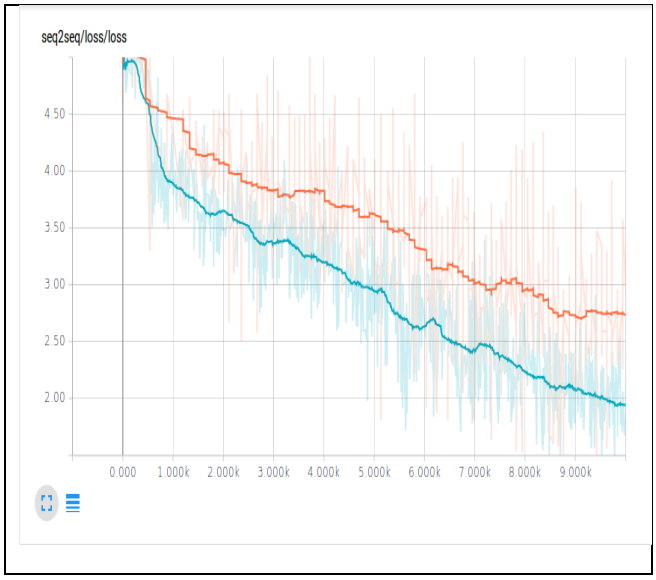


Fig. 2. Loss Graph showing the progress of training

VII. Results

Recall Oriented Understudy for Gisting Evaluation (ROUGE) is one of a metric which is used for the evaluation of the computer-generated summaries with the reference summaries generated by the humans. It is an important package in Natural Language Processing (NLP). Total five metrics of ROUGE are available for evaluating the decoded summaries; ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. These evaluate the automatically generated summaries of various models by counting the total number of overlapping words, word sequences or word pairs with the reference summaries.

Recall, precision and F-scores for Rouge-1, Rouge-2 and Rouge-L are calculated to evaluate the decoded summaries of our model. Rouge scores of both the models for both Urdu and English datasets are shown in Table 1. Rouge-1 score of our rewriter model is 42.65 % which is higher than Rouge-2 and Rouge-L as it refers to 1-gram overlap; single word for decoded and reference summary. The comparison shows that our model gives 3.21% more accurate results than the baseline model which was used by extending it for Urdu word embeddings and vocabulary which can be seen by comparing the summaries of both models given in Table II.

TABLE I. Rewriter Model Rouge Scores

	Rouge-1	**Rouge-2**	**Rouge-L**
Baseline Model (Urdu dataset)	39.44	17.66	36.37
Proposed Rewriter Model (Urdu dataset)	**42.65**	**19.37**	**38.97**
Proposed Rewriter Model (English dataset)	**39.68**	**17.29**	**36.42**

TABLE II. Article, Reference and Decoded Summary of both English and Urdu

Article 1:

سماعت چیف جسٹس پاکستان ثاقب نثار کی سربراہی میں سپریم کورٹ کے 3 رکنی بنچ نے کی سماعت کے دوران ڈائریکٹر جنرل فوڈ نے عدالت کو بتایا کہ حلیب اینگرو فوڈز نیسلے ایوری ڈے فوجی شکر گنج سمیت فیصد کمپنیاں ٹی وائٹنر بنا رہی ہیں گزشتہ سماعت کے دوران عدالت نے ڈبے پر جلی حروف میں 90 یہ دودھ نہیں لکھنے کا حکم دیا تھا تاہم عدالتی احکامات پر ٹی وائٹنر بنانے والی کمپنی ترنگ نے ڈبے کا نیا ڈیزائن پیش کیا جسے عدالت نے مسترد کر دیا عدالت نے ٹی وائٹنر کمپنی کے وکیل سے استفسار کیا کہ اس میں کہاں لکھا ہے کہ یہ دودھ نہیں ہے جس پر ٹی وائٹنر کمپنی کے وکیل کا کہنا تھا کہ ڈیزائن پر لکھا ہے کہ بچوں کے دودھ کے متبادل نہیں سپریم کورٹ لاہور رجسٹری میں ملاوٹ شدہ دودھ کی فروخت کے خلاف ازخود نوٹس کیس کی سماعت کے دوران چیف جسٹس آف پاکستان کا کہنا تھا کہ ملک میں ڈبے کے تمام دودھ جعلی اور مضر صحت ہیں سماعت کے دوران سپریم کورٹ نے ملک بھر میں دودھ کی پیداوار کیلئے بھینسوں کو لگائے جانے والے ٹیکوں پر پابندی عائد کر دی چیف جسٹس آف پاکستان ثاقب نثار کا کہنا تھا کہ دودھ کی پیداوار بڑھانے کے لیے بھینسوں کو لگائے جانے والے انجیکشن سے کینسر جیسی بیماریاں پھیل رہیں ہیں جبکہ ڈبے کے دودھ میں کینسر کا سبب بننے والا فارملین کیمکل کی موجودگی بھی پائی گئی ہیں ان کا کہنا تھا کہ بچے اور بڑے کینسر زدہ بھینسوں کا دودھ پینے پر مجبور ہیں انہوں نے کہا کہ پاکستان میں ڈبے کے تمام دودھ جعلی اور مضر صحت ہیں کیونکہ تمام ڈبہ پیک دودھ میں فارملین کیمکل موجود ہے جو کہ انسانی صحت کے لیے انتہائی خطرناک ہے انہوں نے مزید کہا کہ اس حوالے سے شہریوں میں سخت تشویش پائی جاتی ہے بعد ازاں عدالت نے ملاوٹ شدہ دودھ کی فروخت کیخلاف ازخود نوٹس کیس کی مزید سماعت دو ہفتوں کے لئے ملتوی کر دی سپریم کورٹ لاہور رجسٹری میں غیر رجسٹرڈ شادی ہالز سے متعلق ازخود کیس کی سماعت کے دوران چیف جسٹس پاکستان

ثاقب نثار نے ملک بھر کی عدالتوں کو شادی ہالز کو حکم امتناعی دینے سے روک دیا سماعت کے دوران لاہور ڈویلپمنٹ اتھارٹی ایل ڈی اے کے ڈائریکٹر جنرل ڈی جی عدالت میں پیش ہوئے ڈی جی ایل ڈی اے نے عدالت کو بتایا کہ شہر کے 186 شادی ہالز کا سروے کر لیا ہے ان کا کہنا تھا کہ جب بھی کسی شادی ہال کیخلاف کارروائی کرتے ہیں تو وہ عدالتوں سے حکم امتناعی لے آتے ہیں جس پر سپریم کورٹ نے ملک بھر کی عدالتوں کو شادی ہالز کو حکم امتناعی دینے سے روک دیا عدالت نے ڈی جی ایل ڈی اے کو احکامات جاری کرتے ہوئے کہا کہ تمام شادی ہالز کے مالکان کو نوٹس جاری کیا جائے اور جو شادی ہالز غیر قانونی ہیں انکے خلاف کارروائی کر کے رپورٹ جمع کروائیں سماعت کے دوران الجنت شادی ہال کے مینیجر کا کہنا تھا کہ جب 1997 میں قانون بنا تو شادی ہالز پہلے ہی بن چکے تھے جس پر چیف جسٹس نے ریمارکس دیے کہ ایل ڈی اے الجنت شادی ہال کو سب سے پہلے شنوائی کا نوٹس دیکر مسمار کرے چیف جسٹس ثاقب نثار نے پنجاب حکومت کی جانب سے اشتہارات کی بندر بانٹ پر از خود نوٹس لے لیا چیف جسٹس نے حکومت پنجاب کی تشہیر کے لیے کیے گئے اخراجات کا ریکارڈ طلب کرتے ہوئے کہا کہ اشتہاری مہم کم کر دیں

Reference Summary:

سپریم کورٹ لاہور رجسٹری میں ٹی وائلٹز کی فروخت کیخلاف سماعت کے دوران چیف جسٹس آف پاکستان ثاقب نثار کا کہنا تھا کہ پاکستان میں ڈبے کے تمام دودھ جعلی اور مضر صحت ہیں

Decoded Summary by Baseline Model:

سپریم کورٹ لاہور رجسٹری میں ملاوٹ شدہ دودھ کی فروخت کی سماعت کے دوران چیف جسٹس آف پاکستان کا کہنا تھا کہ ملک میں ڈبے کے تمام دودھ جعلی اور مضر صحت ہیں سماعت کے دوران سپریم کورٹ نے ملک بھر میں دودھ کی پیداوار کیلئے بھینسوں کو لگائے جانے والے ٹیکوں پر پابندی عائد کر دی

Decoded Summary by Rewriter Model:

سپریم کورٹ لاہور رجسٹری میں ٹی وائلٹز کی فروخت کیخلاف ازخود کیس کی سماعت کے دوران چیف جسٹس آف پاکستان ثاقب نثار کا کہنا تھا کہ پاکستان میں ڈبے کے تمام دودھ جعلی اور مضر صحت ہیں

Article 2:

Feeling sleepy at your desk? Well, this docile dog could just sum up how you feel. Shaheen Pirouz from Denton, Texas, filmed her tiny pet canine being propped up and repeatedly falling forwards. Footage shows him being positioned on his back legs, with his eyes immediately starting to close. He then flops over to one side as he falls into a deep sleep. Pirouz is heard cooing in the background as she watches the sweet moment unfold. `Puppy had been awake for a few hours and couldn't help but fall asleep even in the funniest positions, ' the filmer later wrote online. Ready for bed: Shaheen Pirouz from Denton, Texas, filmed her tiny pet canine being propped up and repeatedly falling forwards. Good night! Footage shows him being positioned on his back legs, with his eyes immediately starting to close.

Reference Summary:

Shaheen Pirouz from Denton, Texas, filmed her tiny pet canine being propped up and repeatedly falling forwards.

Decoded Summary by Rewriter Model:

Shaheen Pirouz from Denton, Texas, filmed her tiny pet canine being propped up and repeatedly falling forwards. Footage shows him being positioned on his back legs, with his eyes immediately starting to close. He then flops over to one side as he falls into a deep sleep.

The summaries generated by our Rewriter Model can be compared with human generated reference summaries as depicted in Table II. The "Article 1" tag shows the sample of Urdu article, under the "Reference Summary" is the human generated summary of that article, below the tag of "Decoded Summary by the baseline model" is the summary generated by the previous baseline model and "Decoded Summary by Rewriter Model" shows the summary generated by our newly proposed Rewriter Model. Under the tag of "Article 2" is the CNN news article along with its reference summary (human generated) and the decoded summary generated by the baseline model. The two summaries have been taken for comparison. The rouge scores of the two summaries have been shown in Table I which delineates the difference between the rouge scores of the two summaries.

The content of the article that is closely associated with the reference summary is highlighted with the brownish green color. As shown in the decoded summary of the Article 1 by the Rewriter Model, word "ٹی وائلٹز" highlighted by pink color is generated from the vocabulary as it was not present in the line that was used in the summary. The word "ازخود کیس" highlighted by blue shows that it is exactly copied from the source text because it is not in reference summary. Whereas, the summary generated by the baseline model is exactly copied from the source text. This illustrates that the summaries generated by our Rewriter model are closer to human generated summaries as it does not merely copy the text from the source article but generates the unique words too. This model also generates more concise and compact summaries as compared to the baseline model. Thus, our model outperforms state-of-the-art model proposed for English datasets when we tested them after extending for Urdu word embeddings. Article 2 shows that our model can also generate the summaries of the English news articles as we have also trained our system on the CNN/Daily Mail dataset.

VIII. Conclusion

In this paper, we proposed a Rewriter model for the abstractive summarization of Urdu news articles using word embeddings. Introduction of coverage mechanism shows that repetition of words and inaccuracies has been minimized. We experimented this model on the articles of Urdu Dawn news and compared the results with the previously proposed Baseline abstractive model. The comparison showed that our model outperformed the state-of-art baseline model.

ACKNOWLEDGMENT

We would like to express our earnest tribute to the National ICT R&D Fund for aiding our research work. The authors would also like to acknowledge the organization (KICS) and full team (colleagues and management) for their support, dedication, technical sessions and knowledge sharing.

REFERENCES

[1] Goldstein, Jade, and Jaime Carbonell. "Summarization:(1) using MMR for diversity-based reranking and (2) evaluating summaries." *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*. Association for Computational Linguistics, 1998.

[2] Erkan, Günes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of artificial intelligence research* 22 (2004): 457-479.

[3] McDonald, Ryan. "A study of global inference algorithms in multi-document summarization." *European Conference on Information Retrieval*. Springer, Berlin, Heidelberg, 2007.

[4] Nallapati, Ramesh, Bowen Zhou, and Mingbo Ma. "Classify or select: Neural architectures for extractive document summarization." *arXiv preprint arXiv:1611.04244* (2016).

[5] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 452–462.

[6] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." *arXiv preprint arXiv:1509.00685* (2015).

[7] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016b. Abstractive text summarization using sequence-tosequence rnns and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290.

[8] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." *arXiv preprint arXiv:1704.04368* (2017).

[9] Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. "Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors."

[10] Humayoun, Muhammad, Rao Muhammad Adeel Nawab, Muhammad Uzair, Saba Aslam, and Omer Farzand. "Urdu Summary Corpus." In *LREC*. 2016.

[11] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Stock Price Forecast Using Recurrent Neural Network

Shakir Ullah
Department of computer Science
Namal College
Mainwali, Pakistan
shakir201650@gmail.com

Noman Javed
Department of computer Science
Namal College
Mainwali, Pakistan
Noman.javed@namal.edu.pk

Ambreen Hanif
Department of computer Science
Namal College
Mainwali, Pakistan
Ambreen.hanif@superior.edu.pk

Ali Abdullah
Department of computer Science
Namal College
Mainwali, Pakistan
mianaliabdullah@gmail.com

***Abstract*—Investors and researchers have continuously been trying to predict the behavior of the stock market. The accurate predictions can be helpful in taking timely and correct investment decisions. Many statistical and machine learning based techniques are proposed. Neural Networks are among the ones having the potential to model the nonlinear behavior of the market. Since there are many different types of neural networks available and a number of factors can influence the stock market, the choice of network and the choice of data is extremely important. These can have a drastic impact on the accuracy of the forecast. We report the findings of employing a systematic approach for the design of neural network and selection of relevant data features. Recurrent neural networks are found to outperform others when tested over four stocks of Pakistan Stock Exchange.**

***Keywords*—Pakistan Stock Market, Machine Learning, Feedforward Neural Network, Deep Neural Network, Recurrent Neural Network, PSX, Multi day Forecasting, stock price**

I. INTRODUCTION

Due to its dynamic and non-linear nature, forecasting the stock market is extremely challenging. Yet it presents itself as an attractive venue for investments. To maximize the return on investment (ROI), investors and researchers have been involved in devising strategies for timely and right decision making. The accurate forecast of the market movement is a key to these investment decisions. The accuracy of the forecast is heavily dependent on the choice of forecasting model and data. There is a strong opinion that stock prices proceed stochastically. Random Walk Theory [1], [2] and Efficient Market Hypothesis [3] suggest that current price of the stock is a reflection of all the available information, thus making it unpredictable. On the other hand, a number of studies [4, 5, 6, 7, 8] suggest the contrary by providing the empirical evidence. During the last decade, there is a resurgence in stock market forecast, primarily because of increase in computational ability to deal with high volume and high-speed data as well as due to adaptability of the historical data [9, 10]. A variety of machine learning models have been investigated for stock market prediction. Support vector machine [11], evolutionary computing [12, 13], random forest [14], fuzzy systems [15] and artificial neural networks (ANN) [16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. These techniques are not only experimented independently, but hybrid approaches are investigated as well. Mostly these hybrid approaches apply ANN in combination with evolutionary computing [26, 27, 28] and/or fuzzy systems [29, 30] or support vector machine [31, 14].

The stock market is under influence of so many factors ranging from past data of a stock to the data of other stocks, from political situation of a country to situation of its economy, from macroeconomic indicators to company level information. Careful choice of data to forecast the future price of a stock has become extremely important. A number of approaches restrict themselves to just daily stock price data [11, 32, 33] while others use data from diverse sources [34, 35]. Some approaches investigate the use of technical indicators [34, 36, 35] while some suggest minimizing the input data without sacrificing the prediction accuracy. It, however, has been observed that this choice of data is mostly a matter of personal preference. If one opts for using a huge amount of diverse data, feature selection and dimensionality reduction [26, 37, 38, 39] techniques are required to filter out irrelevant attributes and finding the appropriate combination of input features.

Since ANN is the most widely used machine learning techniques in stock market prediction and yields better results [35, 40, 20, 9]. Our objective in this paper is to employ a systematic approach for finding the optimal network architecture and the most relevant features from input data for the stocks of Pakistan Stock Exchange (PSX). The secondary objective of the work is to study the usability of this model and feature set for the multi-day forecast. Rest of the article is organized as follow: Next two sections presents the related work and methodology respectively. Section IV describes the systematic selection of Neural network model. Section V describes the criteria for selecting the appropriate features.

Section VI presents the details of the experiments conducted and analysis of the results. We conclude by presenting the prospects in section VII.

II. RELATED WORK

Scientists and researchers have conducted numerous studies using machine learning techniques to predict the stock market. These studies are not limited to the forecast of stock prices. These studies range from the generation of trading strategies to the optimal portfolio building. But we are limiting this section for the review of stock price prediction.

