April-2019 SPECIAL ISSUE ISSN: 2278-4187

Anusandhan

Science Technology & Management Journal

A special issue for
Publication of Selected Research Papers Presented in

National Conference on Advances in Computer Science and Information Technology

at

Dr. C.V. Raman University, Bilaspur (C.G.)

on

22nd & 23rd February 2019

Approved by UGC indexed by Copernicus
Online version (e-ISSN 2457-0656)
http://aujournals.ipublisher.in/Default

Proceedings- NCACSIT-2019

Published By

Rabindranath TAGORE UNIVERSITY
MADHYA PRADESH, BHOPAL
AN AISECT GROUP UNIVERSITY

Village- Mandua, Post- Bhojpur, Distt. Raisen (M.P) India
Pin- 464993, Ph. 0755-2700400, 07480-295707
E-mail: aisectjournal@rediffmail.com, vcau.bhopal@rediffmail.com
Website: www.rntu.ac.in

Glimpses – NCACSIT held of CVRU Bilaspur (CG)

Vol.VIII / Special Issue-XVI April-2019 ISSN : 2278-4187

Anuṣandhan

Science Technology & Management Journal

Approved by UGC Indexed by Copernicus
Online version (eISSN 2457-0656) http://aujournals.ipublisher.in/Default

Proceedings- NCACSIT-2019

Rabindranath
TAGORE UNIVERSITY™
MADHYA PRADESH (Formerly known as AISECT University)

Rabindranath TAGORE University: Village-Mandua, Post-Bhojpur, Distt.-Raisen (M.P.)
India Pin-464993, Ph-0755-2700400, 07480-295707
E-mail-aisectjournal@rediffmail.com, vcau.bhopal@rediffmail.com
Web site : www.rntu.ac.in

A SCIENCE TECHNOLGY & MANAGEMENT JOURNAL OF RNTU

Editorial Board

Dr RN Yadav
(Enviroment Sc.)

Dr Sangeeta Jauhari
(Management)

Dr SR Nigam
(Techology)

Dr SR Awasthi
(Energy)

Dr Suryanshu
(Science)

Dr Deepti M
(Entrepreneurship)

Craft & Compile

Mangesh Jadhav
Graphics designer

Manoj K. Kanade
Compilation

Guest Editor

Dear Readers

Phenomenal development of new techniques, innovations, systems and knowledge in the field of computer science and information technology, during last one decade, has taken the world by surprise. By virtue of their significant contribution in the fast growth of technology, the computing system and information technology have now occupied the driver's seat, to take the world economy to a new level. Computer and Information Technology have entered all the disciplines of studies. They have truly become growth engines for today's world progress. Computing system has become an essential and vital element in most of the projects which are opening new avenues for progress.

Dr. C.V. Raman University from its inception has focused on research and innovation. Considering the strong potential of innovation and research in Computer Science and Information Technology, the University has paid great attention, to develop advanced research facilities in its labs and build an environment, to encourage research in the Computer Science and Information Technology departments. In an effort to promote research and motivate students and faculty, the University also organizes large number of events related to research and innovation. This year in the beginning itself a national conference was organized on **"Advances in Computer Science and Information Technology – NCACSIT-2019,"** by the dept of CS & IT which has set the agenda on research for rest of the year. More than 150 papers were received for the conference from nooks and corners of the country. It was great pleasure to note that all the experts who were invited in the 2 days conference commended the excellent content and high quality of papers and presentations made during the conference. The conference provided a vibrant platform for sharing of knowledge, research findings and experience, among delegate researchers, faculty, students and experts. A jury of experts selected only 30 papers, to publish in special issue of UGC approved journal **ANUSANDHAN** as proceedings of the conference, due to journal's limited capacity which it offered for the special issue. In this special issue all those 30 selected papers have been published. I am sure readers will find them quite useful and thought provoking.

I congratulate all the authors and members of the organizing team of **NCACSIT-2019,** for the great efforts put in and honour earned of getting the proceedings published in a renound UGC approved journal.

Dr. R.P. Dubey
Vice Chancellor,
Dr. C.V. Raman University, Bilaspur (C.G.)

In This Special Issue

Vol VIII/Issue XVI Special Issue April- 2019 ISSN: 2278-4187

A Prediction Model for Improving Accuracy of Students Performance Using Big IoT Data Analytics

Abdul Alim[1], Diwakar Shukla[2]
[1]Research Scholar, Dept of CSA, Dr. Harisingh Gour Vishwavidyalaya, Sagar (M.P.) India.
[2]Prof. & HoD, CSA, Dr. Harisingh Gour Vishwavidyalaya, Sagar (M.P.) India.

ABSTRACT

Today the big data analytics are playing a very important role in various field like business, organizations, health care, weather forecast because the huge amount of big datasets have been produced since the past some years via educational apps, organization's websites, and social media. There are a lot of educational big dataset: are available which have generated through the online courses. Institution's report like student academic result, student placement activity performance records. Those types of datasets have stored in form of unstructured data. Thus big IoT data analytics has the capability to process unstructured data and predict the valuable information insight into large datasets and this information we can use for the next coming student in the campus for improving their performance. In this paper, we have given a prediction model to improve the accuracy of student's performance such as job placement by using big IoT data analytics. The Internet of Things (IoT) devices has enabled to connect all digital devices to each other and share the information among them. The proposed prediction model can be improved existing learning process and will find out which students are not performing well then further the management can provide extra facility to that type of poor students.

Keywords: Big Educational Data, Big IoT Data Analytics, Clustering, Sampling Technique.

I INTRODUCTION

Due to rapid growth of technologies in education sectors for learning online by using various tools and framework. Almost every institutions and organizations are moving towards online administration process such as admission process, online result display, placement activities etc. in the online process the educational data has been stored in digital form within databases. Once the student has completed their course then his academic data has stored only for records but here we can previous students' performance data for improvement of future student's performance. The Internet of Things is the interlinking of heterogeneous devices with each other together via internet. By the using IoT technologies we can improve the existing learning systems and to make smart education system. The smart education system, the student continuously monitored by using the sensors devices which will measure the student's activity in daily life and to be stored information on the servers. If the student is not able to come and attend the lecture in classroom then he can join from his home also because IoT technologies have ability to connect each and every device with each other and share the information to all over the devices [1]. After the came information and communication technologies especially the internet becoming over more ubiquitous within the education system and its play key role in Improving education quality. At present there are different seven types of technologies which are bring numerous innovation and benefits in the education sectors like consumer technology, visualization technology, learning technology, enabling technology, social media technology, digital strategies, and internet technologies. In which IoT technology to supports education in various ways. The following figure1 has explored how IoT technologies transforms knowledge delivering and obtaining creating an intelligent experimental teaching and learning environment. The IoT can help solve diverse challenges across an educational sectors, it is used to reduce cost and resource usages. These technologies allow enabling to remote access to equipment, virtual meetings and learning sessions, use cloud computing and big data analytics for shared services and solutions. Furthermore improve design of organizations and universities buildings, designing ICT- intense smart building which has smart doors, locks, radio frequency-Identification, Cameras and connected devices with the help of smart devices monitoring and surveillance of entire buildings for secure and safe learning environment [2].

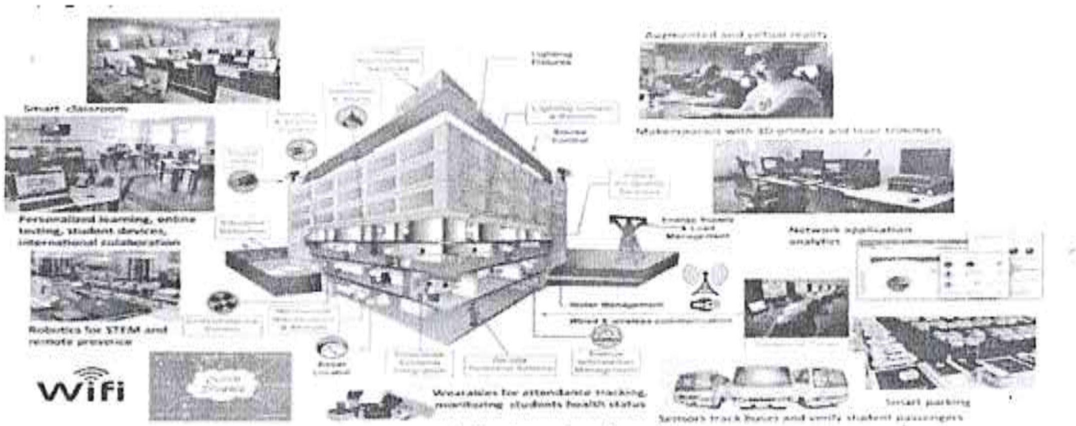

Smart classroom
Augmented and virtual reality
Lighting fixtures
Makerspaces with 3D printers and laser trimmers
Personalized learning, online testing, student devices, international collaboration
Energy supply & Load Management
Network application analytics
Robotics for STEM and remote presence
Water Management
Wired & wireless communication
Wifi
Wearables for attendance tracking, monitoring students health status
Smart parking
Sensors track buses and verify student passengers

Fig.1 The smart IoT educational environment

II INTERNET OF THINGS IN EDUCATION SECTORS

The Indian education system is mostly moving around the reading books, attending class, exam and grades, where the creating learning lies far away. Even teachers also teaches within the syllabus and the students focused on that only part but today an existing technology development in education that is known as cloud computing. The students and administration have opportunities to quickly access different types of application platform and resources by the web pages on their need without leaving the classroom they can take notes, textbooks could be scanned to be received instant additional resources. The IoT and cloud computing technologies enable users to access and control data on internet. Teachers can identify problem area in which students tend to make mistakes by analyzing student records which are storing every day and will allow teachers to improve teaching material and methods [3]. Big data analytics is rapidly emerging as a key IoT initiative to continuously improve the education system. The IoT data analytics will perform lightning-fast analytics with large queries to allow colleges and universities to gain rapid insights and make quick decision in the case of subject or stream selection. The following figure 2 shows the relationship between IoT and big data analytics.

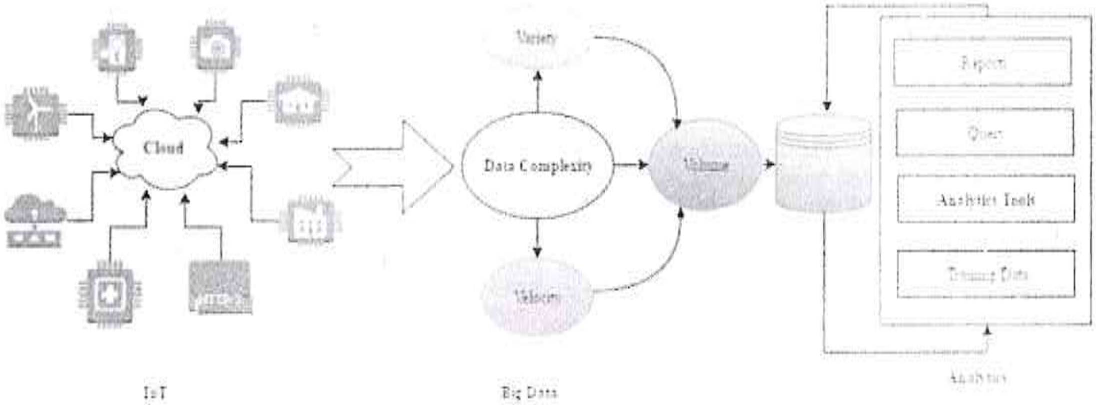


Fig. 2 The Relationship Between IoT and Big Data Analytics

The IoT technology can be divided into three parts and to enable the management of IoT data. The first part managing IoT data sources which are connected through sensor devices such as CCTV cameras by using web applications. The second part, the data generated from different senso, devices are called big data because which are based on 3Vs- Volume (collection of heterogeneous data sets), variety (data in various format structure, unstructured or semi-structured) and velocity (the speed of data generation). These large amounts of data are stored in big data files in shared types of databases. For processing these huge amount of data, we need a powerful analytics tools such as Hadoop MapReduce, spark etc. which is the last part. The figure3 is the complete architecture of Internet of Things (IoT)

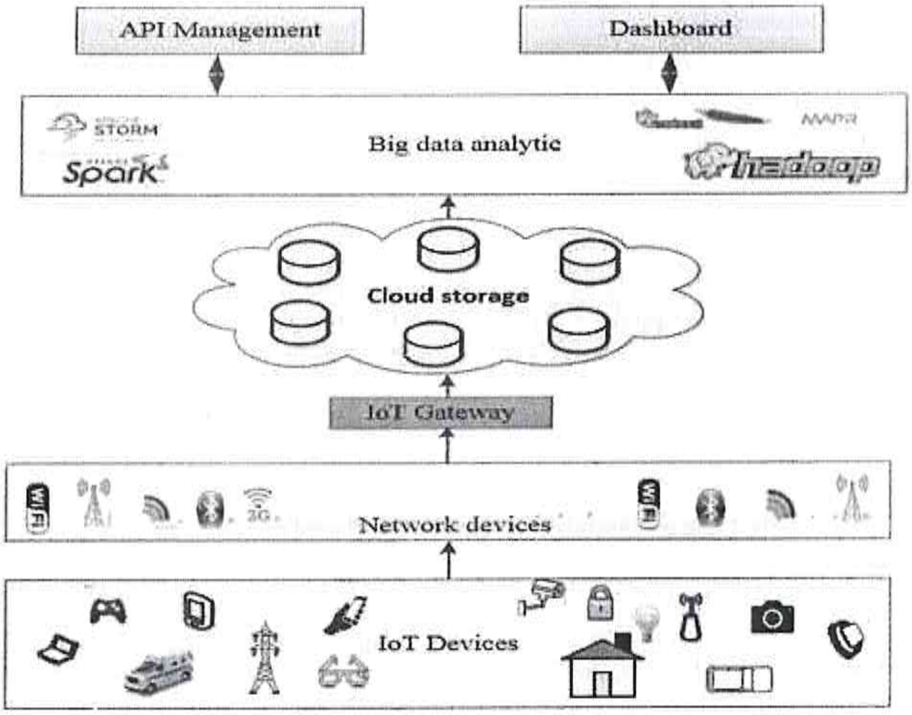


Fig. 3 The architecture of Internet of Things

In the figure3 IoT device layer, connected to all object to each other's via network devices and store the all datasets on cloud through the IoT gateway. The multiple queries will process through the big data analytics layer which will provide information to the users [4]. The IoT inclusions in education provide facility for the student to become active participants accessing to their courses, laboratories and exercises at any time from many places. The following table1 has explained the major concept of IoT in education [5].

Table 1
The collaboration Learning Benefits

Social Benefits	➢ A social support system for learners. ➢ Diversity understanding among students and staff. ➢ A positive atmosphere for modeling and practicing cooperation. ➢ Learning communities.
Psychological Benefits	➢ Increased students' self-esteem. ➢ Reduced anxiety. ➢ Positive attitudes towards teachers.
Academic Benefits	➢ Critical thinking skill. ➢ Active participation in the learning process. ➢ Improved classroom results. ➢ Appropriate problem-solving techniques. ➢ Personalization of learning lectures. ➢ Increase students' motivation.
Assessment Advantages	➢ Utilization of variety of assessment like observation of the group. self-assessment of the group and individual assessment of the members in a group.

Technologies have played a significant role in the field of education for connecting and the students. It will also improve the education infrastructure and administration management with the IoT applications. The following figure has shown the IoT based smart environment.

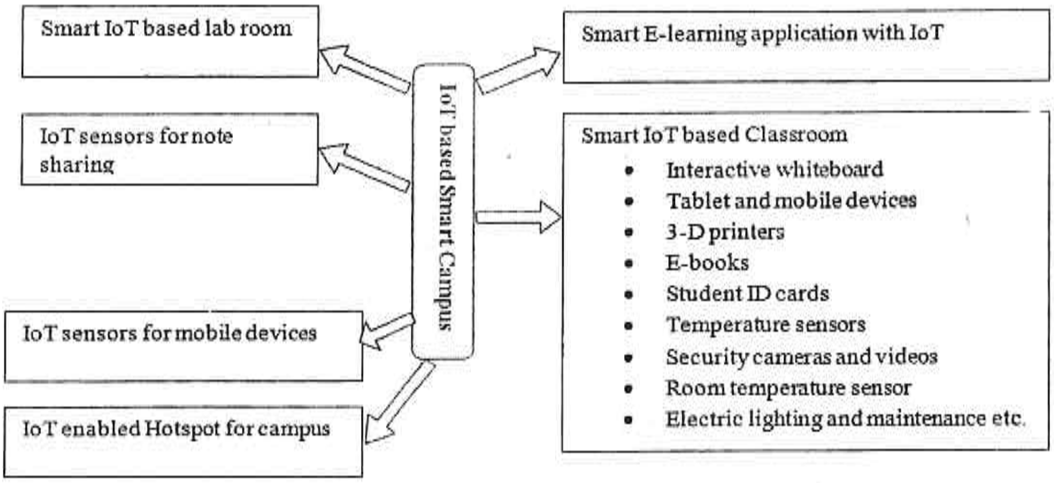

Smart IoT based lab room
IoT sensors for note sharing
IoT sensors for mobile devices
IoT enabled Hotspot for campus
IoT based Smart Campus
Smart E-learning application with IoT
Smart IoT based Classroom
Interactive whiteboard
Tablet and mobile devices
3-D printers
E-books
Student ID cards
Temperature sensors
Security cameras and videos
Room temperature sensor
Electric lighting and maintenance etc.

Fig.4 The architecture of Smart Campus based on IoT

There are also some key points like automatic attendance tracking system, wireless door lock and real-time feedback on lecture quality [6]. After the migration of IoT in education sector then IoT brings tremendous challenges and opportunities also. Some of the IoT challenges in education when IoT will implement, the challenges includes: cloud computing, instructional technologies, security and privacy, research computing, quality and ethics and finance [7].

III ADVANTAGES AND FUTURE SCOPE OF INTERNET OF THINGS

The IoT technology will open door for new and innovative smart education system. School and colleges can improve their campuses and enhance access to information. It creates smart lesion plan for all student where they can access anytime and anywhere. It is also optimizes cost of lighting on the room occupancy, cost of heating, ventilation and air condition system and automatically opening and closing the windows. Furthermore the IoT based technology will helps to enhance and improve the performance of each student. In which the student has facility to customize their courses according to needs because some students does not want to all things. Thus they can customize and adopt the education for the improvement in accountability and performance. One of the most important advantages of IoT based education system is to enhance collaboration among educators and learners [8].

IV PROPOSED PREDICTION MODEL

Data mining is a process to discover and extract novel and useful patterns from the huge amount of datasets through the mining algorithms. The data mining algorithms are enable to extract the hidden pattern from large datasets, through the classification process we can predict the class of object whose class label is unknown. The association analysis is able to frequent items occur together in a given datasets. Data mining enables in various sectors like in industry, healthcare, city governance, education etc. the following figure 5 shows the overview of data mining process [9].

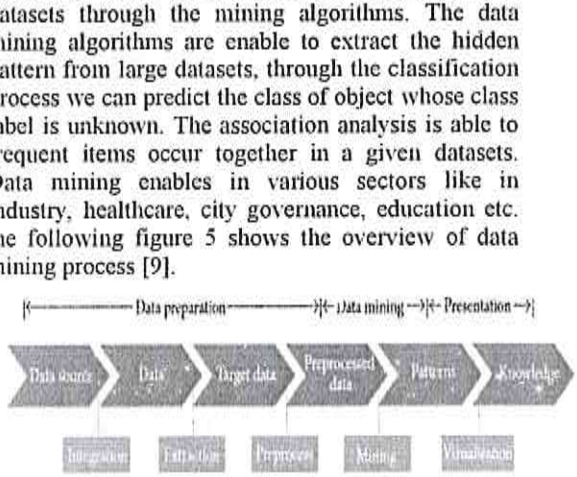


Fig. 5 the Overview of Data Mining Process

The k-nearest neighbor algorithm has been frequently used in pattern recognition because of it enables composed image features in three types: color feature, shape feature, and texture feature. The single types of feature cannot explain the feature of any object with complex background [10]. The application of Internet of Things in education, we can get lots of monitor image data from the student which they have done in their daily academic activity. Through the image data we can identify the satisfactory of student in classroom and also in other academic activity like stress. This image data analysis will help to improve student performance.

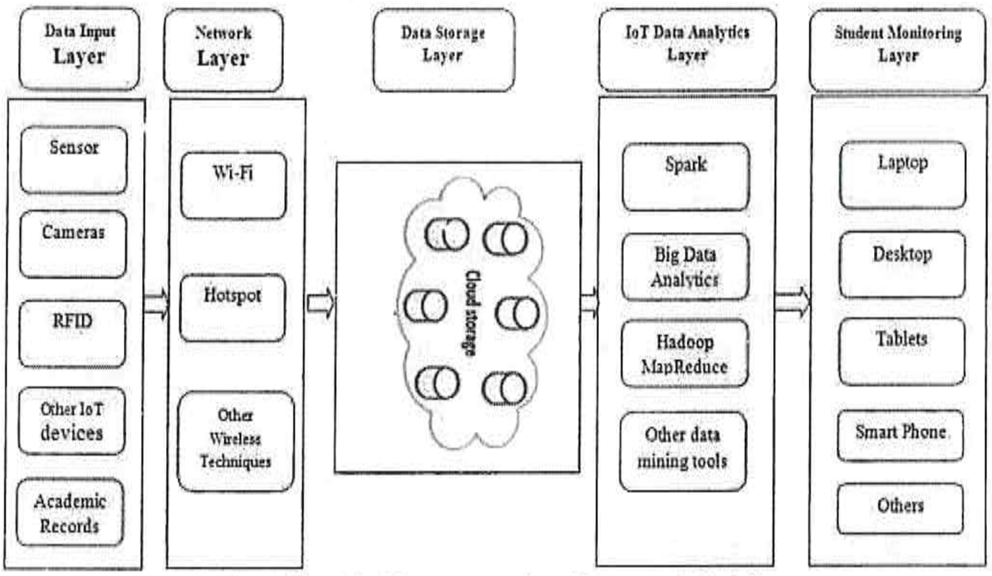


Fig. 6 Graphical Representation of Proposed Model

The proposed model will provide scale teachers and best quality of instruction, convert ad-hoc decision making into data-driven decision. Smart classroom with scale content recordable and replication instruction at any time from anywhere. The student can access to crowd-sourced content and they can also customize curriculum. This model has five layers, the data input layers will collect student activities and teachers also from various IoT devices like sensor, cameras, RFID and so on. After the collection of datasets the devices will send to the cloud via network layer for storage these datasets. The IoT data analytics layer will process those datasets and the result sends to the monitoring layer. The institution management will monitor from anywhere at any time. Furthermore the management can decide which type of facility is needed to improve the student's performance.

V CONCLUSION

This paper demonstrates the potential value of Internet of Things in education sectors for improving the student's performance through the monitoring daily activities. The IoT can change education policies and give potential impact on making secure campus with IoT devices. In this paper, we have discussed about the potential value of IoT in education and have given a prediction model for improved the student's academic performance. This paper also has explained the advantages and challenges of Internet of Things in education and realizes the benefits from the connected people or collaboration learning.

REFERENCES

[1] Rathore, Mazhar, M., Ahmad, Awais, Paul, Rho, Seungmin, (2016), Urban planning and building smart cities based on the Internet of Things using Big Data analytics, *Computer Networks*, Vol. 101 ,PP. 63-80.

[2] Maksimovic, Mirjana. (2017). Green Internet of Things (G-IoT) at engineering education institution: the classroom of tomorrow, 1(3). PP. 1-4.

[3] Venna, Ambica, Manjulatha, B. , Soumya, K., (2016), A Study on the Integration of IOT and Cloud Computing for Education System, *International Journal of Innovative Research in Computer and Communication Engineering*, 4(6), PP. 12279-12284.

[4] Marjani, Mohsen & Nasaruddin, Fariza & Gani, Abdullah & Karim, Ahmad & Abaker Targio Hashem, Ibrahim & Siddiqa, Aisha & Yaqoob, Ibrar. (2017). Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. *IEEE Access*, vol. 5, PP. 5247-5261.

[5] Maksimovic, Mirjana (2017), Iot Concept Application In Educational Sector Using Collaboration, *Teaching, Learning and Teacher Education*, 1(2). PP. 137 – 150.

[6] Gul, Shahla, Asif, Muhammad, Ahmad, Shahbaz, Yasir, Muhammad, Majid, Muhammad, Arshad, Sheraz and Malik, M. (2017). A Survey on Role of Internet of Things in Education, *International Journal of Computer Science and Network Security*, 17(5), PP. 159-165.

[7] Aldowah, Hanan, Rehman, Ul, Shafiq, Ghazal, Samar and, Umar, Naufal, Irfan (2017). Internet of Things in Higher Education: A Study on Future Learning, IOP Conf. Series: *Journal of Physics*: Conf. Series 892, PP. 1-10.

[8] Bajracharya, Biju and Blackford, Cody (2016). Prospects of Internet of Things in Education System. *The CTE Journal*, 6(1), PP. 1-7.

[9] Chen, Feng & Deng, Pan & Wan, Jiafu & Zhang, Daqiang & Vasilakos, Athanasios & Rong, Xiaohui. (2015). Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks*, PP. 1-14.

[10] Hu, Guoxiong & Yang, Zhong & Zhu, Maohu & Huang, Li & Xiong, Naixue. (2018). Automatic classification of insulator by combining k-nearest neighbor algorithm with multi-type feature for the Internet of Things, *EURASIP Journal on Wireless Communications and Networking*, 2018(177), PP. 1-10.

Text Preprocessing and Classification Using Machine Learning Technique

Amit Kumar Dewangan[1], S. M. Ghosh[2], A. K.Shrivas[3]
[1,2,3]Dr. C. V. Raman University, Bilaspur (C. G.) India.

ABSTRACT

Sentimental analysis is the method of finding sentiment such as positive or negative from a text data. In this paper we have used some feature selection techniques such as Mutual information, Information gain and TF-IDF to select features from high dimensionality data set. These methods are evaluated over dataset consists of 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal and is known as "Sentiment Polarity Dataset version 2.0". The reviews are equally partitioned into a positive set and a negative set (1000+1000). The machine learning play very important role for preprocessing and classification of data. The classification is performed using support vector machine(SVM), Random Forest, Random Tree, Naïve Bayes, Bayes Net and J48. We have used ensemble model to achieve high accuracy provided by WEKA tool.

Keywords: Classification, Feature Selection, Cornell Movie Dataset.

I INTRODUCTION

In this modern scenario data is part of human life. It is a necessary element of human daily life. Our environment is bounded from the data and information. Data is any type like structured, semi-structure or unstructured. Structured data like numerical values of attributes and unstructured data is like audio, video, image, text, text with numbers etc. Due to these reasons internet is one of the essential part of human life. The information in it covers a wide range of areas such as academic information, feedback or opinion about products, comments about social issues etc. It helps people to think and make decision in many things. Majority of people always listen to others opinion before taking a final decision. Sentimental analysis is one of the research areas [11].

The analysis of sentiments may be document based where the sentiment in the entire document is summarized as positive or negative. It can be sentence based where each and every sentence, having sentiments, in the text is classified. Sentiment analysis can be phrase based where the phrases in a sentence are classified according to the polarity based on some patterns of their occurrence. Sentiments are classified as positive (denotes a state of happiness, bliss or satisfaction on part of the writer) or negative (denotes a state of sorrow, dejection or disappointment on part of the writer)[7].

II RELATED WORK

The papers are basically focused on encapsulating the movie reviews at characteristic level so that user can find easily that which character of the movie they liked or disliked. In this paper, the author has two different methods are implemented for finding subjectivity of sentences and then rule based system is used to find feature-opinion pair and finally the orientation of extracted opinion is revealed using two different methods. Initially the proposed system uses SentiWordNet approach to find out orientation of extracted opinion and then it uses the method which is based on lexicon consisting list of positive and negative words [1].

It illustrates that comparison of the efficiency of the different classifiers focusing on numeric and text data. Datasets from IMDb and 20newsgroups have been used for the purpose. Current work mainly focuses on comparing different algorithms such as Decision Stump, Decision Table, K-Star, REPTree and ZeroR in the area of numeric classification, and evaluation of the efficiency of Naive Bayes classifier for text classification. In this paper, we have used WEKA tool to evaluate and analysis of datasets [2].

In this paper, authors have analyzed the Movie reviews using various techniques like Naïve Bayes, K-Nearest Neighbor and Random Forest [3].

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 [4].

The other authors have focused on the classification of opinion mining techniques that conveys user's opinion i.e. positive or negative at various levels. The precise method for predicting opinions enable us, to extract sentiments from the web and foretell online customer's preferences, which could prove valuable for marketing research. Much of the research work had been done on the processing of opinions or sentiments recently because opinions are so important that whenever we need to make a decision we want to know others' opinions [5].

This research concerns on binary classification which is classified into two classes. The classes are positive and negative. The positive class shows good message opinion while the negative class shows the bad message opinion of certain movies. This justification is based on the accuracy level of SVM with the 10-Fold cross validation and confusion matrix. The hybrid Partical Swarm Optimization (PSO) is used to improve the election of best parameter in order to solve the dual optimization problem [6].

This paper extends our ideas pertaining to Sentiment Analysis to the regional language Kannada, spoken mainly in Karnataka, a state in southern part of India. They have explored the usefulness of semantic approaches and machine learning approaches, used predominately on English language data set, from Kannada web documents. They found the average accuracy of machine learning approaches to be better than the average accuracy of semantic learning approaches for Kannada data set [7].

In this paper, they proposed an approach to understand situations in the real world with the sentiment analysis of Twitter data base on deep learning techniques. With the proposed method, it is possible to predict user satisfaction of a product, happiness with some particular environment or destroy situation after disasters. Recently, deep learning is able to solve problems in computer vision or voice recognition, and convolutional neural network (CNN) works good for image analysis and image classification. The biggest reason to adopt CNN in image analysis and classification is due to CNN can extract an area of features from global information, and it is able to consider the relationship among these features. The above solution can achieve a higher accuracy in case of analysis and classification [9].

III PROPOSED ARCHITECTURE

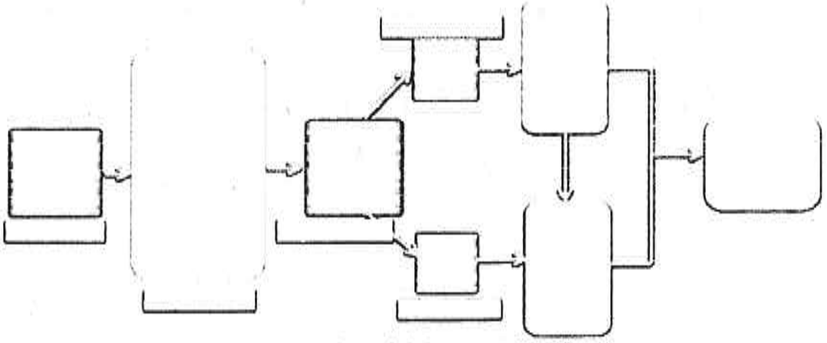


Fig.1 Proposed model for text classification

Figure 1 shows that proposed architecture of research work.In this work, we are using movie review Sentiment Polarity Dataset version 2.0" (http://www.cs.cornell.edu/People/pabo/movie-review-data)[9] dataset. In first step. applied different preprocessing techniques like stemmer. tokenizer, stopwords remover and pruning on the movie review data se to remove the noise and inconsistent data and prepared the smooth dataset. Now. We have divided the dataset into training and testing where training dataset is used for trained the classifier and testing data is used for test the trained classifier. Finally calculated the various performances of classifiers like accuracy, precision, recall and F-measures.

IV METHOD AND MATERIAL

(a) **Dataset-** The dataset consists of 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal at http://reviews.imdb.com/Reviews and is known as "Sentiment Polarity Dataset version 2.0"(http://www.cs.cornell.edu/People/pabo/movie-review-data)[9]. The reviews are equally partitioned into a positive set and a negative set (1000+1000).

(b) **WEKA Tool-** WEKA stands for Waikato Environment Knowledge Analysis(http://www.cs.waikato, ac.nz/~ml/weka/) [9] , it is a collection of various data mining algorithms and tools for in depth analysis.The programming language of WEKA is Java and its distribution is based on GNU (General Public License). There are mainly three uses of WEKA. First the analysis of data mining algorithm; second for generation of model; and last for comparison of various data mining algorithm in order to choose best as predictor.

First thing is to import the dataset from database to weka. Text data is import using TextDirectoryLoader component. To perform the preprocessing in WEKA, we used the StringToWordVector filter from the package weka.filters.unsupervised.attribute. This filter allows to configure the different stages of the term extraction. Configure the tokenizer (term separators), specify a stop-words list and choose a stemmer.

V RESULT AND DISCUSSION

In this experiment first we have applied the StringToWordVector to calculate attribute and we got 47163 in the dataset. After that we apply stremmer, StopWords and Tokenizer for preprocessing. We have used WEKA data mining tool [10] for analysis of Sentiment Polarity Dataset version 2.0" (http://www.cs.cornell.edu/People/pabo/movie-review-data)[9].WEKA is an open source software tool which contains various classification techniques used in this research work. This research work have used decision tree techniques like Naïve Bayes, SMO, Bayes Net, Random Forest, Random Tree and C4.5/J48 for analysis and classification of movie review with 70-30%, 75-25% and 80%-20% training-testing data partition. The accuracy of Naïve Bayes, SMO, Bayes Net, Random Forest, Random Tree and C4.5/J48 as shown in table 1 where Naïve Bayes gives better classificationaccuracy as 79.725%. To achieve the better classification accuracy, we have ensemble the individuals trained classifiers and achieved the best accuracy 81.50% of accuracy with proposed ensemble of Naïve Bayes, BayesNet and Random Forest as shown in table 1. Fig. 5 and Fig. 6 show that accuracy graph with 70-30% and 80-20% training and testing partition of ensemble models. We have also calculated Precision, Recall and F-Measure for the best ensemble model.In case of 70-30% training-testing partition the average Precision , recall and F-measure are 0.815, 0.185 and 0.185 respectively while in case of 80-20% training testing partition , the average Precision, recall and F-measure are 0.816, 0.185 and 0.185. Finally we concluded the our proposed ensemble model gives better perforce of movie review for users.