Devadoss, A. V. et.al [20], Yang, B. et.al [32], Moghaddam, A. H, et.al [25], S. M. A., Burney et.al [41] and Zia, S. N, et.al [24] used ANN with back propagation to predict the Bombay Stock Exchange (BSE), Shanghai Stock Exchange Composite index, NASDAQ and Karachi Stock Exchange (KSX) respectively. They found ANN as the more general and flexible modeling tool for forecasting. It is reported that ANN perform better in stock prediction as compared to the other techniques [35]. The paper [35] focuses on the profitability of ANN models. Their idea is to focus on profitability and directional prediction accuracy. Stock prediction model will be of importance when they will be increasing profit. Multiple input parameters have shown varying results on the same ANN model. Four categories are used for classification model, predicted time interval, input variables and result. ANN performed better than all the other benchmarked strategies.

The studies [26] and [42] used hybrid model of ANN and genetic algorithm (GA) for feature discretization. Since the number of input features can grow quite a lot, GA is employed to filter out the less relevant ones, a process called dimensionality reduction. Debashish, D, et.al [43] used the hybridization of data mining and ANN techniques for the Pharmaceutical sector stock predictions of Dhaka Stock Exchange (DSE). They found that the hybridization of data mining and neural network technique give positive performance improvement of prediction and this study has potential to predict stock forecasts for pharmaceutical sector. Another common hybrid approach is the neural networks and fuzzy logic as reported by [44].

The above mentioned studies established the usability of neural networks for Stock Exchange price prediction with better accuracy. Our study focuses on evaluating different ANN (Feed Forward Neural Network (FFNN), Deep Neural Network (DNN) and Recurrent Neural Network (RNN) using daily prices and some of the technical indicators as feature set for PSX to find out which technique(s) and feature set is better for prediction.

Another significant difference between our work and the others is the multi-day ahead forecast. Moghaddam, A. H, et.al [25] used multi-day forecasting to forecast the daily NASDAQ stock exchange rate with four and nine prior days. They used R-squared as the evaluative measure. [45] also reported 5 day and 10- day ahead stock price prediction. They employed self-evolving –recurrent_–neuro-fuzzy system optimized through harmony search. The results of their work for 5 days and 10 days ahead are not significantly different from ours, although they did not report RMSE or other measures specifically.

III. METHODOLOGY

Pakistan Stock Exchange (PSX) is among the leading markets in recent years. No significant work has been reported regarding the analysis and prediction of PSX to the best of our knowledge. In this work, we are interested in studying the extent to which neural networks can be used to predict stock prices listed in PSX. Instead of picking any neural network, we applied a systematic approach for the selection of best model and best input feature set. Our methodology is presented in figure 1 and can be listed in the following steps:

1) Empirical evaluation of neural networks using daily stock price data.
2) Feature Selection: Selecting the relevant technical indicators
3) Multiday stock price forecast

A. Dataset

As we are focusing on PSX, data of four companies listed in PSX from January 2004 to February 2018 is used. The data is imported through Quandl, a famous marketplace for financial and economic data. These companies remain the part of KSE100 index during this period of evaluation

- Pakistan State Oil (PSO) 3401 (19/01/2004 to 19/02/2018)
- Oil and Gas Development Company (OGDC) 3408 (19/01/2004 to 19/02/2018)
- Engro Corporation (ENGRO) 4358 (03/01/2000 to 19/02/2018)
- Pakistan Petroleum limited (PPL) 3323 (16/05/2004 to 19/02/2018)

Each row of the data corresponds to these five features, daily values of opening price, lowest price of the day, the highest price of the day, closing price and volume of traded shares.

B. Data Pre-processing and Distribution

Daily stock quote data is normalized using the following formula before fed into the model.

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X denotes the value which should be normalized, X_n denotes the normalized value, X_{min} is the minimum value of the closing price in the provided data and X_{max} is its maximum value.

The data is distributed in training and test period on a ratio of 75 to 25. The data is arranged temporally where the first portion is the training period while the later one is the testing period. 20 percent part of the training set is used as the validation set.

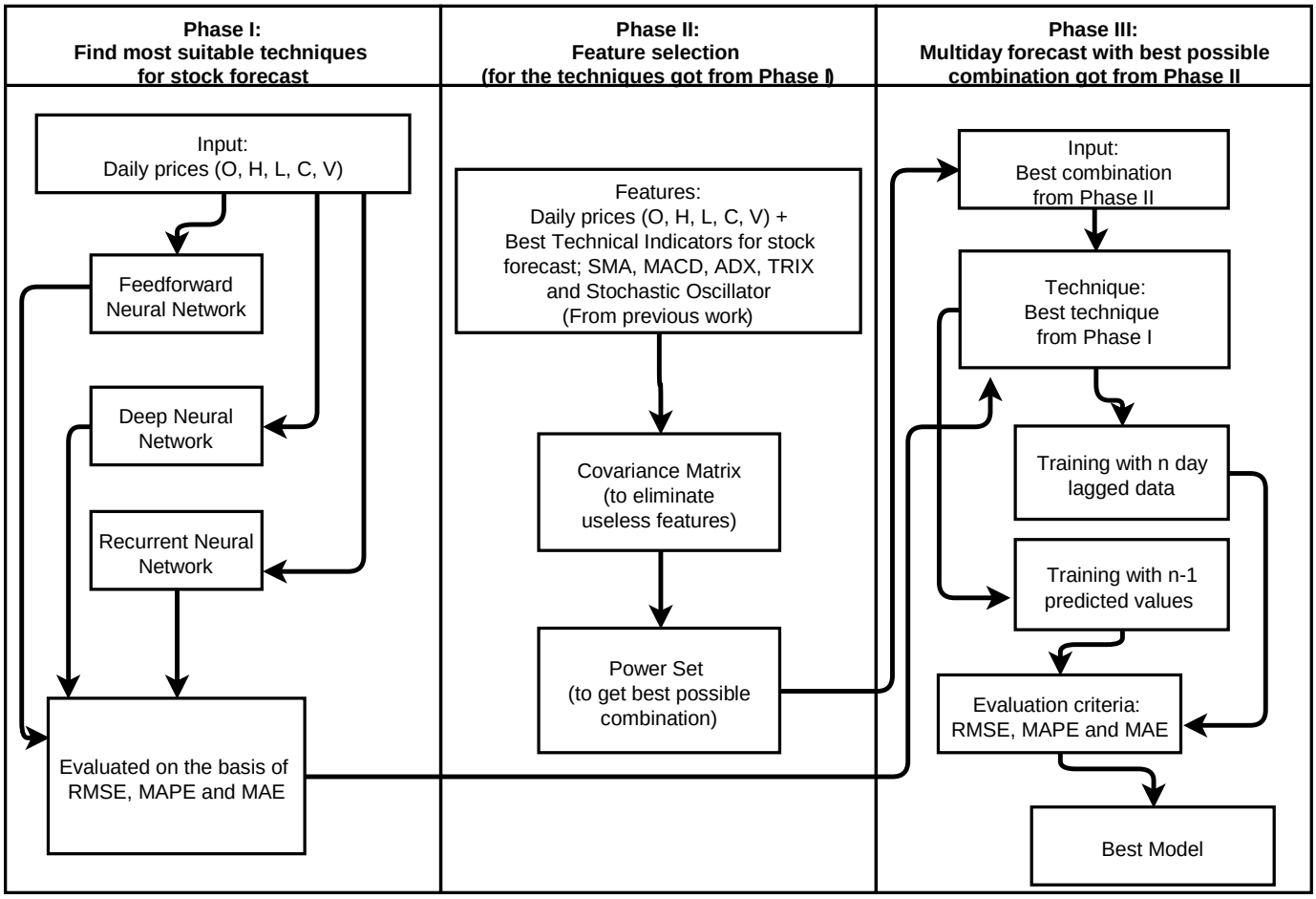


Fig. 1. Overview of the proposed system.

C. Technical Indicators

Technical indicators are the fundamental part of the technical analysis (short-term trading) and are mathematical ways to interpret the future direction and trends of stocks using the historical data. These indicators are mostly used by short-term traders in identifying trading strategies and in turn taking timely investment decisions. There are hundreds of these indicators available in literature and they are mostly used in combinations to form a trading strategy. The choice of technical indicators is thus extremely important. For this study, the choice of these indicators is inspired by the results of [12]. The details of these technical indicators are presented below:

1) *Simple Moving Average (SMA)* is a trend indicator calculated as an average price over a certain period.
2) *Moving Average Convergence Divergence (MACD)* is a trend-following momentum indicator that shows the relationship between two moving averages of prices. The MACD is calculated by subtracting the 26-day exponential moving average (EMA) from the 12-day EMA. A nine-day EMA of the MACD called the "signal line", is then plotted on top of the MACD, functioning as a trigger for buy and sell signals.
3) *Stochastic Oscillator* is a momentum indicator comparing the closing price of a security to the range of its prices over a certain period of time [46].
4) *The Triple Exponential Average (TRIX)* indicator is an oscillator used to identify oversold and overbought markets, and it can also be used as a momentum indicator [47].
5) *The Average Directional Index (ADX)* is an indicator used in technical analysis as an objective value for the strength of a trend. ADX is non-directional, so it quantifies a trend's strength regardless of whether it is up or down [48].

D. Evaluation Criteria

The models developed in the following sections are evaluated by computing the Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These errors are computed using the following formulas respectively:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{a_t - p_t}{a_t}\right|$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|a_t - p_t|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(a_t - p_t)^2}$$

where a is the actual value and p is the predicted value.

IV. EMPIRICAL EVALUATION OF NEURAL NETWORKS

Instead of choosing any random network for multiday forecast, we rely on empirical evaluation of multiple models. For this purpose, we developed the following three networks

1) Feed Forward Neural Network
2) Recurrent Neural Network
3) Deep Neural Networks

To automate the process of selecting optimal parameters for a particular network, we relied on hyper-parameters optimization capability of hyperopt library [49]. To start the optimization process, the library is supplied with the options for learning rate, the number of neurons, activation functions, and the loss functions. The optimized model obtained is then supplied with the data for training.

A. Feedforward Neural Network

In FFNN all layers are forward connected i.e. neurons from input layer are connected to hidden layer and neurons from hidden layer are connected to that of the output layer. There are no backward connections in these layers. FFNN is widely used in the stock prediction problems [20, 25, 50, 24, 26] The architecture of FFNN as generated by Hyperopt is given in table I.

TABLE I
FFNN ARCHITECTURE

Parameters	Parameter Space	Output architecture
Hidden Layer Neurons	5, 10, 20, 40	10
Activation Function	sigmoid, relu, tanh, linear	relu
Loss Function	mae, logcash, mape, mse	mse
Learning Rate	0.01, 0.001, 0.0001, 0.0002	0.001

B. Deep Neural Network

Deep Neural Network (DNN) is rapidly gaining attraction of machine learning community. They are producing better results especially in the fields of speech recognition, image processing and computer vision, natural language processing, and social network filtering etc. In deep neural network instead of neurons, there are networks acting as layers. The output of one network becomes the input of another network and thus forms a deep network. DNN is gaining attraction in financial market but the available studies [9, 51] are still very limited. The best architecture of DNN as prescribed by Hyperopt is presented in table II.

TABLE II
DNN ARCHITECTURE

Parameters	Input parameter Space	Output Architecture
Hidden Layer 1 Neurons	128, 256, 512, 1024	256
Hidden Layer 2 Neurons	128, 256, 512, 1024	256
Hidden Layer 3 Neurons	128, 256, 512, 1024	512
Hidden Layer 4 Neurons	128, 256, 512, 1024	512
Activation Function	sigmoid, relu, tanh, linear	relu
Loss Function	mae, logcash, mape, mse	mse
Learning Rate	0.01, 0.001, 0.0001, 0.0002	0.0001

C. Recurrent Neural Network

Recurrent Neural Network (RNN) make use of sequential data because each neuron can make use of its internal memory to preserve information about the previous data. RNN has loops which allow information to be passed through the neurons while reading in an input. In RNN there is a backpropagation among the layers. Even a neuron can have back connection to itself. RNN can be single layer or multilayer network. Studies show that RNN is used to predict Stock market [52], [28]. The best architecture of RNN as recommended by hyperopt is presented in table III.

TABLE III
RNN ARCHITECTURE

Parameters	Input parameter Space	Output Architecture
LSTM 1 Neurons	128, 256, 512	256
LSTM 2 Neurons	128, 256, 512	128
Activation Function	sigmoid, relu, tanh, linear	relu
Loss Function	mae, logcash, mape, mse	mse
Learning Rate	0.01, 0.001, 0.0001, 0.0002	0.0002

D. Comparison

All these networks are trained using the historical data of four stocks of PSX. Results presented in figure 2 suggest that RNN outperforms the other two by minimizing the evaluation errors.

V. FEATURE SELECTION: SELECTING THE MOST RELEVANT TECHNICAL INDICATORS

The next step is to find the optimal combination of the historical stock data and technical indicators for the prediction of the stock price. The subset of relevant features is computed by calculating the covariance matrix. A power set of the selected features is then computed. The network is trained using the elements of the power set as input features. Each network is then evaluated based on RMSE, MAPE, and MAE. The results of the table IV suggest the most suitable combination of the input features. These results indicate that technical indicators do not play a significant role in the price prediction of a stock.

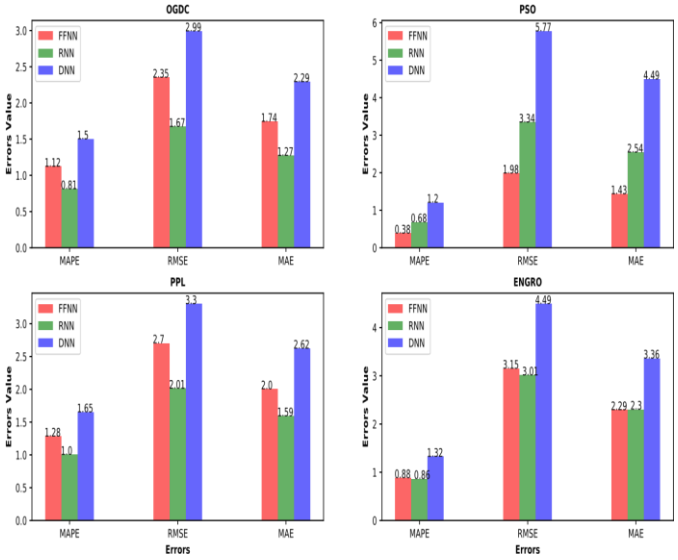


TABLE IV

BEST THREE COMBINATIONS USING POWER SET

Stock	Best Three feature sets	RMSE, MAPE, MAE
OGDC	1.Close 2.Open, Low, Close 3.Open, High, Low, Close	1. 1.64,0.95,0.57 2. 1.62,1.20,0.78 SMA3.1.66, 1.22,0.80
PPL	1. Close 2. Open, Close 3. Open, Close, SMA	1. 1.27, 1.02, 0.70 2. 1.54, 1.13, 0.78 3. 1.66, 1.35, 0.91
PSO	1. Close 2. Open, Close 3. Open, Low, Close	1. 2.35, 1.91, 0.53 2. 2.65, 2.16, 0.59 3. 2.88, 2.12, 0.58
ENGRO	1. Close 2. Open, Close 3. Open, Low, Close	1. 0.86, 0.80, 0.29 2. 1.08, 0.68, 0.24 3. 1.65, 1.12, 0.39

VI. MULTIDAY FORECAST OF CLOSING PRICE

As a result of last two steps, we now have most suitable network architecture and most relevant feature set. To achieve our sub-objective of the multi-day forecast of closing price, two strategies are developed.

1) Train the network with n-day lagged data
2) Train the network using n-1 days predicted price

The results of figure 3 indicates no significant differences between both these approaches. It also shows a significant increase in errors from the second day onwards suggesting the decrease in prediction performance as compared to next day forecast. Further investigation is required to get insights into this behavior and is a matter of future work proposed in next section.

VII. CONCLUSIONS AND FUTURE WORK

A step by step approach for the selection of appropriate neural network model and relevant input features is adopted. Results suggest that recurrent neural network outperforms the feedforward and deep neural networks. It is further observed that while technical indicators are important in taking timely investment decision, their role in predicting future stock price is not apparent. The results also report the limitation of these models for the multi-day forecast. Further experimentation is required to investigate this behavior. Use of other network models, the combination of both n-day lagged, and n-1 day predicted values, use of other macroeconomic indicators can be explored for the said purpose. We also want to evolve neural network architectures using evolutionary algorithms. The neural networks can thus be trained as auto-traders in this way and their performance with human traders can be compared. Since evolving such networks can be computationally expensive, exploring parallel and distributed approaches becomes an important prospect.

To the best of our knowledge, this is the first study that explores the usability of neural networks for predicting Pakistan Stock Exchange stocks prices. Next day stock price forecast can reliably be made using the neural networks but a lot more is required to predict multiple days ahead stock movements with accuracy.