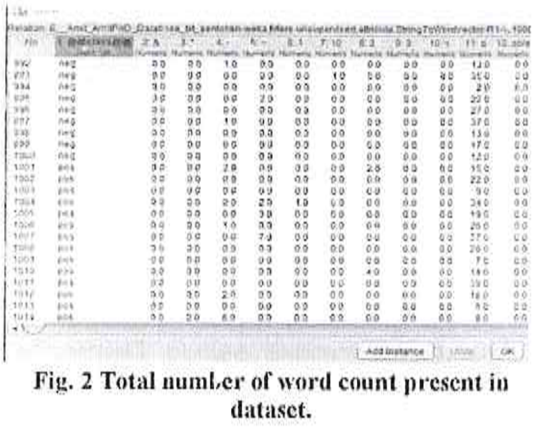


Fig. 2 Total numl.er of word count present in dataset.

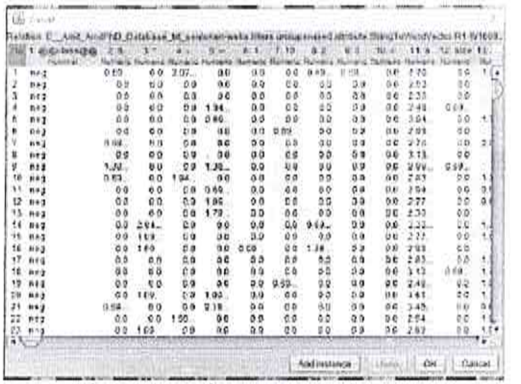


Fig. 3 Term frequency in the dataset.

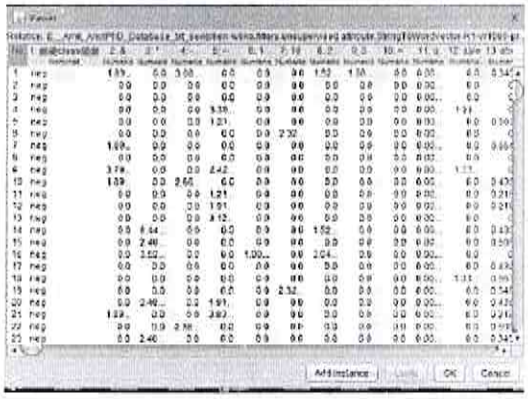


Fig. 4 TF-Idf of the dataset.

Table 1
Accuracy of individuals and ensemble models (in percentage).

Sno.	Model name	split		
		70-30 %	75-25 %	80-20 %
1	J48	65.67	65.00	64.25
2	RF	78.83	76.40	75.50
3	RT	56.50	59.00	57.75
4	SMO	76.00	76.20	75.50
5	NAIVABAYSE	78.67	78.00	79.25
6	BAYSENET	77.17	78.60	77.75
7	NB+BN	81.00	79.80	80.50
8	NB+RF	79.17	77.80	79.50
9	BN+RF	78.83	78.80	77.75
10	NB+BN+RF	81.50	80.60	81.50

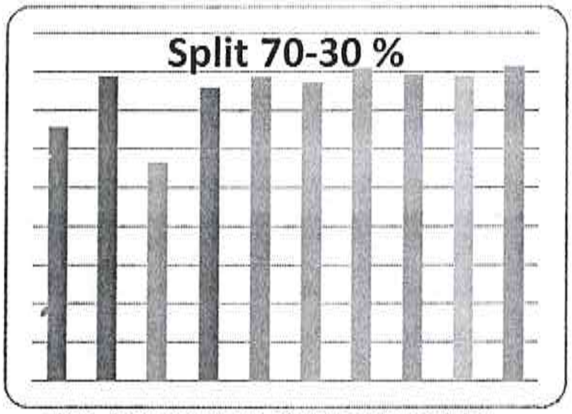


Fig 5. Accuracy table of Ensemble model at 70-30%.

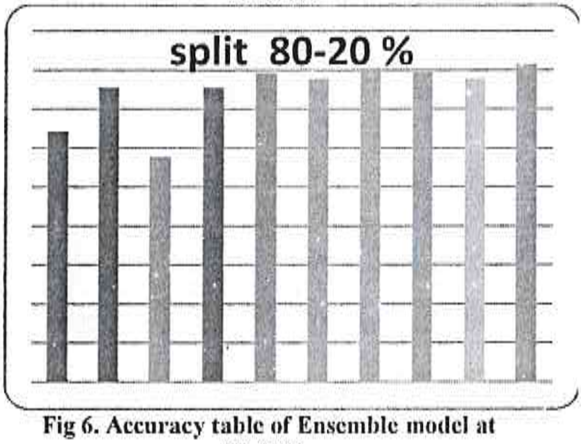


Fig 6. Accuracy table of Ensemble model at 80-20%.

VI CONCLUSION

Text recognition is the very important tasks to get conclude opinion of document, social media post, any script etc. Classification techniques play major role to identify and categorize the sentiments about document, social media post, any script etc. In this research work, proposed ensemble models likeensemble of Naive Bayes, Bayes Net and Random Forest which gives better results compares to individuals and other ensemble model. An optimization of features play important role to develop computationally efficient model. The proposed ensemble model gives satisfactory results with few numbers of features and recommended as classifier for classification of sentiments analysis (positive or negative).

REFERENCES

[1] Sharma S. and Kaur G (2016) . Review Paper On Sentiment Classification Of Movies Review , International Journal of Engineering Applied Sciences and Technology , 2 (1) , PP 67 – 72.

[2] Kumar N., Mitra S, Bhattacharjee M and Mandal L. (2018), Comparison of Different Classification Techniques Using Different Datasets , Proceedings of International Ethical Hacking Conference, PP 261-272.

[3] PalakBaid P, Gupta A. and Chaplot N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques , International Journal of Computer Applications, 179 (7), PP 45-49.

[4] Bilal M., Israr H., Shahid M. and Khan A. (2015) , Sentiment classification of Roman-Urdu opinions using Nai"ve Bayesian, Decision Tree and KNN classification techniques , Journal of King Saud University Computer and Information Sciences , 28 , PP 330-344.

[5] Mishra M. and Jha C. K. (2012), Classification of Opinion Mining Techniques , International Journal of Computer Applications , 56 (13), PP 1-6.

[6] Basari A.S.H., Hussin B., Ananta I.G.P and Zeniarja J. (2013) , Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization , Elsevier , Procedia Engineering 53 , PP 453 – 462.

[7] Kumar K. M. A., Rajasimha N., Reddy M. , Rajanarayana A. and Nadgir K. (2015), Analysis of Users' Sentiments from KannadaWeb Documents , Elsevier, ScienceDirect , Procedia Computer Science 54 , PP 247 – 256.

[8] Liaoa S. ,Wangb J. , Yua R. , Satob K. and Chengb Z. (2017) , CNN for situations understanding based on sentiment analysis of twitter data , Elsevier , Procedia Engineering , 111 ,PP 376–381.

[9] www.cs.cornell.edu/people/pabo/movie-review-data/

[10] http://www.cs.waikato.ac.nz/ml/weka/

[11] Shahana P.H, Bini Omman B. (2015). Evaluation of Features on Sentimental Analysis. Elsevier, Science Direct, Procedia Computer Science 46, PP 1585 – 1592.

Energy Efficient and Secure Offloading using Mobile Cloud Computing

Aradhana[1], Samarendra Mohan Ghosh[2], Praveen Shrivastav[3]
[1, 3]Research Scholar, Dept. of CSE, Dr. C. V. Raman University, Bilaspur (C.G.) India.
[2]Professor, Dept of CSE, Dr. C. V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

Advancements in mobile computing technology have changed the user preferences. Smart mobile devices have limited memory capacity, CPU speed and battery back-up time. The mobile cloud computing provides computational offloading to migrate the intensive task from smart mobile devices to cloud. Task offloading is vulnerable to certain risks. To enhance the security on task offloading, we propose a hybrid approach that renders the request generated by mobile application and change the structure of original code of data using code mixing and dead code insertion method. The approach makes the original code harder, unreadable and difficult to reverse-engineer. Based on our analysis, the obfuscation is able to increase the complexity of the code, confidentiality and integrity during computational offloading on clouds via mobiles agent. As an outcome the proposed frame work will be able to save battery consumption and noticeable amount of execution time in a secure channel.

Keywords: Code Obfuscation, Encryption, Computation Offloading, Reverse Engineering, Mobile Agent

I INTRODUCTION

Mobile cloud computing migrates task of the process to other server that executes some events on the behalf of user and or applications .Protecting intellectual property of any code is must and is not easily achievable; unless the understanding of the intellect present in the system is unidentified. One way of securing the intellectual property is by obfuscating the system. Many types of obfuscation can be done while the functionality of the system remains same. The obfuscation provides higher level of security that makes reverse-engineering a program more difficult and economically unfeasible [1]. Other advantages of obfuscation method on distributed environment include helping to protect unauthorized access, code optimizing, hiding vulnerabilities and shrinking the size of the programme code.

II OBFUSCATION TECHNIQUES

Here are some techniques for java byte code obfuscation techniques. They can help to create a complex defence against reverse engineering and code tampering and renaming. This alters the name of class, methods and local and global variables, identifiers to make the decompiled source much harder for a human to understand.

(a) **Control Flow Obfuscation**: This method changes the sequence of execution of code and flow of data.

(b) **String Encryption**: It transforms strings using encryption and only calls their original value when required.

(c) **Structural obfuscation**: This approach convert common instructions to other form of construct potential for confusing the decompilers.

(d) **Dummy Code Insertion**: This method inserts null code inside the source code. It breaks decompilers to makes reverse-engineered code harder to analyze and extract the original form of code.

(e) **Unused Code and Metadata Removal**: This technique reduces the unreachable code and meta data of the source code for preventing the information available to an attacker.

(f) **Binary Linking/Merging**: It combines many library file and executable lie into one or more than file.

(g) **Code Mixing**:-In this approach we scrambled the code to harden the reverse engineering reverse-engineered code harder to analyze and extract the original form of code.

III PROBLEM STATEMENT

Security, privacy, integrity and trust issues exist since the evolution of mobile cloud computing. It is a big issue of trust between mobile client/user and cloud provider. The cloud provider ensures only security regarding user authentication and data security and the environment he/she is running is not malicious but its monitoring is not in clients control. Hence the necessity for developing trust models/protocols comes in consideration [2]. Data are placed in the form of various packets on different locations on cloud. This causes the security breach thus intrusions are easily affecting the environment. It is necessary to set a security protocol on the data packet. [3]. cloud providers do not provide facility of monitoring security, location monitoring, authenticity and integrity of hardware, software and data. It requires a security model to monitors the running environment [4]. In clouds user cannot identify the location of data so the issued that who controls the integrity of data on virtual machine and difficult to maintain the consistency of security and ensure audit ability of records [5]. This agent is worked like security agent by implementing trusted computing infrastructure through authenticating hardware/software integrity. [6] Author has analysed some security threats, risks and vulnerabilities associated with cloud computing. Providers do not provide facility of monitoring security, location monitoring, and authenticity.

IV BUILDING BLOCKS OF THE PROPOSED FRAMEWORK

We illustrate the main building blocks of the proposed framework using a schematic model consists of major components of each building block in Figure 1. We have implemented a mobile agent platform to provide an agent API with the proposed security mechanism for code migration and communication in the form of JAVA packages to share workload management and reduce the Battery consumption of mobile devices.

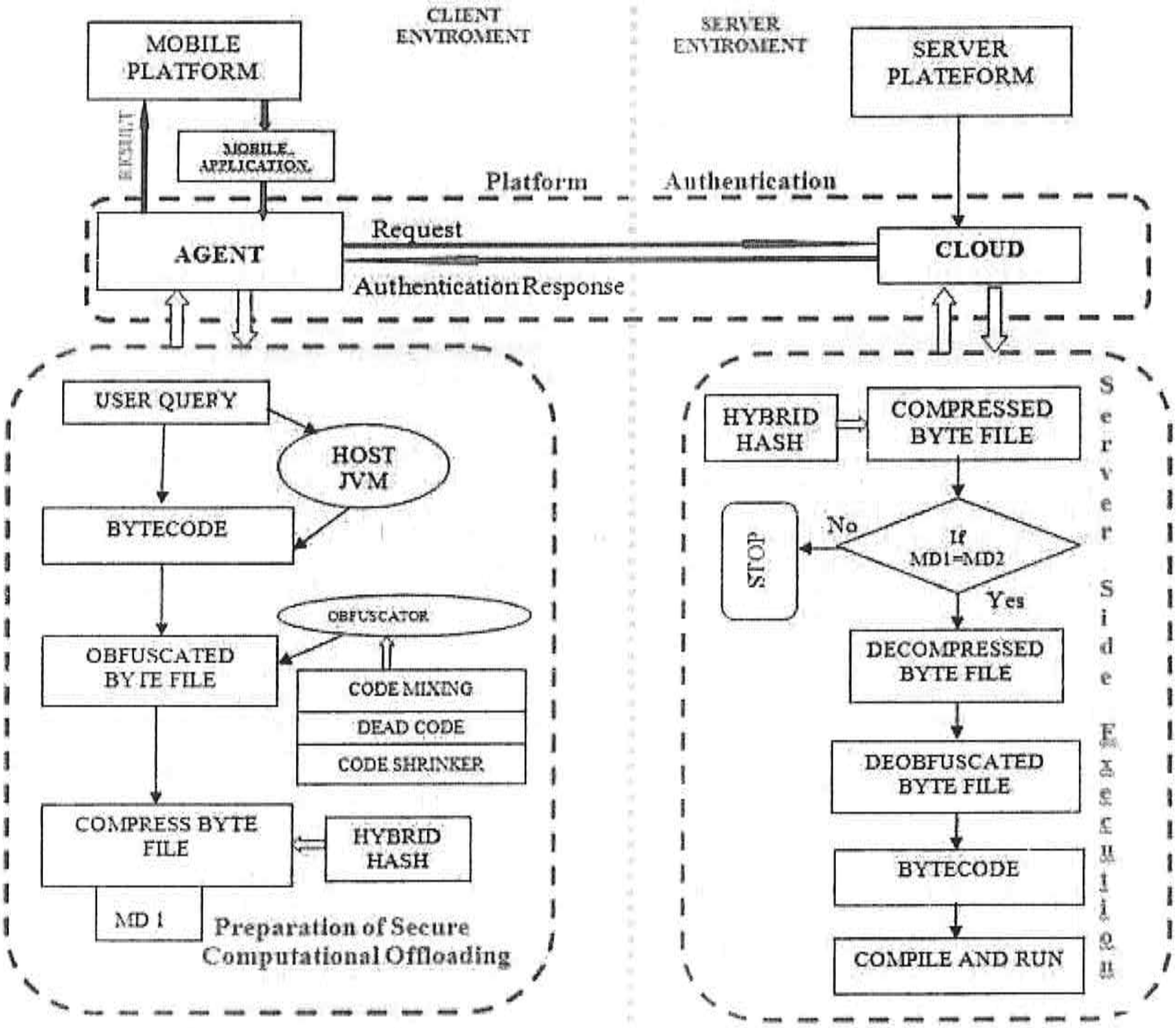


Fig. 1 : The systematic work flow of proposed Framework

V METHODOLOGY

(a) The proposed framework can be partitioned in to three parts:-

(i) **Platform Authentication**: - We have implemented an agent to protect the Client and Server Environment using JSP. The agent should have dedicated lines for individual clients. It only respond to the authenticate clients for this we apply set of protocols to give the permission for platform. The agent system often uses cookies to maintain the authentication states between agents to cloud. Every subsequent request will generate a cookie that will automatically be allowed by internet based application on agent platform. The attackers easily identify the user details using cookies. Cross site scripting (XXS) is a common attack technique that theft cookies related to subsequent request from mobile agent or web browser's databases. We proposed a new method for cookie rewriting that encrypts the cookies and protect from Cross Site Scripting attack. This method is implemented for mobile agents it automatically rewrites the cookies therefore it protects against XXS attack.

(ii) **Code Integrity and Confidentially**:-To protect the bytecode which is off-loadable using hybrid approach of code mixing and dead code insertion techniques with cryptographic advance hash function to make our code more robust and complex and we have also applied the lossless compression technique to reduce the size of code so that it will travel faster on network.

(iii) **Sever Side Execution for Computation offloading**:- Secured code of user query will run on server environment and server will be responsible for deobfuscation process to find the original code to run and sends back the results on client environments.

VI METHOD DESIGN

In this section, we describe the methods used to enhance the security performance of this framework. To increase the security we applied the hybrid approach of obfuscation method using code mixing and dead code insertion method to make it more confidential and for the integrity security we apply cryptographic hash function that generates the message digest which detect the threat present in the transferred code.

(a) **Evolution Matrix-** In this work we have implemented Advance Cryptographic hash Algorithm with hybrid approach on obfuscation methods for code integrity and confidentially for development of secure environment with virtualization technique, malware detection and informal behavioural of monitoring. This approach makes harder for reverse engineering the code. The Framework may affect the size of program and cost of execution time.

(i) **Code Obfuscation**: - Byte code obfuscation is method of modifying the byte code or instance of executable file so that it become harder to read and understand for an attackers but it remains fully function. The methods are Renaming Control, Flow Obfuscation, String Encryption, Structural obfuscation, Dummy Code Insertion, Unused Code and Metadata, Removal Binary Linking/Merging, Code Mixing, Anti-Tamper, Class file Encryption. We apply code mixing and dead code insertion .We identified that code mixing and dead code insertion methods are the best obfuscation for byte code.

(ii) **Code Mixing**: In this approach we scrambled the code to harden the reverse engineering. In this approach we partition the code in to equal chunks and randomize the part of code with a complex manner to harden the identification of the original code.

The Pseudo code to for Code Mixing Algorithm

```
• Original bytecode are divide into equal chunks of SIZE=COUNT
• Chunks array are chunks[i ] =1 to COUNT
• IF the code division original bytecode % SIZE ! = 0 then
• Drop the last chunk i.e. SIZE=SIZE-1
• End IF
• Circularly rotate the divided parts up to last chunks
• chunk[i]=chunk[i+2]
• End
```

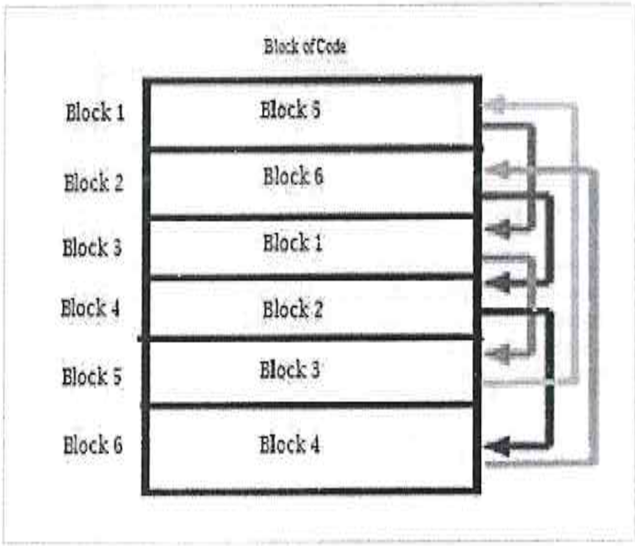


Fig. 2: Graphical Representation of Code Mixing

(iii) **Dead Code Insertion**: This method inserts null code inside the source code. It breaks decompilers to makes reverse-engineered code harder to analyze and extract the original form of code.

(iv) **The Pseudo code for the Dead code insertion:**

```
• Let us assume the L and U are Lower Bound and Upper Bound L=1 and U=SIZE
  MID = (L+U) / 2
• Insert the dead code at MID Position
  Chunks [MID] = Dead Code
• Adjust the Upper Bound of chunk array
  U=MID-1
• Inset the dead code on First partition of the code
  MID = (L+U) / 2
• Adjust the Lower Bound of chunk array
  L= MID+1
• Inset the dead code on Second partition of the code
  MID = (L+U) / 2
• End
```

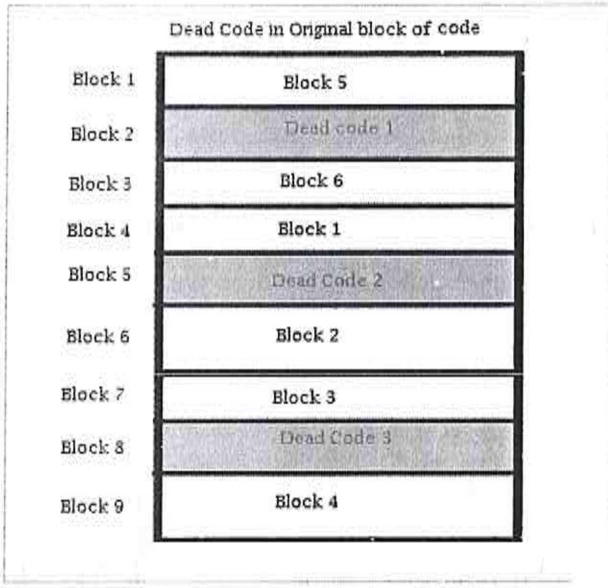


Fig..3: Graphical Representation of Dead Code Insertion

(v) **Lossless Compression**: Run Length Encoding is a loss less compression technique which runs on the basis of the occurrence of data and replace it with single value and count[9].

Algorithm for RLE

```
Step 1   set the count value to zero
         Loop: count = 0
        Step 2 REPEAT the Steps and get next
symbol
Step 3 Increment the count value UNTIL
(symbol unequal to next one)
        count = count + 1
         Step 4 IF count > 1
Step 5 output count
Step 6 REPEAT FROM STEP 1 GOTO Loop
```

(vi) **Cryptographic hash SHA 3**: cryptographic hash function generates the message digest in which data is first absorbed into sponge and then result is squeezed. In the absorbing phase, message blocks are XORed into a subset of the state, which is then transformed as a whole using a permutation function *f*. In the "squeeze" phase, output blocks are read from the same subset of the state, alternated with the state transformation function *f*. The size of the part of the state that is written and read is called the "rate" (denoted *r*), and the size of the part that is untouched by input/output is called the "capacity" (denoted *c*). The capacity determines the security of the scheme. The maximum security level is half the capacity. Given an input bit string *N*, a padding function pad, a permutation function *f* that operates on bit blocks of width *b*, a rate *r* and an output length *d*, we have capacity $c = b - r$ and the sponge construction Z = sponge[*f*,pad,*r*] (*N*,*d*), yielding a bit string *Z* of length *d*, works as follows[10]:

- pad the input *N* using the pad function, yielding a padded bit string *P* with a length divisible by *r* (such that n = len $(P)/r$ is integer)
- break *P* into *n* consecutive *r*-bit pieces P_0, ..., P_{n-1}
- initialize the state *S* to a string of *b* zero bits
- absorb the input into the state: for each block P_i:
- extend P_i at the end by a string of *c* zero bits, yielding one of length *b*
- XOR that with *S*
- apply the block permutation *f* to the result, yielding a new state *S*
- initialize *Z* to be the empty string
- while the length of *Z* is less than *d*:
- append the first *r* bits of *S* to *Z*
- if *Z* is still less than *d* bits long, apply *f* to *S*, yielding a new state *S*
- truncate *Z* to *d* bits

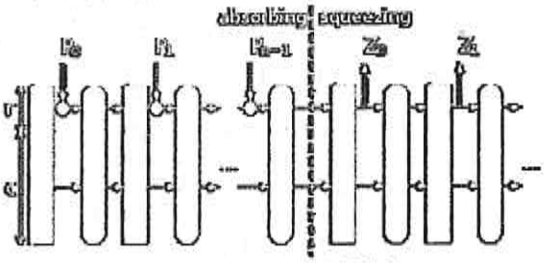


Fig..4: Function of SHA 3

(b) **Data Design-**In this section; we describe the methods used to evaluate the performance of this framework. We also identify the performance evaluation metrics that are identified for evaluation of the framework.

(i) **Evaluation Metrics-**We identify application execution time and consumed energy as two metrics of evaluation that are presented in Table 1 and explained as follows. These two metrics help us to obtain our aim and objectives in this study. Consumed energy and execution time are the most important established metrics to evaluate the lightweight properties of a typical MCC framework.

- **Battery Consumption**: Total battery consumption will be calculated using storing battery power values on session variables. Find the difference between Current state of battery level with last state of battery level
- **Execution Time**: Execution time is the amount of time taken to complete the entire life cycle of one prototype MA for one workload which is measured and stated in millisecond (ms). Different methods of generating execution time exist including manual and automatic data collection [8]. Similar to the energy, in order to avoid man made mistake and to obtain accuracy.

$$T_{total} = T_{mobile} + T_{rt} + T_{cloud}$$

Where T_{mobile}: - The time taken in local Machine, T_{rt}:-The time taken in delay (Round Trip) in network, T_{cloud}:-The times measure the computer latency of cloud computation.

Table 1
Unit and Symbols of Execution Time and Consumed Energy

Metric	Definition	Unit
Execution Time	Total Time takes to complete one task	nano Sec.
Consumed Energy	Total amount of energy consumed on the device to complete one task	Joule

VII RESULT AND DISSCUSSION

This section discusses experimentation findings of evaluating proposed by employing the prototype application. It presents analysis of Total Execution Time and Battery Consumption Cost of the mobile device for different web based application from the perspective of local and remote execution via Agent. In this experiment, we compare the finding values of Total Execution Time and Battery Consumption Cost with three different scenarios within same framework 1) Application runs under local machine, 2) Application runs on cloud without security parameters, 3) Application runs on cloud with three levels of security parameters. Results of performance evaluation generated via offload the five different web based applications are presented in this section in two parts. In the first part, data related to Execution Time Analysis is presented followed by Battery Consumed Energy in part .Experimental finding values are shown in Table 2 and Table 3 and Comparison of results are shown in Figure 4 and 5.

Table 2
Execution Time Analysis of Web Based Application

Total Execution Time in nano-second				
No. of Work load	Type of Mobile Application	Execution in Local Machine	Normal Offloading Execution	Secure Offloading Execution on Cloud
1	Google Home Page	125	180	190
2	Facebook Signup	171	202	214
3	Google Search	190	287	294
4	Song Online	310	402	418
5	Navigation	100	265	280

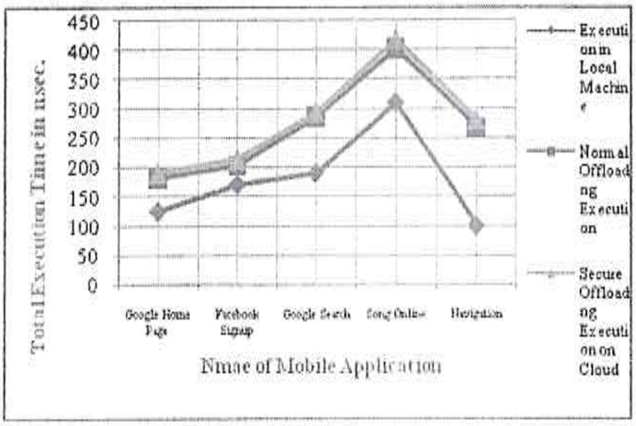


Fig. 4 Comparison of Total Execution Time

Table 3
Battery Consumption Analysis of Web Based Application

Total Battery Consumed in Joule				
No. o: Work load	Type of Mobile Application	Execution in Local Machine	Normal Offloading Execution	Secure Offloading Execution on Cloud
1	Google Home Page	5	4.1	4.4
2	Facebook Signup	6.12	5.2	5.7
3	Google Search	5.99	3.72	3.9
4	Song Online	7.33	5.91	6.1
5	Navigation	18.2	16.77	16.9

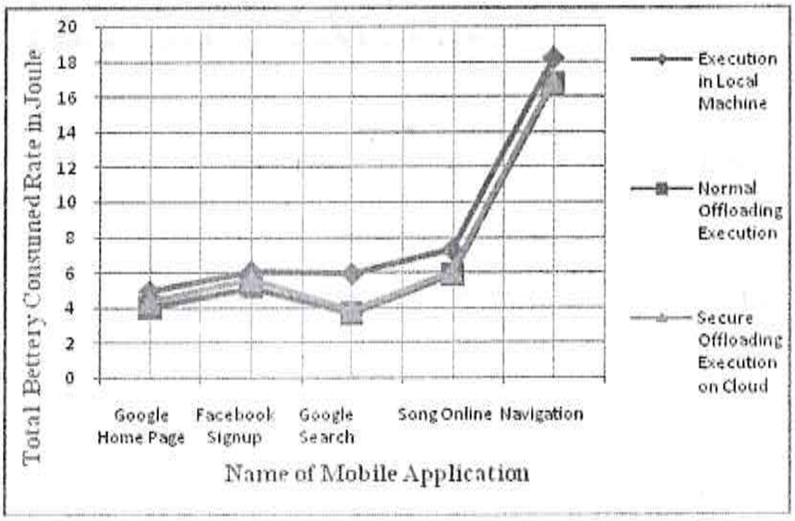


Fig. 4 Comparison of Total Battery Consumption

VIII CONCLUSION

In this work we have implemented energy efficient and secure framework for MCC so that user can trust on mobile cloud computing and enjoy the outcome without any worry. Here we have discussed the different types of obfuscation method and selected two methods to enhance security and privacy requirement, threats, concerns issues, risk associated to cloud and provide multiple securities in framework which fulfil complete cloud security requirements. As an experimental outcome, the secure framework affects the size of task, execution time, and reduces battery consumption. Finally it improves the overall function of the computation offloading using mobile.

REFERENCES

[1] Vasudevan M. (2001) ,"An Architecture of Class Loader System in Java Bytecode Obfuscatio", Indian Journal of Science and Technology. Vol 8(S2),PP. 291–294.

[2] Cachin, C., Keidar I., and Shraer.(2009) A. Trusting the cloud. ACM SIGACT News.PP. 81-86.

[3] Popovic K. & Hocenski, Z.(2010) "Cloud computing security issues and challenges", of the 33rd International Convention, PP .344-349.

[4] Grobauer, B., Walloschek, T., Stocker, E.(2011) "Understanding Cloud Computing Vulnerabilities", Security & Privacy, IEEE, Vol. 9, No. 2, PP.50-57.

Analysis of Quality Improvement through ICT Based Online Admission Process in Open and Distance Mode of Education

Arvind Tiwari[1], Vaibhav Sharma[2]

[1,2]Dr. C. V. Raman University, Bilaspur (Chhatisgarh) India.

ABSTRACT

The Open and Distance Learning (ODL) system has shown a tremendous growth during the past few decades due to its unique feature of being a user-friendly system. In this system, the students are free to learn at their own pace and convenience while being away from the institution. This uniqueness and the ease of gaining knowledge have a pivotal role to play in facilitating today's emerging knowledge society. Information and communication technology (ICT) has significantly affected the process of admission in higher educational institutions. ICT also improves the interaction between the students and the institution as well as academician too. ICT based tools and technology not only useful for the regular mode of education it also plays important role in open and distance learning mode and improves the facilities of the institution for the learners. It has provided an edge to ways and means of traditional form of admission process and has given new dimensions to online admission. The article aims to measure the impact of ICT on online admission process in higher educational institutions in the state of Chhattisgarh. The framework of the study aims to find out the inter-effect of ICT among the academic institutions for the online admission process in the state of Chhattisgarh. The outcome suggests ICT adoption in the process of admission among academic institutions in the state of Chhattisgarh for the open and distance education.

Keywords: ICT, Online admission, satisfaction, higher educational institutions.

I INTRODUCTION

Open and Distance Learning (ODL) programmes are very important and significant part of modern days of higher education and it provides a medium to acquire knowledge with the degree for society. It is more flexible and cost effective education system which provides the education for every age group of students, working professionals and traditional Indian women's who works as housewives as well as also works in many different sectors. New age technologies can help to improve the quality, reliability of education and it also distribute education from the best sources to all the people who are in need of education and it can also be used to provide many sources so the education for our learners can reach in remote areas. The evolution of information and communication technology (ICT) based education has initiated a new revolution in open and distance learning mode of education environment that changed the traditional methods and procedures of teaching and learning. ICT is group of technologies by which various support services shall be provided at different phases of student learning life cycle in distance learning. The various phases of open and distance learning programmes are: the admission phase (pre admission counselling, programme details, online admission, fee structure, online fee payment, online solution of payment issues, admission procedure and registration & re-registration), the learning phase (learning schedule, programme delivery(lectures through video conferencing, webinars, audio & video programmes, multimedia presentations and case studies), the evaluation phase (examination schedule, internal &external assessment, examinations, check your progress report, valuation, revaluation and result declaration) and the certification phase (marks/grades updates, E-certification, certificate printing & issuing and convocation schedule).

ICT not just playing its role in academic fields it also involve in administration area and its role start from the beginning when a learner wants to take admission in any programmes offered by the university. Taking admission is the very first process for the learner to join the open and distance learning institution for any particular programme and learners from different areas are not able to reach the institution/university or take the part in admission process because lack of facilities, resources and their busy schedule for this issues ICT plays a part through online pre counselling and admission process for learner. Process of online admission is very convenient and easy so the learners from any areas can easily enrolled their selves for any particular programme offered by the university. Here in this paper, the role of ICT in online admission and is impact has been analyzed and defined which describe the importance of ICT and online admission. This paper concludes the effects of ICT based online admission pattern in every dimensions of it in the institution which are formed in Chhattisgarh state. The outcome of this includes nature of work; ICT based online admission pattern and the impact of online admission in various fields. The outcomes also describe the advantages of online admission process, limitations and its future in open and distance learning mode of education:

II PROCESS of ONLINE ADMISSION and ICT:

In the modern age of technology in which all kind of institutions are getting hi-tech, computerized and available online with their all kind of services and so the educational institution too. The various facilities and services like fee payment, e-content, e-library has been provided by the educational institution through online mode and now online admission is also one of the facilities by the institution for the student. The online admission process replaces the traditional admission process i.e. offline mode very soon and many institution already started the process of online admission in our country. The online admission process is the process which automates the admission procedure for the regular mode or distance learning mode of education of schools, colleges and universities. The whole procedure based on an internet based application which can be accessed from anywhere, anytime basis and provides the facility of online registration for the students without any use of paper material. In open and distance learning the learners are not able to come to the university campus directly because of various reasons like institution of open and distance learning is not in the reach of the learner so he/she can participate in offline admission process and many more of their personal reasons but they still wants to get enrolled theirs selves for further education and here the facility of online admission works. It provides a convenient way through ICT based application for the learners to get enrolled their selves for any particular programme from open and distance learning mode of university. The overall procedure of online admission is also known as E- admission. Following are the main steps of online admission process:

(a) Pre Admission Counselling: In these process counsellors provides the information about open and distance learning mode and its overall concept with its examination and study structure. Counsellors clear all kinds of doubts and myths of the learners and describe them the benefits of open and distance learning. Pre admission counselling process involves following steps:

(i) At first counsellors defines the aim and objective of the open and distance learning as well as related programme.