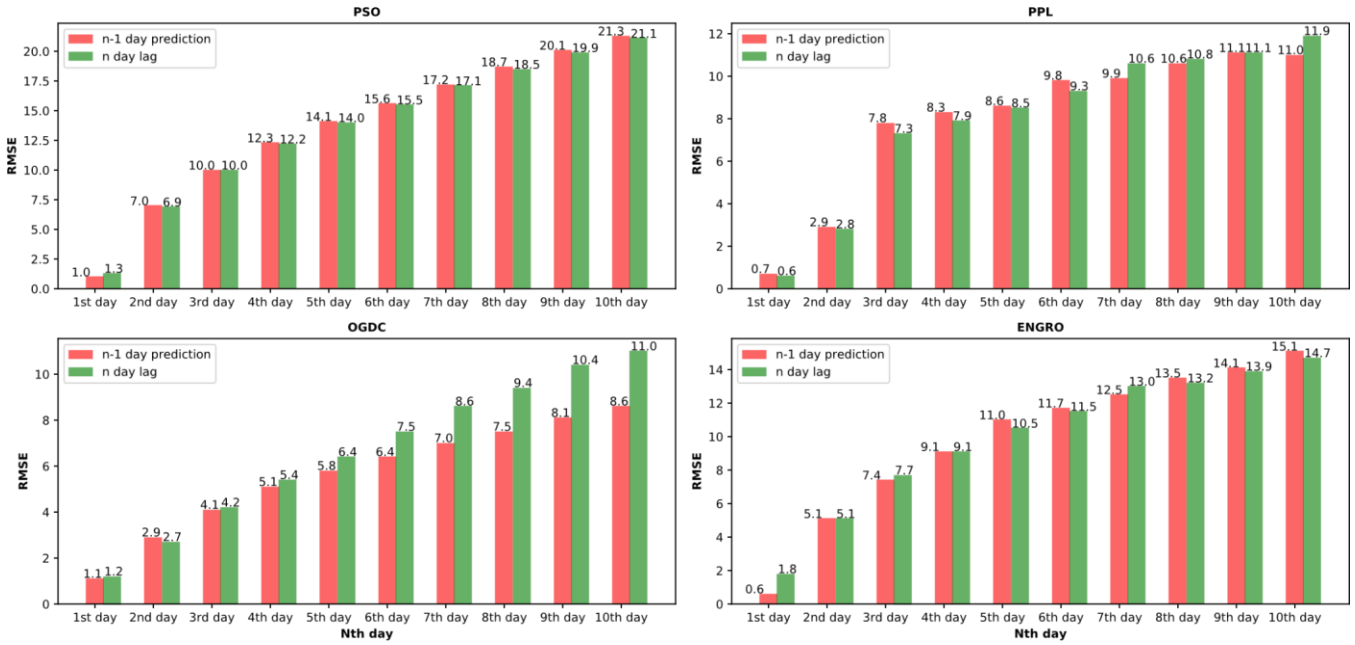


Fig. 3. N-lagged vs N-1 Predicted Comparison.

REFERENCES

[1] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25 (2):383–417, 1970.

[2] Eugene F Fama. Efficient capital markets: Ii. *The journal of finance*, 46(5):1575–1617, 1991.

[3] Eugene F. Fama. Random walks in stock-market prices. *Financial Analysts Journal*, 21:55–59, 1965.

[4] Yaser S Abu-Mostafa and Amir F Atiya. Introduction to financial forecasting. *Applied Intelligence*, 6(3):205–213, 1996.

[5] Ramon Lawrence. Using neural networks to forecast stock market prices. *University of Manitoba*, 333, 1997.

[6] Massimo Santini and Andrea Tettamanzi. Genetic programming for financial time series prediction. In *Proceedings of the 4th European Conference on Genetic Programming*, EuroGP '01, pages 361–370, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-41899-7. URL http://dl.acm.org/citation.cfm?id=646809.704093.

[7] Kun Guo, Yi Sun, and Xin Qian. Can investor sentiment be used to predict the stock price? dynamic analysis based on china stock market. *Physica A: Statistical Mechanics and its Applications*, 469:390 – 396, 2017. ISSN 0378-4371. doi: http://dx.doi.org/10.1016/j.physa.2016.11.114. URL http://www.sciencedirect.com/science/article/pii/S0378437116309384.

[8] Reza Hafezi, Jamal Shahrabi, and Esmaeil Hadavandi. A bat-neural network multi-agent system (bnnmas) for stock price prediction. *Appl. Soft Comput.*, 29(C):196–210, April 2015. ISSN 1568-4946. doi: 10.1016/j.asoc.2014.12.028. URL http://dx.doi.org/10.1016/j.asoc.2014.12.028.

[9] Eunsuk Chong, Chulwoo Han, and Frank C Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205, 2017.

[10] Depei Bao. A generalized model for financial time series representation and prediction. *Applied Intelligence*, 29 (1):1–11, 2008.

[11] Vatsal H Shah. Machine learning techniques for stock prediction. *Foundations of Machine Learning— Spring*, pages 1–19, 2007.

[12] Basit Tanvir Khan, Noman Javed, Ambreen Hanif, and Muhammad Adil Raja. Evolving technical trading strategies using genetic algorithms: A case about pakistan stock exchange. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 335–344. Springer, 2017.

[13] Manuel E Fernandez Garcia, A Enrique, and Raquel Quiroga Garcia. Improving return using risk-return adjustment and incremental training in technical trading rules with gaps. *Applied Intelligence*, 33(2):93–106, 2010.

[14] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4):2162–2172, 2015.

[15] Babita Majhi and CM Anish. Multiobjective optimization based adaptive models with fuzzy decision making for stock market forecasting. *Neurocomputing*, 167:502–511, 2015.

[16] Erkam Guresen, Gulgun Kayakutlu, and Tugrul U Daim.

Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38 (8):10389–10397, 2011.

[17] Zabir Haider Khan, Tasnim Sharmin Alin, and Md Akter Hussain. Price prediction of share market using artificial neural network (ann). *International Journal of Computer Applications*, 22(2):42–47, 2011.

[18] Sunday Olusanya Olatunji, Mohammad Saad Al-Ahmadi, Moustafa Elshafei, and Yaser Ahmed Fallatah. Saudi arabia stock prices forecasting using artificial neural networks. In *Applications of Digital Information and Web Technologies (ICADIWT), 2011 Fourth International Conference on the*, pages 81–86. IEEE, 2011.

[19] AA Adebiyi, CK Ayo, Marion O Adebiyi, and SO Otokiti. Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, 3(1):1–9, 2012. [20] A Victor Devadoss and T Antony Alphonnse Ligori. Stock prediction using artificial neural networks. *International Journal of Data Mining Techniques and Applications*, 2:283–291, 2013.

[21] Chun-Teh Lee and Jia-Shiang Tzeng. Trend-oriented training for neural networks to forecast stock markets. *Asia Pacific Management Review*, 2013.

[22] Hakob Grigoryan et al. Stock market prediction using artificial neural networks. case study of tal1t, nasdaq omx baltic stock. *Database Systems Journal*, 6(2):14–23, 2015.

[23] Gaurav Kshirsagar, Mohit Chandel, Shantanu Kakade, and Rukshad Amaria. Stock market prediction using artificial neural networks. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(5), 2016.

[24] Shams Naveed Zia and Muhammad Zia. Stock price prediction using artificial neural networks: Case study–karachi stock exchange. *Journal of Independent Studies and Research (JISR)*, 3(2):1, 2005.

[25] Amin Hedayati Moghaddam, Moein Hedayati Moghaddam, and Morteza Esfandyari. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, 21(41): 89–93, 2016.

[26] Kyoung-jae Kim and Ingoo Han. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, 19(2):125–132, 2000.

[27] Shahrokh Asadi, Esmaeil Hadavandi, Farhad Mehmanpazir, and Mohammad Masoud Nakhostin. Hybridization of evolutionary levenberg–marquardt neural networks and data pre-processing for stock market prediction. *Knowledge-Based Systems*, 35:245–258, 2012.

[28] Akhter Mohiuddin Rather, Arun Agarwal, and VN Sastry. Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications*, 42(6):3234–3241, 2015.

[29] David Enke and Nijat Mehdiyev. Stock market prediction using a combination of stepwise regression analysis, differential evolution-based fuzzy clustering, and a fuzzy inference neural network. *Intelligent Automation & Soft Computing*, 19(4):636–648, 2013.

[30] Esmaeil Hadavandi, Hassan Shavandi, and Arash Ghanbari. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *KnowledgeBased Systems*, 23(8):800–808, 2010.

[31] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259– 268, 2015.

[32] Yang Bing, Jian Kun Hao, and Si Chang Zhang. Stock market prediction using artificial neural networks. In *Advanced Engineering Forum*, volume 6, pages 1055–1060. Trans Tech Publ, 2012.

[33] Chien-Jen Huang, Peng-Wen Chen, and Wen-Tsao Pan. Using multi-stage data mining technique to build forecast model for taiwan stocks. *Neural Computing and Applications*, 21(8):2057–2063, 2012.

[34] Bin Weng, Mohamed A Ahmed, and Fadel M Megahed. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, 79:153–163, 2017.

[35] Ozg¨ ur Ican and Taha Bugra C¸elik. Stock market predic-¨ tion performance of neural networks: A literature review. *International Journal of Economics and Finance*, 9(11): 100, 2017.

[36] Yauheniya Shynkevich, TM McGinnity, Sonya A Coleman, Ammar Belatreche, and Yuhua Li. Forecasting price movements using technical indicators: Investigating the impact of varying input window length. *Neurocomputing*, 264:71–88, 2017.

[37] Jie Wang and Jun Wang. Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks. *Neurocomputing*, 2015.

[38] Xiao Zhong and David Enke. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, 67:126–139, 2017.

[39] Xiao Zhong and David Enke. A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, 267:152–168, 2017.

[40] Jingtao Yao, Chew Lim Tan, and Hean-Lee Poh. Neural networks for technical analysis: a study on klci. *International journal of theoretical and applied finance*, 2(02): 221–241, 1999.

[41] Syed Muhammad Aqil Burney, Tahseen Ahmed Jilani, and Cemal Ardil. Levenberg-marquardt algorithm for

karachi stock exchange share rates forecasting. *World Academy of Science, Engineering and Technology*, 3: 171–176, 2005.

[42] Mingyue Qiu, Yu Song, and Fumio Akagi. Application of artificial neural network for the prediction of stock market returns: The case of the japanese stock market. *Chaos, Solitons & Fractals*, 85:1–7, 2016.

[43] Das Debashish, Sadiq Ali Safa, and A. Noraziah. An efficient time series analysis for pharmaceutical sector stock prediction by applying hybridization of data mining and neural network technique. *Indian Journal of Science and Technology*, 9(21), 2016.

[44] W. Leigh, C.J. Frohlich, S. Hornik, R.L. Purvis, and T.L. Roberts. Trading with a stock chart heuristic. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1):93–104, 2008.

[45] Rajashree Dash and Pradipta Kishore Dash. Efficient stock price prediction using a self evolving recurrent neuro-fuzzy inference system optimized through a modified technique. *Expert Syst. Appl.*, 52:75–90, 2016. doi: 10.1016/j.eswa.2016.01.016. URL https://doi.org/10. 1016/j.eswa.2016.01.016.

[46] Investopedia Staff. Stochastic oscillator, 2018. URL https://www.investopedia.com/terms/s/ stochasticoscillator.asp.

[47] Cory Mitchell, Sneha Shah, Sneha Shah, Sneha Shah, Joel Kranc, and Bob Ciura. Ultimate guide to the trix indicator, 2018. URL https://traderhq.com/trix-indicatorultimate-guide/.

[48] Investopedia Staff. Average directional index (adx), 2018. URL https://www.investopedia.com/terms/a/adx. asp.

[49] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, 8(1), 2015.

[50] Efstathios Kalyvas. Using neural networks and genetic algorithms to predict stock market returns. *University of Manchester Master of Science thesis*, 2001.

[51] Andres Ar´ evalo, Jaime Ni´ no, German Hern˜ andez, and´ Javier Sandoval. High-frequency trading strategy based on deep neural networks. In *International conference on intelligent computing*, pages 424–436. Springer, 2016.

[52] Mahdi Pakdaman Naeini, Hamidreza Taremian, and Homa Baradaran Hashemi. Stock market value prediction using neural networks. In *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, pages 132–136. IEEE, 2010.

Google play store app ranking prediction using machine learning algorithm

Muhammad Suleman
College of Computer Science and Information Systems- CCSIS
Institute of Business Management
Karachi, Pakistan
std_19230@iobm.edu.pk

Ahsan Malik
College of Computer Science and Information Systems- CCSIS
Institute of Business Management
Karachi, Pakistan
std_19231@iobm.edu.pk

Syed Sajjad Hussain
Faculty of Engineering Sciences and Technology
Hamdard University
Karachi, Pakistan
sshussainr@gmail.com

Abstract—Smartphones has a key role in our routine life. These devices are available in multiple brands with their built-in operating systems mainly known as Android and iOS. Although both platforms provide similar functionality mechanism and output response even an app has same logo and design interface for all stores but the difference in file system emerges a need to establish their own identical app store where they have unique representation of apps searching criterion privacy policy review rating system and app algorithms that helps an app to display in top list with respect to region culture and trends that vary with respect to location and time. This paper study consists of different machine learning algorithms used to predict an app rating on Google play store utilizing real-time dataset of more than 10,000 play store apps.

These results are obtained by collection cleansing training and testing data to evaluate each regression model furthermore alter data to get desired results. Finally after implementation it concluded that linear regression fine tree algorithm provides best app rating prediction results.

Keywords—Smart devices, mobile, app rating, prediction, regression algorithm, machine learning

I. INTRODUCTION

The number of available apps in the Google play store was most recently placed at 2.6 million in December 2018 and these numbers are still growing dramatically [1]. By the time, to fulfil the user needs researcher has to bring new ideas that can solve daily routine problem via smart devices according to competitive market where apps counter growing day by day. As a result, it has been observed that to facilitate a particular problem we may have multiple apps providing same solution with difference in feature or functionality. Here it becomes a problem for end user to select a desired one from the bunch of apps. To minimize this issue app store manage a recommender system that recommend most popular app to the user [2]. This recommendation system works upon app ranking criteria by considering some parameters that includes app category, number of installs, rating, reviews, version compatibility and app Annie analytics.

Study has proven that features like rating, bug report comment or review increases possibilities to engaged user community with developers. It is very useful information for developers to improve their product in meaningful manners [3]. Usually users feedback comprises on app functionality, feature improvement, compatibility issues or app crashes. These reviews have been evaluated in different ways, including general exploratory studies, classification [4], feature extraction [5], review filtering [6], and summarization [7]. However, it's a little information for users to identify top quality app on the basis of rating and reviews only [8]. App ranking is not only dependent on user interface, reviews, downloads and rating but many other parameters can play a significant impact on results. It remains unclear that how long particular attributes can affects on app ranking because app stores modify their 'discovery algorithm' on regular basis [9].

To keep in mind all above factors we apply all regression analysis techniques on dataset containing app category, number of reviews, downloads, size, type, android version and content rating as input fields to predict app ranking as a response field. The goal of this study to evaluate all machine learning algorithm that trained a model and helps to find out app ranking on Google play store.

II. LITERETURE REVIEW

As far as we know, there has been very little work on app ranking system utilizing benchmark datasets. Mostly app related dataset work focused on the app security, version control, performance and user feedbacks. There are several research work related to review and rating. Here we highlighted few of them. For example, D. E. Krutz [10] given a dataset that reports results obtained by a few analytical tools on 4,416 unique versions of 1,179 open-source android applications. Also some analytical research in the domain of app review has been published, For instance, a dataset for mobile app retrieval includes 1,385,607 user feedbacks consists of 43,041 apps that have been used to enhance accuracy mobile app retrieval [11]. Similarly, M. Frank works on Google play store dataset of 188,389 instances with objective to uncover pattern request pattern supporting Boolean matrix factorization [12]. The software marketplace analysis dataset consists of 1,132,373 reviews from 15,094 apps to detect spam or fake reviews [13].

Other relevant research work includes, Hu and Liu [14] provided an extracted sentiment analysis on customer review for a particular product. Implementation of word level regression on movie reviews to predict movie first week revenue [15] and to correlate food menu prices [16]. These text-mining techniques can't be effective to app store reviews, since it has Unicode supported language with very limited number of words as compared to web.

Apart fom above mention datasets analysis some researchers give opinion based sentiment results on user reviews according to expressed feeling (i.e. emoji, negative and positive or anger and excitement emoticons) in reviews [17]. As per our knowledge, up till now no previous work has been done on any comprehensive dataset to train a model for prediction of mobile app rating on the basis of machine learning algorithm.

III. Background Information

Machine Learning is a sub domain of artificial intelligence (AI) that provides machine an ability to learn from experience derived from data without being programmed. The goal is to allow machines to learn automatically [18]. It is divided into three basic categorize supervised, unsupervised and reinforcement learning. Supervised learning use to predict future event on the basis of previous data by training a model with both input and correct labelled output via classification or regression. In unsupervised learning machine try to find unknown pattern or structure in input data without having output response. It works on clustering mechanism that divides input data into their groups accordingly. These clustering techniques mainly used in K-Means, Gaussian methods and artificial neural network (ANN) algorithms for object recognition and market analysis.

It is widely being used in various applications such as: Text classification, speech recognition, computer vision, image recognition, pattern matching, face detection, vehicle self-driving and medical treatments etc.

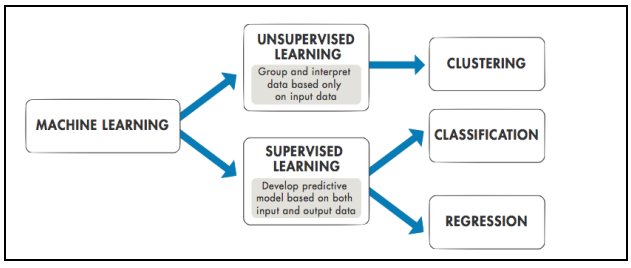


Figure 1: Machine learning technique classification

IV. Theory

Supervised machine learning further divided into two categorizes classification and regression. Classification used to predict future value when data is discrete while regression analysis used when response variable is real or continues value. Their algorithms includes decision tree, linear regression, logistic regression, support vector machines (SVM), naive Bayesian, k means clustering, k-nearest neighbour, ensemble methods, apriori algorithm, principal component analysis (PCA) and artificial neural networks (ANN). Every method has unique characteristic on which they trained a model where more than one algorithm may suitable for a particular machine.

A. *Decision Tree*

Decision tree is the simplest and useful algorithm, as it name describes it crates a tree of decisions. It is used in both classification and regression analysis. This mechanism consists of splitting dataset into tree structure from root to leaf node where it grows downward direction. Every leaf node condition depends upon root node it grows if leaf node condition becomes true. The objective this algorithm is to predict the future value by learning decision rules generated by given dataset. For classification algorithm built a decision tree by entropy and information gain. Entropy also known as Shannon entropy is denoted by H(S) for a finite set S. It measures homogenous data (uncertainty or randomness). Output value of entropy exists between 0 to 1 where 0 represents completely homogenous data and 1 represent identical data that cab divided equally.

B. *Linear Regression*

Linear regression used to find a statistical relationship between target and dependent variables. Where predictor is dependent and response in an independent variable and error is distance from plane. The objective is to plot a line that best fits the data through the points. It applies on non-deterministic relationship where one variable can't be accurately expressed into another variable. Like relationship between height and temperature.

C. *Logistic Regression*

Logistic regression is used to determine a binary output. The output has only two possible outcomes. True or False, Happy or Sad, Email or Spam and Positive or Negative. Logistic regression algorithm also uses a linear equation with independent variables to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. To convert the output into a range of 0 No and 1 Yes sigmoid function implies.

D. *Support Vector Machine*

Support vector machine is supervised algorithm that used for both classification and regression. It is based on decision planes that define decision boundaries. It differentiates objects on the basis of their class members by plot points in n-dimensional space. The goal is to classify all distinct data points on the basis of drawn hyper plane. To are many possibilities to separate two classes with hyper planes boundaries.

E. Naïve Bayesian

Naïve Bayesian classification technique is a simple probabilistic classifier based on Bayes theorem with strong independent assumption between the features. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

F. K-Mean Clustering

K-mean clustering is an unsupervised algorithm that makes partition of datasets into clusters to discover pattern in it. Cluster can be grouped on the basis of full or partially similar characteristics. In other words, k –mean algorithm defines k number of centre points to allocate every data instance to the nearest one. It is also known as an averaging of data.

G. K-Nearest Neighbour

K-nearest neighbor is a non-parametric, lazy method used in both classification and regression. However, it is mostly used in statistical estimation and pattern recognition. The objective is to predict the classification of new sample points from classes. Where different data points represent a class. Calculating the distance between the new sample and all existing sample dataset does this task. Once the k-nearest points are found, the most common class among these points will be the classification for the new sample.

H. Artificial Neural Network

Artificial neural network (ANN) are computing system inspired by human biological neural network. As human behaviour our brain learn from past experiences likewise in ANN each neuron receive input signal and in response of this signal it send another signal to the network as output result. ANN divided into layers pattern where first layer is input that received signals and the last layer is an output while hidden layers are function perform to get desired output response.

V. Methodology

To apply prediction model and analyse data these basic steps has been followed.

1. Data collection: Find a suitable dataset that fulfil the requirement to apply prediction models.
2. Pre-process data: To make data in correct format some filtration functions have been applied.
3. Explore data: Fix if there is any irrelevant value, sparsity, null, repetition and error.
4. Filter data: Remove extra columns that effects computation time and memory utilization.
5. Distribution data: Divide data into training and testing module.
6. Train data: Train the algorithm with training data until a correct model with minimum errors is obtained.
7. Data Evaluation: Compare the model with the testing data
8. Observe data: Analyse results.

All these techniques are implemented in MATLAB 2018, detail of their tests and mechanism described in result and analysis section.

A. Dataset

Dataset used in this research work is authentic android platform user data. It consists of more than ten thousand instances collected in 2018. As it says on dataset repository site, it is hypothetical data available for analyst, mobile developer and university researchers who perform algorithm to estimate mobile app performance. However, in this thesis the dataset will be used as historical data in order to predict the future app ranking that will be produced by a regression analysis. The data from the different parameters will be put together in this dataset, as all the parameters are essential for app analysis. To create data helpful for the machine learning model we divided it into two parts (i) testing and (ii) training data.

The total dataset consists on a matrix of 10839 rows and 8 columns, the first column is the app name, second one is number of reviews, third one is size of app in megabytes, fourth one is number of downloads, fifth one is content rating, sixth one is app categories, and the seventh one is android version and the last one (output) is rating. In addition, as the values range in each column is different, that's why data has been normalized in order to improve the efficiency on the different models.

B. Machenism

To implement machine-learning algorithms we must have two types of identical data testing and training. Some dataset repositories provide different dataset files for training and testing. Testing data is obtained from actual data that always be less than to the training data without response field. In this case it was a single data file. To extract training and testing data it has been randomly selected 8,125 instances for training and 2,728 for testing data also removed the output column from testing data that is rating in this case.

After normalization, to train the dataset all machine-learning regression technique has been applied on training dataset one by one. The algorithms will be first trained with the training data, it is provided a series of input and the known output and the model will work with this data to find a relationship between the predictors and the results. Once the relation has discovered verify it. If it is inaccurate, model can be retrained to achieve the desired result. After a desired relation all algorithms are tested with the testing data. In this step, only input data provided to the algorithm to test if the model could correctly make predictions. Once the predictions are done, evaluate prediction function of each model with test data and then compare the results with real output data in order to see training accuracy.

The results of every trained model will be discussed in details with RMSE, R-Squared, MSE, MAE, prediction speed, training time and training accuracy output as compared to real data.

C. Development Tools

This research work has been done on windows-based operating system with explicitly support of MATLAB 2018

software and Microsoft Excel for data manipulation. MATLAB design to implement all machine learning algorithm, MATLAB provides a series of tools and functions apps to analyse data. It supports both supervised and unsupervised machine learning methods that includes classification learner, regression learner and artificial neural network that used for prediction models [19].

Also, it helps to create, train, visualize, and simulate both shallow and deep neural networks, clustering, dimensionally reduction, time-series forecasting, and dynamic system modelling [20]. The app regression learner will be used in this thesis. This app allows you to predict value by training models. After training the model, evaluate its performance using mean squared error and regression analysis by comparing test data.

After training the models, the algorithm can be extracted in MATLAB script format, so it can be tested with new data by some coding. While Microsoft Excel used to filter, cleanse, align and format correction. It can also help for dividing data into training and testing parts.

VI. Result and Analysis

Given dataset have these attributes.

TABLE I. ATTRIBUTE TABLE

Attributes	Description
Review	Total number of review
Size	Space allocation
Install	Number of downloads
Type	Free or Paid
Content rating	Age distribution categories: General Mature and Adult
Android Version	Android operating system version number
Rating	Rating from 1.0 to 5.0

The evaluation results are following.

TABLE II.

Performance Parameter	Regression				Tree		
	Linear Regression	Interaction Linear Regression	Robust Linear Regression	Stepwise Linear Regression	Fine Tree	Medium Tree	Coarse Tree
RMSE	0.48	4.01	0.49	0.48	0.33	0.40	0.43
R-Squared	0.01	-65.96	-0.01	0.01	0.52	0.31	0.21
MSE	0.23	16.12	0.24	0.23	0.11	0.16	0.19
MAE	0.34	3.93	0.33	0.34	0.21	0.26	0.28
Prediction Speed (K-obs/sec)	160	100	480	280	520	440	880
Training Time	8.0051	6.54	8.97	8.56	10.28	9.17	8.87
Average Test Accuracy %	57.00	97.87	55.24	56.56	61.62	62.97	64.99

TABLE III.

Performance Parameter	SVM					
	Linear SVM	Quadratic SVM	Cubic SVM	Fine Gaussian SVM	Medium Gaussian SVM	Coarse Gaussian SVM
RMSE	0.49	1.81	289.18	0.47	0.47	0.48
R-Squared	-0.01	-12.71	-347221	0.08	0.06	0.03
MSE	0.24	3.30	83622	0.22	0.22	0.23
MAE	0.33	1.64	219.44	0.30	0.31	0.32
Prediction Speed (K-obs/sec)	11	11	360	5	5.2	4.8
Training Time	117.58	499.68	404.16	411.57	418.54	425.47

TABLE IV.

Performance Parameter	Ensemble		Gaussian Process Regression			
	Ensemble Boosted Tree	Ensemble Bagged Tree	Exponential GPR	Squared Exponential GPR	Matern 5/2 GPR	Rational Quadratic GPR
RMSE	0.48	0.39	0.46	0.47	0.47	0.47
R-Squared	0.03	0.34	0.08	0.07	0.05	0.06
MSE	0.23	0.15	0.22	0.22	0.22	0.22
MAE	0.35	0.25	0.31	0.32	0.32	0.32
Prediction Speed (K-obs/sec)	170	45	2.6	3.5	2.5	1.9
Training Time	428.44	428.73	561.77	497.37	556.79	634.63
Average Test Accuracy %	83.02	63.19	58.13	58.17	58.21	58.21

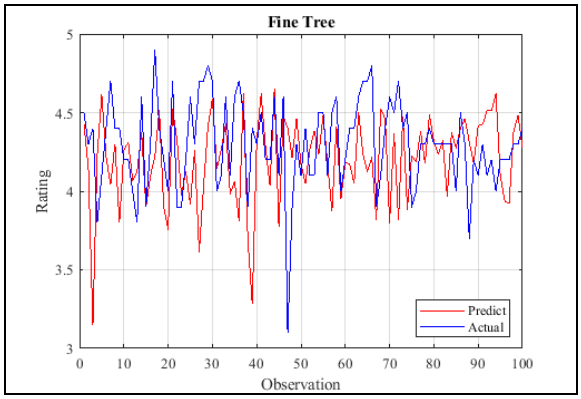


Figure 2: Fine tree value prediction graph

Conclusion

In this paper, Machine learning algorithms have been evaluates to predict app ranking on given dataset. The results show that, despite the wide variety of techniques and complex algorithms, which could be improved using a different dataset or adding app features in order to make more accurate predictions, one of the most simple techniques, Fine tree, have provided the best results in making predictions from play store historical data.

That concluded that, it could be possible and not a hard task to implement tree algorithm on dataset to predict app ranking, in order to forecast the suitable rating against an app, helping to improve app positioning, manage trends in app store meeting the demand of the app stores optimization and making rating systems more accurate.

REFRENCES

[1] Number of available applications in the Google Play Store from December 2009 to December 2018, https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/

[2] Hengshu Zhu, Hui Xiong, Yong Ge, Enhong Chen, "Mobile App Recommendations with Security and Privacy Awareness" Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014.

[3] D. Pagano and W. Maalej. User feedback in the appstore: An empirical study. In 21st IEEE International Requirements Engineering Conference,, pages 125–134. IEEE, 2013.

[4] W. Maalej and H. Nabil. Bug report, feature request, or simply praise? on automatically classifying app reviews. In Requirements Engineering Conference (RE), 2015 IEEE 23rd International, pages 116–125. IEEE, 2015.

[5] E. Guzman and W. Maalej. How do users like this feature? a fine-grained sentiment analysis of app reviews. In Proceedings of the 22nd RE Conference, 2014.

[6] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang. Ar-miner: Mining informative reviews for developers from mobile app marketplace. In Proceedings of the 36th International Conference on Software Engineering, ICSE 2014, pages 767–778, New York, NY, USA, 2014. ACM.

[7] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh. Why people hate your app: Making sense of user feedback in a mobile app store. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 1276–1284, New York, NY, USA, 2013. ACM.

[8] Kuehnhausen M, Frost VS. Trusting smartphone apps? To install or not to install, that is the question. Cognitive Methods in Situation Awareness and Decision Support; IEEE International Multi-Disciplinary Conference; 2013 Feb 25-28; San Diego, CA, USA. IEEE; 2013. pp. 30–37.

[9] Google play store algorithm https://www.kumulos.com/2018/07/06/google-play-store-algorithm/

[10] D. E. Krutz, M. Mirakhorli, S. A. Malachowsky, A. Ruiz, J. Peterson, A. Filipski, and J. Smith. A dataset of open-source android applications. In Proceedings of the 12th Working Conference on Mining Software Repositories, MSR '15, pages 522–525, Piscataway, NJ, USA, 2015. IEEE Press.

[11] D. H. Park, M. Liu, C. Zhai, and H. Wang. Leveraging user reviews to improve accuracy for mobile app retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pages 533–542, New York, NY, USA, 2015. ACM.

[12] M. Frank, B. Dong, A. P. Felt, and D. Song. Mining permission request patterns from android and facebook applications. ICDM, 0:870–875, 2012.

[13] J. Ye, S. Kumar, and L. Akoglu. Temporal opinion spam detection by multivariate indicative signals. In Tenth International AAAI Conference on Web and Social Media, 2016.

[14] M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD, pages 168–177, 2004.

[15] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: an experiment in text regression. In HLT, pages 293–296, 2010.

[16] V. Chahuneau, K. Gimpel, B. R. Routledge, L. Scherlis, and N. A. Smith. Word salad: Relating food prices and descriptions. In EMNLP-CoNLL, pages 1357–1367, 2012.

[17] Daniel Martens and Timo Johann On the Emotion of Users in App Reviews 7th March 2017

[18] What is Machine Learning? A definition https://www.expertsystem.com/machine-learning-definition/

[19] Mathworks, "Statistics and Machine Learning Toolbox, https://se.mathworks.com/products/statistics.html

[20] Mathworks, "Neural Network Toolbox", https://se.mathworks.com/products/neural-network.html

A New Segmentation-Scribble Generation Method for Image Colorization

Humaira Fatima
Department of Computer Science
FAST NU
Lahore, Pakistan
humaira.nu@gmail.com

Mehreen Tahir
Department of Computer Science
FAST NU
Lahore, Pakistan
mehreentahir18@gmail.com

Saad Abdullah
Department of Computer Science
FAST NU
Lahore, Pakistan
saad.1114441@gmail.com

Aamir Wali
Department of Computer Science
FAST NU
Lahore, Pakistan
aamir.wali@ne.edu.pk

Anwar Malik
Department of Computer Science
FAST NU
Lahore, Pakistan
anwaarmalik@outlook.com

***Abstract*—Besides image enhancement and restoration, another area of image quality enhancement that has been of interest to researchers is image colorization. Image colorization is the process of adding colors to a gray scale image. Furthermore, an image may be composed of multiple objects, with multiple components. Each of these must be identified before colorization. In this regard, we propose a novel segmentation-scribble generation (SSG) colorization model. This model not only segments image into components, but also determines and places the scribbles automatically on each component at the best possible location. Using this approach, the user would only have to select the color for each component instead of manually drawing the scribbles. Once the colored scribble is placed, it is propagated on the whole component efficiently. SSG is computationally efficient and it offer user more control to obtain optimal colorization results. Our proposed method showed excellent results on a variety of gray scale images, and performs better than the current state of the art colorization using optimization method.**

Keywords—image colorization, scribbles, color propagation, image segmentation, color transfer

I. INTRODUCTION

With the advancement of technology, trends are shifted from the monochrome images to colored images. Today, modern photography has compact digital cameras with high optical zoom and color detection lenses due to which artists can easily adjust the color and tone of images and videos. With such evolution in field of digital photography now every effect is merely a deal of few clicks. But this wasn't the case when the technology was emerging.

In 1980, we had the cameras which were merely able to click monochrome images. In order to add color to them artist would had to manually delineate the picture. But those were time consuming and expensive methods of colorization. That is why throughout the history many colorization techniques have been proposed to automate the entire process of colorizing massive black and white pictures and footages which were captured in older times. Due to commercial profits and development in this field the colorization techniques haves always been of great interest for both researcher and photo editing artists. It's not very difficult to understand why colorization process is not cheap and of great interest. For instance, consider this, we have found that to colorize one-minute frame of a movie costs three thousand dollars and colorizing a complete film cost on average around three hundred thousand dollars due to manual or physical colorization. Hence, colorization techniques carry a lot of incentive for its followers and researchers.

The goal of our project is to automate colorization by using well known techniques to colorize the grayscale images given example images or scribbles. User intervention would be required for choosing the desired colors in the image.

Fundamentally colorization is a technique of adding color to monochrome images and movie by assistance of computer and user. In order to understand colorization technique, we need to keep in view few major factors involved in colorizing an image that are described below in detail.