(ii) Defines the scheme structure and concept of open and distance learning.

(iii) Introduce programmes offered by the Institution for the learners of different streams and background so they can select the right one for their future studies.

(iv) Clarify the patterns of admission, course work, assignment and process of examination.

(v) In this process counsellors explain the structure of programmes and its credit structure to the learners so they can understands the learning process of their programme.

(vi) In the process of pre admission counselling counsellors also provides the acknowledgement of the goals, outcome and the purpose of the programme so the learner from related fields can decide that in which kind of programme he/she wants to get enrolled.

(vii) At the end of the process of pre admission counselling e-programme guide has been provided to the learners by the counsellors.

(viii) Submission of Online Admission Form: If the learners are satisfied with the nature of programme from their fields with their interest than he/she will go for the online admission process after the process of pre admission counselling. The process of submission an application form is the second step of online admission process in which learners have to submit their application form for the particular programme in which they want to get enrolled through online admission process. For the process of submitting online admission form university will provide online application form and its submission link into their official web portal. Learners just have to open the university's web portal and have to click the link of online admission form and then they have to full the application form with required details of the learner along with his/her educational qualification as per rules and regulation of the university and then have to submit the form for further process.

(ix) Online Fee Payments and its Acknowledgement: After the submission process of the form the next process of the online admission process will occur that is the fee payment and its acknowledgement for any particular programme. It is the most convenient way to pay the fee of any particular programme to the university. After the process of form submission another page automatically pops up in which overall details related to the fee structure of selected programme has been defined along with its duration of fee payments so the learner can understand the fee payment procedure. In the next step of this process learners will enter the amount of their programme as per fee structure and then they have to click the payment option where multiple choices are there like debit card, credit card, internet banking and other payment option from which the learners can select the suitable one to pay the asking fee amount just after this an acknowledgement generated automatically for the learners to their contact detail and email for the confirmation of fee payment along with the printable receipt of the fee.

(x) Approval Process and the Generation of Enrolment Number: After the process of fee payment learners just have to wait for the approval of their submission from the university officials. In this process university verifies the details of the learner as per the rules and regulation of the university to get enrolled in a particular programme that the learner is eligible for the programme or not. After the verification and approval from the university officials the registration number along with enrolment number of the learner has been automatically generated and it will send to the learner in his/her contact detail as well as in their email with the information of their eligibility confirmation massage that the fulfil the eligibility criteria or not in the form of YES and NO.

(xi) Confirmation to the Learner: After the above all process the online admission in ODL has comes to its end and the final message of confirmation of learners successfully registration along with their enrolment number has been sent into their mobile number as well as into their email id. This provides the acknowledgement to the learner that he/she has been enrolled in the programme they have chosen to learn.

Above all the process involves in the online admission process required ICT tools software and digital devices to complete the process. ICT based tools and device provides easy access of the computer, web world/Internet and related software so the process of online admission can become easy so it can be accessible for the learners who are not aware from the computational word. ICT is a potentially powerful tool for online admission process by it can improve the performance, accuracy and reliability of online admission process in ODL mode of education.

III ADVANTAGES AND LIMITATION OF ICT IN ONLINE ADMISSION PROCESS FOR ODL

Online admission process in open and distance learning provides the accessibility to the learners for the admission in selected programme from anywhere and from anyplace. It gives the command directly to the learners to get registered their selves in ODL mode of education through online admission. The implementation of ICT in online admission process will save the time, cost and extra expenses on papers. It also provides the easiest way to access the technology even by the learner who is not friendly with technology. The transparency of the online admission process over offline admission process is more desirable so the admission process for the learners is unbiased.

The effective integration of ICT tools into the ODL education system for online admission process is complex, lengthy and time consuming that involves not just technology but also pedagogy, institutional readiness and implementation of complete academic structure with full administrative command for the online admission process. It also noticeable that the terms and condition for the admission of any programme regarding to institutional guidelines needs to be implement for proper process so the reliability of this process will increase. The biggest problem in implementation of ICT tools is connectivity because internet is not available or having a poor connection in ruler areas so it is not possible to access all devices in online admission process for the learners of remote areas.

IV CONCLUSION

Information and communication technologies (ICT) are potentially powerful enabling tools for educational change and reform. It is true that ICT is playing a vital role not just in regular mode of education but as well as in open distance learning mode. ICT has enormous potential to help countries address issues of access to learning, quality of the teaching-learning process and management of education systems but at same time there are many issues and challenges that are to be addressed for smooth functioning of various online services that are to be implemented for the learners and other public. In this paper the role and importance of ICT implementation in online admission process in open and distance education has been described. The research paper concludes the benefits of online admission system in open and distance mode of education over the offline admission process. Lack of systematic approach in the implementation of ICTs in distance education is also a challenging task so here in this paper, limitation of ICT based tools for the online admission process has been described. The paper defines the value, importance, advantages, benefits, drawback and overall process of online admission in open and distance mode of education.

REFERENCES

[1] Calsoft Labs. "IT/ICT Adoption in Indian Higher Education". A Alten Group Company. 2012.

[2] Farrel, G. M. "The Development of Virtual Education: A Global Perspective". Vancouver, 1999.

[3] Government of India. "Higher Education in India: Twelfth Five Year Plan (2012–2017) and beyond". Planning Commission. New Delhi: Government of India, 2012.

[4] Government of India. "Reports of the Committees under National Mission on Education through ICT". Ministry of Human Resource Development. 2012. Retrieved from www.sakshat.ac.in.

[5] Gandhar H., and Saini V., "Integration of ICT in ODL System: Ongoing Projects and Challanges", IJMRME, Vol. 02, 2016.

A Study of Clustering Approaches and Validation Measures for Big Data Mining

Kamlesh Kumar Pandey[1], Diwakar Shukla[2]

[1]Research Scholar, Dept. of CSA, Dr. Hari Singh Gour Vishwavidyalaya, Sagar (M.P.) India.

[2]Prof & HOD, Dept. of CSA, Dr. Hari Singh Gour Vishwavidyalaya, Sagar (M.P.) India.

ABSTRACT

Present time natures of data are totally changed to big data with respect to volume, variety, and velocity. Clustering is one of the unsupervised approaches of data mining and data mining is one of the approaches for big data analysis and known as big data mining. Clustering is a very helpful technique under big data mining because it discovers distribution of patterns, hidden relations, self-noise and outlines management, class label predication and interesting correlations in large data sets and high dimensional data. Every traditional clustering algorithm works under specific criteria and these criteria define the cluster and validation measure validates this cluster as the requirements. From a theoretically, practically and the existing research perspective, this paper study seven clustering taxonomies such as partition, hierarchical, density, grid, model, fuzzy and graph based clustering taxonomy and their validation measures for the big data mining.

Keywords: Big Data, Big Data Mining, Clustering Taxonomy, Clustering Validation Measures, Cluster Validation Taxonomy

I INTRODUCTION

Nowadays, large scale data are generated from a different type of sources such as health, social network, government, cloud computing, e-marketing, internet of the thighs, financial, sensor network and so on. IDC predicts data volume reaches 44 Zetta-bytes (44 billion terabytes) per day till 2020, which is ten times double by 2003. In general, the big data refer to the large datasets that are collected from heterogeneous sources and these sources are continually growing. Due to the changing the nature data to big data some government establishes a big data department. In March 2012, the Obama Administration launched the Big Data Research and Development Initiative and July 2012, the Japan government established the Big Data development under the national technological strategy. The United Nations issued a report entitled "Big Data for Development: Opportunities and Challenges" which describes outlining the Big Data challenges and their dialogue (Oussous et al., 2018). In Big Data perspective, traditional data related techniques such as data management, process management, analysis technique have touched a bottleneck and cannot finish the data processing in fixed time with accuracy, gives the slow responsiveness, and problems with scalability because big data needs to high capacity for data storages, low-value density, complex dynamic relation between various data types, high processing speed, high scalability, availability and reliability (Weichen et al., 2016).

Clustering is one of the techniques for data analysis which is finding a similar relation or pattern for unlabeled objects. In big data mining, various applications such as social network analysis, customer segmentation, scientific data analysis, bioinformatics, and target marketing used to cluster analysis. The traditional clustering algorithms are not suitable under large-scale data because it takes high computation time. Consequently, computational efficiency is the biggest and most important challenge of large-scale data and need to how to improve the clustering algorithm. At the present time parallel and distributed computation are solutions to these types of the problem (Zhao et al., 2019). A good clustering algorithm in a big data environment need to cluster is more accurate and reliable. The clustering accuracy, reliability and algorithm efficiency is measured by cluster validation. This paper is mainly focused on the clustering algorithm and validation approach in the big data mining environment through five sections.

II BACKGROUND

(a) Big Data

Laney (2011) described big data through three dimensions such as volume, velocity, and variety.This dimension defines a common framework of big data.

The volume describes the size of the data, variety describes the heterogeneous sources and types of data and velocity describes the speed of data generation, analysis, and processed. Social Media is the best example of these 3V's because social media is generati..g high volume data in the form of high variety in the form of high velocity. Gartner et al., 2012, summarized these dimensions of big data as "Big data have high volume, high velocity, and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." After that various, research origination gives some support dimension for the big data framework. IBM describes veracity as a fourth dimension, which is related to unreliability and uncertainty latent in data origin for improving data accuracy and quality with the trustworthiness of data. After that SAS added variability dimensions to big data, which represents the variation in data flow rates. In addition, the variation of data flow rates is related to increasing the variety and velocity. Oracle introduced six supportable dimensions as value, which hold to hidden value or attribute during the big data mining. The last dimension is related to the visualization of all dimensions according to the user needs. Visualization is visualized the data mine and analysis results according to user expectation, such as the graph, table, or another format (Gandomi et al., 2015, Lee et al., 2017, Sivarajah et al., 2017). These all dimensions are known as 7 V's of big data. Volume, Variety, and Velocity are known as basic dimensions of big data creation and Veracity, Variability, Value, and Visualization are known as supportable dimensions of big data accession. Figure 1 summarized all dimensions of the big data.

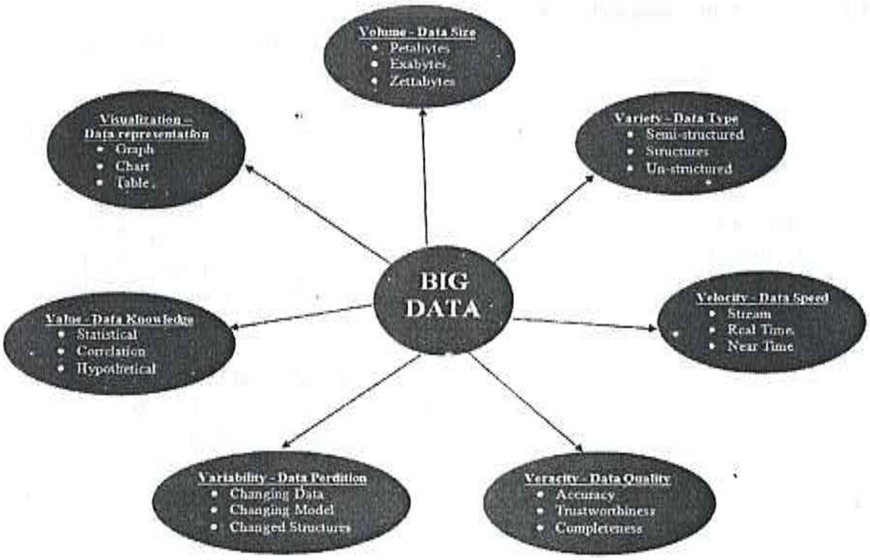


Fig.1 7V's of Big Data

(b) Big Data Mining

The big data mining approach is totally differs to the traditional data mining approach because traditional data mining algorithm based on centralized databases, but the big data mining algorithm based on distributed or multiple sources, traditional data mining algorithm can't be able to handle huge scale data sets, high dimensional, heterogeneous data format and source and scalability. In Big Data perspective, the data mining technique must be deal high volume, high variety and high velocity with scalability.

When big data mining algorithm applied in heterogeneous sources, the method of mining is divided into four groups as the heterogeneous source with pattern analysis, classification, clustering, and fusion (Wang et al., 2018). The purpose of big data mining is not only mining the interesting knowledge's and summarizing the data, but it is also mine the consistent patterns, systematic and complex relationships among the data through classification, clustering, association rule learning, regression and other data mining techniques. Big data mining used machine learning and statistical methods for evolution, extended to utilize traditional data mining techniques (Siddiqa et al., 2016).

Big data mining does not support to the relational model because the relational model handles only structured data, but the big data needs to handle structured, unstructured and semi-structured data with the high scale and distributed natures. At the present time, NoSQL databases are more popular for data storages because it stores the heterogeneous data format. In this database, the data are stored in the distributed file and graph format. In the distributed file format model, the data are stored in the form of a key/value pair and the graph-based model has data organized in the form of a vertex/edge pair. The Graph data model is useful for evaluating the various problems such as distances computing, finding relationships, community detection, determining connectivity and so on (Weichen et al., 2016). The distributed and parallel architecture is very helpful for achieving performance and accuracy reduces the computation time and scalability for the requirements of clustering in big data mining. The MapReduce programming model is one of the approaches for distributed and parallel computing in big data mining (Chong et al., 2015, Sardar et al., 2018).

III CLUSTERING TAXONOMY

Clustering is one of the approaches for analysis and discovering the complex relation, pattern, and data in the form of underlying groups for the unlabeled data objects. The data objects of each group share the same similarity and the other group data objects totally different from another group (Zhao et al., 2019). Data clustering is the most important technique and produces high-quality analytical results for data reduction. Data clustering also increases the efficiency and accuracy of data mining under unclassified problems (Chen et al., 2018). In Big Data perspective, the clustering algorithm must be deal high volume, high variety and high velocity with scalability.

Similarity and dissimilarity (distance) are two basic functions for cluster creation. Every clustering taxonomy is used to these functions for cluster construction on the bases of own cluster creation behaviours and natures.In nowadays various Similarity and dissimilarity measure are available for clustering under the Minkowski, L(1), L(2), Inner product, Shannon's entropy, Combination, Intersection and Fidelity family (Cha et al.,2007, Kocher et al., 2017, Manning et al., 2008).

The design of the clustering algorithm under the big data mining is fulfilled the volume, velocity, and variety related criteria. Fahad et al., 2014 and Pandove et al., 2015 describes Volume related criteria such as cluster is must be dealt huge size, high dim^nsional and noisy of the dataset, Variety related criteria such as cluster is must be recognized as dataset categorization and cluster shape, and Velocity related criteria define the complexity, scalability, and performance of the clustering algorithm during the execution of real dataset.

In general, Clustering algorithm divided into seven groups such as partition, hierarchical, density, grid, model, fuzzy and graph based clustering taxonomy based on their cluster creation, working process, behaviors and cluster nature. The basic concept of the theses clustering algorithm is described under section a to g (Chen et al., 2018, Fahad et al., 2014, Gan et al., 2007, Pandove et al., 2015, Shirkhorshidi et al., 2014)

(a) Partitioning based method

This clustering method creates a cluster on the bases of the center data point, then reason it's known as centroid-based clustering. This method partitions the dataset using a K number of user define the cluster and randomly assigned data object in each K cluster after that it finds the center point and assigns the objects to the nearest center of the cluster. K-Mean, K-Medoids, K-parameter, PAM, CLARA, and CLARANS are the most popular clustering algorithm under this clustering approach.

(b) Hierarchical based Clustering

This technique is also known as Connectivity-based clustering because the cluster is created by using tree (hierarchy of clusters) concept. Agglomerative and Divisive are two basic methods for cluster creation under this clustering taxonomy. The agglomerative method starts from the individual data cluster because each data object has own cluster at the beginning and after that, each cluster pair of data objects is moved up to the hierarchy form until the needed cluster is not found. Divisive starts from the top cluster because all data objects belong into one cluster at the beginning and after that, the cluster is divided into different cluster pairs move down the hierarchy until the needed cluster is not found. Here cluster is shown as dendrogram format. BIRCH, CURE, ROCK, Chameleon, ECHIDNA, WARDS, and SNN are the most popular clustering algorithm under this clustering approach.

(c) Density-based method

This clustering method is creating a cluster on the bases of data object density, connectivity and borderline. These clustering techniques provide the barriers against the noisy and determine the clusters to an arbitrary shape. To find out density it used to Mean-shift concept, firstly calculate the mean of current data point and figure out them and this iteration will be continued until the desired cluster is not founded. DBSCAN, OPTICS, DENCLUE, and GDBSCAN are the most popular clustering algorithm under this clustering approach.

(d) Model-basedclustering

This clustering method creates a cluster on the bases of some existing model such as mathematically model; statically model, probability model and other distribution model, and finding the best data object fit into the model. The model is depending on the existing model, then that reason it is known as distribution-based clustering. Here each model needs to minimize according to their distribution, but the use of multiple parameters it takes high time complexity. COBWEB, SLINK, SOM, ART, and EM are the most popular clustering algorithm under this clustering approach.

(e) Grid-based clustering

This clustering method used for utilizing the space for data sets in multidimensional data. Here each original data space is separated into grids or calls structure for defining the size of the cluster. Here data space is divided into many rectangular cells by using the hierarchical structure for parallel processing for fast processing time, and the data is organized within the different cell levels. This clustering algorithm generally used for statistical techniques. STING, CLIQUE, Wave Cluster, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, and STIRR arethe most popular clustering algorithm under this clustering approach.

(f) Fuzzy based clustering

This type of clustering is based on fuzzy or soft computing or hard computing for real data clustering then reason it's called soft computing based clustering. In the hard clustering, data object must belong to only one cluster, but the soft clustering data object belongs to multiple clusters on the bases of the membership function. Fuzzy clustering provides more insight and knowledge about the data objects to all clusters. FCM, FCS, and MM are the most popular clustering algorithm under this clustering approach.

(g) Graph based clustering

This clustering algorithm is realized on the graph, where the node refers to a data point and the edge is referred to as the relationship betweenall data points. During this cluster,the analysis graph must be in the form of a minimum spanning tree. This clustering algorithm is based on traditional and spectral graph theory. CLICK and MST is the most popular clustering algorithm under this clustering approach.

IV CLUSTER VALIDATION APPROACH

Clustering validation measures evaluate the goodness of clustering results and its validity for the success of the clustering algorithm. The clustering validation measures are categorized into External clustering validation and Internal clustering validation groups (Aggarwal et al., 2014). These groups are measures which clustering algorithm is suitable for volume, variety and velocity criteria.

(a) External clustering validation

External clustering validation measures evaluate "purity" of the clusters respect to given class labels. External validation measures helpful for finding the exact number of the cluster in the advance and given the framework for choosing an optimal clustering algorithm on the particular dataset (Liu et al., 2010). The overall this approach test the points of the data set are randomly organized as pre-specified structured or not and this analysis is done by using the Null Hypothesis (Halkidi et al., 2002). Table 1 shows some popular external clustering validation measures with their equation definition (Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Liu et al., 2010).

Entropy and purity measures are given the purity and accuracy of the class labels with respect to the cluster. F-measure is given the precision and recall value together. F-measure is useful for the information retrieval community. The Mutual Information (MI) measures depended to the random variable, this measure compares how much information is depended to the random variable and compared to another random variable.

The Variation of Information (VI) measures the quantity of information that is lost or grown in changing the form of the class set to the cluster. The Rand statistic, Jaccard coefficient, Fowlkes & Mallows index, and Hubert's statistics I and II evaluate the clustering quality of the pair of data point by using the agreements and/or disagreements in different partitions. The Minkowski score measures the difference between the clusters, which is obtained by the clustering algorithm and also given the disagreements of the data point in different partitions. The classification error is given a total misclassification rate in each class to a different cluster. It is very helpful for minimizing the total misclassification rate. The Van Dongen criterion is related to evaluating the graph clustering measures and it is given the measures of majority objects in each class and each cluster.

Table I shows external clustering validation measures where some mathematically formulation defines as Data set as D with n objects, assume that there is a partition A = {A1, • • • ,AK} of D.

Ai j = ni j/n,
Ai = ni•/n,
Aj = n• j/n.

(b) Internal clustering validation

Internal clustering validation measures evaluate the goodness of a clustering structure with respect to trustiness on information in the data inside of the cluster without external information. Internal validation measures are useful for finding the best clustering algorithms and the optimal number of the cluster without any other information (Liu et al., 2010). Overall, this approach evaluates the clustering algorithms and their resultsby using the quantities and features to the used dataset (Halkidi et al., 2002). Table 2 shows some popular internal clustering validation measures with their equation definition (Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Liu et al., 2010).

Table 2 shows Internal clustering validation measures where some mathematically formulation defines as D use as data set, n use as number of objects in D, c use as center of D, Ci use as the ith cluster,ni use as number of objects in Ci, ci use as center of Ci, A use as attributes number of D,k use as number of nearest neighbors, d(x, y) use as distance between x and y, NC use as number of clusters, qj use as number of Ci's jth object's nearest neighbors which are not in cluster Ci;

Compactness and separation are two basic measures for internal validation. Compactness is measured, how closely related the data objects in the cluster area by lower variance indicates and numerous measures on the base of distance, such as maximum, minimum or average for pairwise and center-based distance. Separation is measured, how distinct cluster is from other clusters for using compare cluster density, centers, and minimum distances. Such internal validate index as Davies–Bouldin (DB), Xie-Beni index (XB), and Silhouette index (S) consider both evaluation criteria compactness and separation and such as Root-mean-square standard deviation (RMSSTD), R-squared (RS), and Modified Hubert statistic (considers describe only one internal validation aspect.

The root-mean-square standard deviation (RMSSTD) validates the homogeneity of the shaped clusters by the square root pooling sample attributes. R-squared (RS) validates the degree of difference between the clusters of the sum of squares. Here, the sum of squares between clusters to validates the total sum of squares of the whole data set. The Modified Hubert statistic validates the difference between clusters by counting the disagreements of data point pairs in the partitions. The Calinski–Harabasz index (CH) validates the cluster validity by average between-and within-cluster on the basis of the sum of squares. Index I (I) validates the separation by the using maximum distance between cluster centers and given the compactness by the sum of distances between data points and their cluster center. Dunn's index (D) is given a minimum pairwise distance between different clusters data point for intercluster and the maximum diameter among different clusters data point for intracluster compactness. The Silhouette index (S) validates the clustering performance by using the pairwise difference of the cluster distances within and between the clusters. The Davies–Bouldin (DB) validity index is useful for cluster similarity obtained by averaging all the cluster similarities. The smaller index is indicating to the better clustering result and the max index is indicating the clusters are most distinct from. The Xie-Beni index (XB) validate the inter-cluster separation by the minimum square of the distance between cluster centers and also validate the inter compactness by the mean square of the distance between all data points and its cluster center. SD index (SD) validates average scattering and the total separation of clusters. It validates compactness by variables of cluster objects and also validates the separation, difference by distances between cluster centers.

Table 1
External Clustering Validation Measures
(Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Halkidi et al., 2002, Liu et al., 2010)

No	External Clustering Validation Measures	Definition
1.	Entropy (*E*)	$-\sum A_i\left(\sum_j \frac{A_{ij}}{A_i}\log\frac{A_{ij}}{A_i}\right)$
2.	Purity (*P*)	$\sum_i A_i\left(max_j \frac{A_{ij}}{A_i}\right)$
3.	F-measure (*F*)	$\sum_j A_i max_i\left[2\frac{\frac{A_{ij}}{A_i}\frac{A_{ij}}{A_i}}{\frac{A_{ij}}{A_i}+\frac{A_{ij}}{A_i}}\right]$
4.	Variation of Information (*VI*)	$\sum_i A_i log_{A_i} - A_j log_{A_j} - 2\sum_i\sum_j A_{ij}\log\frac{A_{ij}}{A_iA_j}$
5.	Mutual Information (*MI*)	$\sum_i\sum_j A_{ij}\log\frac{A_{ij}}{A_iA_j}$
6.	Rand statistic (*R*)	$\frac{\left[\binom{n}{2}-\Sigma_i\binom{n_{i.}}{2}-\Sigma_i\binom{n_{.j}}{2}-\Sigma_{ij}\binom{n_{ij}}{2}\right]}{\binom{n}{2}}$
7.	Jaccard coefficient (*J*)	$\frac{\Sigma_{ij}\binom{n_{ij}}{2}}{\Sigma_i\binom{n_{i.}}{2}+\Sigma_i\binom{n_{.j}}{2}-\Sigma_{ij}\binom{n_{ij}}{2}}$
8.	Fowlkes & Mallows index (*FM*)	$\frac{\Sigma_{ij}\binom{n_{ij}}{2}}{\sqrt{\Sigma_i\binom{n_{i.}}{2}+\Sigma_i\binom{n_{.j}}{2}}}$
9.	Hubert Γ statistic I (Γ)	$\frac{\left[\binom{n}{2}-\Sigma_{ij}\binom{n_{ij}}{2}-\Sigma_i\binom{n_i}{2}-\Sigma_j\binom{n_{.j}}{2}\right]}{\sqrt{\Sigma_i\binom{n_{i.}}{2}+\Sigma_j\binom{n_{.j}}{2}}\left[\binom{n}{2}-\Sigma_i\binom{n_{i.}}{2}\right]\left[\binom{n}{2}-\Sigma_j\binom{n_{.j}}{2}\right]}$
10.	Hubert Γ' statistic II (Γ')	$\frac{\binom{n}{2}-2\Sigma_i\binom{n_{i.}}{2}-2\Sigma_j\binom{n_{.j}}{2}-4\Sigma_{ij}\binom{n_{ij}}{2}}{\binom{n}{2}}$
11.	Minkowski score (*MS*)	$\frac{\sqrt{\sqrt{\Sigma_i\binom{n_{i.}}{2}+\Sigma_j\binom{n_{.j}}{2}}-2\Sigma_{ij}\binom{n_{ij}}{2}}}{\sqrt{\Sigma_j\binom{n_{.j}}{2}}}$
12.	Classification error (ε)	$1-\frac{1}{n}max_\sigma\sum_j n_\sigma(j)j$
13.	Van Dongen criterion (*VD*)	$\frac{2n-\Sigma_i max_j n_{ij}-max_j n_{ij}}{2n}$
14.	Micro-average precision (*MAP*)	$\sum_i A_i\left(max_j\frac{A_{ij}}{A_i}\right)$
15.	Goodman-Kruskalcoeff (*GK*)	$\sum_i A_i\left(1-max_j\frac{A_{ij}}{A_i}\right)$

16.	Mirkin metric (*M*)	$\sum_i n_{i.}^2 + \sum_j n_{.j}^2 - 2\sum_i \sum_j n_{ij}^2$

Table 2
Internal Clustering Validation Measures
(Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Halkidi et al., 2002, Liu et al., 2010)

No	Internal Clustering Validation Measures	Definition
1.	Root-mean-square standard deviation	$\left\{\frac{\sum_i \sum_{x\in Ci} \lvert\lvert x - Ci\rvert\rvert^2}{[A\sum_i (n_i - 1)]}\right\}^{\frac{1}{2}}$
2.	R-squared (*RS*)	$\frac{\sum_{x\in D} \lvert\lvert x - Ci\rvert\rvert^2 - \sum_i \sum_{x\in Ci} \lvert\lvert x - Ci\rvert\rvert^2}{\sum_{x\in D} \lvert\lvert x - Ci\rvert\rvert^2}$
3.	Modified Hubert Γstatistic (Γ)	$\frac{2}{n(n-1)} \sum_{x\in D} \sum_{x\in D} \sum_{y\in D} d_{(x,y)} d_{x\in Ci, y\in Cj\,(Ci,Cj)}$
4.	Calinski-Harabasz index (*CH*)	$\frac{\sum_i n_i d^2(Ci, C) / NC - 1}{\sum_i \sum_{x\in Ci} d^2(x - Ci) / n - NC}$
5.	*I* index (*I*)	$\left(\frac{1}{NC} \cdot \frac{\sum_{x\in D} d\,(x, c)}{\sum_i \sum_{x\in Ci} d(x,\ Ci)} . max_{ij} d(Ci, Cj)\right)^p$
6.	Dunn's indices (*D*)	$min_i \left\{ min_j \frac{(max_{x\in Ci, y\in Cj}\, d(x,y))}{max_k \{max_{x,y\in Ck}\, d(x,y)\}} \right\}$
7.	Silhouette index (*S*)	$\frac{1}{NC} \sum_i \left\{ \frac{1}{n_i} \sum_{x\in Ci} \frac{b(x) - a(x)}{\max[b(x),\ a(x)]} \right\}$
8.	Davies-Bouldin index (*DB*)	$\frac{1}{NC} \sum_i max_{j,j\neq i} \left\{ \left[\frac{1}{n_i} \sum_{x\in Ci} d(x, Ci) + \frac{1}{n_j} \sum_{x\in Cj} d(x, Cj)\right] / d(Ci, Cj) \right\}$
9.	Xie-Beni index (*XB*)	$\sum_i \sum_{x\in Ci} d^2(x, Ci) / n. min_{i,j\neq i} d^2(Ci, Cj)$
10.	SD validity index (*SD*)	$Dis(NC_{max}) Scat(NC) + Dis(NC)$ $Scat(NC) = \frac{1}{NC} \sum_i \lvert\lvert\sigma(Ci)\rvert\rvert / \lvert\lvert\sigma(D)\rvert\rvert$ $Dis(NC) = \frac{max_{ij}\, d(Ci, Cj)}{min_{ij}\, d(Ci, Cj)} \sum_i \left(\sum_j d(Ci, Cj)\right)^{-1}$

V CONCLUSION

This paper reviews the core idea of big data, big data mining, big data storage structures, clustering and its taxonomy with the validation measures taxonomy. This paper also defined how to validate the traditional clustering taxonomy by using internal and external validation measures under big data mining. These validation measures define clustering algorithm given a more accurate and reliable cluster under high volume, heterogeneous data and sources with scalability. The first and second section of this paper gives the basic background of evaluation of traditional data to big data, big data dimensions, big data mining, and clustering approach and define how to clustering approach scalable under the big data mining. The third section presents the concept of all existing clustertaxonomies by their creation process and behaviors. The fourth section defines various cluster validation measures for cluster accuracy under the cluster internal and external natures. The overall this paper presents various clustering approaches for the availabilityof big data mining and their validation approach for cluster reliabilities and accuracy.

REFERENCES

[1] Aggarwal, C. C., & Reddy, C. (2014). Data Clustering Algorithms and Applications. CRC Press Taylor & Francis Group.ISBN 978-1-4665-5822-9.

[2] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., &Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243-256. doi:10.1016/j.patcog.2012.07.021.

[3] Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, 4(1), 300-307. doi:10.1109/icpr.2000.906010.

[4] Chen, W., Oliverio, J., Kim, J. H., & Shen, J. (2018). The Modeling and Simulation of Data Clustering Algorithms in Data Mining with Big Data. Journal of Industrial Integration and Management, 12(4), 1-16. doi:10.1142/s2424862218500173.

[5] Chong, D., & Shi, H. (2015). Big data analytics: A literature review. Journal of Management Analytics, 2(3), 175-201. doi:10.1080/23270012.2015.1082449.

[6] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., . . .Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), 267-279. doi:10.1109/tetc.2014.2330519.

[7] Gan, G., Ma. C.. & Wu, J. (2007). Data clustering: Theory, algorithms, and applications. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics.

[8] Gandomi, A., &Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007.

[9] Halkidi, M., Batistakis, Y., &Vazirgiannis, M. (2002). Cluster validity methods. ACM SIGMOD Record, 31(2), 40. doi:10.1145/565117.565124.

[10] Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. Information Processing & Management, 53(5), 1103-1119. doi:10.1016/j.ipm.2017.04.004.

[11] Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303. doi:10.1016/j.bushor.2017.01.004.

[12] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining. doi:10.1109/icdm.2010.35.

[13] Manning, C. D. ,Raghavan, P. , &Schütze, H. (2008). Introduction to information retrieval . Cambridge: Cambridge University Press .

[14] Oussous, A., Benjelloun, F., Lahcen, A. A., &Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431-448. doi:10.1016/j.jksuci.2017.06.001.

[15] Pandove, D., &Goel, S. (2015). A comprehensive study on clustering approaches for big data mining. In Proceedings of IEEE 2nd International Conference on Electronics and Communication Systems (pp. 1333-1338). IEEE Xplore Digital Library. doi:10.1109/ecs.2015.7124801.

[16] Sardar, T. H., & Ansari, Z. (2018). An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm. Future Computing and Informatics Journal, 3(2), 200-209. doi:10.1016/j.fcij.2018.03.003.

[17] Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., &Herawan, T. (2014). Big Data Clustering: A Review. In Murgante B. et al. (eds), International Conference on Computational Science and Its Applications (Vol. 8583, Lecture Notes in Computer Science, pp. 707-720). Springer. doi:10.1007/978-3-319-09156-3_49.

[18] Siddiqa, A., Hashem, I. A., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., &Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71, 151-166. doi:10.1016/j.jnca.2016.04.008.

Factors Influencing ICT Usages on HRM Practices in different Manufacturing Industries of Western Odisha

Manini Nanda[1], **D.K Mahalik**[2]

[1,2]Dept. of Business Administration, Sambalpur University, Jyoti Vihar (Odisha) India.

ABSTRACT

The aim of this research paper is to find out factors influencing ICT usages on HRM practices in different manufacturing Industries of Western Odisha and rank them according to the most influential factor to less influential factor of ICT usages. The present study also determines the association between different factors of ICT usages on HRM practices of Industries of Western Odisha. The sample size was 8 HR managers from different Manufacturing Industries of Western Odisha Data collection done through questionnaire by ranking of different factors like giving 1 to most influential factor, 2 given to 2nd most influential factors on like so on to 7th influential factor. Fried man ANOVA technique used for ranking of different influential factors of different manufacturing Industries of western Odisha. Also Kendall's Coefficient of Concordance statistic is used to find the correlation between the ranks of factors influencing ICT usages on HRM practices. The result showed that there is a correlation between different factors influencing ICT usages on HRM practices as W^a found to be .741 and also finds there is a Significant difference among the factors influencing the ICT implementation on HRM practices in manufacturing industries of western Odisha as Chi square value was found to be 35.571which is greater than tabulated value.

Keywords: ICT, HRM practices, Fried man ANOVA technique, Kendall's Coefficient of Concordance, Chi square value, Western Odisha

I INTRODUCTION

In this age of information, every organization wants to develop their competitiveness and be the market leader. The use of internet and different types of technology advancement brought revolutionary changes around all the sphere of business world. All organizations need to use new innovations to get a vital position in this competitive global market. Their new technology like ICT (information communication technology) plays a major role in the management of various industries. Also the human resource department also forced to use this ICT innovation in their operations to achieve competitive advantage. In fact due to use of HRIS (Human Resource information system) or ICT on HRM practices reduces the overall cost of HR activities which keep them to advance the HR system (Wiblen, Grant and pery, 2010). More over the use of ICT in HRM practices reduce manpower; cost, time and more importantly can bring accuracy and efficiency in work. It can also eliminate human error and one we can easily assess it sitting at the place of its own site (obeng,, 2004). ICT revolution in the area of human resources has served to all its stakeholders including organizations, employs and the society. There are various factors influencing these ICT usages in HRM practices in various industries. The researchers attempted to identify those factors influencing ICT usages in HRM practices of selected manufacturing industries of western Odisha and also ranking the most influential factor of ICT usages for HRM practices.

II STUDY AREA

HRIS is a systematic procedure for collecting, storing, maintaining, retrieving and validating the data needed by an organization for its human resources, personnel activities and organization unit characteristics **Walker (1999)**. It can support long-term planning in relation to manpower **(Kovach et al., 2002)** including supply and demand forecasts, staffing, separations and development with information on training program costs and work performance of trainee. It can also support compensation programs, salary forecasts, pay budgets, employee relations, contract negotiations etc. Communication and information technologies have added value to HR applications which helped in developing a human resource information system (HRIS). Human Resource Information System (HRIS) as a computerized system used to collect, record, and store, analyze and retrieve data pertaining to organization's human resources **Alwis (2010)**. He also defines Human Resource Management System (HRMS) as a tool designed to ensure that the organization's human resources are recruited, selected, developed, employed, deployed, and supported effectively. Emergence of HRIS and its journey from mere administrative worker to strategic partner to new age relationship builder and knowledge facilitator explores by **Kashive, N.(2011)** .Also he studied the challenges and benefits of HRIS and also compares three model for HRIS with ^{their} advantages and limitations. HRIS has high impact on Human Resource Management strategies in Jordanian commercial Banks **Mohmoud Khaled,(2014)**.By using simple regression analysis he found that HRIS had a significant effect on Human resource management strategies. Human Resource Information System (HRIS) have used for more strategic purpose because it do more faster than any other method with less time and with less manpower **Rakib & Bhuiyaan, (2013)**. He studied the use of Human Resource Information System in both the manufacturing sector and service sector. For this they used Simple Non Probability Sampling method .Data Collection done through Semi Structured Questionnaires and used correlation analysis for establishing relationship between variables. HRIS provides an organisation more flexibility administratively and strategically **Srivastav Shefali (2014)**.She studied the importance of the Human Resource Information System in the current Scenario. She used Simple Random Sampling for her study and Descriptive Statistics and Chi Square Test for data analysis. The use of IT in HRM in organizations helped a lot to human resource staffs to free from routine job roles and enable them to concentrate on strategic planning in human resource development in this globalization era **Pinsonneault (1993)**. Organizational environments have become increasingly complex. Managers in these organizations face growing difficulties in coping with workforces as they are spread across a variety of countries, cultures and political systems. Managers can utilize IT as a tool in general as well as in human resource functions to increase the capabilities of the organization **Tansley and Watson (2000)**. Hong Kong industries perceived that the greatest benefits to the implementation of HRIS were the quick response and access to information that it brought, and the greatest barrier was insufficient financial support **Ngai, E.W. et al. (2004)**.Also they present a comprehensive literature review of human resource information systems (HRIS) and to report the results of a survey on the implementation of HRIS in Hong Kong. Moreover, there was a statistically significant difference between HRIS adopters and non-adopters, and between small, medium, and large companies, regarding some potential benefits and barriers to the implementation of HRIS. **Altarawneh and Shqaira, (2010)** studied the extent to which public Jordanian universities have adopted Human Resource Information System (HRIS) and examine the HRIS uses, benefits and HRIS barriers in Jordanian universities. They constructed a structured questionnaire to get data from HRIS users in Jordanian universities. **HRIS** have quick work and quick access to information done. But there were some HRIS implementation barriers like the insufficient financial support, difficulty in changing the organization's culture and lack of commitment from top managers studied the applications of HRIS in human resource management (HRM) in Indian companies."Technical and strategic HRM" and "performance and reward management" are the most important factors for HRIS applications. The most common application of HRIS in organizations working in India was found to be in "employee record", and "pay roll" system. Also "technical and strategic HRM", "performance and reward management" and "corporate communication" were also used in Indian companies. By the results of ANOVA they found that manufacturing and service companies differed significantly on all sophisticated HRIS applications and by Mean scores they got on all the sophisticated HRIS applications, service companies had significant edge over the manufacturing companies. Also Indian and multinational companies did not differ significantly on any of the HRIS applications **(Kundu and Kadian, 2010)**. HRIS increases the efficiency of HR functions like payroll, time, and attendance, appraisal performance, recruiting, learning management, training system, performance record, employee self-service, scheduling, absence management, systems, styles, reduced HR cost, increased motivation of the HR personnel etc. **(Shiri, 2012)**. The impact of IT on HRM and it lead to the emergence of HRMS. It merged all HRM activities and processes with the information technology field while the programming of data processing systems evolved into standardized routines and packages of enterprise resource planning software **(Adewoye 2012)**. There is a influence of Information Technology on Human Resource

Management of SME. A Study on the use of Applicant Tracking System (ATS) suggested a model of Human Resource Information System with reference to application tracking system in a small seized enterprise **(Mukherjee, Bhattacharya and Bera 2014)**. **Mamoudou,S., Joshi,G.P.,(2014)**, in their paper "Impact of Information Technology in Human Resources Management" give a brief overview about possibilities of IT usage in HR field for measuring and tracking human capital and using the HR information system. There is an impact of (HRIS) on Human Resource Management Strategies in Jordanian Commercial Banks (Shawabkeh, 2014). By using simple regression analysis found that HRIS had a significant effect on Human Resource Management Strategies in Jordanian Commercial Banks. **Shahzadi and Lodhi ,(2014)** studied the impact of Enterprise Resource Planning system implementation in Human Resource management practices .They used Convenient sampling method and sample size was 300 They used Questionnaires method consist of different items related to ERP product, HRM activities and organizational productivity. They used Structural equation modelling and regression analysis for data analysis. They found the ERP implementation have negative impact on recruitment and selection and also not showing relation with compensation and benefits but having positive impact on training and development of employees. **V. K. Jain (2014)** studied the impact of ICT on human resource management practices. According to him Human Resource Management (HRM) is not limited to recruitment and training. It has become an inseparable part of every organization. ICT and HRM both are closely related to each other. ICT has significant impact on increasing the efficiency of recruitment, development and decision-making, maintenance, functions. **Khoualdi &**

Basahel, (2014) explored the benefits of using the SAP system to manage the human resource at the Saudi Electricity Company (SECO) and find out the challenges for implementation of SAP. They used employee feedback to find results. They used Pearson's Coefficient of Correlation and One-sample T-Test. They found that by using SAP all functions should be easy like faster data retrieval and getting data, reduced cost, speedy transaction, increasing customer satisfaction etc. **Ejaz Ali et al. (2015)** analyze the moderating impact of ERP module (HRIS/HCM) on the relationships between human resource management practices and organizational performance. Data were collected through structured questionnaire method after establishing reliability and validity. The SPSS were used to assess the model fitness, hypotheses testing and to establish validity of the instruments through Pearson Inter-correlation Matrix. A total 220 employees of 25 firms from corporate sector were sampled through simple random sampling technique. He found that HRM practices (selection, compensation) and HRIS have significant impact on organizational performance. The results further showed that HRIS moderated the relationships between selection, compensation and organizational performance. ICT was found to have significant and positive effect on Human resource management practices **Elhazzam, (2015)**. In his paper "The effect of ICT on Human Resources Management Practices (case of number of organizations in southwest Algeria: Bechar city) explores the effect of ICT on HRM practices. Information communication technology (ICT) improves the efficiency; innovation reduces the time help in easier functioning of the organisation. It improves the performance of the employee. It helps to reduce the work time **Vohra et al. (2015)**. In his human resource management and development functions like recruitment, maintenance and development, management development, skills development and organizational development have strong association with acceptance level of ICT in human resource management and development among different private, public private partnership and government institutions in Bangladesh. **David et al. (2013)** studied employees' perception in order to explore the Information technology enabled human resource functions across the globe. The study revealed seven factors namely "Compensation", "Overall Development", "Efficiency", "Reliable features", "Motivation", "Employee benefits", "Commitment". They found human resource information system (HRIS) practices will have a vital influence in this area and HRIS has become a key for developing and improving organizational effectiveness. **Muriithi, et al. (2014)** in his paper "Effects of Human Resource Management Practices and firm performance in listed commercial Banks at Nairobi Securities Exchange" explores the factors affecting the success of HRIS adoption in the listed companies at the Nairobi Securities Exchange and how the use of HRIS strategically and positively paper" Impact of Information and Communication Technology in HRM" study the Impact of Technology Advancement on Human Resource Performance, Challenges in human resource management from technological advancement and Importance of ICT in human resource performance for this he taken factor analysis and found the conclusion that ICT improves organizational efficiency. Positive impact of ICT are so high that it over shadow the negative impacts but they also state that quantity of communication increased but quality of conversation decreased. level of the usage of HRIS and Electronic recruitment within medium-size and large Croatian companies and to evaluate the relationship among the usage of these modern managerial tools and the overall success of human resource management within these companies are studied by **Wihan and Eileen (2016)**. **Pivac, S., Tadic, I., Marasovic, B., (2014)**. The usage and acceptance level of ICT in human resource management and development at different institutions in Bangladesh was studied by **Faroque, O., Amin.R., (2016)** . Frequency distribution, descriptive statistics, chi-square analysis have been used in the study. The data were collected from 67 private, government and Public Private Partnership (PPP) organizations from different areas in Gopalganj, Bangladesh by purposive sampling technique. From Descriptive statistics they found the usage level of ICT in most of the areas of human resource management and development in the private institutions is much more than that of government institutions. On the other hand. the usage level of ICT in most of the areas of human resource management and development in the banking and financial organizations are much more than that of other organizations. By using x2-test, they found different impacts on firm performance. **Karanja, (2014)** studied how ICT can make academic management of Kenyan public universities more effective and efficient and also find out the challenges that require to achieve this effectiveness through ICT. He found ICT has became an important tool in management of universities .As compared to private universities Public universities are lacking in some field of ICT uses in Kenya. **Khashman, (2016)** explores the impact of human resource information system (HRIS) on organizational performance in Jordanian private hospitals; by examine the HRIS components like job analysis, recruitment, selection, performance appraisal applications, and communications have a significant impact on organizational performance like efficiency and effectiveness. The data was collected through questionnaire method. The population of the research included all private hospitals located in Amman city and the sample size of the research was 170 employees working in HR departments from the private hospitals. He found that there is a positive impact of the HRIS applications on organizational performance and of employees has positive attitudes towards all human resource information system applications. **Piabuo et al. (2017)** in their paper "The impact of ICT on the efficiency of HRM in Cameroonian enterprises: Case of the Mobile telephone industry" study the impact of Information and Communication Technology on the efficiency of Human Resource Management in the Cameroon mobile Telecommunication Sector. They used exploratory research design for the study and the sample size was 120.They used Pearson correlation coefficient was used to establish the relationship between the variables and regression analysis for establish the combined effect of study variables on the dependent variable. They found a significant positive relationship between the use of ICT in selection and recruitment, training and development, Human resource planning, evaluation and compensation and human resource management efficiency. **Khera.S.N and Gulati.K (2013)** focuses on the role of HRIS in HRP. The research_was empirical in nature as 127 respondents from top 7 IT companies (as per their market share) were taken. The research was done with the help of the questionnaire. After analysis they found that HRIS has various benefits but the foremost is HRIS stores vital data about the employees of the organizations that helps in human resource planning. HRIS also helps in the strategic activities of HR managers and more in training and development, succession planning, applicant tracking in recruitment and selection and manpower planning. While analyzing the overall contribution of HRIS in HRP it was concluded that HRIS identifies filled and unfilled positions in an organization very effectively and accurately. **Kavanagh, et. al. (1990)** also found that HRIS works interactively with human resources management functions such as human resource planning, staffing, training and career development, performance management, and compensation management. They further concluded human resource information system in a sequence called management information system, decision support system and executive support system, transactional processing system, **J. Anitha and M. Aruna, (2013)** in their paper "Adoption of Human Resource Information System in Organisations" identify various variables that influence adoption of HRIS or any Information Systems through a thorough literature study and consolidate them under four major factors namely Technological, Organisational, Environmental and Psychological factors. Validating this model would help the organisations to understand the essential focus areas for successful adoption of HRIS. **Hanif MI, Yunfei S, Hanif MS, Muhammad ZU, Ahmed KT, et al. (2014)** explored the factors that influence the decision making regarding HRIS adoption in the telecom sector of Pakistan particularly PTCL. The study has great significance as it contributes to the existing body of knowledge by providing improved understanding of the various technological, organizational and environmental factors which facilitate or prohibit the HRIS adoption decision in the telecom sector of Pakistan. There are some influential factors of HRIS adoption in Bangladesh. Most of the variables are retrieved from the literature review. Primary data has been collected through survey by distributing structured questionnaire to the employees of the organizations. Some factors such as organization, Technology and Environment of the organization are found to be most influential factors for HRIS adoption in the organizations. The study indicates that the practice of HRIS has positive effects on the organizational performance **(Khan,A.R., et.al. 2015)**. There are some critical factors that influencing the decision of to adopt HRIS in hospitals management in Bangladesh **(Alam et al. 2016)** .

III OBJECTIVES, HYPOTHESI & METHODOLOGY

(a) Objectives

(i) To identify the factors that are most likely to influence the ICT implementation in HRM practices in the manufacturing industries if Western Odisha.

(ii) To rank the factors influencing ICT usages for HRM practices across different manufacturing sectors of Western Odisha.

(iii) To determine whether there is any correlation between the ranks obtained on factors influencing ICT usages on HRM practices of Industries of Western Odisha.

(b) Hypothesis

(i) H_{01}: There is no significant difference among the factors influencing the ICT implementation on HRM practices in manufacturing industries of western Odisha.

(ii) H_{a1}: There is a significant difference among the factors influencing the ICT implementation on HRM practices in manufacturing industries of western Odisha.

(iii) $H_{o}2$: Different Organizations wise the ranking of factors influencing ICT usages are independent of the type of organizations.

(iv) $H_{a}2$: Different Organizations wise the ranking of factors influencing ICT usages are not independent of the type of organizations.

(c) Methodology

Sample size was 8 Human Resource managers from different Industries of western Odisha like Acc cement limited Bargarh, Bhusan steel Plant ,SMC ,Viraj steel ,Aryan steel , Espl, Shyam Dry from Jharguguda and Sambalpur region.Data collection done through questionnaire by ranking of different factors like giving 1 to most influential factor ,2 given to 2nd most influential factors on like so on to 7th influential factor.

The non-parametric test such as Fried man ANOVA technique used for ranking of different influential factors of different manufacturing Industries of western Odisha. Also Kendall's Coefficient of Concordance statistic is used to find the correlation between the ranks of factors influencing ICT usages on HRM practices.

IV DATA ANALYSIS

Table1

Ranks given for the Factors

ORGANIZATIONS	EASY OF USE	REDUCED COST AND TIME	SAFETY AND SECURITY	COMPETITIVE PRESSURE	INCREASED PRODUCTIVITY	IMPROVE QUALITY OF WORK	FACILITATING CONDITION
Bhusan steel	1	2	3	4	5	7	6
ACC cement	2	1	6	3	7	5	4
Hindalco FRP	3	1	5	2	6	4	7
Aryan steel	3	2	4	1	7	6	5
SMC	2	1	4	3	5	7	6
Viraj	2	1	3	4	6	5	7
ESPL	1	2	3	5	4	7	6
Shyam Dry	3	2	5	1	6	4	7
Rj	17	12	33	23	46	45	48
Rj^2	289	144	1089	529	2116	2025	2304

Table2

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Easy of use	8	2.13	.835	1	3
Reduced cost & time	8	1.50	.535	1	2
Safety and security	8	4.13	1.126	3	6
Competitive pressure	8	2.88	1.458	1	5
Increased productivity	8	5.75	1.035	4	7
Improve quality of work	8	5.63	1.302	4	7
Facilitating condition	8	6.00	1.069	4	7

Table 3
Ranks

	Mean Rank
Easy of use	2.13
Reduced cost & time	1.50
Safety and security	4.13
Competitive pressure	2.88
Increased productivity	5.75
Improve quality of work	5.63
Facilitating condition	6.00

Table2
Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Easy of use	8	2.13	.835	1	3
Reduced cost & time	8	1.50	.535	1	2
Safety and security	8	4.13	1.126	3	6
Competitive pressure	8	2.88	1.458	1	5
Increased productivity	8	5.75	1.035	4	7
Improve quality of work	8	5.63	1.302	4	7
Facilitating condition	8	6.00	1.069	4	7

Table 3
Ranks

	Mean Rank
Easy of use	2.13
Reduced cost & time	1.50
Safety and security	4.13
Competitive pressure	2.88
Increased productivity	5.75
Improve quality of work	5.63
Facilitating condition	6.00

Table 4

Test Statistics[a]

N	8
Chi-Square	35.571
df	6
Asymp. Sig.	.000

a. Friedman Test

So, here the calculate value of fried man statistic value that is x^2 is 35.571 .Here the number of columns(k) is 7, The table value for 5 percent level of significance for (k-1) degrees of freedom (that is 6) is 12.592. So, calculated Value of Chi square (35.571) is greater than tabulated value of chi square (12.592).
So ,we reject the null hypothesis of no significant difference among the factors influencing ICT usages on HRM Practices of different manufacturing industries of western Odisha. So, alternate hypothesis is accepted that there is a Significant difference among the factors influencing the ICT implementation on HRM practices in manufacturing Industries of western Odisha. Looking at the R_j score, it finds that reduced cost and time is the most influential Factor for ICT usages on HRM practices for manufacturing industries of western Odisha.

Ho: Different Organizations wise the ranking of factors influencing ICT usages are independent of the type of Organizations.

Table 5

Ranks

	Mean Rank
Easy of use	2.13
Reduced cost time	1.50
Safety and security	4.13
Competitive pressure	2.88
Increased productivity	5.75
Improve quality of work	5.63
Facilitating condition	6.00

Table 6

Test Statistics	
N	8
Kendall's W[a]	.741
Chi-Square	35.571
df	6
Asymp. Sig.	.000
a. Kendall's Coefficient of Concordance	

Ha: Different Organizations wise the ranking of factors influencing ICT usages are not independent of the type of Organizations.

Here the Kendall's Coefficient of Concordance is found to be .741 which is more than 0.05.
So, here the null hypothesis the ranking of factors influencing ICT usages are independent of the type of organizations was rejected and we found that there is a correlation between the ranks of factors influencing ICT usages on HRM practices of Industries of Western Odisha

V CONCLUSION

The usage of ICT in HRM practices enhances the effectiveness of industries of Western Odisha. This paper takes a step forward towards effectively identify the factors influencing ICT in these industries and ranking of different factors most influential to less influential factors. Its seven fold dimension covered under seven questions namely ease of use, competitive pressure, reduced cost and time, safety and security, improve quality of work, facilitating condition and increased productivity attempts to cover almost every area pertaining to organization while choosing ICT. It will also help other stakeholders of the industries of Western Odisha and to the society. This paper identifies most influential factor is the cost and time reduction then ease of use then competitive pressure then safety and security then improve quality of work ,then improve productivity and last the facilitating condition. Further the study can also help in undertaking research work in other similar areas of ICT usages on HRM practices.

REFERENCES

[1] A.Ejaz, T.Saeed, A.I.Soomro, R.M. Aslam. (2015), "Human resource information system: role in hrm practices and organizational performance", Cambridge Business & Economics Conference, p.p.1-11.

[2] A.M.I.,Khashman.(2016), "The Impact of Human Resource Information System (HRIS) Applications on Organizational Performance (Efficiency and Effectiveness) in Jordanian Private Hospitals", *Journal of Management Research*, Vol. 8, No. 3. 31-44.\

[3] ,S.M. Khaled.(2014), "Human Resource Information Systems and Their Impact on Human Resource Management Strategies: A Field Study in Jordanian Commercial Banks", *Journal of Management Research*, Vol. 6, No. 4, p.p.99-108,.

[4] A.Vohra, A.Shrivastava, et.al.(2015) "Impact of Information and Communication Technology in HRM", *International Journal of Computer Science and Information Technology Research*, Vol.3, Issue.2, pp. 511-516. 2015.

[5] Broderick, R., & Boudreau, J. (1992). "Human resource management, information technology, and the competitive edge".*CAHRS Working Paper* #91-19. http://dx.doi.org/10.5465/amc.1992.4274391

[6] C. Tansley., T.Watson.(2000) "Strategic exchange in the development of human resource information systems (HRIS)". New Technology. Work and Employment. Vol. 15. No. 2. pp. 108-22.

[7] D. Shine. J. Jain, K. Jain. (2013). "Study of Human Resource Information System in Global Milieu", International Journal of Management and Social Sciences Research, Vol.2 (11), p.p.35-38.

[8] D.Karanj.(2014) "The Role of ICT in Kenya's Public Universities Academic Management", Journal of Education and Literature. Vol. 1, No. 1, 39-48.

[9] De Alwis. C. A.(2010). "The Impact of Electronic Human Resource Management on the Role of Human Resource Managers". E + M Ekonomie A Management, Vol. 4, p.p. 47-60.

[10] E.W. Ngai. and F.K. Wat.(2004), "Human resource information systems: a review and empirical analysis". Emerald Group Publishing Limited, Vol. 35. No. 3, pp. 297-314.2006, 00483486.DOI.10.1108/00483480610656702

[11] Faroque.O., Amin.R.(2016),"Usage and Acceptance Level of ICT in Human Resource Management and Development by Private and Government Institutions in Bangladesh" ,The International Journal Of Business & Management, Vol. 4, Issue 6, PP.248-256.

[12] H. Gachung, K. Mburugu.(2014), "Effects of Human Resource Information Systems on Human Resource Management Practices and Firm Performance in Listed Commercial Banks at Nairobi Securities Exchange". European Journal of Business and Management, Volume.6 (29), P.p-55.

[13] Hanif MI, Yunfei S, Hanif MS, Muhammad ZU, Ahmed KT, et al.(2014), "Explore the Adoption of HRIS in Telecom Sector in Pakistan", International Journal of Economics and Management Sciences, Volume 3 , Issue 1 , 162: doi:10.4172/2162-6359.1000162.

[14] I. Altarawneh, Z. AShqairat.(2010), "Human Resource Information Systems in Jordanian Universities", International Journal of Business and Management. Vol. 5. No. 10. p.p.113-127.

[15] J. O. Adewoye, A.K. Obasan.(2012) "The Impact of Information Technology (IT) on Human Resource Management (HRM): Empirical evidence from Nigeria Banking Sector. Case Study of Selected Banks from Lagos State and Oyo State in Southwest Nigeria". European Journal of Business and Management, Vol 4, No.6, p.p.28-38.

[16] J.Anitha, and M. Aruna.(2013) "Adoption of Human Resource Information System in Organisations", SDMIMD Journal of Management. vol-4, p.p.5-15.

[17] K. Khoualdi, & A. Basahe.(2014). "The Impact of Implementing SAP System on Human Resource Management: Application to Saudi Electricity Company". International Journal of Business and Management, Vol. 9, No. 12, 2014.

[18] K.A. Kovach, A.A. Hughes, P. Fagan, & P.G. Maggitti. (2002) "Administrative and strategic advantages of HRIS". Employment Relations Today, 29. 43-8.

[19] Kashive,N.(2011), "Managing today's workforce: human resource information system (hris), its challenge and opportunities", International Journal of Research in Finance & Marketing, Volume 1, Issue 6.

[20] Kavanagh. M.J.. Gueutal, H.G. & Tannenbaum, S.1. (1990). "Human resource information systems: Development and application", Boston: PWS-Kent Publishing Co.

[21] Khan.A.R.. Hasan,N., Md. Rubel.(2015). "Factors Affecting Organizations Adopting Human Resource Information Systems: A Study in Bangladesh" ,IOSR Journal of Business and Management (IOSR-JBM) . Volume 17. Issue 11 .Ver. II, PP 45-54.

[22] M. Hitt, J. Wu and X. Zhou. (2002), "Investment in Enterprise Resource Planning: Business Impact and Productivity Mea- sures," Journal of Management Information Systems, Vol. 19, No. 1, pp. 71-98.

[23] M.Elhazzam.(2015) ,"The Effect of ICT on Human Resources Management Practices Case of Number of Organizations in Southwest Algeria (Bechar City)", International Journal of Innovative Research in Engineering & Management (IJIREM),vol-2,p.p.35-38.

[24] M.Rakib, U.Bhuiyan.(2013),"Application of Human Resource Information System (HRIS) in the Firms of Bangladesh and Its Strategic Importance, Australia" Journal of Economic Literature (JEL) Classification System, Volume 6(ISSN 9781-9920). Pp 1 10.

[25] M.Shahzadi, S.Muhammad, and N.R. Lodhi.(2014), "Impact Study of Enterprise Resource Planning (ERP) in HRM Practices", Middle-East Journal of Scientific Research,21 (1): 218-222,DOI:10.5829/idosi.mejsr.2014.21.01.2114.

[26] Mamoudou. Joshi. (2014), "Impact of Information Technology in Human Resources Management", Global Journal of Business Management and Information Technology, Vol. 4. No.1. pp.33-41.

[27] Mishra. A; and Akhman I.(2010), "Information Technology in Human Resource Management: An Empirical Assessment", Public Personnel Management. Vol. 39, No.3, pp 271-290.

[28] Mukherjee, A; Bhattacharya, S; and Bera, R. (2014). "Role of Information Technology in Human Resource Management of SME: A Study on the use of Applicant Tracking System", IBMRD's Journal of Management and Research. Vol. 3 No.1. pp 1-22.

[29] Pivac,S., Tadić.I.. Marasović,B.(2014), "The level of the usage of the human resource information system and electronic recruitment in Croatian companies", Croatian Operational Research Review. CRORR 5.291–304.

[30] Ruël, H., Bondarouk. T.. & Looise. J. K. (2004). "E-HRM: innovation or irritation: an explorative empirical study in five large companies on web-based HRM". Management Revue. 15(3). 364–380.

[31] S. Shiri.(2012) "Effectiveness of Human Resource Information System on HR Functions of the Organization; A Cross Sectional Study", US-China Education Review, A 9, p.p.830-839.

[32] S. Srivastav, (2014), "A Study on the Scope of Human Resource Information System in Present Scenario". Kanpur (U.P.). Journal of Indian Stream Research Journal, Vol.4, Issue-6, P.p.-9.

[33] S.C Kundu, R. Kadian.(2012),"Applications of HRIS in Human Resource Management in India: A Study", European Journal of Business and Management , Vol 4, No.21, 34-41.

[34] S.M. Piabuo, E.N.Piendiah, N.L.Namshi, T.Guhong. (2017) "impact of ICT on the efficiency of HRM in Cameroonian enterprises: Case of the Mobile telephone industry", Journal of Global Entrepreneurship Research 7:7 DOI 10.1186/s40497-017-0063-5.

[35] S.Mamoudou, G.P Joshi. (2014), "Impact of Information Technology in Human Resources Management", Global Journal of Business Management and Information Technology. Vol. 4, Number 1, pp. 33-41.

[36] Shikha N. Khera,S.N., Gulati.K. (2012), "Human Resource Information System and its impact on Human Resource Planning: A perceptual analysis of Information Technology companies", IOSR Journal of Business and Management (IOSRJBM). Volume 3, Issue 6 ,PP 06-13.

[37] V. K. Jain.(2014),"Impact of Technology on HR Practices", International journal of informative & futuristic research, Vol.1, Issue -10.

[38] W.D.Wihan, K.Eileen. (2016). "Exploring the Impact of Information Communication Technology on Employees' Work and Personal Lives". A Journal of Industrial Psychology, Vol.42, No.1.

[39] Walker. (1999)."Information for Human Resource Management, School of public Health". University of the Western Cape Report.pp19-31.

[40] Alam, M.G.R., Masum. A.K.M., Beh L.-S., Hong. C.S. "Critical Factors Influencing Decision to adopt HRIS in Hospitals". PLoS ONE. 11. 1--22 (2016)

[41] Mamun M.A.A.. Islam. M.S. "Perception of Management on Outcomes of Human Resource Information System (HRIS)". International Journal of Business and Social Research. 6, 29-37 (2016)

A Survey on Multi Oriented Text Recognition

Nisarg Gandhewar[1], **S. R. Tandan**[2]
[1]S. B. Jain Institute of Technology, Management & Research, Nagpur (M.S.) India.
[2]Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT
Increasing use of smart phone in our day to day life to capture images initiates a need to recognize text from natural images which is nowadays a hot research topic in the field of computer vision due to its various applications. Text in natural scenes exists in almost every phase of our daily life. From the facade of the buildings in our city to the cover of a book in our library. Undoubtedly, text is among the most brilliant and influential creations of humankind. Despite the enduring research of several decades on optical character recognition (OCR), recognizing texts from natural images is still a difficult task because of series of grand challenges which is still be encountered when detecting and recognizing text. Scene texts are often found in irregular shape (curved, or arbitrarily oriented) and its recognition not yet been well addressed in the literature. Most of the existing methods on text recognition work with regular (horizontal) texts and not generalized to handle irregular texts. This survey is aimed at summarizing and analyzing the major changes and significant progress of multi oriented text recognition in the deep learning era.

Keywords: Scene text recognition, optical character recognition, deep learning. Smart phone.

I INTRODUCTION

Scene text recognition is an essential process in computer vision tasks. Many practical applications such as traffic sign reading, product recognition, intelligent inspection, and image searching, benefit from the rich semantic information of scene text. With the development of scene text detection methods, scene character recognition has emerged at the forefront of this research topic and is regarded as an open and very challenging research problem.

Nowadays, regular text recognition methods have achieved notable success. More-over, methods based on convolution neural networks have been broadly applied. Integrating recognition models with recurrent neural networks and attention mechanisms yields better performance for these models. Nevertheless, most current recognition models remain too unstable to handle multiple disturbances from the environment. Furthermore, the various shapes and distorted patterns of irregular text cause additional challenges in recognition. As illustrated in Fig. 1, scene text with irregular shapes, such as perspective and curved text, is still very challenging to recognize.(Liu et al. 2019; Luo et al. 2019).Curved text detection is a difficult problem that has not been addressed sufficiently.

Reading text is naturally regarded as a multi classification task involving sequence-like objects. Usually, the characters in one text are of the same size. However, characters in different scene texts can vary in size.

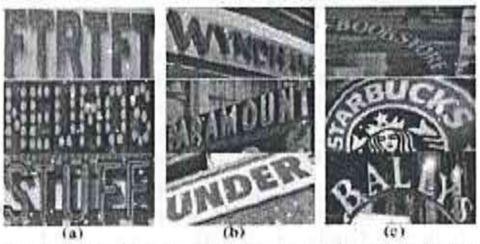


Fig No.1 Examples of regular and irregular scene text. (a)Regular text. (b) Slanted and perspective text. (c) Curved text.

II REVIEW OF LITERATURE

Scene text understanding essentially includes two tasks: text detection and word recognition. Here we present a brief introduction to related works on text detection, word recognition, and text spotting systems that combine both.

(a) Text Detection:
Text detection aims to localize text in images and generate bounding boxes for words. Existing approaches can be roughly classified into three categories: character based, text-line based and word based methods (Li et al. 2017).

(i) **Character based methods:** It firstly finds characters in images, and then group them into words. They can be further divided into sliding window based (Jaderberg M et al, 2014; Wang T et al., 2018) and Connected Components (CC) based methods. Sliding window based approaches use a trained classifier to detect characters across the image in a multi-scale sliding window fashion.CC based methods segment pixels with consistent region properties (i.e., color, stroke width, density, etc.) into characters. The detected characters are further grouped into text regions by morphological operations. conditional random fields or other graph models.

(ii) **Text-line based methods:** It detects text lines firstly and then separates each line into multiple words. The motivation is that people usually distinguish text regions initially even if characters are not recognized. Based on the observation that text region usually exhibits high self-similarity to itself and strong contrast to its local background, (Zhang et al, 2016)propose to extract text lines by exploiting symmetry property and localize text lines via salient maps that are calculated by fully convolution networks. Post processing techniques are also proposed in to extract text lines in multiple orientations.

(iii) **Word Based Methods**: More recently, a number of approaches are proposed to **detect words** directly using DNN based techniques. such as Faster R-CNN, YOLO, SSD (Li et al, 2017). By extending Faster R-CNN, (Zhong et al. 2017) design a text detector witha multi-scale Region Proposal Network (RPN) and a multi-level ROI pooling layer. (Tian et al., 2016) Develop a vertical anchor mechanism, and propose a Connectionist Text Proposal Network (CTPN) to accurately localize text lines in images. (Gupta et al. 2016) use a Fully Convolution Regression Network (FCRN) for efficient text detection and bounding box regression, motivated by YOLO. Similar to SSD, (Liao et al. 2017) propose "Text Boxes" by combining predictions from multiple feature maps with different resolutions, and achieve the best-reported text detection performance on datasets.