The one is color of each object. During colorization, color of each object in an images is important to take care of. As same object can carry different colors depending upon it material and mood when picture was captured. For instance, tree leaves in summer have different gradient of green and without any extensive change in shape turn into yellowish brown in autumn. In our project, in order to meet this requirement user would have to give the color marks himself. Automating this process would involve the machine learning which is out of our project's scope.

Second factor is boundary of each object. We need to keep track of boundaries of each objects in order to prevent color mixing and leakages with other objects in an image. For instance, to correctly colorize a grayscale image of a person's face we are required to identify boundaries between his eyes, face and hair. Today a lot of work has been done to achieve this task in colorization process. Some colorization techniques involve specifically segmenting the image into different identical region and then colorizing the image. Where as many technique does not practically segment the image before

colorization. Lastly propagation of color over the different objects of image is required to be done very carefully for good results.

So far a number of techniques have been developed which handles above mentioned factors very well which includes both user assisted (automatic) and computer assisted (manual) techniques. The research we have conducted so far concludes that, colorizing a grayscale image can be accomplished by various methods which can be broadly categorized as:

In this study we will be discussing our problem in the light of research we have done so far on the above mentioned categories of colorization. A complete literature review is provided below which will be giving overview of the research we have done so far and will be supporting our implemented colorization techniques. Results of each technique is highlighted along with their detailed analysis which involves pros and cons and physical limitation of those techniques. The major complexities in this field involve boundary detection, prevention of color mixing and leakages across different object and lastly proper color propagation. During our research we have also come across many techniques which although produces good results but are very time consuming and expensive due to above mention difficulties.

II. LITERATURE REVIEW

The process of colorizing a grey scale image can be largely categorized as example-based or computer assisted and stroke-based or user assisted.

A. Example based Methodology

In the example based approach, there are two input images, one is gray scale source image, and the other is the colored reference image. In it, the entire color "mood" of the reference image is transferred onto the source. This can be done pixel-by-pixel or by matching luminance and texture information between the images. These procedures work reasonably well only when corresponding color regions between the two images have the same color or luminance values. Otherwise the results produced by this method are not natural. In this situation, swatches based approach is used that allows user to match areas of the two images with rectangular swatches.

Various algorithms have been proposed for color transfer. Reference [3] uses the orthogonal color space developed by Runderman called as lαβ that is used to reduce the correlation between the three color channels of RGB. Once the image is converted, the statistical computation and color corrections are applied by scaling and shifting the spaces of both images. In this way, the appearance and feel of source is transferred to reference image. The results of this algorithm are not up to the standard especially when two images, source and reference, of varying compositions are used.

In another approach, both the source, and the reference images are converted into HSV components and then intensity comparison is taken over both images. The V component of each pixel of source image is compared to each pixel of target image. If a match occurs, then the H and S component of reference image are transferred to the source image [4]. This method is also known as global image matching. One of the limitations of this method is that it is very high running time.

B. Stroke based Methodology

In scribble-based methods the user drives the colorization by defining colored strokes onto the gray scale image. Unlike color transfer it doesn't require two images for colorizing, only one input image is required on which the scribbles are assigned and later the colors of the scribbles are propagated [5-9].

Since an image can constitute of more than one object and each of them may have distinct colors, so image need to be segmented. After the image objects are identified it is highly likely that user is promoted to add color patches called as scribble to the image based on those input color patches(scribbles) their color is propagated onto the entire object and progressively to the entire image based on intensity points.

One popular technique is colorization using optimization [5]. This technique did not require any image segmentation algorithm. The algorithm colorizes gray scale images by simply using the color information which is annotated by an artist or user as scribbles over the image, and then annotated colors are automatically propagated across the image. To do this, the color is converted to YUV space, and then the UV channels are assigned to the rest of the area. This is done in a progressive manner. User adds the scribbles, and they are propagated. If any region doesn't have the desired goal or there is leakage of color of one object onto the next, more scribbles are added on this output image. Hence in this algorithm the image is repeatedly scribbled and colorized.

Another method similarly propagates the colors of the scribbles, but it uses an array where the boundary pixels of the scribbled region are saved [6].
The algorithms follow the approach that color is to be spread from the scribbles outwards towards the other image ends. A boundary pixel of the scribble is chosen and the distance of its eight adjacent pixels in the window is calculated by absolute value of intensity difference between the two adjacent pixels (current pixel(s) and neighboring pixel(t)). The pixel with minimum distance will be taken further, and its window of eight is taken into account.

Reference [7] proposes a method that unlike others, doesn't propagate the color only on the basis of intensity difference but rather it also takes into account the pattern continuous regions. This proposed method works efficiently for black and white manga's that exhibit not only intensity continuity but also pattern continuity. The places of similar pattern or intensity in the image are intelligently detected using the Level Set method. This methods obvious drawback is that it works well for manga images, but does not suit natural images like sceneries, human body with more details. Reference [8] the scribbles are diffused across spatiotemporally smooth regions. The leakage is prevented by spatiotemporal discontinuities Other methods of propagation involve using gray-scale

image information such as edge and gradient [9], and propagation using isolines on geographical or distance map [10].

Since, scribble based approach requires image segmentation, in the next subsection we explore some image segmentation algorithms.

C. Image Segmentation

Normally, real world images consist of some objects and a background. Separating different objects, and foreground from background is known as segmentation of objects in a gray scale image.

There are numerous methods of segmenting a gray scale image i.e. by using gradient, by histogram, using k-means clustering, fuzzy c-mean, Watershed and texture filter, all proposed by [11]. Segmentation by Histogram is a technique which converts the images in histogram dataset i.e. finding the frequency of gray scale image and then applying threshold on the dataset. The algorithm produces accurate results for simple images (i.e. high contrast images) but fails on complex images (i.e. low or blur images). Fuzzy C-Mean performs better than all others in terms of time and result. Fuzzy c-mean is an AI based algorithm that makes clusters of image pixels by their intensity.

A novel graph based segmentation approach was proposed by [12]. The basic approach used is to measure the evidence of boundary between two regions by comparing intensities across boundaries and by comparison of the intensities of neighboring pixels within each region. Zahn's method, uraqulant algorithm, splitting and merging methodology are the basic roots on which this algorithm is proposed upon. The "pair wise region comparison predicate" is used by the algorithm proposed in this paper for segmenting an image.

III. PROPOSED SOLUTION

In light of the literature review given in the previous section, we propose a segmentation-scribble generation (SSG) colorization model. In this model, we will be using Gradient based image segmentation algorithm slightly tailored for our problem, our automatic scribble generation and placement algorithm, and our more efficient colorization using optimization algorithm.

In it we will be following a sequential approach of first segmenting an image into segments by preserving the perceptual information of each object. These segmented images are them feed to the scribbling module that determines the various locations within each segmented component where the scribble must be placed. The location of scribbles is the function of the centroid of the component, and the luminance value. If the luminance value > 225, then another centroid value is calculated between this pixel and the component boundary. The scribble is normally the length of 10 pixels.

Determining the position of each scribble automatically will ease task of user, as the user would only have to select the appropriate color for the generated scribbles instead of manually drawing the scribbles. In the third step, the annotated colors will be propagated across the image to colorize the segment. Propagation of scribbles is done using the method similar to the one proposed in [5]. Our color propagation is simply a flood fill method. In it, first we have changed the image into YUV color space and then U, and V component of the scribbled area is fetched and assigned to non-scribble area. At the end all the colorized segments are attached and whole image is created.

IV. RESULTS AND ANALYSIS

In this chapter we will be providing the experimental results. Since there is no standard data set for image colorization, we conducted the testing on some typical images used in various papers in image colorization such as the fruits image, and the pepper image. In total, the experimentation involved a small set of images and the colorized images were observed manually. Each output was tested for two things: the quality of color and leakage. Our proposed algorithm worked well for almost all images. Some results are shown and discussed next.

A. Comparison of original and our modified Colorization using Optimization algorithm

Consider fig. 1 that has a gray scale image before and after applying scribbles. The result of original colorization using optimization is given in fig. 2. Even after drawing 50 scribbles, the image is not properly colorized. Carefully observe the corners of eyes and mouth in fig. 2. Color leakage is very evident. In order to prevent color leakage and color mixing, user would be required to add more scribble as per the iterative working of the algorithm. This progressive colorization is tiresome and annoying for professional users.

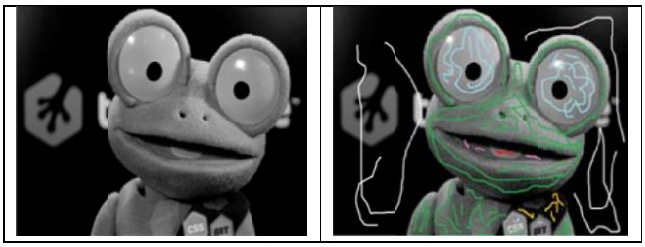


Fig. 1. Gary scale image before (left) , and after (right) applying the scribbles.

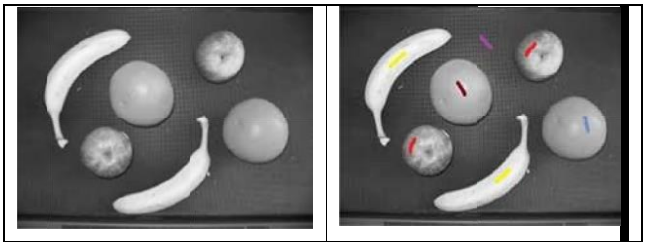
Fig. 2. Colorized image using the original colorization using optimization method.

Another major drawback of this technique is color propagation over pattern continuous region and high intensity images. Images which includes pattern, hatching

and screening effects are difficult to colorize properly because the intensity difference become very high and the weight assigned to the neighboring pixels become very small as a result no color is propagated.

As we know that this algorithm colorizes different pixels of image by looking at the intensity difference of neighboring pixels. Consider images in which the grayscale level is very intense and change very suddenly even with in objects. In such scenario where the intensity is changing diversely the color of scribbles are also not propagated properly across that region and in some situations they are not propagated at all because the intensity difference is observed so high, because of which that region is not consider part of that scribble.

In order to overcome this problem, we tried tweaking



around the underlined approach of algorithm and we get to learn that if we increase scribbled area a little farther then original places of scribbles we can actually overcome this problem. The improved result using this approach is shown in fig. 3 and fig. 4. Fig. 3 shows the gray scale image before and after applying scribbles. The results of the original (left) and our modified (right) version of the colorization using optimization algorithm are given in fig. 4.

Fig. 3. Gray scale image of fruits, before (left), and after (right) applying the scribbles.

In fig. 4, in case of the original algorithm, few of yellow, red and purple scribbles are not propagated at all and other are propagated but very less amount of original color is actually assigned to neighboring pixels. Our proposed tweaking actually works because high-intensity difference areas are now explicitly considered. But once the scribbled area is increased for high intensity images it would become difficult to proper colorize the light intensity images.

Fig. 4. Results of colorization using optimization (left) and of our proposed tweaking (right).

B. Image Segmentation

The image segmentation results by gradient based segmentation algorithms are presented in fig. 5 and fig. 6. The boundaries are identified based upon the intensity difference. This image segmentation works only on certain type of images. In our approach, we have catered most simple images which do not contain hatches, repeated intensity patterns and screening effects.

Fig. 5. Results of image segmentation method

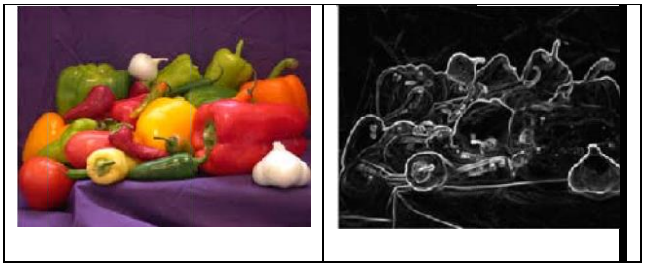


Fig. 6. Results of image segmentation method

C. Colorization using SSG

Finally, we present the results of our SSG model after plugging-in the modified color propagation, and the image segmentation algorithm given in previous sections. Consider

the gray scale image in fig. 7.

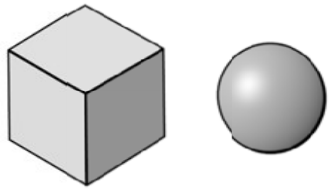


Fig. 7. Gr ay scale image of 3 D shapes

After segmentation, the different components of the image are shown in fig. 8.

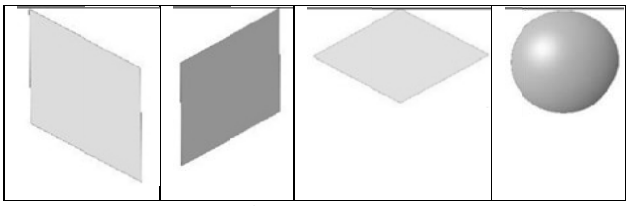


Fig. 8. Various components of the input image.

The scribble generator places the selected color scribbles within each component automatically as shown in fig. 9.

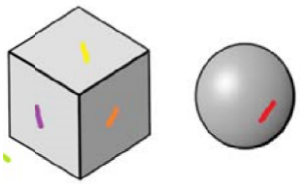


Fig. 9. Scribbles placed automatically by the scibble generation module.

Propagation of color on each segment is done and after that they are combined to form one complete image – which is shown in fig. 10.

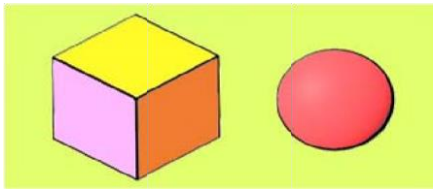


Fig. 10. Resulting colorized image.

The experimental results and analysis implemented algorithm are illustrated above. After all the research that we have conducted on colorization algorithm, we can conclude that each algorithm has it wonders as well as drawbacks. In order to improve results we have observed that underlying approaches used for implementation of colorization are required to be modified. Time consumption is one the greatest issue in colorization algorithms. But we can easily change lesson the running time of algorithms by using such platform and languages which execute the given task very efficiently, little tweaks while implementation also do the job.

V. CONCLUSION

Trends have shifting with the advent of technology. Few years back cameras were merely able to click monochrome images. In order to add color to them artist used to manually delineate the picture, it was not only time consuming but rather expensive too. Many colorization techniques have been proposed to automate the entire process of colorizing massive black and white pictures and footages which were captured in older times.

In stroke based methodologies we came across many algorithms that can serve the purpose few were dependent on the core image segmentation few were handling the boundaries by merely calculation of intensities. We implemented and catered various methods in this domain also. Based upon the literature review we worked on the underlying methodologies of Optimization, Chrominance blending, Image Segmentation and automated scribble generation. An image may have multiple objects inside it, each of them should be identified as separate distinct object by preserving the perceptual information. So our first major task was to search for such an algorithm / methodology that can help us in detection of distinct objects. After that our next target was to make the location of scribble automated by our algorithm that can facilitate the user as he would have to now select only the color instead of manually scribbling the image. Meanwhile, we are also researching how to spread an annotated color on the entire image. We started from exploring the algorithmic details of each of them followed by their implementation.

We came up with idea that why do we take the best of all the algorithms and come with a hybrid approach that is neither entirely user-assisted nor entirely computer-assisted.

In it, we will be following a sequential approach of first segmenting an image into segments by preserving the perceptual information of each object by using image Segmentation algorithm. Followed by automated scribble generation and propagating the annotated colors across the image.

We will be empowering the user to select his images, scribble it from the color panel and colorize it by any algorithm of his choice. Or he can leave it to application to let it choose the algorithm that will facilitate him for scribble positions and now the user task would be mere color selection else will be handled by our application. He would be able to see his previous images and can improve them at any time he wishes too because of progressive colorization feature of the application.

We did a rigorous testing of the SSG model, and it produced excellent results as compared to colorization using optimization method.

REFERENCES

[1] A.A. Shah, G. Mikita and K.M. Shah, “Medical image colorization using optimization technique”, International Journal of Scientific and Research Publications, vol. 3, 2013.

[2] M. Yang, “Still image colorization,” ECE Department journals of Northwestern University, March, 2005.

[3] E. Reinhard, M. Adhikhmin, B. Gooch and P. Shirley, “Color transfer between images,” IEEE Computer graphics and applications, vol. *21*, pp.34-41, September 2001.

[4] T. Welsh, M. Ashikhmin and K. Mueller, “Transferring color to greyscale images,” in ACM Transactions on Graphics, vol. 21, pp. 277-280, ACM, July 2002.

[5] A. Levin, D. Lischinski and Y. Weiss, “Colorization using optimization,” ACM transactions on graphics, Vol. 23, pp. 689694, ACM, August 2004.

[6] L.Yatziv and G. Sapiro, “Fast image and video colorization using chrominance blending,” IEEE Transactions on Image Processing, vol. 15, 2004.

[7] Y. Qu, T.T. Wong and P.A. Heng, “July. Manga colorization,” In ACM Transactions on Graphics, vol. 25, pp. 1214-1220, ACM, July 2006.

[8] S. Paul, S. Bhattacharya, and S. Gupta, “Spatiotemporal Colorization of Video Using 3D Steerable Pyramids,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, pp.1605-1619, August 2017.

[9] Y. Li, M. Lizhuang and W. Di, “Fast colorization using edge and gradient constrains,” Proceedings of WSCG'07, pp. 309315, 2007.

[10] A. Popowicz and B. Smolka, "Isoline Based Image Colorization," in UKSim, pp. 280-285 March 2014.

[11] Y. Yang and S. Huang, "Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term," Computing and Informatics, vol. 26, pp.17-31, January 2012.

[12] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graphbased image segmentation," International journal of computer vision, vol. 59, pp.167-181, September 2004.

Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms

Syed Adnan Ali Zaidi
Department of Computer Science
Muhammad Ali Jinnah University
Karachi, Pakistan
FA17PHCS0003@maju.edu.pk

Syed Muhammad Hassan
Department of Computer Science
Muhammad Ali Jinnah University
Karachi, Pakistan
FA17PHCS0004@maju.edu.pk

***Abstract* — Classifying text in Urdu is very challenging task, especially when we have multiple classes to implement in multiple machine learning algorithms. In this paper, we are actively scraping Urdu news heading from news agencies including BBC Urdu and Urdupoint. Our corpus contains 141289 new words of eight categories with Armed forces, International, Entertainment, Education, Accident, Local, Sports, and Weather. The machine learning algorithms were not able to work directly on the data, so we applied the preprocessing techniques like stop words removal and a feature vector. Our model classify text into multiple category and after comparing different machine learning algorithms, Ridge Classifier is the best predictor and we achieved up to 87% accuracy.**

Keywords — Urdu; Text Classification; Tf-Idf; Linear SVC; Ridge Classifier; Random Forest; Multinomial NB; SGD;

I. INTRODUCTION

Nowadays data on Internet is available in all type of spoken languages which is easy to understand, and people can grab data accordingly. Urdu language is considered as national language in Pakistan and most spoken language when we focus on Indo-Pak regions. More than 100 million people speak Urdu widely throughout world. There is a lot of work still required in the Urdu language specially in the area of classifying text into different categories which is one of the most common and useful techniques used to solve problem like spam filtering which predicts data as spam or not. Another area in which this technique is mostly used is sentiment analysis where we can identify class especially negative and positive using dataset. Text classification is helpful in article tagging where we want to assign category tag to the articles.

We propose a model on Urdu News text classification which gives best result on our own created dataset by applying different machine learning algorithms and identify which algorithm is best to predict our dataset. Here we use eight pre-defined classes which are Army, International, Entertainment, Education, Accident, Local, Sports, and Weather news as an input into ten different machine learning techniques. Our method contains five primary processes: stop words removal, stemming, feature vector, applying the machine learning algorithms and assign the class to the sentence.

This paper is divided into multiple sections, Section II contains literature review which describe few of pervious work has been done on Text Classification. Section III describes methodology that has complete process starting from how to collect corpus using web crawler, then applying stop words and Tf-Idf feature vector to filter our data. Section IV describes results that predicts our class on any given sentence and last section V contains conclusion. After preprocessing we applied different machine learning algorithms like Linear SVC, Multinomial Naïve Bayes, Logistic Regression, Ridge Classifier, Passive Aggressive Classifier, Perceptron, K-Neighbors Classifier, SGD Classifier and Random Forest Classifier to train our corpus. Section IV explains results and the last section concludes the summary of work.

II. RELATED WORKS

M. Ikonomakis et al. [1] is a text classification technique and compare different machine learning algorithms for training model and describe pre-processing step of data.

Muhammad Bilal et al. [2] use three classification models are used for text classification using Waikato Environment for Knowledge Analysis (WEKA). Opinions written in Roman-Urdu and English are extracted from a blog. These extracted opinions are documented in text files to prepare a training dataset containing 150 positive and 150 negative opinions, as labeled examples. Testing data set is supplied to three different models and the results in each case are analyzed. The results show that Naive Bayesian outperformed Decision Tree and KNN in terms of more accuracy, precision, recall and F-measure.

Mehreen Alam et al. [3] address this problem and transform Roman-Urdu to Urdu transliteration into sequence to sequence learning problem. Roman-Urdu to Urdu corpora was created and passed it to neural machine translation model that predicted sentences up to length 10 while achieving BLEU score of 48.6 on the test set.

Neelam Mukhtar et al. [4] resource focus poor languages such as Urdu are mostly ignored by the research community. After acquiring data from various blogs of about 14 different genres, the data is being annotated with the help of human annotators. Three main well-known machine learning algorithm Support Vector Machine, Decision tree and k-Nearest Neighbor (k-NN) are tested for comparison which concluded that k-NN is performing better than Support Vector Machine and Decision tree in terms of accuracy, precision, recall and f-measure.

Muhammad Usman et al. [5] use five well-known classification techniques on Urdu language corpus and assigned a class to the documents using majority voting. The corpus contains 21769 news documents of seven categories (Business, Entertainment, Culture, Health, Sports, and Weird). After preprocessing 93400 features are extracted from the data to apply machine learning algorithms up to 94% precision and recall using majority voting.

III. METHODOLOGY

Our methodology contains a step-wise procedure. We started from the Urdu language corpus collection and then used some preprocessing techniques for features selection to apply actual classification algorithms. The flow chart in Fig-1 summarizes the process which we followed for our technique.

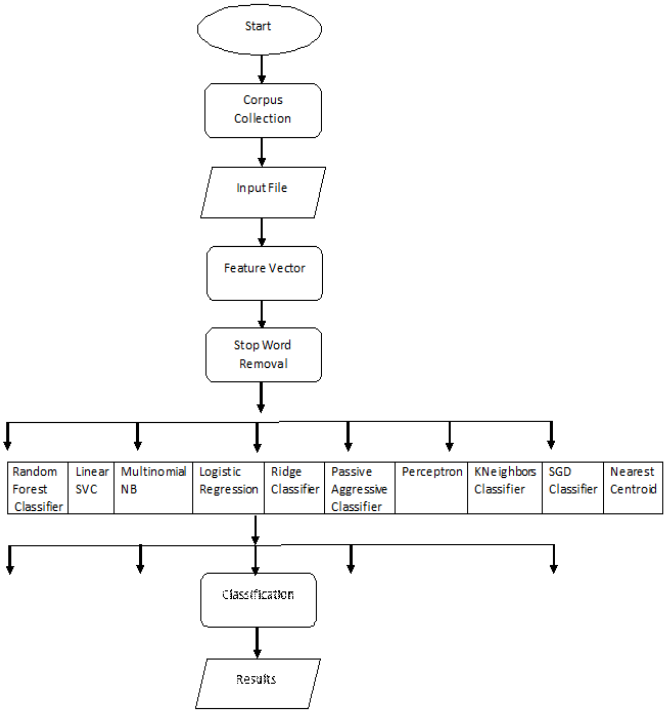


Fig-1 summarizes the process

A. Corpus Collection

Extensive training data plays a vital role in the development of a model that uses supervised learning algorithm. For this purpose, we write multiple crawlers to collect data from different news websites, e.g., bbcurdu.com, and urdupoint.com. In total, we collected 141289 words. Data is collected category-wise in the text files, and categories are as follow: Army, International, Entertainment, Education, Accident, Local, Sports, and Weather.

```
[('7 ,'افواج),
 ('3 ,'بین الاقوامی),
 ('5 ,'تعلیم),
 ('2 ,'تفریح),
 ('6 ,'حادثات),
 ('0 ,'قومی),
 ('4 ,'موسم),
 ('1 ,'کھیل)]
```

B. Stop Words Removal

The words which are either not useful for the proposed classification models or used as prepositions are included in the stop words list. In our case, we maintained a list of stop words (total 265 major) to omit from our text to extract meaningful data for the classifiers.

sw= ["آ", "آئی", "آئیں", "آئے", "آتا", "آتی",.....]

C. Feature Selection

Feature selection is an important part of building machine learning models. We use the chi square test of independence to identify the important features.

	Most correlated uni-grams:	Most correlated bi grams:	Top uni-grams:	Top bi-grams
# 'افواج':	. ایس	. جی ایس	. پاک	. جی پنجاب
	. پاک	. پاک فوج	. جی	. پاک فوج
	. آرمی	. ایس پی	. پولیس	. کراچی پولیس
	. جی	. آرمی چیف	. آرمی	. جی سندھ
	. آر	. پی آر	. ایس	. آرمی چیف
# 'بین الاقوامی':	. ایران	. ترک حکمران	. سعودی	. بھارتی فوج
	. امریکی	. دنیا طاقتورقوم	. اسرائیلی	. افراد ہلاک
	. عرب	. قسط نمبر	. امریکی	. بیت المقدس
	. ہلاک	. افراد ہلاک	. اسرائیل	. ولی عہد
	. سعودی	. سعودی عرب	. ہلاک	. اقوام متحدہ
# 'تعلیم':	. طلبہ	. پنجاب یونیورسٹی	. تعلیم	. جامعہ کراچی
	. تعلیم	. پرچہ آؤٹ	. اسکول	. تعلیمی اداروں
	. امتحانات	. تعلیم سندھ	. تعلیمی	. نتائج اعلان
	. نقل	. نقل مافیا	. یونیورسٹی	. پنجاب یونیورسٹی
	. اسکول	. محکمہ تعلیم	. اسکولوں	. کراچی میٹرک
# 'تفریح':	. اداکار	. علی خان	. اداکارہ	. سلمان خان
	. فلم	. ادبی شخصیات	. اداکار	. علی خان
	. دیہکا	. شخصیات سکینڈلز	. معروف	. امیتابھ بچن
	. شادی	. فلمی ادبی	. گلوکار	. رخ خان
	. اداکارہ	. سکینڈلز قسط	. فلم	. دیپیکا پڈوکون
# 'حادثات':	. گر	. روڈ ٹریفک	. حادثہ	. چھت گر
	. روڈ	. چھت گر	. لٹ	. افراد زخمی
	. لٹ	. الم ناک	. گر	. الم ناک
	. ٹریفک	. مسافر بس	. ٹریفک	. کھائی گری
	. حادثہ	. ٹریفک حادثہ	. ٹکر	. ٹریفک حادثات
# 'قومی':	. کورٹ	. وزیر اعظم	. پی ٹی	. وزیراعظم
	. جسٹس	. سپریم کورٹ	. نوازشریف	. وزیراعلی پنجاب
	. نیب	. چیف جسٹس	. بہبازشریف	. عمران خان
	. عمران	. وزیراعظم عمران	. نیب	. چیف جسٹس
	. وزیراعظم	. عمران خان	. عمران	. چینی قونصلیٹ
# 'موسم':	. سورج	. کراچی موسم	. بارش	. گرم ترین
	. کراچی	. کراچی گرمی	. موسم	. درجہ حرارت
	. گرمی	. علاقوں بارش	. گرمی	. مون سون
	. موسم	. دھار بارش	. سورج	. کالی گھٹائیں
	. بارش	. موصلا دھار	. سردی	. بیٹ اسٹروک
# 'کھیل':	. کرکٹ	. قومی ٹیم	. کرکٹ	. شعیب ملک
	. لینڈ	. سرفراز احمد	. کھلاڑی	. پی پی
	. میچ	. شعیب ملک	. ٹیم	. شاہد آفریدی
	. نیوزی	. ایس ایل	. صدھو	. نجم سیٹھی
	. ٹیم	. نیوزی لینڈ	. کپتان	. لاہور قلندرز

D. Term Frequency- Inverse Document Frequency.

Specifically, for each term in our dataset, we calculate a measure called Term Frequency, Inverse Document Frequency, abbreviated to tf-idf.

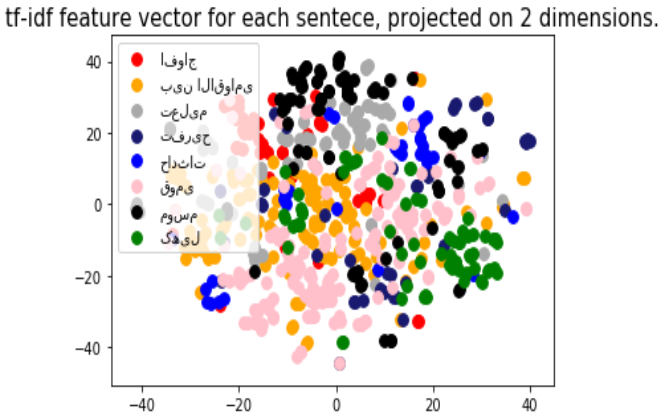


Fig-2 TF-IDF Feature Vector

IV. RESULTS

After all preprocessing techniques, we fed our dataset to machine learning algorithm. We divided our dataset into training 80% and testing 20%.

Random Forest Classifier, Linear SVC, Multinomial NB, Logistic Regression, Ridge Classifier, Perceptron, K-Neighbors Classifier, SGD Classifier, Nearest Centroid are used to train our data. For all above machine learning models, the comparisons have been given by Model Name, Algorithm accuracy, mean predicted values and confusion matrix.

Fig-3 Basic Model Comparison

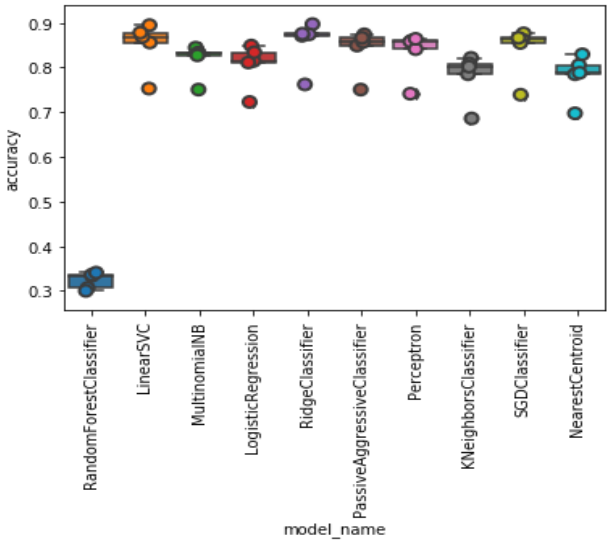


Fig-4 Algorithm Accuracy

Classifier	Accuracy	Result
RandomForestClassifier	32.367651	Failure
GaussianNB	75.303729	Average
KNeighborsClassifier	78.058373	Average
NearestCentroid	78.107202	Average
LogisticRegression	80.632812	Good
MultinomialNB	81.738182	Good
PassiveAggressiveClassifier	83.348770	V Good
Perceptron	83.298882	V Good
SGDClassifier	83.803790	V Good
LinearSVC	85.000054	Excellent
RidgeClassifier	85.505178	Excellent

Fig-5 Mean Predicted Values

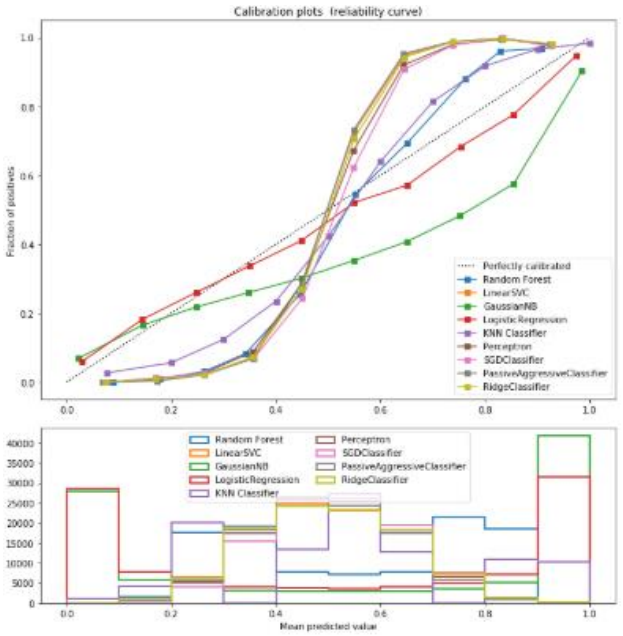
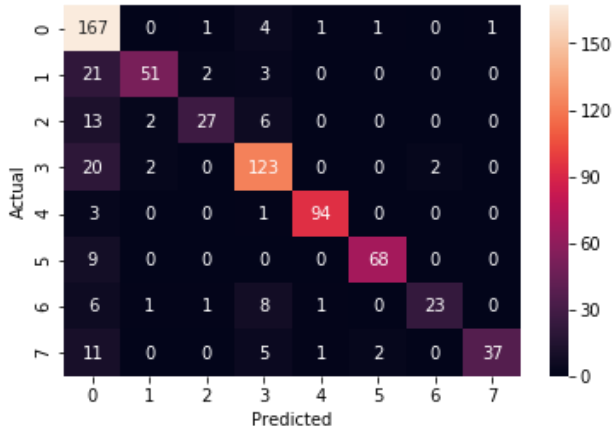


Fig-6 Confusion Matrix



The following are some test samples and corresponding predicton by our model.