(b) **Text Recognition:** Traditional approaches to text recognition usually perform in a bottom-up fashion, which recognize individual characters firstly and then integrate them into words by means of beam search. dynamic programming. etc. In contrast. (Jaderberg et al.2014) consider word recognition as a multi-class classification problem, and categorize each word over a large dictionary (about 90Kwords) using a deep convolutional neural network (CNN).With the success of RNNs on and writing recognition, He et al. and Shi et al. treat word recognition as a sequence labeling problem. RNNs are employed to generate sequential labels of arbitrary length without character segmentation, and Connectionist Temporal Classification (CTC) is adopted to decode the sequence. (Shi X. et al, 2016) propose to recognize text using an attention-based sequence-to-sequence learn-ing structure. In this manner, RNNs automatically learn the character-level language model presented in word strings from the training data. The soft-attention mechanism allows the model to selectively exploit local image features. These networks can be trained end-to-end with cropped word patches as inputs. Moreover, Shi et al. insert a Spatial Transformer Network (STN) to handle words with irregular shapes.

(i) **Multi-Oriented Text Recognition:** Compared with regular text recognition work, irregular text recognition is more difficult. One kind of irregular text recognition method is the bottom-up approach which searches for the position of each character and then connects them. Another is the top-down approach. This type of approach matches the shape of the text, attempts to rectify it, and reduces the degree of recognition difficulty.

In the bottom-up manner, a two-dimensional attention mechanism for irregular text was proposed by (Yang et al. 2017)Based on the sliced Wasserstein distance. the attention alignment loss is adopted in the training phase, which enables the attention model to accurately extract the character features while ignoring the redundant background information. (Cheng et al. 2018), proposed an arbitrary-orientation text recognition network, which uses more direct information of the position to instruct the network to identify characters in special locations.

In the top-down manner, STAR-Net used an affine transformation network that transforms the rotated and differently scaled text into more regular text. Meanwhile, a ResNet is used to extract features and handle more complex background noise. RARE regresses the fiducially transformation points on sloped text and even curved text. thereby mapping the corresponding points onto standard positions of the new image. Using thin plate spline to back propagate the gradients. RARE is end-to-end optimized. (Luo et al. 2019) proposed a MORAN model which uses the top-down approach. The MORAN consists of a multi-object rectification network and an attention-based sequence recognition network. The multi-object rectification network is designed for rectifying images that contain irregular text. It decreases the difficulty of recognition and enables the attention-based sequence recognition network to more easily read irregular text. But the MORAN will fail when the curve angle in text is too large.

(c) **Text Spotting:** Text spotting needs to handle both text detection and word recognition. (Wang et al. 2017) take the locations and scores of detected characters as input and try to find an optimal configuration of a particular word in a given lexicon, based on a pictorial structures formulation. (Neumann et al. 2013) use a CC based method for character detection. These characters are then agglomerated into text lines based on heuristic rules. Optimal sequences are finally found in each text line using dynamic programming. which are the recognized words. These recognition-based pipelines lack explicit word detection. Some text spotting systems firstly generate text proposals with a high recall and a low precision, and then refine them using a separate recognition model it is expected that a strong recognizer can reject false positives, especially when a lexicon is given. (Jaderberg et al. 2014) use an ensemble model to generate text proposals, and then adopt the word classifier in for recognition. (Gupta et al. 2016) em-ploy FCRN for text detection and the word classifier in for recognition. (Liao et al. 2017) combine "Text Boxes" and "CRNN", which yield state-of-the-art text spotting performance on datasets.

III BENCHMARK DATASETS

Despite the success in Scene text detection & recognition, current methods are only evaluated on single datasets after being trained on them separately. Naturally, a new dataset representing several challenges would also provide extra momentum for this field. Evaluation of cross dataset generalization ability is also preferable, where the model is trained only on one dataset and then tested of another. We have collected existing datasets supporting multi oriented text summarized in Tab.1.

The CTW1500 dataset contains 1500 images, with 10,751 bounding boxes (3,530 are curved bounding boxes) and at least one curved text per image. The images were manually extracted from the Internet, image libraries, such as Google Open-Image, and data collected via phone cameras, which also contain a large amount of horizontal and multi-oriented text. The distribution contains indoor, outdoor, born digital, blurred, perspective distortion text and other text. This dataset is multilingual, containing mostly Chinese and English text. Some of the images from this dataset is shown in fig 6.

Table 1
Datasets with support for multi oriented text

Data Set	Images Training / Testing	Orientation	Language	Remark
CTW1500 (2019)	1000/500	Multioriented	English, Chinese	
ICDAR 2015	1000/500	Multioriented	English	Blur Images
ICDAR RCTW(2017)	8034/4229	Multioriented	Chinese	
Total-Text (2017)	1255/300	Curved	English, Chinese	
CTW (2017)	25K/6K	Multioriented	Chinese	
COCO-TEXT (2017)	63686/43686	Multioriented	English	Some images do not contain text
MSRA-TD500	300/200	Multioriented	English, Chinese	long text

In the same way Fig 2, Fig 3, Fig 4 and Fig 5 shows sample images from MSRA-TD500, ICDAR2015, ICDAR2013, COCO TEXT dataset. (Liu Y, et al. 2019;Ma J et al. 2018).

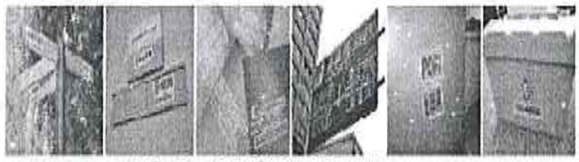


Fig No.2 MSRA-TD500

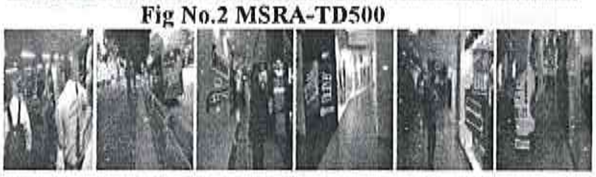


Fig No.3 ICDAR2015

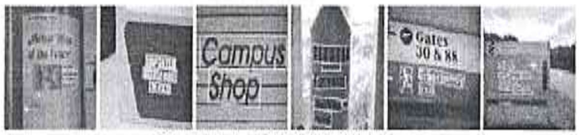


Fig No.4 ICDAR2013

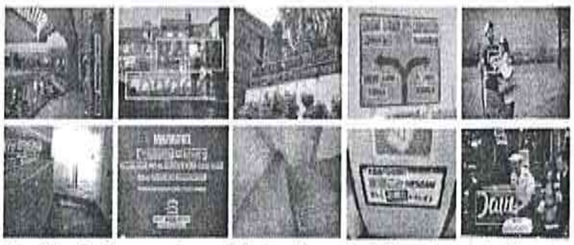


Fig No.5. Examples of detection results. From left to right in columns: ICDAR2015, ICDAR2013, MSRA-TD500, MLT, COCO-Text.

Fig No.6 Examples of annotations in the CTW1500 dataset [1]

IV SIGNIFICANT WORK FOR MULTI ORIENTED TEXT RECOGNITION

Table 2
Significant Work on Multi Oriented Text Recognition

Sr No	Author Name	Year	Proposed Work	Dataset Used	Remark
1	Yuliang Liu et al [1]	5 Feb 19	(Liu et al. 2019) proposed a new dataset called CTW1500, comprising mainly English and Chinese curved text. They also proposed a polygon-based curved text detector that can detect curved text without using an empirical combination	CTW1500, Total-text, MSRA-TD500	Proposed dataset could be enlarged as a curved-text-based recognition dataset
2	Canjie Luo et al. [2]	10 Jan 19	(Luo et al. 2019) Proposed a multi-object rectified attention network (MORAN) for scene text recognition. The proposed framework involves two stages: rectification and recognition. rectification transform an image containing irregular text into a more readable One. For recognition attention-based sequence recognition network was designed to recognize the rectified image and outputs the characters in sequence	IIIT5K, SVT, ICDAR2003, ICDAR2013, ICDAR2015, SVT-Perspective and CUTE80	The MORAN will fail when the curve angle is too large
3	Zhanzhan Cheng et al. [3]	22 Mar 18	(Cheng et al. 2018) develop the arbitrary orientation network (AON) to directly capture the deep features of irregular texts, which are combined into an attention-based decoder to generate character sequence. develop an arbitrary orientation network consisting of the horizontal network (HN), the vertical network (VN) and the character placement clue network (CN) for extracting horizontal, vertical and placement features respectively.	CUTE80, SVT-Perspective, IIIT5k, SVT and ICDAR	Computational cost and the number of parameters are major concerns in resource-constrained scenarios such as embedded computer Systems.
4	Xiao Yang et al [4]	19 Aug 17	(Yang et al. 2017) Developed a robust end-to-end neural-based model which includes two learning components: (1) an auxiliary dense character detection task using a FCN that helps to learn text specific visual patterns, (2) an alignment loss that provides guidance to the training of an attention model. Also generated a large-scale synthetic dataset containing perspectively distorted and curved text.	SVT-Perspective ,CUTE80, ICDAR03, III5K	Future directions would be to combine the proposed text recognition model with a text detection method for a full end-to-end system.

5	Jianqi Ma et al[5]	15 Mar 18	(Ma et al. 2018) present the Rotation Region Proposal Networks (RRPN), which are designed to generate inclined proposals with text orientation angle information. The angle information is then adapted for bounding box regression to make the proposals more accurately fit into the text region in terms of the orientation.	MSRA-TD500, ICDAR 2015, ICDAR 2013,	consider only three datasets
6	Pengyuan Lyu et al [6]	27 Feb 18	(Lyu et al. 2018)propose a model to detect scene text by localizing corner points of text bounding boxes and segmenting text regions in relative positions. In inference stage, candidate boxes are generated by sampling and grouping corner points, which are further scored by segmentation maps and suppressed by NMS.	ICDAR2013, ICDAR2015, MSRA-TD500, MLT and COCO-Text	When two text instances are extremely close, it may predict the twoinstances as one.
7	Baoguang Shi et al [7]	12 Mar 16	(Shi et al 2016), propose RARE (Robust text recognizer with Automatic rectification) RARE is a specially designed deep neural network, which consists of a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN)	IIIT 5K-Words,Street View Text, ICDAR 2003, ICDAR 2013	It did not address the end-to-end scene text reading problem.

V CONCLUSION

The past several years have witness the significant development of methods for detecting & recognizing a text. Despite the success so far, methods for text detection and recognition are still confronted with several challenges. Existing methods on text recognition mainly work with regular (horizontal) texts and not generalized to handle irregular texts. While human have barely no difficulties localizing and recognizing text, current methods are not designed and trained effortlessly.

The recent work on multi oriented text recognition address the problem of handling irregular text but still they have not yet reached human-level performance & still there is scope for improvement. Few datasets are available which supports irregular text but still we require more to train and evaluate current and future methods efficiently.

REFERENCES

[1] Liu Y, Jin L, (2019), Curved scene text detection via transverse and longitudinal sequence connection, *Elsevier Journal of Pattern Recognition* (Vol 90), PP 337-345.

[2] Luo C, Lianwen J, (2019), MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition. *Journal of Pattern Recognition, Science Direct.* (Vol 90), Pages 109-118.

[3] Cheng Z, Bai F, (2018), AON: Towards arbitrarily-oriented text recognition, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5571–5579.

[4] Yang X, He D, (2017), Learning to Read Irregular Text with Attention Mechanisms, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.

[5] Ma J, Shao W, (2018), Arbitrary-Oriented Scene Text Detection via Rotation Proposals, Draft submitted to cornell university.

[6] Lyu P ,Yao C, (2018), Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation, Draft submitted to Cornell University.

[7] Shi B, Wang X, (2016), Robust Scene Text Recognition with Automatic Rectification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[8] Li Hui, Wang P, (2017), Towards End-to-end Text Spotting with Convolutional Recurrent Neural Networks *IEEE International Conference on Computer Vision (ICCV).*

[9] Q. Ye and. Doermann D, (2015). Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.,37(7):1480–1500.*

[10] Zhu Y. Yao C. (2016). Scene text detection and recognition: recent advances and future trends. *Frontiers of Computer Science, 10(1):19–36.*

[11] Jaderberg M, Vedaldi A,(2014), Deep features for text spotting, *In Proc. European. Conference of Computer Vision.*

[12] Wang T. (2012), End-to-end text recognition with convolutional neural networks, *In Proc. IEEE Int. Conf. Pattern. Recognition.*

[13] Zhu S, Zanibbi R,(2016), A text detection system for natural scenes with convolutional feature learning and cascaded classification, *In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[14] Huang W. Lin Z. (2013), Text localizationin natural images using stroke feature transform and text co-variance descriptors. *In Proceeding IEEE International. Conference. Computer. Vision.*

[15] Zhang, Z., Zhang, C., Shen, W., Yao,C., and Bai,X. (2016), Multi-oriented text detection with fully convolutional networks", In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

[16] Tian Z.,Huang W.,He T.,He P., and Qiao Y., (2016), Detecting text in natural image with connectionist text proposal network", In Proc. Eur. Conf. Comp. Vis..

[17] Neumann L, Matas J, (2013), Scene text localization and recognition with oriented stroke detection, In Proc. IEEE Int. Conf. Comp. Vis., 2013.

[18] Gupta A, Zisserman A, (2016), Synthetic data for text localisation in natural images", In Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

[19] Liao M, Shi B, Bai B, Wang X, and Liu W, (2017) Textboxes: A fast text detector with a single deep neural network, In Proc.National Conf. Artificial Intell.

[20] Tian, Z, Huang W. He T, He P, and Qiao Y,(2016), Detecting text in natural image with connectionist text proposal network, In Proc. Eur. Conf. Comp. Vis.

A Survey: Chronic Diseases Bigdata Mangement

Nitin Chopde[1], Dr.Rohit Miri[2]

[1,2]Dept. of Comp. Science & Engineering, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

chronic diseases are the main grounds of human death and disability worldwide. Chronic diseases are main threat that affects million people in future, with many young age people as well as average age. The people dying by chronic diseases are more as compared to that of all infectious diseases maternal and nutritional deficiencies combined. The most of chronic disease deaths occur in poor and average income countries and half are in women. The basic chronic diseases are heart disease diabetes; stroke; cancer and chronic respiratory diseases. This proposed paper survey some of the presented activities and opportunities related to big data for health management prediction, outlining some of the key primary risk that need to be manage and tackled.

Keywords—Big-data, chronic disease.

I INTRODUCTION

As world health organization diabetes is caused by raised blood glucose (sugar) levels as well as be deficient in hormone insulin. The blood glucose levels in human body controlled. The excessive weight and physical inactivity cause the most common type of diabetes is type. An absolute be deficient in insulin from childhood causes Type1 diabetic

Stroke is cardiovascular diseases this disease of the brain caused by obstruction to the blood supply is Stroke. The heart disease is causes heart attack. The leading cause of death globally is coronary heart disease. The disease of the blood vessels of the heart turns into coronary heart disease.

Chronic respiratory diseases are related to lung. The chronic obstructive respiratory disease and asthma are due to lungs problem. That is permanent obstruction of the larger airways in the lung. The reversible barrier of the smaller airways in the lung is turn into asthma disease.

The abnormal cells proliferate and extend out of control can turn into Cancer disease. In cancer all organs of the body can starts to damage and body can convert into cancerous.

II CHRHONIC DISEASE AND ITS RISK FACTOR

The most important risk factors related to chronic disease age and heredity nonchangeable risk factors that clarify the majority of new actions of chronic respiratory diseases, heart disease, stroke, and some important cancers. The major modifiable risk factors and modifiable risk factor are similar in many ways around the world.

(a) infancy risk

There are many evidences found where most of country, before birth where in early childhood persuadehealth in adult life situations. In some situation increased rates of high blood pressure at time of birth which is symptoms of, heart disease. The stroke and diabetes are also having same feature.

(b) Risk amassing

Age is important risk factor old age signals to increase of modifiable risks for chronic disease: the brunt of risk factors increases over the life course.

(c) primary determinants

The primary determinants of chronic diseases are impact of globalization and in additional effect of urbanization and population. The major forces driving social change, economic conditions and cultural change also effect on human.

(d) supplementary risk factors

A smaller proportion of disease is disease related to risk factor of chronic disease. Use of Harmful alcohol is significant provider to the global burden of disease.

(e) The foremost alcohol association to chronic disease is more complex as compare to other risk. Therisk factor of liver cancers is infections. The environmental factors like air pollution, it leads to a rangeof chronic diseases. the chronic disease like asthma and other chronic respiratory diseases caused due air pollution. Another factor such as genetic factors participates in chronic disease.

III REVIEW OF LITERATURE

Due to unhealthy diet, physical inactivity and use of tobacco increase blood pressure. It also increased glucose levels. There are some abnormal blood lipids, overweight and obesity. **C. Friedmanet al. (2004)**proposed de a method based on natural language processing (NLP) that automatically maps an entire clinical document to codes with modifiers and to quantitatively evaluate the method. M. **Marcos et al. (2013)** they introduce a complete approach, with a set of tools as well as methodological guidelines, to deal with the interoperability of CDSSs and EHRs based on archetypes. **R. Boland et al. (2013)** contributes a highthroughput method for associating epoch with universal diseases usinglinked electronic records.

SA S. (2013) build model for heart disease prediction system is developed using three data mining classification modeling techniques. **J. Sun et al. (2014)** developed personalized hypertension management plans. **R. Eriksson (2014)** used techniques for temporal data mining of EPRs in order to detect ADRs ina patient- and dose specific manner. **D. W. Bates et al. (2014)** big data, including analytics, is a controlling tool that willbe as valuable in health care. EHRs describing patient treatments and outcomes are rich but underused information. Traditional health data centres capture and store an enormous amount of structured data concerning a wide range of information including diagnostics, laboratory tests, medication, and ancillary clinical data. **M. R. Eriksson, et al .(2014)** the systematic analysis to individual patient reports, the use of natural language processing plays an essential role

TABLE 1
DATA OF VARIOUS VARIABLES EXTRACTED FOR THE OVERVIEW OF CHRONIC DISEASE MODELS

Study location	Data was collected on the location of the studies including U.S. versus non U.S. based and whether or not the studies were done in rural or urban settings.
Study design	Data was also collected if the studies were randomized controlled clinical trials and if they were interventional or not.
Studies follow up	The period of the studies was also accessed to examine the force of the chronic models on longitudinal
Disease studied	The survey is focused on diseases including Diabetes, Chronic Obstructive Pulmonary Disease and Cardiovascular diseases because of their predominance in resulting death and disability worldwide.
Chronic disease model and its elements	Data was recorded on the specific chronic disease models and their elements that were described and evaluated across all these studies
Outcomes assessed	Data was also recorded about the various outcomes that were measured in these studies.

Fig.1 Discovered by Ashoo Grover1 and Ashish Joshi(2014)

IV CHRONIC DISEASES BIGDATA MANAGEMENT

According to **WHR (2015)** BigData can be transformed an actionable asset rather than a siloed bureaucratic nightmare with its potential for solving big healthcare problems at both the population and personal level.

(a) verdict what's working
Medicitynews(2018) Big Data opens up real opportunities for solving big healthcare problems at both the population and personal level. When someone is diagnosed with diabetes, they are asked to check their blood sugar at least once a day. This simple act creates hundreds of data points per year, per person. This data has the potential to provide some of the richest insights about specific changes to behaviour and therapies that can improve quality of life and overall health outcomes.

(b) Summarizing, and presenting results automatically
Savvy algorithms can rapidly analyze patient glucose data, for example, to determine a patient's best day. Combining results from Big Data sets can go further, providing the content and context to both summarize and present information in terms best suited for individual physicians and patients. Rather than spend precious time on data analysis, people can focus on making the most of what they've learned.

(c) Enabling decision support
Data drives the creation of valuable decision support tools. It answers to questions like when to changesettings on a medical device, if a medication dose should be changed, or if a patient needs moral support can all be garnered from machine learning algorithms that use past behaviors and personal data to recommend decisions. Doctors are starting to access algorithms results and use them during appointments more frequently.

(d) Predicting when changes are coming
Is a person's glycemic control, blood pressure, or weight trending in the wrong direction? Is a personlikely to decline in health? Big Data forms the basis for predictive algorithms that can give patients and their doctors' early warning about real issues that may be on the horizon.

(e) Knowing where to ask more questions
Often, data analysis uncovers more questions. In diabetes, for example, knowing a person's glucose level is only step one. This data uncovers questions about diet and exercise habits, stress, or other healthproblems. With new, patientgenerated health data (PGHD) flowing in from wearable and smart phone, gathering the data required to answerth

ese questions is closer than it has ever been before

V CONCLUSION

The Big data can provide to boost theapplicability of clinical research studies into realworld scenarios .where population heterogeneity is problem. It equ ally provides the opportunity to enableeffective and precision medicine by performing patient strati fication. This is indeed a key task toward persona lized healthcare. A better use of medical resources by means of personalization can lead to wellmana ged health services that can overcome the challen ges of a rapidly increasing and aging population. Thus, advances in big data processing for chronic disease model input data, health informatics, bioinf ormatics and sensing will have a great impact on future clinical research.

REFERENCES

[1] M. Marcos, J. A. Maldonado, B. Mart'inezSal vador, D. Bosca, and M.Robles(2013), "Intero perability of clinical decisionsupport systems a nd electronic health records using archetypes: A case study in n clinical trial eligibility," J. Biomed. Informat., vol. 46, pp. 676–689.

[2] SA, S. (2013). Intelligent heart disease predict ion system using data mining techniques. Inte rnational Journal of Healthcare & Biomedical Research, 1, 94-101.

[3] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak(2004), "Automated encoding of clin ical documents based on natural language pro cessing," J. Amer. Med. Informat. Assoc., vol . 11, pp. 392–402.

[4] J. Sun, C. D. McNaughton, P. Zhang, A. Per er, A. GkoulalasDivanis, J. C. Denny, J. Kirb y, T. Lasko, A. Saip, and B. A. Malin(2014) , "Predicting changes in hypertension control using electronic health records from a chronic disease management program," J. Amer. Me d. Informat. Assoc., vol. 21, pp. 337–344.

[5] R. Eriksson, T. Werge, L. J. Jensen, and S. Brunak(2014). "Dosespecific adverse drug reac tion identification in electronic patient records: Temporal data mining in an inpatient psychi atric population." Drug Saf., vol. 37, pp. 237 –247.

[6] D. W. Bates, S. Saria, L. OhnoMachado, A. Shah, and G. Escobar(2014), "Big data in he alth care: Using analytics to identify and ma nage highrisk and highcost patients." Health Affairs, vol. 33, pp. 1123–1131.

[7] https://medcitynews.com

[8] M. R. Boland, G. Hripcsak, D. J. Albers, Y. Wei, A. B. Wilcox, J. Wei, J. Li, S. Lin, M. Breene, and R. Myers(2013), "Discovering medical conditions associated with period on titis using linkedelectronic health records," J. Clin. Periodontol., vol. 40, pp. 474–482.

[9] Ashoo Grover1 and Ashish Joshi(2014). An Overview of Chronic Disease Models: A Syst ematic Literature Review," Global Journal of Health Science; Vol. 7, No. 2,1916-9736.

[10] https://www.whrhealth.com

[11] Javier AndreuPerez, Carmen C. Y. Poon, Rob ert D. Merrifield, Stephen T. C. Wong, and GuangZhong Yang(2015), "Big Data for Healt h Fellow," IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 4,pp.1193-1208.

Proposing PSO Based Algorithm for Classifying Breast Cancer Data Effectively

Pranjali Dewangan[1], Neelam Sahu[2]
[1,2]Dr. C.V.Raman University, Bilaspur (C.G.) India.

ABSTRACT

Digital Image Processing is processing of images that are digital in nature by a digital computer. Thresholding is the most commonly used intensity-based image segmentation technique which converts gray scale images into binary image. The performance of thresholding algorithms mainly depends on selection of threshold value. Various statistical properties such as maximum likelihood, moment, entropy and between class-variance has been utilized for selecting a proper threshold. This study takestwo objectives in account. The first to develop an efficient segmentation algorithm based on PSO and second to test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently. The proposed hybrid approach for data mining has included two phases. In the first phase, we adopted the statistical method in pre- processing. It can eliminate the insignificant features in order to reduce the complexity for next data mining stage. In the second phase, we proposed the data mining methodology that based on the standard PSO which called discrete PSO. In this study, we have used the Wisconsin breast cancer data set to test our proposed DPSO algorithm. In this study, a new hybrid approach of using both integrated statistical method and DPSO is proposed and successfully applied to the classification risk of Wisconsin- breast-cancer data set. According to our testing results, the proposed hybrid approach can improve the accuracy to 96.25%, sensitivity to 100% and specificity to 96.32%. These results are very promising compared to the previously reported classification techniques for mining breast cancer data.

Keywords:-Particle swarm Optimization, Optimization, Medical application

I INTRODUCTION

Digital Image Processing is processing of images that are digital in nature by a digital computer. Image Processing is motivated by three major applications; 1. improvement of pictorial info for human perception, 2. image processing for autonomous m/c application, 3. efficient storage and transmission. Segmentation is characterized as the process of dividing an image into distinct parts. Segmentation is utilized to identify an object or extract relevant information from digital images. Thresholding is a simple image segmentation methodology which yields a binary output from gray scale input images. Segmentation is the procedure by which an image is grouped into various units that are homogeneous with respect to one or more characteristics. It is an important task in image processing applications. The main objective of segmentation is to simplify the image into a form which is more meaningful and easier to analyse. Segmentation allows us to analyse an object or region in a more meaningful manner. Image segmentation can be done by three distinct methodologies. They are Intensity based segmentation, Edge-based segmentation, Region-based segmentation. Thresholding is the most commonly used intensity based image segmentation technique which converts gray scale images into binary image [Rutuparna Panda , Sanjay Agrawal, Sudipta huyan.2011].The performance of thresholding algorithms mainly depends on selection of threshold value. Various statistical properties such as maximum likelihood, moment, entropy and between class-variance has been utilized for selecting a proper threshold. It is defined as the partitioning of an image into non-overlapping regions that are homogeneous with respect to some visual feature, such as intensity or texture [Elsevier,2013]. This study taken into account following objectives.

(a) To develop an efficient segmentation algorithm based on PSO.
(b) To test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently.

II LITERATURE REVIEW

Segmentation algorithms are involved in virtually all computer vision systems, at least in a pre-processing stage, up to practical applications in which segmentation plays a most central role: they range from medical imaging to object detection, traffic control system and video surveillance, among many others. The importance of developing automated methods to accurately perform segmentation is obvious if one is aware about how tedious, time-consuming, subjective and error-prone manual segmentation can be according to the general principle on which the segmentation is based, we can build a taxonomy of the different segmentation algorithms distinguishing the following categories [M. Sonka, V. Hlavac, R. Boyle,R. Klette,2007]: thresholding techniques (based on pixel intensity), edge-based methods (boundary localization), region-based approaches (region detection), and deformable models (shape). Metaheuristic are general-purpose stochastic procedures designed to solve complex optimization problems[F. Glover,2003]. They are approximate and usually non-deterministic algorithms that guide a search process over the solution space. Unlike methods designed specifically for particular types of optimization tasks, they are general purpose algorithms and require no particular knowledge about the problem structure other than the objective function itself, when defined, or a sampling of it (training set) when the optimization process is based only on empirical observations. Metaheuristic are characterized by their robustness and ability to exploit the information they accumulate about an initially unknown search space in order to bias the subsequent search towards useful subspaces. They provide an effective approach to manage large, complex and poorly understood search spaces where enumerative or heuristic search methods are inappropriate. Despite their importance and the number of scientific publications on the use of metaheuristic for deformable model optimization.

Liqiang Liu and Haijiao Ren et al(2010) in this paper, space contraction transformations are introduces into standard Ant Colony System algorithm to increase the speed and to improve the search ability of algorithm.

Myung-Eun Lee1 et al(2009) in this paper reviewer discuss about the ACO algorithm for the segmentation of brain MR images can effectively segments the fine details.

H. Shah-Hosseini (2011) in this paper, a novel metaheuristic called ―Galaxy based Search Algorithm‖ or GbSA is introduced for multilevel thresholding. The proposed GbSA may be viewed as a variable neighbourhood search algorithm or as an Iterative local Search algorithm

This exhaustive Literature survey includes all relevant papers related with the hybridization of metaheuristic and segmentation models. At the same time, it aims at drawing some guidelines to help those who are willing to incorporate the advantages, and ease of use, of metaheuristic into the design of new segmentation approaches based on the same principles.

III RESEARCH METHODOLOGY

(a) The Proposed Hybrid Approach- The proposed hybrid approach for data mining has included two phases. In the first phase, we adopted the statistical method in pre- processing. It can eliminate the insignificant features in order to reduce the complexity for next data mining stage. In the second phase, we proposed the data mining methodology that based on the standard PSO which called discrete PSO. In this study, we have used the Wisconsin breast cancer data set to test our proposed DPSO algorithm. The data set included 9 features and 1 Order variable. We substituted the missing data by filling the values which appear the most frequently in that feature. Beside the Order variable, the value of 9 features is between 1 and 10, the higher value corresponding to a more unusual situation of the tumor such as the data in Table 1. The data set contains 699 points, 461 were diagnosed to be benign (Order = 2) and 238 to be metastatic (Order = 4). We divided the training data set which contains 459 patients' re- cords and validation data set which contains 240 patients' records from original data set randomly.

Table 1
The feature variable of dataset

Featurevariable	Domain	Simplifiedexpress
LumpViscosity	1–10	Z1
Cell Size Uniformities	1–10	Z2
Cell Shape Uniformities	1–10	Z3
Fringe Cohesion	1–10	Z4
SingleDeciduaCell Size	1–10	Z5
BasicCore	1–10	Z6
MildChromatin	1–10	Z7
RegularCore	1–10	Z8
Mitospore	1–10	Z9
Order	2,4	Z10
2:benign,4:metastatic		

IV DATA ANALYSIS

(a) Objectives - 1

- To develop an efficient segmentation algorithm based on PSO.

V DPSO FOR DATA MINING

Theparticleswarmoptimization(PSO)techniqueisapopulationbasedstochasticoptimizationtechniquefirstintroducedby Gordan. B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016). It belongs tothecategoryofSwarm

Intelligencemethods;itisalso anevolutionarycomputationmethodinspiredby themetaphorofsocialinteractionandcommunicationsuchasbirdflockingandfish schooling.Thedetailshavebeen giveninthefollowing.

InPSO,asolutionisencodedas afinite-lengthstringcalleda particle(Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016); Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., ... & Chang, K. S. (2014); Du, K. L., & Swamy, M. N. S. (2016); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Gonzalez-Vidal, A., Barnaghi, P., & Skarmeta, A. F. (2018); Shabbir, F., & Omenzetter, P. (2015)).Alloftheparticleshavefitnessvalues which are evaluatedby thefitnessfunctionto be optimized,andhave velocitieswhichdirectthe flyingof theparticles (Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016); Du, K. L., & Swamy, M. N. S. (2016); Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., ... & Chang, K. S. (2014); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Dubey, A. K., Gupta, U., & Jain, S. (2015); Shabbir, F., & Omenzetter, P. (2015)).PSOis initializedwitha populationofrandomparticles,withrandompositionsandvelocities insidethe problemspace,and thensearchesfor optima byupdatinggenerations.Itcombineslocalsearchandglobalsearch yieldinginhighsearchefficiency.Each particlemovestowardsits bestpreviouspositionandtowardsthebestparticleinthewhole swarmineveryiteration.Theformerisalocal bestanditsvalueis calledpbest, andthelatterisaglobalbestandits valueiscalled gbest intheliterature(Allahverdi, A. (2015); Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017); Shen, X., Chen, J. G., Zhu, X. C., Liu, P. Y., & Du, Z. H. (2015); Gonzalez-Vidal, A., Barnaghi, P., & Skarmeta, A. F. (2018); Shabbir, F., & Omenzetter, P. (2015)).After findingthetwobestvalues,theparticle updates itsvelocityandpositionwiththefollowingequationincontinuous PSO:

v^t $\quad t-1 \qquad t-1 \quad t-1$

$t-1 \quad t-1$

The values c1q1 and c2q2 determine the weights of two parts, and the value of (c1 þ c2) is usually limited to 4 (Gordan, B., Armaghani, D. J., Hajihassani, M., & Monjezi, M. (2016)).

Fig.6 The flowdiagramoftheproposedhybridapproach .

To apply PSO, several parameters including the number of population (m), cognition learning

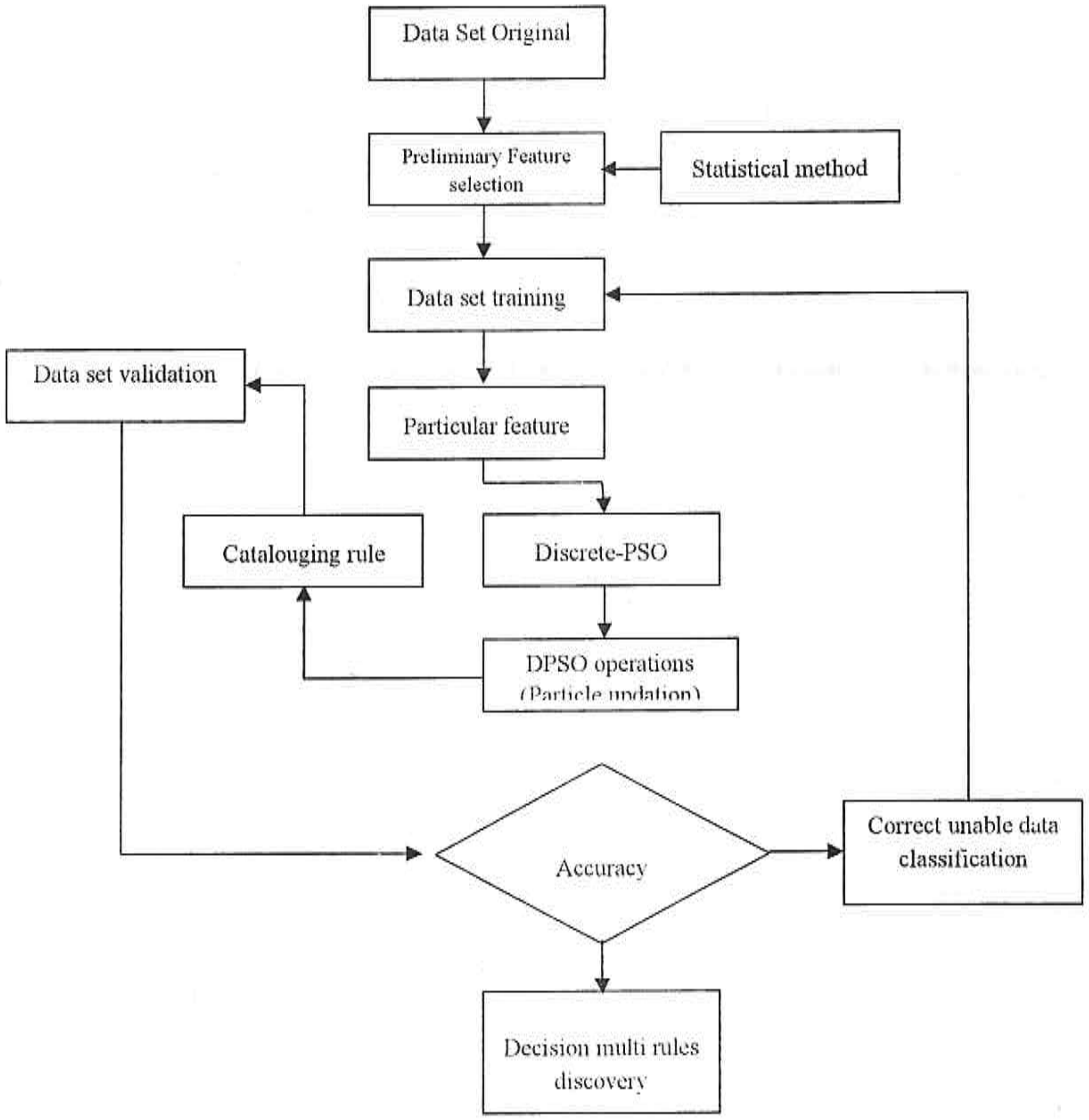

Data Set Original
Preliminary Feature selection
Statistical method
Data set training
Data set validation
Particular feature
Catalouging rule
Discrete-PSO
DPSO operations (Particle updation)
Accuracy
Correct unable data classification
Decision multi rules discovery

Fig. 1

factor (c1), social learning factor (c2), inertia weight (w), and the number of iterations or CPU time should be determined properly. We conducted the preliminary experiments, and the complete computational procedure of the PSO algorithm can be summarized as follows:

(i) Initialize:Initializeparametersandpopulationwithrandom positionandvelocities.

(ii) Evaluation:Evaluatethefitnessvalue(thedesiredobjective function)for eachparticle.

(iii) Findthe pbest:Ifthefitnessvalueofparticleiisbetterthan its bestfitnessvalue(pbest)in history,thenset currentfit-nessvalueasthenewpbest to particlei.

(iv) Findthegbest:If any pbestisupdatedanditisbetterthanthe currentgbest, thenset gbestto thecurrentvalue.

(v) Updatevelocity andposition:Updatevelocity andmovetothe nextpositionaccordingto Eqs.(1) and(2).

(vi) Stoppingcriterion:IfthenumberofiterationsorCPUtimeare met,thenstop:otherwisegoback tostep2.

(a) **Objectives - 2**

To test proposed algorithm in classifying risk thereby classifying breast cancer data more effectively and efficiently.

Pre-processing using statistical method

From correlation and regression analysis, we could eliminate the insignificant features. In this phase, the feature ''Order'' would be taken into a dependent variable and the remaining features would be taken into independent variables. Table 2 shows the experimental results from SPSS statistics package software. Obviously, we could find the three insignificant features including ''Fringe Cohesion'', ''Single Decidua Cell Size'' and ''Mitospore''. In addition, the adjusted R2 has already reached 0.837 which represent the intersection between independent variables was insignificant. Hence, the intersection could be ignored in this study.

Table 3 shows that we had the same results from correlation and regression analysis so we only need to keep these 6 features to the next phase which included ''Lump Viscosity'', ''Cell Size Uniformities'', ''Cell Shape Uniformities'', ''Basic Core'', ''Mild Chromatin and Regular Nucleoli.

Table 2
The experimental results from SPSS statistics package software

Modelsummary[b]

Model	R	R^2	AdjustedR^2	Std.
1	.925[a]	0.841	0.840	0.3844

Coefficients[b]

Model		Unstandardized		Standardized	t	Sig.	95%confidenceinterval	
							Lowerbound	Upperbound
		B	Std. error	Beta				
1	(Constant)	1.509	0.038		45.942	0	1.449	1.577
	Z1	0.060	0.011	0.24	8.998	0	0.09	0.085
	Z2	0.048	0.018	0.149	3.499	0.007	0.07	0.16
	Z3	0.039	0.017	0.108	2.681	0.013	0.011	0.068
	Z4	0.020	0.013	0.040	1.421	0.163	−0.009	0.034
	Z5	0.017	0.08	0.039	1.406	0.169	−0.011	0.043
	Z6	0.097	0.014	0.358	14.495	0	0.15	0.115
	Z7	0.049	0.06	0.110	4.082	0	0.028	0.069

a Predictors: (constant), Z_9, Z_6, Z_1, Z_8, Z_5, Z_4, Z_7, Z_3, Z_2.

b Dependent variable: Z_{10}.

VI RESULT ANALYSIS AND CONCLUSION

The results of DPSO for mining the Wisconsin breast cancer data are tested. To perform the robustness of methodology in this study, we have presented the 10 results of experiment and the relative parameters of the algorithm below. The number of particle was 30, the number of generation was 50, Cw ¼ 0:1; Cp ¼ 0:4 and Cg ¼ 0:9. The setting of parameters in DPSO was case dependent, and we can do the research on them in the future study. In rule 1, we could derive the best accuracy to be 0.9528 that Cell Shape Uniformities > 2 and Mild

Chromatin > 1". In our study, the data could not be classified correctly by rule 1, we adopted the method of Nahar, J., Imam, T., Tickle, K. S., Ali, A. S., & Chen, Y. P. P. (2012) that new decision rule is to be explored. In this process, both the selected feature of training data not being classified correctly and all the unselected feature of data are preserved for mining in an additional rule (Nahar, J., Imam, T., Tickle, K. S., Ali, A. S., & Chen, Y. P. P. (2012)). After the repeated process, we found that Rule 2 is ''Lump Thick- ness > 3 and Basic Core > 2". So far, this study utilized two rules to improve the accuracy to 98.71%. Table 7 shows the comparison results with Gas. We found that the proposed DPSO can enhance the accuracy by 1.28%. Table shows the sensitivity can be up to 100% and specificity can be up to 98.21%, respectively. The performance of Type I error in GAs and DPSO to be equivalent. According to the above results, the proposed DPSO had shown to be better than the GAs in enhancing the performance of Type II error by 4.58%. Table 11 has compared the results of previous research in Wisconsin breast cancer with the proposed DPSO.The best way to improve the breast cancer victim's chance of long-term survival is to detect it as early as possible. Data mining and statistical analysis is one of the good solutions for searching the valuable information in large volumes of data (Muro, N., Larburu, N., Bouaud, J., Belloso, J., Cajaraville, G., Urruticoechea, A., & Séroussi, B. (2017, June)).

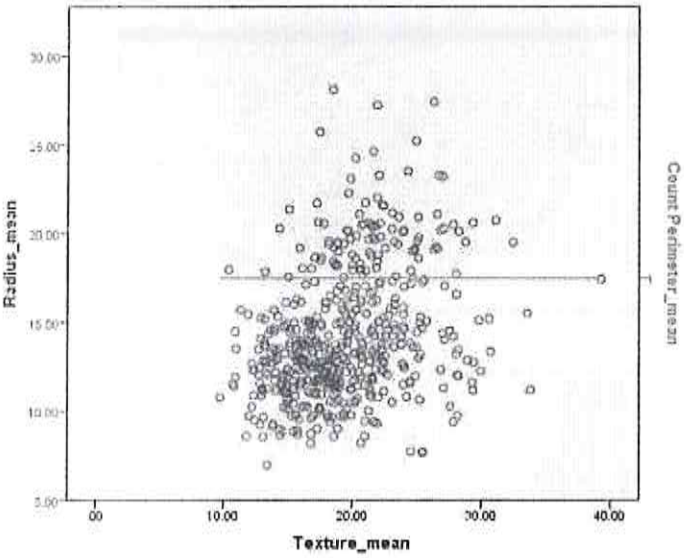


Fig. 2

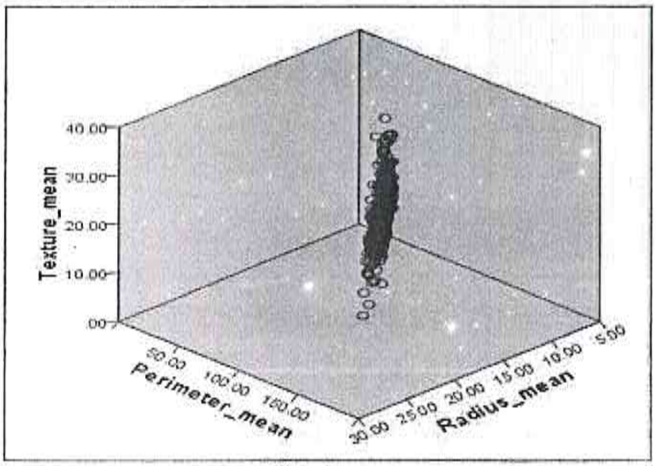


Fig. 3

In this study, a new hybrid approach of using both integrated statistical method and DPSO is proposed and successfully applied to the classification risk of Wisconsin- breast-cancer data set.

VII CONCLUSION

According to our testing results, the proposed hybrid approach can improve the accuracy to 96.25%, sensitivity to 100% and specificity to 96.32%. These results are very promising compared to the previously reported classification techniques for mining breast cancer data.

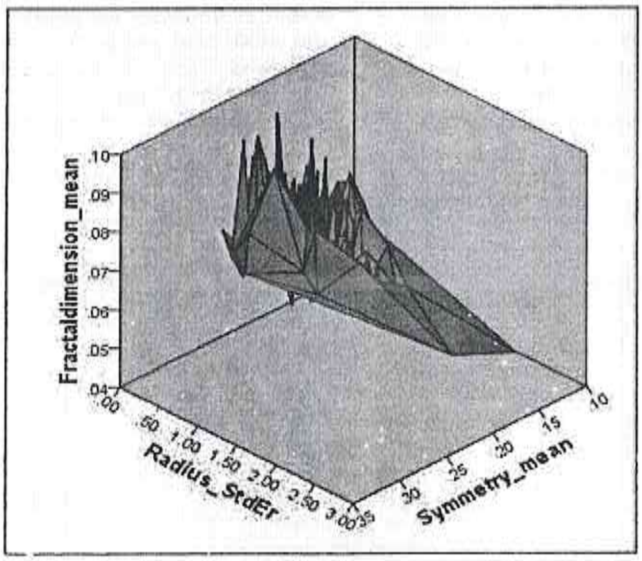


Fig. 4

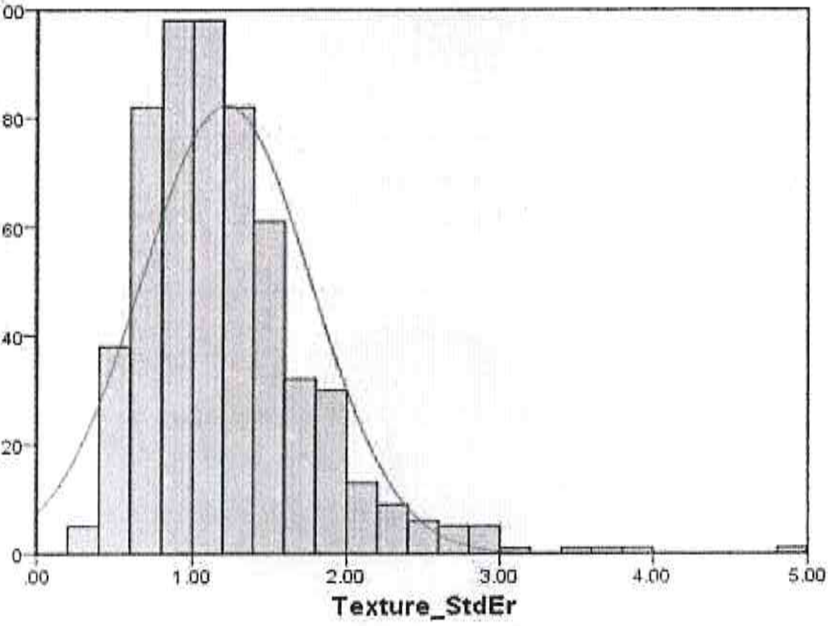


Fig. 5

Furthermore, the advantage of using statistical method to eliminate the insignificant featrres in pre-processing can improve the efficiency of DPSO process when the data set has included many features. Besides, the proposed DPSO can improve the constraint of genetic operation.

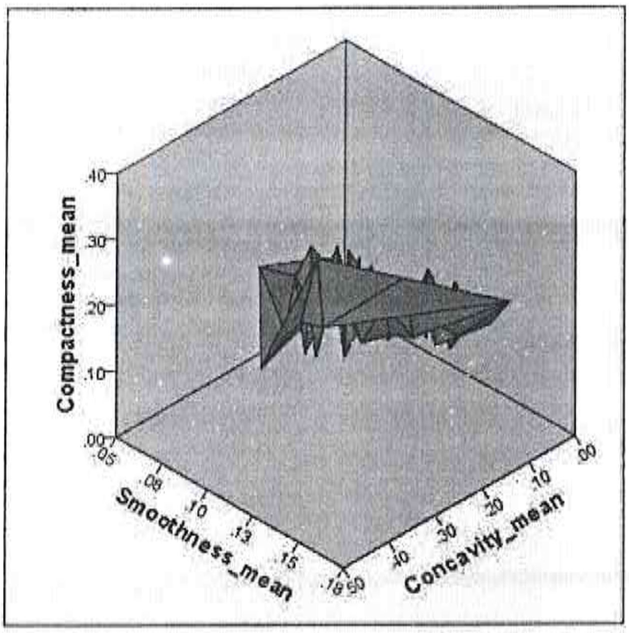


Fig. 6

The high classification accuracy from our proposed algorithm can be used as the reference for decision making in the hospital and the researchers. In future research, we recommend to better the process of data mining and apply it to the various domains in order to improve the medical quality for our lives.

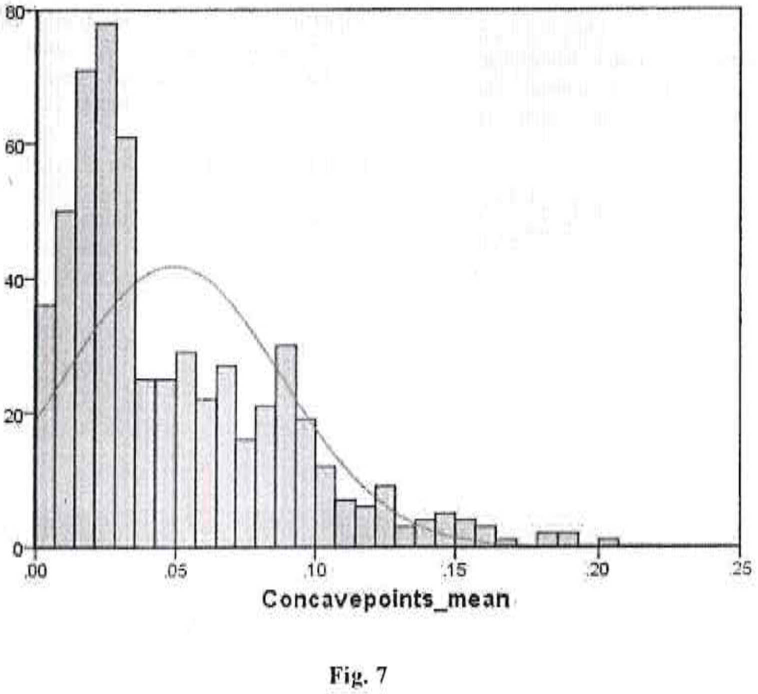


Fig. 7

REFERENCES

[1] Rutuparna Panda, Sanjay Agrawal, Sudipta Bhuyan, "Edge magnitude based multilevel thresholding using Cuckoo search technique.

[2] "Expert Systems with Applications 40 (2013) 7617–7628, 2013, Elsevier.

[3] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis, and Machine Vision, Thomson-Engineering,2007.

[4] R. Klette, Concise Computer Vision - An Introduction into Theory and Algorithms, Springer,2014.

[5] C. A. Floudas, P. M. Pardalos (Eds.), Encyclopedia of Optimization, Springer, 2009.

[6] D.-H. Li, M. Fukushima, On the Global Convergence of the BFGS Method for Nonconvex UnconstrainedOptimization Problems, SIAM Journal on Optimization 11 (4) (2000) 1054,1064.

[7] J. Grefenstette, Optimization of control parameters for genetic algorithms, IEEE Trans. Syst. Man Cybern.16 (1) (1986) 122–128.

[8] V. Zografos, Comparison of Optimisation Algorithms for Deformable Template Matching, in: Procs. Ofthe International Symposium on Advances in Visual Computing: Part II (ISVC),1097–1108, 2009.

[9] F. Glover, G. A. Kochenberger (Eds.), Handbook of metaheuristic, Kluwer Academic Publishers, 2003

[10] Sridevi, M, Mala, C. ; Sivasankar, E, "Optimized Multilevel Threshold Selection Using Evolutionary Computing,"17th International Conference on Network-Based Information Systems, IEEE, 2014.

[11] Paulo S. Rodrigues et al, "Improving a firefly meta-heuristic for multilevel image segmentation using Tsallis entropy", Pattern Anal Applications, 10044-015-0450-x, Springer, January 2015.

[12] Ye zhiwei et al, "Image Segmentation Using Thresholding and Artificial Fish-Swarm Algorithm", 2012 International

[13] Conference on Computer Science and Service System, 978-0-7695-4719-0/12, IEEE, 2012.

[14] V. Rajinikanth, N. Sri Madhava Raja, and K. Latha, "Optimal Multilevel Image Thresholding: An Analysis with PSO and BFO Algorithms", Aust. J. Basic & Appl. Sci., vol. 8, no. 9: pp. 443-454, 2014.

[15] Sathya, P.D. and Kayalvizhi, R. Optimal multilevel thresholding using bacterial foraging algorithm. Expert Systems with Applications, 38:15549 – 15564, 2011.

[16] Ming-Huwi Horng "Multilevel Thresholding selection based on the artificial bee colony algorithm for image segmentation "Expert Systems with applications, 2011.

[17] P.D. Sathya and R. Kayalvizhi, "Optimum Mutilevel Image Thresholding Based on Tsallis Entropy Method with Bacterial Foraging Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

Classification of High Time Resolution Universe Survey 2 Data by Machine Learning Technique

Pratibha Verma[1], Sanat Kumar Sahu[2], A.K. Shrivas[3]
[1]Ph.D. Scholar, Dr. C.V. Raman University, Bilaspur (C.G.) India.
[2]Dept. of Computer Science, Govt. Kaktiya P.G. College, Jagdalpur (C.G.) India.
[3]Dept. of IT, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

Machine learning techniques are useful to finding or discovering the new pulsar. We study the candidate filters used to moderate theseproblems for the period of past few years. Pulsars are types of star of huge scientific interest. We have used two classification techniques like Conditional Inference Tree (CTREE) and Classification and Regression Tree (CART) and their ensemble model (CTREE+CART) for classifying the HTRU2 dataset. The ensemble model (CTREE+CART) give the better performance compared to the Individual models of each classifiers. The ensemble model (CTREE+CART) is useful of the candidates must be classified in to pulsar and non-pulsar classes to aid discovery.

I INTRODUCTION

Magnetic high-speed neutron stars are known as pulsars. Whose linearly transmitted polarized electromagnetic radiation extends along its magnetic poles? While Pulsar, maintains the variance. It radiates from time to time the journeys towards the observer's vision, such as the rotating beacon, which is the result of a periodic train of narrowband radiation pulses, which was detected by a radio telescope[1], [2]. Pulsar studies involve accurate measurements of the arrival time of the pulse, followed by an appropriate modeling of the observed arrival times to study and understand the various phenomena that may influence arrival times. Machine learning techniques are useful to finding or discovering the new pulsar. Here we describe the classification techniques to decide pulsar and non pulsar candidates. Classification techniques CTREE and CART are tree based classifiers and their ensemble model is CTREE, CART. These are useful to predict pulsar candidate selection.

II RELATED WORK

Prior to the discovery of pulsar, many researchers have worked in the past. Some of its precise introductions are as follows.

Each candidate must be inspected by an automatic method like machine learning techniques and a human expert to determine its authenticity[3]. The process for deciding which candidates are worth investigating is known as 'selection of candidates' and called as "pulsar candidates"[4]. The authors [4] have presented a new model it selecting promising candidate using a purpose built in tree–based machine learning classifiers. With the help new approaches they have discovered 20 new pulsars.The authors [5]have explained the discovery of a new pulsar survey by using the Parkes Radio Telescope. The high time and frequency resolution of our digital backend system leads to increased sensitivity for short period, high-DM pulsars compared to previous surveys.

III METHODOLOGY

(a) **Classification:** Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels[6]. We can build a classification model to classify the dataset in to different groups or class. Classification has several types of applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

(i) **Conditional Inference Tree (CTREE):** In addition to traditional decision trees, conditional inference trees are another popular tree-based method. Like traditional decision trees, conditional inference trees also divide the data recursively by performing a univariate division in the dependent variable. However, what makes conditional decision trees have conditional inference trees is that conditional inference trees adapt meaning verification procedures to select variables instead of selecting variables maximizing information measures. In this way, we will present how to adjust a conditional inference tree to construct a classification [7].

(ii) **Classification and Regression Tree (CART):** CART adopt a greedy (i.e., non back tracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built [6], [8]. CART is classification and regression tree uses recursive partitioning to split the training records into subdivision with similar target field ideals using Gini index.

(iii) **Ensemble Model:**When two classification techniques like CTREE Tree and CARTcombined it is called hybrid or ensemble model[9].

IV RESEARCH DATA

The dataset was downloaded from UCI Machine Learning Repository. The HTRU (High Time Resolution Universe Survey) 2 dataset have total number of instance is 17898 with 1639 are positive instances and 16259 are negative instances. The total number of attributes (features) is 8 with an additional class label [10]. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve.

Table 1
Descriptions of HTRU 2 Data Set

Sl. No.	Attributes	Details
1	Profile_mean	Mean of the integrated profile
2	Profile_stdev	Standard deviation of the integrated profile
3	Profile_skewness	Skewness of the integrated profile
4	Profile_kurtosis	Excess kurtosis of the integrated profile
5	DM_mean	Mean of the DM-SNR curve
6	DM_stdev	Standard deviation of the DM-SNR curve
7	DM_skewness	Skewness of the DM-SNR curve
8	DM_kurtosis	Excess kurtosis of the DM-SNR curve
9	Class	Negative and Positive

V CLASSIFICATION FRAMEWORK PROCESS

The Figure 1 shows the process flow the used classifier like CTREE and CART and their ensemble model. The HTRU 2 dataset classify the classifier used under 10 folds cross validation techniques. The finally obtained performance of the classifier is accuracy, sensitivity and specificity.

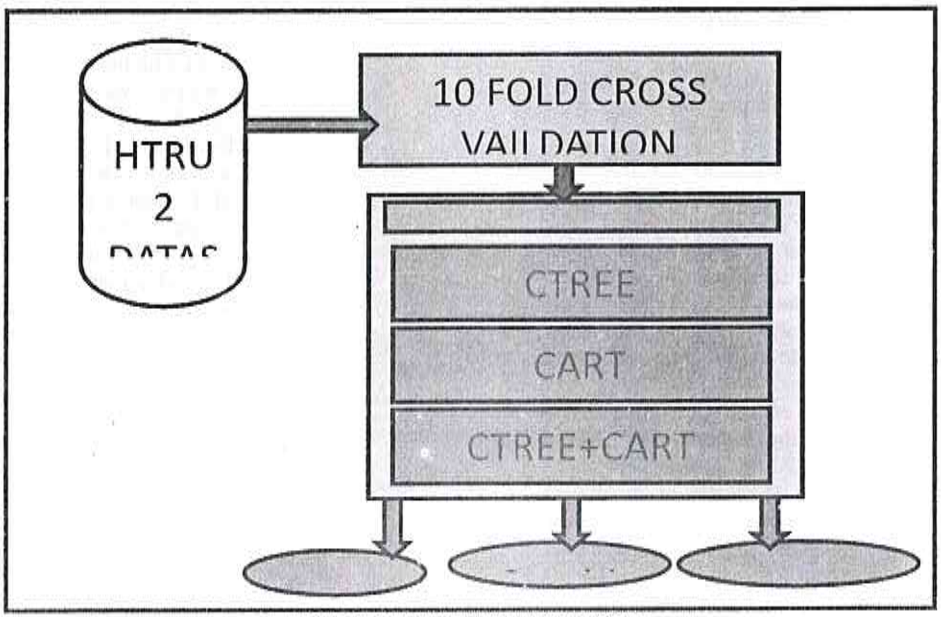


Fig. 1: Process Flow of Classifier

VI RESULT AND DISCUSSION

After loading the HTRU 2 [10]dataset, the input and output attributes are selected and given to the classifier models, one by one, the results are automatically derived and presented in the form of various performances measurement. However, we considered three performances measurement like accuracy, sensitivity and specificity, as these three measures clearly reflect the efficiency of the classification model as shown in Table 2. Among three techniques of the classification ensemble model (CTREE+ CART) is producing remarkable results.

Table 2
Comparison of Classification Performance

Sl. no.	Name of Algorithm	Accuracy	Sensitivity	Specificity
1	CTREE	97.77	84.19	99.13
2	CART	92.96	87.48	84.25
3	CTREE+CART	**97.84**	84.56	99.18

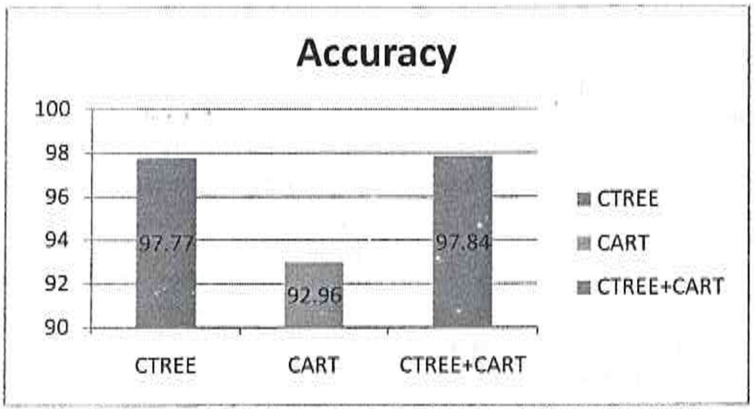


Fig. 2: Comparative Accuracy graphs of the different Classifier

The highest accuracy is obtained by ensemble model (CTREE+CART) compared to the individual models CTREE and CART.

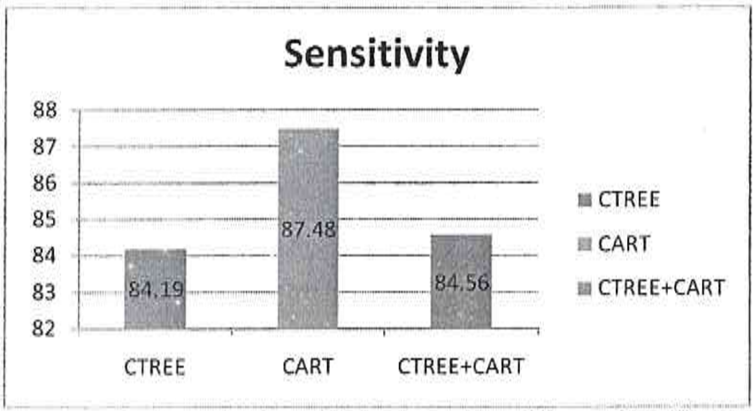


Fig. 3: Comparative Sensitivity graphs of the different Classifier

The figure 3 shows a fluctuation in the sensitivity of performance. CTREE gives 84.19% of sensitivity and CART gives 87.48% whereas ensemble model (CTREE+CART) gives 84.56% of sensitivity.

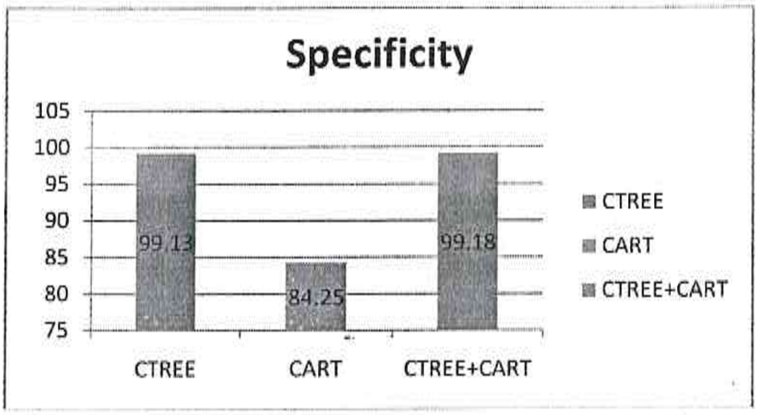


Fig. 4: Comparative graphs of the different Classifier

The maximum Specificity is obtained by ensemble model (CTREE+CART) compared to the individual models CTREE and CART.

VII CONCLUSION

The pulsar candidate can be very well predicted using many classifiers in data mining and machine learning. The purpose of this study was to analyze the application of data mining algorithms and machine learning in the HTRU2 dataset and to predict pulsar and non pulsar. In this paper, the pulsar candidate is predicted using two types of classification techniques CTREE and CART and their ensemble model. In ensemble model obtained the highest accuracy 97.84 % compared to other classification models.

There are other possible techniques for selecting the features that can be used to obtain better results from the ensemble model. Other classification techniques and the techniques for selecting features can be performed in the future.

REFERENCES

[1] S. , Saha, Aperture Synthesis Methods and Applications to Optical Astronomy. springer, 2011.

[2] P. S. Ramkumar and a. a. Deshpande, "Real-time signal processor for pulsar studies," J. Astrophys. Astron., vol. 22, no. 4, pp. 321–342, 2001.

[3] R. J. Lyon, "WHY ARE PULSARS HARD TO FIND ?," 2016.

[4] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles, "Fifty Years of Pulsar Candidate Selection : From simple filters to a new principled real-time classification approach," vol. 22, no. March, pp. 1–22, 2016.

[5] M. J. Keith et al., "The High Time Resolution Universe Pulsar Survey – I , System configuration and initial discoveries 1 I N T RO D U C T I O N 2 S I M U L AT I O N A N D S U RV E Y S T R AT E G Y," vol. 627, pp. 619–627, 2010.

[6] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third. Elsevier, 2012.

[7] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased Recursive Partitioning: A Conditional Inference Framework," J. Comput. Graph. Stat., vol. 15, no. 3, pp. 651–674, 2006.

[8] A. Pujari, Data mining techniques. Third. University press. 2013.

[9] S. Haykin, Neural Networks and Learning Machines. vol. 3. 2008.

[10] R. Lyon, "HTRU2," 2016. [Online]. Available: https://figshare.com/articles/HTRU2/3080389/1. [Accessed: 01-Dec-2017].

Feature Selection of High Dimensional Big Data of Gene Expression for Cancer Dataset

Prem Kumar Chandrakar[1], **A. K. Shrivas**[2]

[1] Dept. of Computer Science, Mahant Laxminarayan Das College, Raipur (C.G.) India.

[2] Dept. of CS & IT, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

Feature selection is an essential data preprocessing technique for such high-dimensional data classification tasks. Traditional dimensionality reduction approach falls into two categories: Feature Extraction (FE) and Feature Selection (FS). The microarray technology has capability to determine the levels of thousands of gene simultaneously in a single experiment. The major challenge to analyze gene expression data, with a large number of genes and small samples, is to extract disease-related information from a massive amount of redundant data and noise. Analysis of gene expression is important in many fields of biological research in order to retrieve the required information. As time progresses, the illness in general and cancer in particular have become more and more complex and complicated, in detecting, analyzing and curing. We know cancer is deadly disease. Cancer research is one of the major area of research in medical field. Predicting precisely of different tumor types is a great challenge and providing accurate prediction will have great value in providing better treatment to the patients. To achieve this, data mining algorithms are important tools and the most extensively used approach to achieve important feature of gene expression data and plays an important role for gene classification. Gene expression profiles, which represent the state of a cell at a molecular level, has greatpotential as a medical diagnosis tool. But compared to the number of genes involved, available trainingdata sets generally have a fairly small sample size for classification. These training data limitationsconstitute a challenge to certain classification methodologies. Feature selections techniques can be usedto extract the marker genes which influence the classification accuracy effectively by eliminating the unwanted noisy and redundant genes. One of major challenges is to discover how to extract useful information from huge datasets. Gene selection, eliminating redundant and irrelevant genes, has been a key step to address this problem.This paper presents a various of feature selection techniques that have been employed in micro array data based cancer classification and presents recent advances in the machine learning based gene expression data analysis with different feature selection algorithms.

Keyword–Gene Expression, Cancer Classification, Feature selection

I INTRODUCTION

Feature selection is an active research area in pattern recognition, statistics, and data miningcommunities. The main idea of feature selection is to choose a subset of input variables byeliminating features with little or no predictive information. Feature selection can significantlyimprove the comprehensibility of the resulting classifier models and often build a model thatgeneralizes better to unseen points. Further, it is often the case that finding the correct subset ofpredictive features is an important problem in its own right. For example, physician may make adecision based on the selected features whether an expensive surgery is necessary for treatmentor not.

II DNA MICROARRAY

Microarray technology is a developing technology used to study the expression of many genes atonce. It involves placing thousands of gene sequences in known locations on a glass slide calleda gene chip. A sample containing DNA or RNA is placed in contact with the gene chip.Complementary base pairing between the sample and the gene sequences on the chip produceslight that is measured. Areas on the chip producing light identify genes that are expressed in the sample.

Microarray technology provided an opportunity for the researchers to analyze thousands of geneexpression profiles simultaneously that are relevant to different fields including medicineespecially cancer. The categorization of patient gene expression profile has become a commonstudy in biomedical research. The real problem is managing microarray data with its dimension.Since the dimension of microarray is large, classifying and handling the algorithms becomes toocomplex to study the gene expression characteristics. Due to the presence of more improperattributes in the dataset, the accuracy of the classification algorithm also gets affectedsignificantly. The aim of feature selection algorithm is to isolate the most important featuresfrom the microarray data to minimize the feature space in order to improve the accuracy of theclassification.

A microarray gene expression data set can be represented in a tabular form. in which each rowrepresents to one particular gene, each column to a sample or time point, and each entry of thematrix is the measured expression level of a particular gene in a sample or time point.respectively.