"تحریک انصاف کا وفد چین روانہ، کس کس کو ساتھ لے گئے ؟ میرٹ کی دھجیاں اڑ گئیں "
- Predicted as: 'قومی'

"عید میلاد النبی ﷺ پر امیتابھ بچن کا مسلمانوں کیلئے پیغام"
- Predicted as: 'قومی'

"قلمی و ادبی شخصیات کے سکیٹلاز . قسط نمبر558"
- Predicted as: 'تفریح'

"ریاست مدینہ کی طرز پر کام شروع ، کوئی سردار ہویا نواب درگزر نہیں کیا جائے گا"
- Predicted as: 'بین الاقوامی'

" جس وقت چینی قونصل خانے پر حملہ ہوا اس وقت کتنے چینی اندر موجود تھے اور وہ اب کہاں"
- Predicted as: 'قومی'

"مدارس اور تعلیمی اداروں کی اصلاحات کی بات کی تھی، آصف غفور "
- Predicted as: 'تعلیم'

"کراچی : آج سے موسلا دھار بارشوں کا امکان"
- Predicted as: 'موسم'

"کراچی: لانڈھی نمبر دو میں تیز رفتار ٹرالر اُلٹ گیا"
- Predicted as: 'حادثات'

"بھارتی فوج کی لائن آف کنٹرول پر بلا اشتعال فائرنگ ، آئی ایس پی آر"
- Predicted as: 'افواج'

"محمد حفیظ نے پی ایس ایل کی فرنچائز پشاور زلمی چھوڑ دی لیکن کیوں ؟ حیران کن وجہ"
- Predicted as: 'کھیل'

V. CONCLUSION

In this paper we created our own dataset on Urdu news heading by capturing data from different sites. After collecting our dataset, we passed it on preprocessing techniques to filter our data.

Then finally we applied different machine learning algorithm to train our Urdu dataset. We found Ridge Classifier is best algorithm for text classification that gives almost 86% accuracy to predict our class.

REFERENCES

[1] M. Ikonomakis, S. Kotsiantis, V. Tampakas. "Text Classification Using Machine Learning Techniques". WSEAS Transactions on Computers, Issue 8, Volume 4, August 2005, pp. 966-974

[2] Muhammad Bilal et al. "Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University – Computer and Information Sciences (2016) 28, 330–344.

[3] Mehreen Alam et al. "Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration." 20th International Multitopic Conference (INMIC' 17).

[4] Neelam Mukhtar et al. "Urdu Sentiment Analysis Using Supervised Machine Learning Approach" International Journal of Pattern Recognition and Artificial Intelligence, Vol. 32, No. 2 (2018) 1851001 (15 pages).

[5] Muhammad Usman et al. "Urdu Text Classification using Majority Voting" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, 2016.

POSTER PAPERS

Diagnosis of Breast Cancer Using Deep Dense Neural Network

Sadia Mushtaq
Department of Computer Science
Muhammad Ali Jinnah University
Karachi, Pakistan
mushtaq_sadia@hotmail.com

Hira Farman
Department of Computer Science
Muhammad Ali Jinnah University
Karachi, Pakistan
hira.farman@jinnah.edu

***Abstract* — Breast Cancer is the commonest cancer among female and the foremost cause of death among women. It is a global issue now and requires public awareness and advancement in detection with modern technology to face the challenges of epidemic disease. Deep Dense Neural Network and optimization techniques are used in this study in order to diagnose the occurrence of breast cancer in patients.**

We worked on Breast Cancer Wisconsin (Diagnostic) Data Set features of Fine Needle Aspiration Cytology for breast mass to train our Neural Network model. It has 569 number of patients with Benign and Malignant diagnosis of breast cancer. In Neural Network, we split our data into training and testing. First, we trained our model by using training data, then we applied this model to predict on test data. Our model provides efficient results on testing data. For this purpose, we distributed our dataset into 75/25 i.e. 75% (426) records for the training and 25% (143) records for testing of our Neural Network.

Our model achieved state-of-the-art accuracy, 97.22%, after applying Activation Function as Rectified linear unit (ReLU) and regularization as dropouts on Deep Neural Network dense layers. As we had two categories of diagnosis so for this purpose, we utilized sigmoid Binary Activation Function on last Output layer to Classify our Breast Cancer data.

Keywords*— Deep Dense Neural Network, Breast Cancer Detection, Wisconsin Dataset, keras Api, Activation Function (ReLU and Sigmoid), Regularization strategy (Dropout).*

I. INTRODUCTION

Breast cancer is the most common cancer in women worldwide. Nearly two million new cases are diagnosed every year that means it is the second most common cancer overall. In Pakistan, the most frequently diagnosed cancer among females is also breast cancer, accounting for nearly one in nine female patients. Its incidence in Pakistan is 2.5 times higher than that in neighboring countries like Iran and India. The risk factors associated with breast cancer are gender, age, family history, early menarche, late menopause, post-menopausal hormonal therapy, late pregnancy, nulliparity, alcohol consumption, physical inactivity, obesity; however, breast feeding is a protective against breast cancer.

When detected in its early stages, there is 90-95% chance that the cancer can be treated effectively, but the late detection of advanced-stage tumors makes the treatment more difficult. Currently, the most used techniques to detect breast cancer in early stages are: mammography (63% to 97% sensitivity), FNAC (Fine Needle Aspiration cytology) with visual interpretation (65% to 98% sensitivity) and surgical biopsy (almost 100% sensitivity).

Therefore, mammography and FNAC with visual interpretation correctness diverges broadly, and the surgical biopsy, even though reliable, is invasive and costly.

Fine needle aspiration cytology (FNAC) of breast lump is an accepted and established method to determine the nature of the lump and it may play an important role when it is difficult to determine the nature of breast lump by clinical examination. It has been shown that FNAC can reduce the number of open breast biopsies. They are helpful in finding of common causes of breast lumps.

The aim of this study is to utilize the most amazing developing technology of Deep Neural Network on the findings of Fine Needle Aspiration cytology (FNAC) and to provide low cost computational technology to the health professionals, as it helps in early, efficient and timely diagnosis of breast cancer besides that it will also ease patients who are in great dilemma not only due to disease but also due to expensive assessments.

Deep Neural Networks is a subset of machine learning in artificial intelligence that got ability to learn unsupervised or unstructured data. Deep Neural Networks, which mimic human brain, emerges as a new world on the horizon of information technology in recent years, and they have demonstrated their ability to learn from image, audio and text data. They perform extremely well and amazing. They are Feed Forward Network with input layer and multiple hidden layers and output layer, so we utilize FNAC data to classify Breast cancer cases using Deep Neural Network techniques and designed a model that achieved high level of accuracy with a low rate of false negatives.

II. RELATED WORK

In the recent past years, several studies were published on the diagnosis of Breast Cancer by utilizing the technology of Neural Network and machine learning.

Street WN et al [1] was the first study on the data set of Wisconsin FNA to diagnose breast cancer published by the dataset authors. They utilized image processing techniques with linear programming classifiers as Multi surface method (MSM) tree and estimated the performance of unseen cases using ten-fold cross-validation.

Arpita Joshi et al [3] focused on Neural Network and Deep Neural Network to classify breast cancer. They also used Wisconsin Diagnostic Breast Cancer dataset and for missing data values, they used interpolation technique mean imputation and also utilized Principal Component Analysis

(PCA) for feature extraction and Linear Discriminant Analysis (LDA) for data compression.

Zejmo M et al [4] is based on deep learning approach to classify breast cancer using Convolution Neural Network (CNN) of two types GoogLeNet and AlexNet. They used images data of 50 patients, and applied Support Vector Machine (SVM) and tuned Neural Network using gradient descent.

Levy D et al [5] worked on Convolutional Neural Network (CNN) to classify pre-segmented breast masses in mammograms as benign or malignant. They used Digital Database for Screening Mammography (DDSM), a collaboratively maintained public dataset at the University of South Florida. It comprised approximately 2500 images. They applied transfer learning and data augmentation as rotation, cropping and mirroring to increase data effectiveness.

Karabatak M et al [6] utilized Association rules for reducing the dimension of breast cancer data set and Neural Network to classify cases. They also used Wisconsin Breast Cancer Dataset. They applied Apiori algorithm for dimension reduction. In order to evaluate the system performance, they used 3-fold cross validation method.

Garud H et al [7] presented a deep convolutional neural network (CNN) classification technique for the diagnosis of the Fine Needle Aspiration Cytology cell samples using their microscopic high-magnification multi-views. They tested their model on GoogLeNet architecture of CNN.

Liu K et al [8] proposed a model which had first layer as fully connected layer then have CNN means (FCLF-CNN). They also used WDbC and WBCD datasets for breast cancer and obtained the results by a fivefold cross validation.

Na Wu et al [9] utilized deep convolutional neural network to classify breast density. They worked on big dataset of over 200,000 breast cancer cases.

III. Dataset information

We used Breast Cancer Wisconsin (Diagnostic) Data Set for this study, which is publicly available at UCI Machine Learning Repository, created by Dr. William H. Wolberg, (General Surgery Dept. University of Wisconsin, Clinical Sciences Centre), W. Nick Street, (Computer Science Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706), Olvi L. Mangasarian, (Computer Sciences Dept. University of Wisconsin, 1210 West Dayton St., Madison, WI 53706).

Features are calculated from a digitized image of a Fine Needle Aspirate Cytology (FNAC) of a Breast mass. They showed features of the cell nuclei found in the image.

A. Attribute Information

It has 32 attributes of information with first column as ID of Patient, second column as diagnosis (M = Malignant, B = Benign), and other rest ten real-valued features are computed for each cell nucleus are:

a) radius (mean of distances from centre to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^A2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Table 1 UCI Dataset Information

Dataset	***Attributes***	***Instances***	***Classes***
Breast Cancer Diagnostic	32	569	2

We exclude two attributes as first attribute is ID of Patient and second attribute is diagnosis information and provide remaining attributes to our Deep Neural Network model. We have some cases with 0 values but we ignore them as they are the cases of Benign Cancer and we are focused towards Malignant cases. We convert our attribute number 2 which contains diagnosis information (M=Malignant and B=Benign) to one hot encoding with 1=Malignant and 0=Benign. It has 357 cases of Benign and 212 cases of malignant cancer.

0	S42302	M	1799	10.38	122.80	1001.00.11840	0.27760	0.30010.14710	
1	842517	M	20 57	17.77	132.90	1326.00.08474	0.07864	0.08690.07017	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974 0.12790
3	84348301	M	11	42	20.38	77.58	386.10.14250	0.28390	0.24140.10520
4	84358402	M	20 29	14.34	135.10	1297.00.10030	0.13280	0.19800.10430	

IV. Neural Network Methodology

For the Generation of Neural Network and to provide fast and accurate diagnosis on test data, we used high level Neural Network open source API keras which was developed in python and run on top of Theano, TensorFlow and CNTK. It was developed with a focus on enabling fast experimentation, being able to go from idea to result with the least possible.

A. *Google Collaboratory*

We also used free cloud service as Google Collaboratory to run our code, which is a jupypter note book environment that needs no setup to start. One can develop deep learning applications using popular libraries such as Keras, TensorFlow, PyTorch, OpenCV, numpy, pandas, and sklearn. Code files are stored in google drive and can easily be shared to other researchers.

B. GPU Computation

The process of training Deep Neural Network is computationally expensive due to huge amount of calculations on arrays of data and weights. CPU is designed for general computing having single-threaded performance with not more than 72 cores available right now, but GPU have 1000’s of cores with parallel computing for expensive operations.

Computational speed is extremely important because training of Deep Neural Networks can range from days to weeks. In fact, many of the successes of Deep Learning may

have not been discovered if it were not for the availability of GPUs.

C. Deep Dense Neural Network

Our model depends on Deep Dense Neural Network layer (linear operation) architecture. Deep Dense Neural Network (DNN) is an Artificial Neural Network (ANN) with multiple hidden layers between input and output layer. Basically DNNs are Feed Forward networks in which data flows from the input layer to the output layer without looping back.

Dense layers are fully connected layers so all the neurons in the preceding layer are fully connected with the neurons in the succeeding layer and so on.

D. Distribution of Dataset for training and testing

We distributed our dataset into 75/25 i.e. 75% (426) records for the training and 25% (143) records for testing of our Neural Network.

During the training phase, we first created sequential model which is a linear stack of layers and provided us facility to take input from single source. It requires list of layer instances to constructor. Our First Neural Network layer which is input plus first hidden layer consists of input array of 30 dimensions, with 64 array of outer dimensions and we applied ReLU R(z)=max(0,z) as an Activation Function on this layer, which is the most used Activation Function right now, ReLU converts values less than zero to zero and other than zero to one. Therefore, its range is zero to infinity.

E. Regularization technique Dropout

After that we applied Regularization technique of Dropout to help reduce over fitting and reduction in our training error of Neural Network, it did some addition in loss function and not learn interdependent set of features weights.

Dropout reduces inter-reliant learning in Neural Network neurons. It actually makes our model features robust that requires in functioning with other neurons random subsets.

Table 4 Confusion Matrix for Breast Cancer Diagnosis

Number of Patients in Test data with Benign and Malignant Cancer=143	***Actual***	***(No)***	***(Yes)***
	(No)	True Negative=88	False Positive=1
	(No)	False Negative=3	True Positive=51

F. Second Hidden Layer and Output Layer

Then we added our Second hidden layer with input from preceding hidden layer and consisted of array of 64 outer dimensions and again we applied ReLU as an Activation Function, and like preceding working, we again applied Dropout. Our last layer with an output dimension of single array had binary classification function Sigmoid applied to it. Sigmoid which has values between 0 and 1and is used to predict the probability of our Deep Dense Neural Network model.

After this step, we compiled our model and used Adam as an Optimization function, which is fast and efficient than classical Stochastic Gradient Descent. Adam is the combination of two Optimization functions Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). Finally we used logarithmic loss function for binary classification (binary_crossentropy) to calculate loss of our model.

After creating model, we executed our Code. In this phase our model performed iterations on our training dataset, with batch size and epochs parameters, at the start of training phase our model has bad accuracy but Neural Networks can achieve good accuracy after adjusting weights on successive iterations and minimize loss, at the end of training loop our Neural network can achieve good efficiency on the cost of decreasing loss.

G. Testing Phase

The last phase of our network is the testing phase. In this phase we first performed evaluation on our testing data set based on the model we designed in the training phase. It provided us the loss value and accuracy value of our model in test mode. Our model achieved 0.1% loss and 97.22% accuracy.

Then we applied prediction on testing data set. We executed in batch with a rule that greater than 90% prediction value belongs to malignant cases and less than 90% prediction value belongs to benign cases.

V. COMPARISION WITH OTHER ACTIVATION FUNCTION

In this study, we also evaluated our model performance with other Neural Network Activation Function like Leaky ReLU which is used to fix the dying ReLU problem, ReLU can "die" if a large sufficient gradient changes the weights so that our neuron never activates on new data. Instead of the function being zero when x < 0, a Leaky ReLU will instead have a small negative slope (of 0.1, 0.01 or so on). Some studies report good results with this form of activation function, but the results are not always consistent.

We use leaky ReLU with the slope of (0.1), with no changes in other model parameters and found 97.20% accuracy, which is also good but (0.02%) less accurate than our model with ReLU as an activation function.

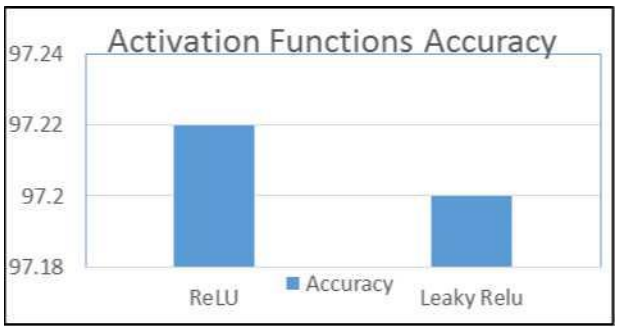


Fig. 1 Graph showing results of Activation Function

VI. MODEL EVALUATION

Here we are using Confusion Matrix to evaluate the performance of our binary classification model on the testing dataset. It gives Visual Representation of our model accuracy. It shows the relationships between Actual and Predicted classes.

Our model accurately classifies 51 Malignant and 88 Benign Cases from testing data in which 4 Cases were predicted wrongly.

Table 4 Confusion Matrix for Breast Cancer Diagnosis

Number of Patients in Test data with Benign and Malignant Cancer=143	***Actual***	***(No)***	***(Yes)***
	(No)	True Negative=88	False Positive=1
	(No)	False Negative=3	True Positive=51

VII. CONCLUSION

Deep Neural Network with dropouts is New and Cost-Effective technique with potential diagnostic value. We achieved 97.22 % accuracy on Breast Cancer Wisconsin (Diagnostic) Data Set features for Fine Needle Aspirate Cytology (FNAC) of a breast mass. For future work foresee more public data sets would be available in the field of medicine in order to perform experiments for such deadly diseases and will provide comfort, care and health to humanity in a cost-effective manner.

We also observed the absence of such computationally intelligent, efficient and robust system in the field of Medicine. At present there are no such software working in the hospitals which can perform computation on Neural Networks.

Limitations of study: Unavailability of public data sets to perform experiments.

REFERENCES

[1] Street WN, Wolberg WH, Mangasarian OL. "Nuclear feature extraction for breast tumor diagnosis". In Biomedical Image Processing and Biomedical Visualization 1993 Jul 29 (Vol. 1905, pp. 861-871). International Society for Optics and Photonics.

[2] LeCun Y, Bengio Y, Hinton G. "Deep learning". Nature. 2015 May; 521(7553): 43 6.