DNA microarrays are created by robotic machines that arrange large amounts of hundreds orthousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellularmachinery begins to copy certain segments of that gene. The resulting product is known asmessenger RNA (mRNA), which is the body's template for creating proteins. The mRNAproduced by the cell is complementary, and therefore will bind to the original portion of theDNA strand from which it was copied.

To determine which genes are turned on and which are turned off in a given cell, a researchermust first collect the messenger RNA molecules present in that cell. The researcher then labelseach mRNA molecule by using a reverse transcriptase enzyme (RT) that generates acomplementary cDNA to the mRNA. During that process fluorescent nucleotides are attached tothe cDNA. The tumor and the normal samples are labeled with different fluorescent dyes. Next, the researcher places the labeled cDNAs onto a DNA microarray slide. The labeledcDNAs that represent mRNAs in the cell will then hybridize – or bind – to their syntheticcomplementary DNAs attached on the microarray slide, leaving its fluorescent tag. A researchermust then use a special scanner to measure the fluorescent intensity for each spot/areas on themicroarray slide.

If a particular gene is very active, it produces many molecules of messenger RNA, thus, morelabeled cDNAs, which hybridize to the DNA on the microarray slide and generate a very brightfluorescent area. Genes that are somewhat less active produce fewer mRNAs, thus, less labeledcDNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of themessenger molecules have hybridized to the DNA, indicating that the gene is inactive.Researchers frequently use this technique to examine the activity of various genes at differenttimes. When co-hybridizing Tumor samples (Red Dye) and Normal sample (Green dye)together, they will compete for the synthetic complementary DNAs on the microarray slide. As aresult, if the spot is red, this means that that specific gene is more expressed in tumor than innormal (up-regulated in cancer). If a spot is Green that means that that gene is more expressed inthe Normal tissue (Down regulated in cancer). If a spot is yellow that means that that specificgene is equally expressed in normal and tumor.

III RELATED WORK

Cancer is one of most deadly dieses and lot of people die world- wide because of cancer. As per the WHO statistics in 2018 more than 20 million new cases were identified and around 9.6million cancer related death occur. Globally, about 1 in 6 deaths is due to cancer.The number of new a case is expected to rise by about 70%over the next2 decades (source: WHO 2018). It has been identified long ago that cancer occurs because of genedisorder. Gene expression is nothing but level of production of protein molecules defined by agene. Monitoring of gene expression is one of most fundamental approach in genetics. Thetechnique for measuring gene expression is to measure the mRNA instead of protein, becausemRNA sequences hybridize with their complementary RNA or DNA sequence while thisproperty lacks in protein. The DNA arrays are novel technologies that are designed to measuregene expression of tens of thousands of genes in a single experiment. Gene expression datausually contain a large number of genes (in thousands) and a small number of experiments (indozens). In machine learning terminology, these data sets are usually of very high dimensionswith undersized samples. The purpose of Gene selection is to find a set of genes that bestdiscriminate biological sample of different types. The selected genes are "biomarkers," and theyform a "marker panel" for analysis. For analyzing the marker panel rank based scheme suchinformation gain was used. It was observed that the information gain with large group was notaccurate, therefore in paper(**Zhu, Wang, Yu, Li, & Gong, 2010**)they proposed model-based approach to estimate the entropy onthe model. instead of on the data themselves. Here, they used multivariate Gaussian generativemodels. which model the data with multivariate normal distributions.

IV FEATURE SELECTION METHOD

There are two types of feature selection methods have been studied: filter methods (**Langley, Flamingo, & Edu, 1994**) andwrapper methods(**Kohavi & John, 1997**).Filter methods are essentially data preprocessing or data filtering methods. Features are selectedbased on the intrinsic characteristics that determine their relevance or discriminative powers withregard to the target classes.In wrapper methods, feature selection is "wrapped" around a learning method: the usefulness ofa feature is directly judged by the estimated accuracy of the learning method. Wrapper methods typically require extensive computation to search for the best features.

(a) Basic featrre selection algorithm

(i) Input:

S - Data sample f with features X, |X| = n

J - Evaluation measure to be maximized

```
GS – successor generation operator
    (ii) Output:
Solution – (weighted) feature subset
L: = Start Point(X);
Solution: = {best of L according to J };
    (iii) Repeat
L: = Search Strategy (L, GS (J), X);
X`:= {best of L according to J};
If J (X') =J (Solution) or (J (X') =J (Solution) and
|X`| < |Solution|) then
Solution: =X';

    (iv) Until Stop (J, L).
```

The discriminating criteria are being used by filter method for feature selection. The correlationcoefficient or statistical test like t-test or f-test is used to filter the features in the filter featureselection method.Many interesting results were obtained by researchers aiming to distinguish between two or moretypes of cells (e.g., diseased versus normal, or cells with different types of cancers), based ongene expression data in the case of DNA microarrays. Since microarray data have large amountof data and attributes, which makes complex for researcher to do analysis. A small subset ofgenes is easier to analyze as opposed to the set of genes available in DNA microarray chips.Therefore it is important to focus on very few genes to give insight into the class association fora microarray sample. It also makes it relatively easier to deduce biological relationships amongthem as well as to study their interactions.In paper **(Shah, Marchand, & Corbeil, 2012)**they obtained feature selection algorithms for classification with tight realizableguarantees on their generalization error. The proposed approaches are a step toward which aremore general learning strategies that combine feature selection with the classification algorithm.and have tight realizable guarantees. They classified microarray data, where the attributes of thedata sample correspond to the expression level measurements of various genes was considered.They chosen decision stumps as learning bias, which is in part been motivated by thisapplication.(Banerjee, Mitra, Member, & Banka, 2007).In this paper they introduced an evolutionary rough feature selection algorithm for classifyingmicroarray gene expression pattern. Microarray data typically consist of large number ofredundant features; therefore an initial redundancy reduction of attributes was done to enablefaster convergence. The main aim was to retain only those genes that play a vital role indiscerning between objects. Rough set theory was employed to generate reducts, which represent the minimal sets of non redundant features capable of discerning between all objects, in amultiobjective framework.

V EXPERIMENTAL ANALYSIS

Lung cancer dataset was used to compare different filter based feature selection methods for the prediction of disease risks. Four classification algorithms reviewed above were considered to evaluate classification accuracy. The feature selection methods are

CSEBT-CfsSubsetEval_BestFirst
CSEGS- CfsSubsetEval_GeneticSearch
CLSEBFDT- ClassifierSubsetEval_BestFirst_
Decision Tree
GS- Greedy_stepwise
GSDT- GreedyStepwise_ Decision Tree
PCA- Principal Component Analysis
TRF- Tree RandomForet
TSC-Tree Simple Cart
TJ48-Tree J48
BBN-Bayes.BayesNet
BNB- Bayes.NaiveBayes
FRBFN-Function.RBFNetwork
FMLP-Function.MultilayerPerceptron

At first, feature selection methods were used to find relevant features in the lung cancer dataset and then, classification algorithms were applied to the selected features to evaluate the algorithms. Same experiment was repeated for four classifiers. WEKA 3.6.8 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is an open source software. The software contains tools for data pre-processing, feature selection, classification, clustering, association rules and visualization. ome performance measures were used for the evaluation of the classification results, where TP/TN is the number of True Positives/Negatives instances, FP/FN is the number of False Positives/Negatives instances. Precision is a proportion of predicted positives which are actual positive:

The following table show the experimental result of gene expression data set . Result show performance of various Attribute selection mode.

Cancer Data Set
Name:- Brain Tumour (Malignant glioma types)
Instances: 50
Attributes: 10368

Table 1
Attribute selection Performance

Sr. No	Evaluation Algorithm	Evaluator	Parameters Tuning	Atrribute Selection Mode	Evaluation mode
1	Attribute Subset Evaluator	CFS Subset Evaluator	Best first Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 764460 Merit of best subset found: 0.996	Including locally predictive attributes	Evaluate on all training data
			Greedy Stepwise (forwards) Start set: no attributes Search direction: forward Merit of best subset found: 0.996	Including locally predictive attributes	Evaluate on all training data
			Genetic search Start set: no attributes Population size: 20 Number of generations: 20 Probability of crossover: 0.6 Probability of mutation: 0.033 Report frequency: 20 Random number seed: 1	Including locally predictive attributes	Evaluate on all training data
			Linear Forward Selection Start set: no attributes Forward selection method: forward selection Stale search after 5 node expansions Linear Forward Selection Type: fixed-set Number of top-ranked attributes that are used: 50 Total number of subsets evaluated: 11148 Merit of best subset found: 0.968	Including locally predictive attributes	Evaluate on all training data
	Attribute Subset Evaluator	Classifier Subset Evaluator	Best first Classifier-ZeroR Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 82914 Merit of best subset found: 152.942	Including locally predictive attributes	Evaluate on all training data
		Classifier Subset Evaluator	Genetic search Classifier-ZeroR Start set: no attributes Population size: 20 Number of generations: 20 Probability of crossover: 0.6 Probability of mutation: 0.033 Report frequency: 20 Random number seed: 1	Including locally predictive attributes	Evaluate on all training data

Sr. No	Algorithm	FST	Total Number of Features Brain Tumour (Malignant glioma types)	Selected Features
1	CFS Subset Evaluator	Best first	10368	99
2	CFS Subset Evaluator	Greedy Stepwise	10368	95
3	CFS Subset Evaluator	Genetic search	10368	4148
4	CFS Subset Evaluator	Linear Forward Selection	10368	39
5	Classifier Subset Evaluator	Best first/Decision Table	10368	04
6	Classifier Subset Evaluator	Genetic search	10368	1484

VI RESULTS

Cancer dataset was used to compare different feature selection methods for the prediction of disease risks. Six feature selection techniques are usedwith classification algorithms. CFS Subset Evaluator with Genetic search is performed better result as compare to other feature selection algorithm.

VII CONCLUSION

This feature selection algorithms shows that the feature selection algorithmconsistently improves the accuracy of the classifier. Each feature selection methodology has itsown advantages and disadvantages. Each algorithm has different behavior which shows thatusing single algorithm for different dataset is infeasible. The feature selection algorithms are onewhich decides the accuracy of the classification of different datasets. The feature selectionalgorithm must select the relevant features and also remove the irrelevant and inconsistentfeatures which cause the degradation of accuracy of the classification algorithms. Featureselection algorithm is playing a major role in accurate classification of large data set like geneexpression. Therefore proper cancer classification can be achieved using feature selectionalgorithms, and on time and accurate treatment may be provided to the patients.

REFERENCES

[1] Banerjee, M., Mitra, S., Member, S., & Banka, H. (2007). Evolutionary Rough Feature Selection in Gene Expression Data, 37(4), 622–632.

[2] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

[3] Langley, P. A. T., Flamingo, L., & Edu, S. (1994). Selection of Relevant Features in Machine Learning, 127–131.

[4] Shah, M., Marchand, M., & Corbeil, J. (2012). Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data, 34(1), 174–186.

[5] World health organization (2018). Cancer (https://www.who.int/en/news-room/fact-sheets/detail/cancer)

[6] Zhu, S., Wang, D., Yu, K., Li, T., & Gong, Y. (2010). Feature Selection for Gene Expression Using Model-Based Entropy, 7(1), 25–36.

A Rule Based Sentiment Analysis System for Hindi Language

Pritendra Kumar Malakar
Research Scholar, Center for Information & Language Engineering,
MGAHV, Wardha (M.S.) India.

ABSTRACT

Hindi is one of the most spoken languages of the world. Today Hindi language users has better input mechanism to express their sentiments easily on Social Media, so a large volume of User Generated Content in Hindi are digitally store on the internet. It is being seen as an important source of information, but no computational system has yet been available to analyzing these contents. A Sentiment Analysis system has been proposed to solve this problem that analyzes these Hindi contents automatically. The basic principles of Software Engineering and Natural Language Processing have been implemented to design this system. It is a rule based system that follows some linguistics rules to classify any input text into Positive, Negative or Neutral. To evaluate this system, a dataset of 4000 sentences has been created by compiling User Generated Content from Twitter and e-Newspapers. The accuracy to Polarity Classification of the system for Known and Unknown dataset has been measured about 69% and 52%, respectively

Keywords: Hindi, User Generated Content, Sentiment Analysis, Bhav Vishleshak, Natural Language Processing

I USER GENERATED CONTENT IN HINDI

Hindi is one of the most spoken languages of the world. Approximately 4.70 % of world's population uses Hindi for their daily life communication [1]. According to a study by KMPG in India and Google, the total number of Hindi language users on the Internet is 254 millions [2]. In view of this scope, encoding standard and linguistic tools have been developed to support Hindi on the internet which has empowered Hindi users to express their sentiments on Social Media. Today Hindi users have been expressing their sentiments towards any subject regularly, so a large volume of User Generated Contents are digitally available on the internet in Hindi. It is being seen as an important source of information by Government, Business Organizations or Individuals to facilitate their decision making processes [3-5].

II PROBLEM STATEMENT

As discussed, a huge amount of User Generated Contents in Hindi available on the Internet, but it brings many serious challenges when it comes to analyze these contents manually. The manual analysis of the contents requires more time and effort that is very complex and tedious task. No computational system has yet been developed to analyzing these Hindi contents. Although a little but important works has been conducted to Sentiment Analysis for Hindi contents. Some of the major work is following:

(a) Akshat Bakliwal and Piyush Arora (IIIT Hyderabad) developed a Hindi Subjective Lexicon of all possible and closely related Synonyms and Antonyms words. They have performed n-Gram Modeling and Machine learning technique to analyze the sentiments from the text [3].

(b) Aditya Joshi, Balamuraly and Pushpak Bhattacharya (IIT Bombay) using **H-SWN** (Hindi-SentiWordNet) in which all sentimental word is classifying into Positive and Negative class with a fixed numerical score [4].

To overcome this problem a Sentiment Analysis system has been developed for Hindi. The detailed description about system development and working procedure has been given below.

III BHAV VISHLESHAK: A SENTIMENT ANALYSIS SYSTEM FOR HINDI

Bhav Vishleshak is a computational system developed to Sentiment Analysis of Hindi contents. The system is mainly designed to classify the given piece of text into Positive, Negative or Neutral automatically. Bhav Vishleshak works only on those Hindi contents that has been written in Unicode based Devanagari Script (Such as-Mangal and Kokila font).

(a) **Principles and Technologies used-** The basic principles of Software Engineering are applied to develop this system. All the functions of Bhav Vishleshak have been defined according to Natural Language Processing. C#.Net is used to design and code the system using Microsoft Visual Studio 2008. To create the database of the system MS-Access 2007 has been used.

(b) **Structure of the system**

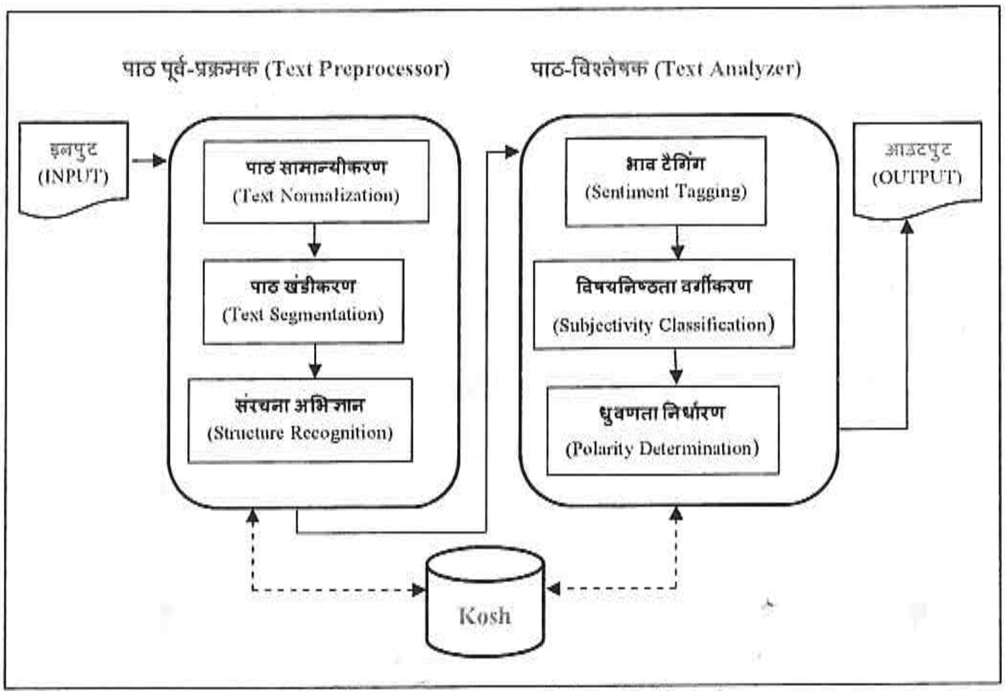

पाठ पूर्व-प्रक्रमक (Text Preprocessor)
पाठ-विश्लेषक (Text Analyzer)
इनपुट (INPUT)
पाठ सामान्यीकरण (Text Normalization)
पाठ खंडीकरण (Text Segmentation)
संरचना अभिज्ञान (Structure Recognition)
भाव टैगिंग (Sentiment Tagging)
विषयनिष्ठता वर्गीकरण (Subjectivity Classification)
धुवणता निर्धारण (Polarity Determination)
आउटपुट (OUTPUT)
Kosh

Fig. 1: System Structure of Bhav Vishleshak

Above figure shows the internal structure of the Bhav Vishleshak. Two components namely Text Preprocessor and Text Analyzer have been used in Bhav Vishleshak. These are depicted by the rounded rectangle in the figure. Some NLP based processes have been followed by these components to execute these functions. These processes are displayed by the rectangle inside the components. The order of the execution of these processes is sequential. After the completion of one process, the next process begins. It is represented by the arrow in the figure. The Database of the system namely **'Kosh'** is represented by the cylinder shape in the figure. Both the components associated bi-directionally with the database that is depicted by dashed lines.

(c) **Text Preprocessor-**It is not necessary that input text always be acceptable by the system. Sometimes system is unable to process the text different from standard format. Therefore input is firstly converted into system understandable format by the Text Preprocessor. Following three processes have been used under the Text Preprocessor:

(i) **Text Normalization-**Text Normalization is a process to remove the undesired elements from the input text in terms of Sentiment Analysis. Various types of symbols, logos, URLs, emoticons, special characters, etc. are used by Social Media users in their expressions which may be affects the accuracy and quality of Sentiment Analysis. Therefore all the unnecessary elements are removed from the text through Text Normalization.

(ii) **Text Segmentation-**Since Bhav Vishleshak is mainly designed for Sentence level Sentiment Analysis so here Text Segmentation has been used to split the text at sentence level. Initially the input text is divided into sentences and each sentence analyzed separately.

(iii) **Structure Recognition-**Structure Recognition is a process defined under the Text Preprocessor to identify the structure of the sentences. The structure of sentences is either simple or non-simple. A sentence having more than one sub-sentence is called non-simple sentence. All the sentences are tagged as simple and non-simple. The purpose of finding the structure of sentences here is to identify the sub-sentences used in the sentences. so that they can be analyzed separately as a sentence.

(d) **Text Analyzer-**This is the second and most important component of the system. To analyze the text it uses following three processes sequentially:

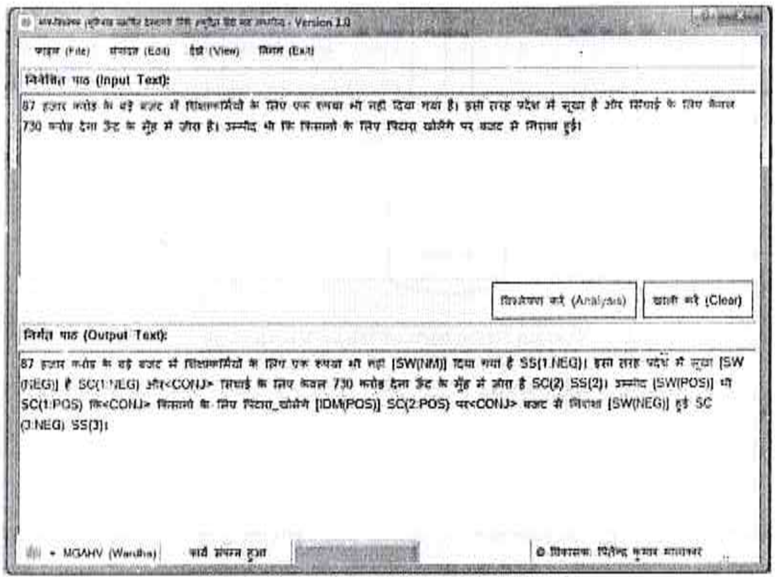
(i) **Sentiment Tagging-**Sentiment Analysis mainly works on the basis of sentimental words and idioms presented in the text. These are the essential elements for doing Sentiment Analysis, so here Sentiment Tagging has been used to identify all the sentimental words and idioms in the text.

(ii) **Subjectivity Classification**-It is a process used to classify the sentences into Subjective or Objective. A sentence that contains any sentimental words and/or idioms is called Subjective otherwise it is called Objective.

(iii) **Polarity Determination**-Polarity Determination is a process used to decide the polarity of each subjective sentence presented in the text. Subjective sentence is classified as Positive, Negative or Neutral by this process. It follows some predefined classification rules to determine the polarity of the text

(e) **Kosh**-It is a database created under the system. It stores all the information necessary for the system.

(f) **Interface of the system**



87 हजार करोड़ के बड़े बजट में शिक्षाकर्मियों के लिए एक रुपया भी नहीं दिया गया है। इसी तरह प्रदेश में सूखा है और सिंचाई के लिए केवल 730 करोड़ देना ऊँट के मुँह में जीरा है। उम्मीद थी कि किसानों के लिए पिटारा खोलेंगे पर बजट से निराशा हुई।

87 हजार करोड़ के बड़े बजट में शिक्षाकर्मियों के लिए एक रुपया भी नहीं [SW(NM)] दिया गया है SS(1:NEG)। इसी तरह प्रदेश में सूखा [SW(NEG)] है SC(1:NEG) और<CONJ> सिंचाई के लिए केवल 730 करोड़ देना ऊँट के मुँह में जीरा है SC(2) SS(2)। उम्मीद [SW(POS)] थी SC(1:POS) कि<CONJ> किसानों के लिए पिटारा_खोलेंगे [IDM(POS)] SC(2:POS) पर<CONJ> बजट से निराशा [SW(NEG)] हुई SC(3:NEG) SS(3)।

Fig. 2: Interface of Bhav Vishleshak

The above figure shows the main interface of the system in which the names of all the controls for the convenience of users are kept in Hindi and English. Different options have been provided to the user through menu displayed below the titlebar in the interface. The interface contains two textboxes; one to receive input from the user and second one to display output to the user. There are two buttons are used in the interface; one for analyze the input text and second to clear the input box.

(g) **Evaluation of the system**-The Bhav Vishleshak has been evaluated on Known and Unknown Dataset.

(i) **Known Dataset**: A Dataset that is referred to develop any system is called Known Dataset. All the sentences are analyzed to Rule Derivation and Database Creation for the system. Since the system is developed on the basis of such dataset, it is treated as Known Dataset.

(ii) **Unknown Dataset**: A Dataset that is not referred to develop any system is called Unknown Dataset. Normally systems produce higher accuracy on the Known Dataset, so to make system more versatile it should also be evaluated on the Unknown Dataset.

To evaluate the Bhav Vishleshak, a dataset of 4000 sentences is created. The Dataset is divided into Known and Unknown dataset. Each dataset is organized into 10 different sets. In order to maintain the balance between sets, the equal number of sentences has been kept in each set. There are 200 sentences in each set, in which there are 100 positive and 100 negative sentences. The following formulas have been used to calculate the accuracy percentage of the Bhav Vishleshak:

Formula 1: (To SUs Identification)

$$= \left(\sum \text{Identified (SUs)} / \sum \text{Total (SUs)}\right) \times 100$$

Where, SUs means Sentimental Units **Formula 2:** (To Polarity Classification)

= (Number of correctly classified Sentences / Total Sentences) × 100

Table 1
Accuracy Measurement Table for Known Dataset

S.No.	SET	Accuracy (%)	
		SUs Identification	Polarity Classification
1.	A	60	60
2.	B	90	90
3.	C	70	80
4.	D	80	90
5.	E	40	30
6.	F	90	60
7.	G	50	70
8.	H	70	80
9.	I	40	90
10.	J	80	100
Avg. Acc.		**61**	**69**

Table 2
Accuracy Measurement Table for Unknown Dataset

S.No.	SET	Accuracy (%)	
		SUs Identification	Polarity Classification
1.	K	70	80
2.	L	60	60
3.	M	60	50
4.	N	50	60
5.	O	80	70
6.	P	90	60
7.	Q	70	80
8.	R	30	60
9.	S	40	40
10.	T	10	40
Avg. Acc.		**49**	**52**

Comparison of the accuracy to Polarity classification on the known and unknown dataset of Bhav Vishleshak is shown in the following graph:

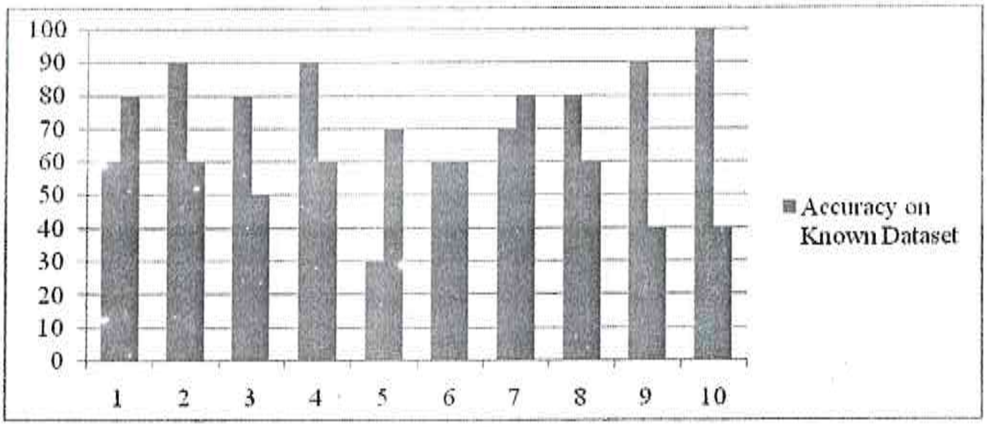


Fig. 3: Comparison graph for known and unknown dataset.

IV CONCLUSION

The above system is an attempt towards Sentiment Analysis of Hindi content. Since it mainly works on predefined rules and knowledge, so its accuracy and quality are limited. If the number of rules and the size of the database are increased, the system will produce better results. To make the system more effective, it has to be customized according to Unknown Datasets. The system is to be also useful for Text Analysis, POS Tagging, Language Teaching, etc.

REFERENCES

[1] Sharma, R., & Bhattacharya, P. (2014). A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon. *ICON 2014*: http://ltrc.iiit.ac.in/icon/2014/proceedings.php.

[2] https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

[3] Arora, Piyush. (2013) Sentiment Analysis for Hindi Language (MS Thesis), IIIT Hyderabad.

[4] Joshi, A., R, Balamuraly A R., & Bhattacharyya, P. (2010). A Fall-back Strategy for Sentiment Analysis in Hindi:a Case Study. *Proceedings of ICON 2010: 8th International Conference on Natural Language Processing.* Hyderabad: Macmillan Publishers.

[5] Sharma, R., Nigam,S. & Jain,R. (2013). Opinion Mining In Hindi Language: A Survey. *IJFCST*, Vol.4, No.2.

[6] Pang B., and Lillian Lee ((2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2): 1–135.

Model Specification on ICT based Online Admission Practices: A Study on Higher Educational Institutions in the State of Chhattisgarh

Pushkar Dubey[1], Deepak Kumar Pandey[2], Neelam Sahu[3]

[1]Asst. Prof. & Head, Dept. of Management, Pandit Sundarlal Sharma (Open) University, Bilaspur (C.G.) India.
[2]Research Scholar, Dept. of IT & Computer Application, Dr. C.V. Raman University, Bilaspur (C.G.) India.
[3]Associate Professor, Dept of IT &Computer Application, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

ICT has wide range of applications in the modern organizations. In terms of University admission system usages of ICT plays a prominent role. The objective of the study is to build a model specification for online admission system in the state of Chhattisgarh. The present study examined the effect of predicting factor resource utilization and value for money, ease of use, student's satisfaction, information access, safety and security and transparency and efficiency of ICT based online admission practices on student's satisfaction (criterion variable). Demographic factors like age, income, gender, occupation and family type were the controlled variable in the study. Hierarchical multiple regression analysis (step wise method) was used for model specification with the help of SPSS v 24 (licensed) and AMOS v24 (licensed). The model specified as per fit signifies that age of the respondents and fathers occupation have a significant effect on the ICT based online admission process, where as easy medium of access of the online admission system, accessibility of information of admission and safety and security features involved in online admission system significantly effects the satisfaction of students admitted through online mode. Result indicates the explaining percentage of all predictors was 42.7%; this total of the variance included 3.3% for age, 1.2% for occupation, 32.1% for ease of use, 4% for information access and 2.1% for safety and security. This model can be generalized for higher educational institutions for ICT based online admission practices in the state of Chhattisgarh.

Key Words: ICT, online admission system, hierarchical multiple regression, for higher educational institutions

I INTRODUCTION

Learning is a never-ending process, which shapes the personality of an individual. Learning occurs in all stages of life beginning from childhood stage, pre-schooling stage to schooling stage than to higher education. Institutions providing learning plays an important role in development of individual's identity. Higher educational institutions in this regards is plays a pivotal role. The enrolment process of higher educational institutions in past has been lengthy and time consuming since it was based on human processes. Of late, there has been transformation in the admission process, which is replaced by ICT based online admission process. It has become more technology centric and minimizes human effort. It is more transparent medium of admission, which eliminates favouritism, increases reliability and enhances fairness. (Turkle, 1997; Terrell &Dringus, 2000; Robinson &Hullinger, 2008; Forsyth, 2014).All the higher educational institutions in the state of Chhattisgarh are practicing online admission practices. ICT based online admission practices are influences by many factors. Some factors, which were identified in the literature, include resource utilization and value for money, ease of use, student's satisfaction, information access, safety and security and transparency and efficiency. The present study deals with proposing a model, which finds the best fit as per the dimensions, considered for the study in the higher educational institutions in the state of Chhattisgarh.

In the present study all students enrolled in higher education in the state of Chhattisgarh through online admission process were considered to be the entire population of the study. Whereas the element of the study constitute of the individual students. A sample is a small percentage of a population selected for study. It is a finite part of a statistical population whose properties are studied to gain information about the whole (Webster, 1985). Since the population is large and is spread across geographical territories the researcher decided to have 400 respondent students from higher educational institutions as the sample size of the study. Used in the present study. The scale contains 26 items with six broad dimensions namely resource

II METHODOLOGY

Correlation design is adopted in the study as it helps researchers to ascertain relationship between two closely connected variables (Fraenkel, Wallen& Hyun, 2011).Purposive sampling technique was used in the study for selecting a sample from the population. The inclusion criteria for identification of the respondent students were based on two criteria. First the participants must have enrolled in the first year diploma, undergraduate and post graduate programme of university and second the respondent student must have some sort of knowledge regarding online admission process. Usage of ICT in admission process of university/ higher educational institution questionnaire developed by Dubey et al. (2018) is utilization and value for money, ease of use, student's satisfaction, information access, safety and security and transparency and efficiency. The scale has five-point rating scale ranging from strongly agree. agree. neutral. disagree and strongly disagree with numerical notation of 5,4,3,2 and 1 respectively.

III PROCEDURE/ METHODS OF DATA COLLECTION

The study was conducted from May to August 2018. For collecting the primary data the pre designed self structured questionnaire was administered to the students and university officials of higher educational institutions in the state of Chhattisgarh. Direct personal interview method was used for obtaining information from the participants under study. During the initial phase of data collection a pilot testing was made to test the adequacy of the questionnaire. The inclusion criteria used for selection of prospective participants includes those students who took admission in first year of diploma, undergraduate and post graduate programmes of the university and were known or familiar to the online process of admission.

IV DEMOGRAPHIC PROFILE OF THE RESPONDENTS

Table 1 presents the demographic profile of the respondent (students) under study. The demographic profile of the respondents (students) which are included in the study comprises of age, gender, fathers occupation, family type, yearly income of the family, area of residence. programme of admission. admission stream. A total of 400 respondent students of eight Universities were grouped into the following categories:

Table 1
Demographic profile of the Respondent Students (N=400)

Variables	Frequency	Percent	Mean	SD	95% CI
Age (Years)					
15-20	159	39.7	24.65	6.286	24.03-25.27
21-30	176	44.0			
31-40	59	14.8			
Above 40	6	1.5			
Gender					
Male	291	72.7			1.23-1.32
Female	109	27.3			
Fathers Occupation					
Service	164	41.0			1.95-2.15
Business	93	23.2			
Agriculture	102	25.5			
Others	41	10.3			
Yearly Income					
High	110	27.5	869439	54655	758323-924375
Medium	167	41.7			
Low	123	30.8			
Family Type					
Nuclear	260	65.0	4.99	1.91	4.80-5.17
Joint	140	35.0			
Area of Residence					
Village	76	19.0			2.10-2.24
Town	179	44.7			
City	145	36.3			
Programme					
Diploma	105	26.3			1.91-2.05
Undergraduate	197	49.2			
Postgraduate	98	24.5			
Stream					
Arts	78	19.5			2.62-2.82
Commerce	43	10.8			
Science	191	47.7			
Others	88	22.0			

V ANALYSIS AND RESLUTS

The objectives of the study which aim to build a model specification for online admission system in the state of Chhattisgarh. Hierarchical multiple regression analysis (step wise method) was used to examine the effect of predicting factor resource utilization and value for money, ease of use. student's satisfaction, information access, safety and security and transparency and efficiency of ICT based online admission practices on students satisfaction (criterion variable).

Demographic factors like age, income, gender, occupation and family type were the controlled variable in the study.