[3] Arpita Joshi and Ashish Mehta, **"**Breast cancer data classification using neural network and deep neural network techniques," International Journal of Recent Scientific Research Vol. 9, Issue, 4(D), pp. 2578825792, April, 2018

[4] Zejmo M, Kowal M, Korbicz J, Monczak R. "Classification of breast cancer cytological specimen using convolutional neural network". InJournal of Physics: Conference Series 2017 Jan (Vol. 783, No. 1, p. 012060). IOP Publishing.

[5] Levy D, Jain A. "Breast mass classification from mammograms using deep convolutional neural networks". arXiv preprint arXiv:1612.00542. 2016 Dec 2.

[6] Karabatak M, Ince MC. "An expert system for detection of breast cancer based on association rules and neural network". Expert systems with Applications. 2009 Mar 1; 36(2):3465-9.

[7] Garud H, Karri SP, Sheet D, Chatterjee J, Mahadevappa M, Ray AK, Ghosh A, Maity AK. "High-Magnification Multi-views Based Classification of Breast Fine Needle Aspiration Cytology Cell Samples Using Fusion of Decisions from Deep Convolutional Networks". InCVPR Workshops 2017 Jul 1 (pp. 828-833).

[8] Liu K, Kang G, Zhang N, Hou B. "Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks". IEEE Access. 2018; 6:23722-32.

[9] Na Wu, Krzysztof J. Geras, Yiqiu Shen, Jingyi Su, S. Gene Kim, Eric Kim, Stacey Wolfson,Linda Moy ,Kyunghyun Cho "Breast density classification with deep convolutional neural networks,"arXiv:1711.03674v1 [cs.CV] 10 Nov 2017

[10] Street WN, "A Neural Network Model for Prognostic Prediction," InICML 1998 Jul 24 (pp. 540-546).

[11] Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen,S. Gene Kim, Linda Moy, and Kyunghyun Cho, "High resolution breast cancer screening with multi-view deep convolutional neural networks," arXiv:1703.07047v3 [cs.CV] 28 Jun 2018

[12] Fonseca P, Castaneda B, Valenzuela R andWainer J. "Breast Density Classification with Convolutional Neural Networks", InIberoamerican Congress on Pattern Recognition 2016 Nov 8 (pp. 101-108). Springer, Cham.

[13] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014 Jan 1; 15(1): 1929-58.

Blood Transfusion Prediction

Hira Farman[1], Sadia Mushtaq[2], Anus-Ur-Rehman[3], Syed Huzaifa Ali[4],Fouiza Naaz[5], Ishrat Fatima[6] , Zain Noreen[7]

hira.farman@jinnah.edu[1], mushtaq_sadia@hotmail.com[2], anasrehman2925@gmail.com[3], huzzishah4@gmail.com[4],fouzia.naaz@jinnah.edu[5], ishrat.fatima@jinnah.edu[6] ,zain.noreen@jinnah.edu[7]

Department of Computer Science
Mohammad Ali Jinnah University

Abstract - Transfusion of blood established a standard way of treating patients, who are lacking in one or more blood constituents and is therefore a necessary part of health care. There is an increase in demand for blood. This can be caused by many reasons like severe accidents, increased medical surgeries etc. Efficient blood transfusion can cater to this increased demand. For increasing future blood donation, medical professionals require necessary details which can be provided by blood donor prediction. This study aims to suggest the provider, efficient prediction and motivates the suitable blood donors.

The focus of the paper is on the analysis of different classification algorithm for the prediction of finest blood donor (SVM, Cross Validation, Decision tree, K-nearest neighbor, Correlation and Regression, Principle component analysis, Gaussian Naive Bayes classifier of machine learning is used in blood transfusion dataset. The suggested methodology will result in higher accuracy and efficiency in selection process when compared with the present one.

Keywords: Binary classification, K-fold cross validation, SVM, blood donation

I. INTRODUCTION

The necessity for blood is gradually increasing day by day as the population is increasing. As we know, American red alert blood endowment is no longer serviceable after 42 days approximately. Blood donation is moral action which involves transfusion of blood to other people who is in dire need of it.

Many blood banks are available, they maintain details of blood collection, issued details etc. They maintain hundreds or thousands of blood donor record. Many applications are available to maintain the record of patient and donor, for this purpose many blood typing device are designed. They identify blood type before any blood transfusion in order to avoid inconsistencies that could be a prime factor of patient's death. The main objective of this research is to present a inclusive methodology to discover the knowledge for selecting targets for direct promoting from database. This paper focuses on analyzing the efficiency of different classification algorithm using blood transfusion dataset.

The present study provides efficient prediction and motivates the suitable blood donors. Basically, we derive out that a person has donated blood at specific time in the past or not. This revision expands RFM model by including four parameters e.g. RFMT. Using binary classification since a binary variable representing with the help of dataset if a person donated blood (1 stand for donating blood; 0 stands for not donating blood).This methodology gives more effective and accurate results than the existing ones.

In this experiment it will be shown that how python is used when there are problems in classification. By using several machine learning classification algorithm, it will be determined, which one gives the best results in the prediction, and which one doesn't. Most of the classification task frequently found in routine life. A classification is a process that works on predefined groups or classes based on a number of observed attributes associated to those entities. While there are some more traditional tools for classification, such as certain K-fold cross validation, statistical procedures have shown to be an effective solution for this type of benefits to use python- they are data driven, they can approximate any function - linear as well as non- linear (which is quite important in this case because groups often cannot be divided by linear functions).

II. RELATED WORK

In the recent years, several studies presenting the prediction of efficient blood donor.

S.Asha Rani et al. [1] used different classification algorithm to calculate better efficiency in blood transfusion dataset. For comparison, they used various classification algorithms to determine & to measure the accuracy with short time intense for prediction of suitable blood donors.

Maryam Ashoori et al. [2] has used four different data mining algorithm including a decision tree algorithm. They assist doctors to determine suitable blood storage and avert from side effects of blood deficiency in an emergency condition and for additional accurate prediction model they also apply attributes set such as hemoglobin, minimum blood pressure, maximum blood pressure, temperature, blood pressure and pulse.

The main purpose of Gaston Godin et al. [3] is to support the idea that separate awareness schemes should be adopted to increase repeated blood donation among experienced versus new donors. By using Logistic regression technique they identify factors predicting repetitive blood donation among new donors and experienced one.

In Wijai Boonyanusith et al. [4] research, the target group of donors may cause adverse effects on higher cost of donation, time lost, and poor quality of blood. Applying information of consumer survey such as, behaviors, feelings and opinions of the donors in blood donation can enhance the analysis of the feasibility of blood donation of each individual and this Classification model and donor database system will contribute greatly for blood arrangement especially when there are emergency needs for blood for uses in the live saving treatments. Decision tree algorithms and neural network used for this purpose.

Arvind Sharmal [5] classifies and guesses the number of blood donors with the help of key factor, their blood group and their age. For the extraction of the knowledge of blood donor, they build a data mining model for classification, dataset has been implemented on WEKA tool and used J48 algorithm for prediction. In this work, clinical databases are also used for prediction.

Maryam Ashoori [6] gives detailed explanation of algorithms on data mining. To predict the future behavior of a healthy blood donor by classification algorithms of data mining. To analyze the data Clementine software version 12.0 technology was used. Four different data mining and classification algorithms used: CHAD, C&R Tree, C5.0 and QUEST.

Shih-Yen Lin [7] targets the dental clinic and try to reallocate resources for those who are severe patients. By using classification and regression method calculating accuracies for lost patients and very severe patients, to classify each patient into appropriate group with rules and to predict the frequency of a particular case. Dental clinic can pay much attention to those who might be considered as very severe patients and try to reallocate resources to those who might be the lost patients.

Van Looy S. et al. [8] works on SVM for classifying dataset .For the prediction of tacrolimus concentration of blood in liver transplant patients from an ICU dataset.

Wen-ChenLee and Bor-Wen Cheng [9] uses machine learning technique e.g. clustering and classification to determine the variation in blood donation behavior among the existing donors and predicts their intentions towards blood donation by understanding the contributing factors, these factors are then consume to design a strategy that would lead to increase the voluntary blood donation frequency.

The main purpose of Cheng Yeh et al. [10] research is to resolve the issue for direct marketers, how to sample targets from the population for a direct marketing campaign. For this, they expand RFM model to RFMTC model by adding extra parameters. By using probability theory and Bernoulli series, they predict that one customer will buy at the next time etc.

M. Musthakahamed [11**]** resolves the issue of communication between donor and recipient. For this, direct communication is designed in android application. It gives notification by using any mobile device. With the help of this application seekers and donors can easily access the blood donors and save various lives easily.

Abja Sapkota [12] focuses on blood transfusion practice targeted in Nepal health care personnel. For this, they include many areas for observation like general surgery, orthopedics, general medicine etc. to increase the training to the healthcare personnel on blood transfusion practices

III. ATTRIBUTE INFORMATION

Data set is taken from UCI machine learning repository. Following is the variable type, variable name, the measurement unit and a short explanation. The "Blood Transfusion Service Center" is a classification problem, the order of this listing resembles to the order of numerical along the rows of the database."R" (Recency - It represents last donation in months), "F" (Frequency - It represents the total number of donation),

"M" (Monetary - It represents total blood donated in c.c.),

"T" (Time - It represents first donation in month)

IV. METHODOLOGY

In this work, cross validation technique K-fold has used because it is a statistical method for analyzing a dataset In which, there are one or more independent variables that outcome is measured with a dichotomous variables (in which there are only two possible outcomes) we are using this for train and test and the library we used: from sklearn.model_selection import cross_val_score.

Secondly, K-fold training and testing splits has used to implement an algorithm called SVM (Support Vector Machine) which is another way of classifying and representing responsive predictions. This has used for train and test and the library we used: from sklearn.model_selection and import svm.

Moreover, three more machine learning algorithms were implemented in order to achieve the accuracy which is best suitable for this data model. Algorithms implemented are, K-nearest neighbor, Gaussian-Naive Bayes and CART (Classification and Regression Trees).

A. Decision Tree

Decision tree is the representation through graph of structure of tree. Ovals is used for leaf nodes and rectangle is used for internal nodes. There are child nodes of internal nodes ,two or more than two , which is used to test the attribute [11].

B. K-Nearest Neighbor

K-Nearest Neighbor used Euclidean distance, Squared Euclidean distance, Manhattan distance to classify new data or for prediction .KNN algorithm applied for clustering purpose. Mostly initial K – vectors is defined. It's implemented in pattern recognition task mostly because of its good performance [14]. KNN is non-parametric method. It's used for classification and regression both. In both belonging, input comprises of the closest K. In short, we can say it checks similarity with their neighbors & on the basis of this classifies data.

Before applying K-nearest neighbor, select suitable value for parameter K. Generally value of K depends on the dataset .Value of K which is too small may contribute to over fitting, while large K value increases complexity and affects in decision [12].

C. *Finding Distance metric*

Performance of KNN based on the distance metric, classification accuracy of KNN improved by selecting a distance metric for dataset. Many formulas used for this e.g.(euclidean distance , cosine distance, city block metric etc,). [14]

Euclidean distance:

$$d^2_{st=} (x_s-y_t)^2 + (x_s-y_t)^{2\ root}$$

D. Naïve Bayes

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

V. RESULT & VISUALIZATION

For visualization, matlplotlib pyplot library was used.

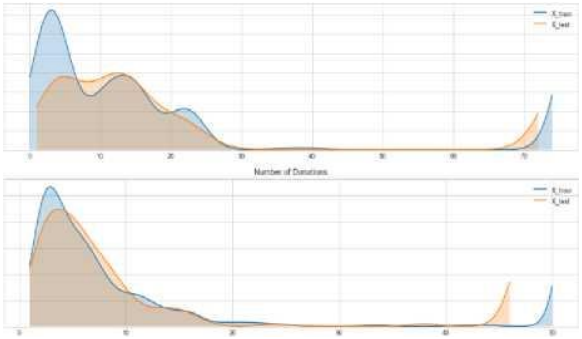


fig 1: months since last donation graph &

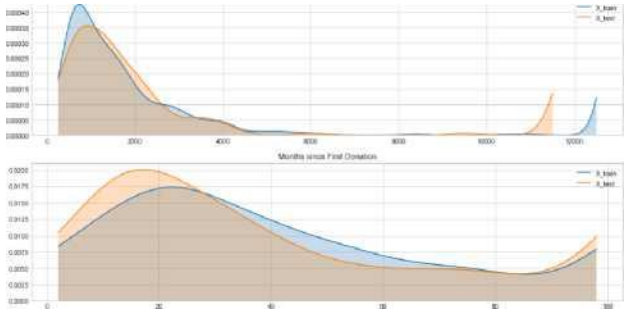


fig 2: Total volume donated and month since first donation

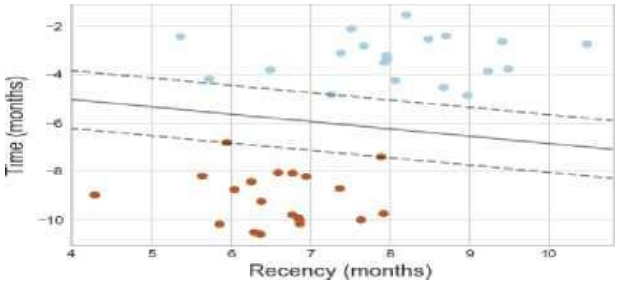


fig 3: Support vector machine for segregates (months) & recency (months)

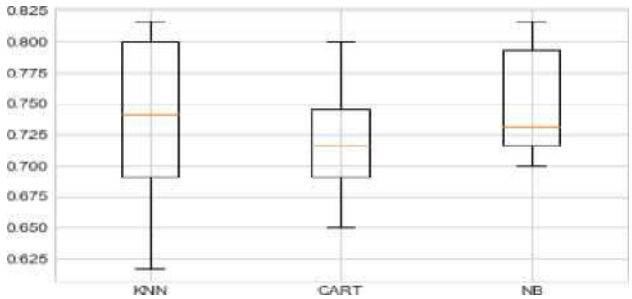


Fig 4: three algorithms (KNN, CART, NB) comparison accuracy range

This study compares the performance of various binary classification algorithms which do not invested previously on clustered data and non-clustered data to see, if we can better predict whether a person is going to donate blood or not. To interpret a result, with the help of Decision tree classifier, nearest neighbor classifier, Gaussian Naive Bayes classifier And the result of our interpretation is 0.7633 is good as compared to other researchers.

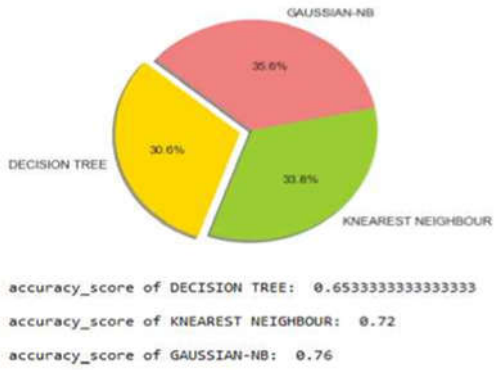


fig 5. Shows the accuracy pie chart of the algorithms

VI. CONCLUSION:

Analysis through python framework first a dataset is selected and secondly, python is usedto load dataset through import pandas and sklearn method and then some machine learning algorithm were chosen for decision to identifywhich algorithm is fit to be used so, we chose logistic regression and then visualized data through matplot library and classified it with the help of machine

REFRENCES

[1] S.AshaRani, Dr.S.Hari Ganesh, “A comparative study of classification algorithm on blood transfusion”, 2014.

[2] Maryam Ashoori, “A model to predict the sequential behavior of healthy blood donors using data mining” , 2015.

[3] Gaston Godin, “Determinants of repeated blood donation among new and experienced blood donors” , 2007.

[4] Wijai Boonyanusith and Phongchai Jittamai,Member, IAENG, “Blood Donor Classification Using Neural Network and Decision Tree Techniques “, 2012.

[5] Arvind Sharma, “Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool”, 2012

[6] Maryam Ashoori, Somaye Alizade, Hoda Sadat Hosseiny Eivary, Saber Rastad, Somaye Sadat Hossieny Eivary, “A model to predict the sequential behavior of healthy blood donors using data mining”, 2015.

[7] Shih-YenLin,Jo-Ting Wei,Chih-Chien Weng and Hsin-Hung Wu, “A Case Study of Using Classification and Regression Tree and LRFM Model in A Pediatric Dental Clinic”, 2011.

[8] Van Looy S,Verplancke T, Benoit D, Hoste E, Van Maele G, De Turck F, Decruyenaere J.”A novel approach for prediction of tacrolimus blood concentration in liver transplantation patients in the intensive care unit through support vector regression”, 2007.

[9] Wen-Chen Lee, Bor-Wen Cheng, “An Intelligent System for Improving Performance of Blood Donation”, 2011.

[10] Cheng Yeh, King-Jang Yang, Tao-Ming Ting, “Knowledge discovery on RFM model using Bernoulli sequence”,2009

[11] M. Musthak Ahamed, R. Rajmohan, S. Mohamed Nizar”, Design and Implementation of Blood Donors Alerting System”, 2017.

[12] Abja Sapkota, Sabitra Poudel, Arun Sedhain, and Niru Khatiwada, “Blood Transfusion Practice among Healthcare Personnel in Nepal: An Observational Study”, 2018.