VI ANALYSIS

Hierarchical multiple regression(Wampold, & Freund, 1987;Scialfa, & Games, 1987;Seibold, &Mcphee, 1979; Schafer, 1991Berry, 1993; Weisberg, 2005; Gelman& Hill, 2006; Cohen, West & Aiken, 2014) analysis was computed taking into consideration, the composite scores of the response on the predicting variables (i.e. resource utilization and value for money, ease of use, student's satisfaction, information access, safety and security and transparency and efficiency) separately with their respective dimensions. The entire controlled variable and the predicting variable was entered into the SPSS v25 with student's satisfaction as the criterion variable. In each hierarchical multiple regression analysis first control variable age, gender, income, occupation and family type were entered in to the equation. Next the respective variables with their composite scores were entered into the equation. The output generated from the analysis showed five different models which were found significant.

VII RESULTS

The results of the hierarchical multiple regression analysis for the composite scores of the independent variables (i.e. i.e. resource utilization and value for money, ease of use, student's satisfaction, information access, safety and security and transparency and efficiency) and controlled variable (i.e. age, gender, income, occupation and family type) are presented in table 2.

In model 1, age made significant contribution in variation of the students satisfaction F (1, 398) = 13.648, p<0.01) and explained 3.3 % of the variance in students satisfaction (R = 0.182, $\Delta R^2 = 0.033$). The standardised beta value ($\beta = 0.182$, $p<0.01$) indicated significant positive association between predictor age of the respondents and students satisfaction level; it means higher age groups have more satisfaction towards online admission process.

In model 2, occupation of father made significant contribution in variation of the students satisfaction (ΔF (1, 397) = 9.339, $p<0.01$). The introduction of factor occupation explained additional 1.2% variance in students satisfaction with overall 4.5% (R = 0.213, $\Delta R^2 = 0.012$). The predictor occupation was found to have significant positive association ($\beta = 0.110$, $p<0.01$) with students satisfaction; which indicates that fathers occupation of the respondent students is linked to students satisfaction level of the students.

In model 3, ease of use made significant contribution in variation of students satisfaction (ΔF (1, 396) = 76.309, $p<0.01$) and explained overall 36.6% of variance in students satisfaction (R = 0.605, $\Delta R^2 =$ 0.321); the model explained additional 32.1% of the variance in students satisfaction. The results indicated significant positive association between predictor ease of use of online admission process on students satisfaction ($\beta = 0.572$, $p<0.01$); that means higher the ease of use of online admission process higher will be the students satisfaction in using online medium of admission.

In model 4, factor information access made significant contribution in variation of the students satisfaction (ΔF (1, 395) = 67.694, $p<0.01$). The introduction of factor information access explained additional 4% variance in students satisfaction with overall 40.7% (R = 0.638, $\Delta R^2 = 0.040$). The predictor occupation was found to have significant positive association ($\beta = 0.229$, $p<0.01$) with students satisfaction: which indicates that with the rise in the information access of online admission process the satisfaction level of the students also increases.

In model 5, factor safety and security made significant contribution in variation of students satisfaction (ΔF (1, 394) = 58.821, $p<0.01$) and the model explained additional 2.1% of the variance in students satisfaction (R = 0.654, $\Delta R^2 = 0.021$). The overall variance of the model was found to be 42.7%. The results indicated significant positive association between predictor safety and security factor of online admission process on students satisfaction ($\beta = 0.164$, $p<0.01$); that means higher the safety and security features of online admission process higher will be the students satisfaction in using online medium of admission.

Findings clearly indicated that the control variable gender income and family type did not make any significant variation in student's satisfaction in online admission process. In addition resource utilization and value for money and transparency and efficiency did not contribute significantly in the variation of student's satisfaction on online admission process.

Result indicates the explaining percentage of all included 3.3% for age, 1.2% for occupation, 32.1% for ease of use, 4% for information access and 2.1% for safety and security.

Variance inflation factor (VIF) found, ranged from 1.000 to 1.474, which was distant from the 1.0 to 3.0,

The model specified as per fit is presented in figure 1. The implications drawn from the model signifies that age of the respondents and fathers occupation have a significant effect on the ICT based online admission process, where as easy medium of access of the online admission system, accessibility of information

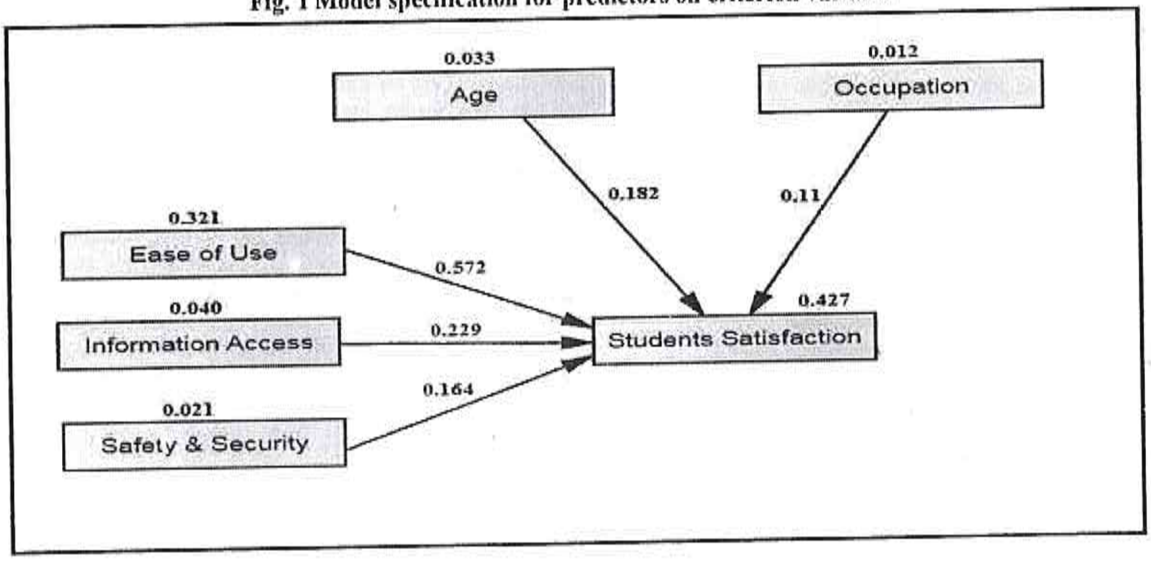
predictors was 42.7%; this total of the variance criteria that may indicate multicollinearity concern (O'brien, 2007). It means that multicollinearity found significant correlation between all predicting variables.

of admission and safety and security features involved in online admission system significantly effects the satisfaction of students admitted through online mode. This model can be generalized for higher educational institutions for ICT based online admission practices in the state of Chhattisgarh.

Table 2
Result of hierarchical multiple regression analysis

Predictors	Model 1			Model 2			Model 3			Model 4			Model 5		
	β	t	VIF	B	t	VIF	β	t	VIF	β	t	VIF	β	t	VIF
Age	.182	3.69	1.000	.192	3.892	1.007	.113	2.791	1.027	.100	2.537	1.031	.084	2.149	1.044
Occupation				0.110	2.239	1.þ07	.085	2.112	1.009	.113	2.873	1.029	.087	2.217	1.061
Ease of Use							.572	14.166	1.020	.468	10.627	1.290	.406	8.776	1.474
Information Access										.229	5.185	1.300	.213	4.877	1.313
Safety & Security													.164	3.775	1.298
R	0.182			0.213			0.605			0.638			0.654		
R^2	0.033			0,045			0.366			0.407			0.427		
ΔR^2	0.033			0.012			0.321			0.040			0.021		
ΔF	F(1,398) = 13.648**			ΔF (1,397) = 9.399**			ΔF (1,396) = 76.309**			ΔF (1,395) = 67.694**			ΔF (1,394) = 58.821**		

Fig. 1 Model specification for predictors on criterion variable



VIII CONCLUSION

ICT based online admission system has replaced the traditional means of admission in the state of Chhattisgarh. The prevailing system of online process of admission in the state satisfies the students at large. The implications drawn from the specified model signifies that age of the respondents and fathers occupation have a significant effect on the ICT based online admission process, where as easy medium of access of the online admission system, accessibility of information of admission and safety and security features involved in online admission system significantly effects the satisfaction of students admitted through online mode. This model can be generalized for higher educational institutions for ICT based online admission practices in the state of Chhattisgarh.

REFERENCES

[1] Berry, W. D. (1993). Understanding regression assumptions (Vol. 92). Sage Publications.

[2] Cohen, P., West, S. G., & Aiken, L. S. (2014). Applied multiple regression/correlation analysis for the behavioral sciences. Psychology Press.

[3] Dubey P., Pandey D.,&Pandey D.K.(2018). Development and validation of Questionnaire on Usage of ICT in University Admission Process. International Journal of Mechanical Engineering & Technology (IJMET), 9 (13),709-719.

[4] Forsyth, I. (2014). Teaching and learning materials and the Internet. Routledge.

[5] Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). How to design and evaluate research in education. New York: mcgraw-Hill Humanities/Social Sciences/Languages.

[6] Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

[7] O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & quantity, 41(5), 673-690.

[8] Robinson, C. C., &Hullinger, H. (2008). New benchmarks in higher education: Student engagement in online learning. Journal of Education for Business, 84(2), 101-109.

[9] Schafer, W. D. (1991). Reporting Hierarchical Regression Results. Measurement and Evaluation in Counseling and Development, 24(3), 98-100.

[10] Scialfa, C. T., & Games, P. A. (1987). Problems with step-wise regression in research on aging and recommended alternatives. Journal of gerontology, 42(6), 579-583.

[11] Seibold, D. R., &Mcphee, R. D. (1979). Commonality analysis: A method for decomposing a explained variance in multiple regression analyses. Human Communication Research, 5(4), 355-365.

[12] Terrell, S. R., &Dringus, L. (2000). An investigation of the effect of learning style on student success in an online learning environment. Journal of Educational Technology Systems, 28(3), 231-238.

[13] Turkle, S. (1997). Life on the screen: Identity in the age of the Internet (New ed.). London: Phoenix.

[14] Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. Journal of Counseling Psychology, 34(4), 372.

[15] Ward, J., &labranche, G. A. (2003). Blended learning: The convergence of e-learning and meetings. Franchising World, 35(4), 22-22.

[16] Webster, R. (1985). Quantitative spatial analysis of soil in the field. In Advances in soil science (pp. 1-70). Springer, New York, NY.

[17] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

A Comprehensive Study of Sentimental Analysis Methods on Social Media Data

Rajat Kumar Yadu[1], Ragini Shukla[2]
[1]Asst. Prof. Dept. of Computer Science, Mahant Laxminarayan Das College, Raipur (C.G.) India.
[2]Asst. Prof. Dept. of CSIT Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT
Sentiments and Opinions are only attributes to express views about attitude, emotions, and sentiments through various social networking sites like twitter, Facebook, Google+ but to categorize accurate positive and negative thoughts of peoples on social media data we have to use various sentimental analysis methods. In this research paper we discuss about various methods related to sentimental analysis and tabulate the accuracy of different techniques and comparing best methods to improve accuracy for social media data. In this research paper we compare various sentimental analysis methods related to classification techniques and create an analysis table for different supervised and unsupervised methods for different social media datasets and accuracy percentage for different techniques. In this research paper we compare various sentimental analysis methods related to classification techniques and create an analysis table for different supervised and unsupervised methods for different social media datasets and accuracy percentage for different techniques. Sentiment analysis relates to the problem of mining the sentiments from online available data and categorizing the opinion expressed by an author towards a particular entity into at most three preset categories: positive, negative and neutral. In this paper, firstly we present the sentiment analysis process to classify highly unstructured data on Twitter. Secondly, we discuss various techniques to carryout sentiment analysis on Twitter data in detail. Moreover, we present the parametric comparison of the discussed techniques based on our identified parameters.

Keywords: Social media, Sentiment Analysis, opinion mining, Decision Tree, Sentiment Methods.

I INTRODUCTION

-Sentiment analysis is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral.

There are various research papers available related to sentimental analysis of social media data where different supervised and unsupervised classification methods used to correct classify the positive and negative comments by using different tools such as Weka, Matlab and Hadoop and have produced different accuracy percentage.

There are two main approaches for sentiment analysis: machine learning based and lexicon based. Machine learning based approach uses classification technique to classify text. Lexicon based method uses sentiment dictionary with opinion words and match them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary.

There are various levels of sentimental analysis such as Document level sentiment analysis, Sentence level sentiment. Sentimental analysis methods for social media data can be categorized can be used in three ways-
From the document, but sometimes nouns or verbs may also express opinion.

II STUDY AREA

In this paper various social media data have used for sentimental analysis where the first step in Supervised Machine learning technique is to collect the training set and then selects the appropriate classifier. Once the classifier is selected, the classifier gets trained using the collected training set.

There various methods used in various sentimental analyses of social media data such as-

(a) **Naive Bayes Classification (NB)** -The statistical Bayesian algorithm characterizes a supervised learning technique, it supposes an originating probabilistic method and it permits us to confine improbability about the method in an honorable way by formative chances of the results. It solves predictive and diagnostic and problems and it used to evaluate learning algorithm and provides a useful perspective for understanding.

(b) **Support Vector Machine (SVM)** -In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

(c) **Maximum Entropy algorithm(MEA)**- ME models named as Gibbs, log linear, multinomial logic models , exponential and present a general purpose machine learning technique for prediction and classification which has been productively practical to fields as different as econometrics and computer vision.

(d) **K Nearest Neighbors (KNN)** -In pattern recognition, the k-nearest neighbors algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples.

(e) **10-fold Cross validation**-Cross-validation is a technique to evaluate predictive models by partitioning the original sample into training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples.

(f) **Decision tree** - It is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

III METHODOLOGY

(a) Sentiment Analysis based on Supervised Machine learning technique.
(b) Sentiment Analysis by using Lexicon based Technique.
(c) Sentiment Analysis by combining the above two approaches.

In Supervised Machine learning techniques, two types of data sets are required: training dataset and test data set. An automatic classifier learns the classification factors of the document from the training set and the accuracy in classification can be evaluated using the test set. Various machine learning algorithms are available that can be used very well to classify the documents. The machine learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used successfully in many research and they performed well in the sentiment classification.

Table 1
Comparative table for accuracy of various Sentiment Analysis methodsusing Different Techniques

Research Paper	Approach	Dataset	Technique	Accuracy
A Study on Sentiment Analysis on Social Media Using Machine Learning Techniques	Supervised	Twitter dataset	Naive Bayes	53%
			SVM	50%
			KNN	44%
Social media metrics and sentiment analysis to evaluate the effectiveness of social media post	Supervised	Facebook dataset	KNN	82.3%
Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach	Supervised & Unsupervised	Twitter dataset	Naïve Bayes	66.66%
			Hybrid Techniques	76.31%
Classification of Facebook News Feeds and Sentiment Analysis	Supervised	Facebook News Feeds dataset	SVM	97%
			logistic regression	60%
Mining Sentiments from Tweets	Supervised	Google+, Twitter, Myspace dataset	SVM	88%
Sentiment Analysis of Twitter Data: A Survey of Techniques	Supervised	Movie Reviews	SVM	86.40%
		Twitter	Co Training SVM	82.52%
		Stanford Sentiment Treebank	Deep Learning	80.70%

Analysis of Various Sentiment Classification Techniques	Supervised & Unsupervised	Movie Reviews	SVM	82.9%
			MaxEnt	81%
			Naïve Bayes	81.5%
Facebook Posts Text Classification to Improve Information Filtering	Supervised	Facebook News Feeds dataset	SVM	78.9%
			K-NN	55.57%
Sentiment analysis in twitter data using data ytic techniques for lictive modelling	Supervised	Twitter dataset	10- ford cross validation	85%
Sentiment Analysis and Classification of Tweets Using Data Mining	Supervised & Unsupervised	Twitter dataset	Decision Tree	84.66%
			K-NN	50.72%
			Naïve Bayes	64.42%

IV CONCLUSION

In this research paper we compare various sentimental analysis methods related to classification techniques and create an analysis table for different supervised and unsupervised methods for different social media datasets and accuracy percentage for different techniques. The final result of this comprehensive study is that the approx 80% accuracy produced by the support vector machine so SVM is the best suitable sentimental analysis method which can be future used as best classifier for social media data analysis.

REFERENCES

[1] Kumar, P. (2018). A sentiment analysis system for social media using machine learning techniques: Social enablement. Digital Scholarship in the Humanities, IJRAET, Volume-5, Issue -4, 5, pp- 2347 – 2812 doi:10.1093/llc/fqy037.

[2] Poecze, F., Ebster, C., & Strauss, C. (2018). Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. Procedia Computer Science, International Conference on Ambient Systems, Networks and Technologies 130, 660-666. doi:10.1016/j.procs.2018.04.117.

[3] Goyal,S., (2016). Sentimental Analysis of Twitter Data using Text Mining and Hybrid Classification Approach, International Journal on Data Science and Technology,Volume2, Issue 5,pp-1-9.

[4] Setty, S., Jadi, R., Shaikh, S., Mattikalli, C., & Mudenagudi, U. (2014). Classification of facebook news feeds and sentiment analysis. International Conference on Advances in Computing, Communications and Informatics (ICACCI). doi:10.1109/icacci.2014.6968447.

[5] Bakliwal, A.,(2012). Mining Sentiments from Tweets. International Institute of Information Technology, Volume 7, Issue 12, pp-11-18.

[6] Kharde A.,(2016). Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications, Volume 139, Issue 5, pp-5-15.

[7] B., V., & M., B. (2016). Analysis of Various Sentiment Classification Techniques. International Journal of Computer Applications, 140(3), 22-27. doi:10.5120/ijca2016909259.

[8] Benkhelifa, R., & Laallam. F. Z. (2016). Facebook Posts Text Classification to Improve Information Filtering. Proceedings of the 12th International Conference on Web Information Systems and Technologies. doi:10.5220/0005907702020207.

[9] Sulthana, A. R., Jaithunbi, A. K., & Ramesh, L. S. (2018). Sentiment analysis in twitter data using data analytic techniques for predictive modelling. Journal of Physics: Conference Series, 1000, 012130. doi:10.1088/1742-6596/1000/1/012130.

[10] A., V., & Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11), 5-15. doi:10.5120/ijca2016908625.

[11] Desai, M., & Mehta, M. A. (2016). Techniques for sentiment analysis of Twitter data: A comprehensive survey. 2016 International Conference on Computing, Communication and Automation (ICCCA). doi:10.1109/ccaa.2016.7813707.

[12] Parveen, H., & Pandey, S. (2016). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). doi:10.1109/icatcct.2016.7912034.

[13] Vairagade, A. S., & Fadnavis, R. A. (2016). Automated content based short text classification for filtering undesired posts on Facebook. 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave). doi:10.1109/startup.2016.7583984

[14] .Kaur, A., & Baghla, S. (2018). Sentiment Analysis of Eng

[15] Shoeb. M.. (2017). Sentiment Analysis and Classification of Tweets Using Data Mining. IRJET, Volume: 04 Issue: 12, pp 1471-1474.

[16] Ahmadzadeh, E., & Chan, P. K. (2017), Mining pros and cons of actions from social media for Decision support, IEEE International Conference on Big Data (Big Data).doi:10.1109/bigdata.2017.8258003,V-12,I-14,pp 877-882.

[17] Shoeb, Md. & Ahmed, J. (2017), Sentiment Analysis and Classification of Tweets Using Data Mining,International research journal of engineering and technology (IRJET),V-04,I-12,pp 1471-1474.

[18] Atzmueller, M. (2012), mining social media: Key players, sentiments, and communities. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, V-2, I-5, pp 411-419,doi:10.1002/widm.1069.

[19] Bifet, A.(2013), Mining Big data in Real time ,Informatics ,V-37,I-1, pp 15-20.

[20] Bratawisnu, M. K., Giri, R. R., & Rinaldi, R. (2017), Association perception customer feedback with text network analysis in social media (case study on internet banking BRI, BCA, Mandiri in Indonesia), International Conference on Information and Communication Technology (ICoIC7). doi:10.1109/icoict.2017.8074694,V-10,I-04,pp 1-6.

[21] Barbier, G., & Liu, H. (2011), Data Mining in Social Media. Social Network Data Analytics, pp 327-352, doi:10.1007/978-1-4419-8462-3_12.

[22] Dunren, C. & Mejdl & S. and Zhiyong, P. (2013), From Big Data to Big Data Mining Challenges, Issues, and Opportunities, Published in the Proc. Of International Conference On Database Systems for Advanced Applications Organized & Published by Springer held in Suzhou,V-7827, issue April 22 ,pp 1-15.

[23] Daniel, B. K. (2016), Overview of Big Data and Analytics in Higher Education. Big Data and Learning Analytics in Higher Education published in Springer international journal Switzerland, V-28, I-08, pp1-4, doi: 10.1007/978-3-319-06520-5_1.

[24] Devakunchari, R. Valliyammai, C. (2016), Big Social Data Analytics: Opportunities, Challenges and Implications on Society in International Conference on Communication, Media, Technology and Design V-6, I-6, pp 25-33.

[25] Felt, M. (2016), Social media and the social sciences: How researchers employ Big Data Analytics. Big Data & Society, V-1, I-15, pp 1-15. doi: 10.1177/2053951716645828.

Performance Evaluation of Different Equalization Techniques for QAM in Wireless Communication

Rashmi Kashyap[1], Saurabh Mitra[2]

[1,2]Asst. Prof. Dept of EC, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

Due to the distortive character of the propagation environment and High bit data rate transmission over wireless channel makes the channel response extend over more than one symbol period transmitted data symbols will spread out in time and will interfere with each other, a phenomenon called Inter Symbol Interference (ISI). This is undesirable and makes the recovery of signal difficult. Equalization is method which is commonly employed to fight with ISI. In this paper different equalizer has been analyzed and compared for 2x2 MIMO channel.

Keywords- DPSK, MIMO, interference, MMSE, Zero forcing AWGN, Rayleigh

I INTRODUCTION

The emergence of internet and mobile technology has enabled us to share video, text, voice and all other information in all over the world. Introduction of wireless and 3G mobile technology has made it possible to transfer the data at very high speed while keeping the high quality of the data intact. To achieve high quality data at very high data rate is a big challenge. These problems can be minimized by applying Orthogonal Frequency division Multiplexing technology. Unlike wired media, in wireless media, signal reach the receiver from different path and hence lead to the inter symbol interference. This inter symbol interference phenomenon causes the increased bit error rate [1].

Generally in designing the communication system, it is assumed that the AWGN Channel or non dispersive channel passes all the frequency which is practically not possible.

For any band-limited or dispersive channel, the impulse response of the channel resembles the impulse response of ideal low pass filter. Due to this the transmitted signal is smeared in time and hence spread the symbols length causing the overlapping of adjacent symbols. The interference caused by this phenomenon is known as inter-symbol interference (ISI). This phenomenon is undesirable in communication system and increased the bit error rate (BER) and hence need to be resolved correctly. This problem of ISI can be overcome by either designing the band-limited pulses otherwise known as nyquist pulses for transmission or by filtering the received signal to suppress the effect of ISI introduced by the impulse response of the channel. The process of mitigating the effect of ISI by using appropriate filtering operation is known as equalization process [2].This paper presents a performance evaluation of some of the equalization techniques like zero forcing, zero forcing with successive interference cancelling (ZF-SIC), MMSE, ZF-SIC with optimal ordering, Maximum likelihood (ML) equalizer, MMSE-SIC with optimal ordering 2x2 MIMO system under Rayleigh fading and noisy channel.

II BASE WORK

(a) **Quadrature Amplitude Modulation (QAM)**- Quadrature amplitude modulation (QAM) is a modulation scheme in which two sinusoidal carriers, one exactly 90 degrees out of phase with respect to the other, are used to transmit data over a given physical channel. Because the orthogonal carriers occupy the same frequency band and differ by a 90 degree phase shift, each can be modulated independently, transmitted over the same frequency band, and separated by demodulation at the receiver. For a given available bandwidth, QAM enables data transmission at twice the rate of standard pulse amplitude modulation (PAM) without any degradation in the bit error rate (BER). QAM and its derivatives are used in both mobile radio and satellite communication systems. Fig1.8 shows the block diagram of QAM and Fig1.9 shows the waveform of QAM.

(b) **Channel Model**-AWGN (additive white Gaussian Noise) is model of channel which produces only white Gaussian noise (having Gaussian distribution) whose spectral density is constant. This channel model does not introduce frequency selectivity, fading dispersion and interference phenomenon. This channel model is sufficient enough to analyze the effect of Gaussian noise coming from various natural sources [3] with the simple mathematical model. Fading is the phenomenon of introducing distortion in carrier modulated signal in some propagation medium [4]. The main reason of fading phenomenon in wireless media is multipath propagation which results in transmitting signal's reaching the receiver by two or more path. These different paths introduce constructive and destructive interference in the signal causing phase shifting of the signal. Rayleigh fading is one of the types of fading which occurs due to the multipath reception. It can be simulated with the help of statistical model for analyzing the effect of propagation environment on a signal [3].

Channel model having the characteristics of multipath environment can be simulated. The impulse response of 3-tap multipath channel model with spacing T is shown below-
$h[k] = [h_1\ h_2\ h_3]$

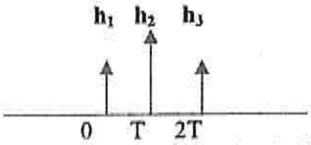


Fig. 1.2 Impulse Response of multipath Channel

Apart from experiencing multipath effect, the transmitted signal is also affected by AWGN

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(Absolute Gaussian noise) noise n. This noise is represented by Gaussian function given by Here the μ represents the mean of distribution and σ is variance.

With the known channel response h(k) and noise n, the signal received at the receiver is given by $y(k)= x(k)\otimes h(k)+n$

here the ⊗ represents the convolution operation[5].

(c) **Equalizer**-Equalization is the process of mitigating the ISI effect by decreasing the error probability which occurs in the communication system when no ISI suppression method is applied. But since the suppression of ISI tends to enhance the noise power therefore the optimum balance between noise power enhancement and suppression of ISI need to catered [4].

(i) **Adaptive equalization[7][8]**- An adaptive equalizer is a type of digital filter or equalization filter which is designed in such a way that it automatically adapts itself to the time varying properties of communication channel. This technique is frequently used to mitigate the distortion produced by multipath effect.

(ii) **Zero Forcing Equalizer[9][10]**- Proposed by Robert lucky, zero forcing method of equalization is a linear equalization method which restores the transmitted signal by inverting the frequency response of the channel. The name zero forcing comes from the fact that it is able to reduce the ISI to zero value in case of noise free environment.

(iii) **MMSE Equalizer [11]**- This type of equalizer uses the squared error as performance measurement [11]. The receiver filter is designed to fulfill the minimum mean square error criterion. Main objective of this method is to minimize the error between target signal and output obtained by filter.

(iv) **Zero Forcing with Successive Interference cancellation (ZF-SIC) Equalizer [12]** - In this method, first of all the zero forcing equalizer find the estimated symbol x_1 and x_2 then one of the estimated symbol is subtracted from received symbol to compute the equalized symbol by applying maximum ration combining(MRC)[36].

(v) **Successive Interference Cancellation using optimal ordering Equalizer [13]**- In the previous successive interference cancellation method, estimation symbol is chosen arbitrarily and then its effect is subtracted from received symbol y1 and y2. A better result can be obtained if we choose estimated symbol whose influence is more than other symbol. For this first of all the power of both the symbol is computed at the receivers and then the symbol having higher power is chosen for subtraction process.

(vi) **MMSE SIC with optimal ordering [14]**- The same concept of successive interference with optimal ordering can also be applied to the MMSE equalizer and the resultant equalizer is known as MMSE SIC with optimal equalizer.

(vii) **ML (Maximum Likelihood) Equalizer** - Let x represent the signal matrix, H represent the channel response and n represent the noise then the signal obtained at the receiver is given by

III METHODOLOGY

In order to analyze how different techniques of equalization perform in MIMO having noisy and rayleigh channel characteristics, a simulation program is designed for all the six method in MATLAB environment.

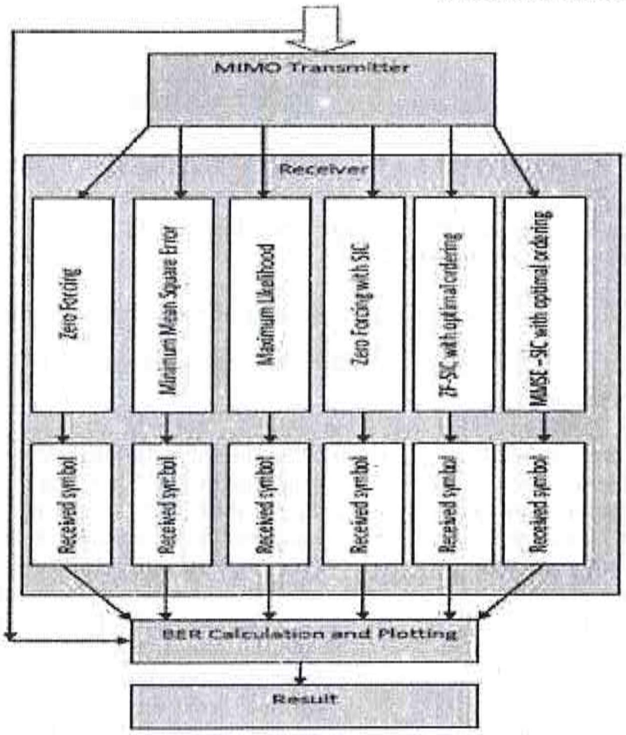


Fig. 4.1 Flow Diagram of Methodology

IV EXPERIMENTAL RESULTS

Here is the performance of all the six equalizer for QAM in 2x2 MIMO System

(a) BER for ZF Equalizer

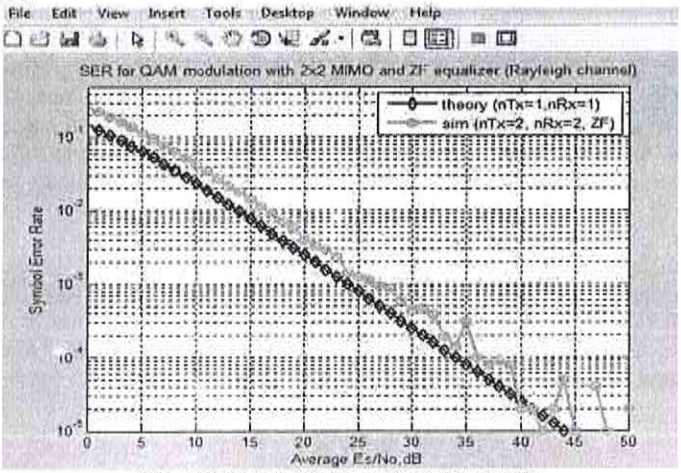


Fig.1 BER for ZF Equalizer for QAM

Table 1

Theoretical and Simulated BER table for ZF equalizer (QAM Modulation)

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC)	Simulated BER for nTx=2,nRx=2(ZF)
0	0.146447	0.058058	0.258012
10	0.023269	0.001599	0.042425
20	0.002481	1.84E-05	0.004732
30	0.00025	1.87E-07	0.000429
40	2.5E-05	1.87E-09	0.000051

Fig.1and table1 shows the BER performance for QAM modulation for 2x2 MIMO system using ZF equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. From the graph we can see that ZF equalizer shows much improvement in SNR. Table shows that as per SNR increases the value of BER decreases for ZF equalizer.

(b) BER for ZF-SIC Equalizer

Table 2

Theoretical and Simulated BER table for ZF-SIC equalizer (QAM Modulation)

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC)	Simulated BER for nTx=2,nRx=2(ZF-SIC)
0	0.146447	0.058058	0.24786
10	0.023269	0.001599	0.08414
20	0.002481	1.84E-05	0.05823
30	0.00025	1.87E-07	0.05447
40	2.5E-05	1.87E-09	0.05456

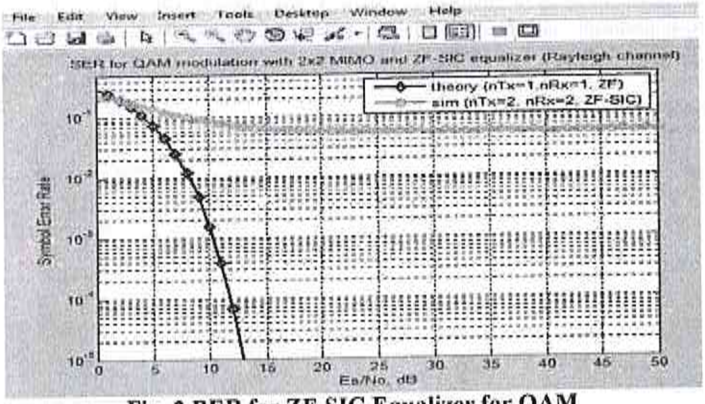


Fig. 2 BER for ZF-SIC Equalizer for QAM

Fig. 2and table 2 shows the BER performance for QAM modulation for 2x2 MIMO system using ZF-SIC equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. from the graph we can see that ZF-SIC does not show the much improvement in SNR.

(c) BER for MMSE Equalizer

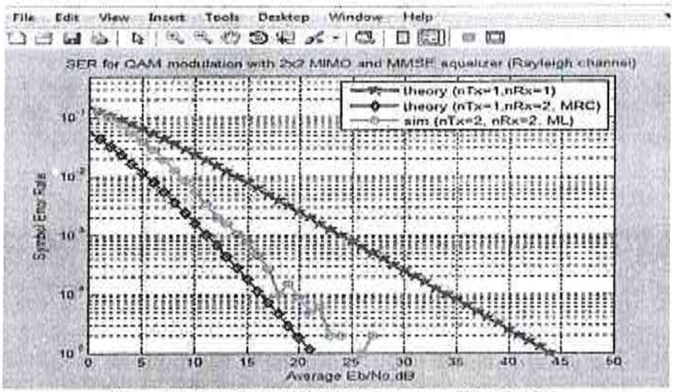


Fig. 3 BER for MMSE Equalizer for QAM

Table 3
Theoretical and Simulated BER table for MMSE equalizer (QAM Modulation)

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC)	Simulated BER for nTx=2,nRx=2(MMSE)
0	0.146447	0.058058	0.205458
10	0.023269	0.001599	0.017327
20	0.002481	1.84E-05	0.001336
30	0.00025	1.87E-07	0.000136
40	2.5E-05	1.87E-09	0.000013

Fig and table 3 shows the BER performance for QAM modulation for 2x2 MIMO system using MMSE equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. From the graph we can see that MMSE shows the much improvement result as compare to ZF and ZF-SIC equalizer.

(d) **BER for ZF-SIC Optimal Order Equalizer**

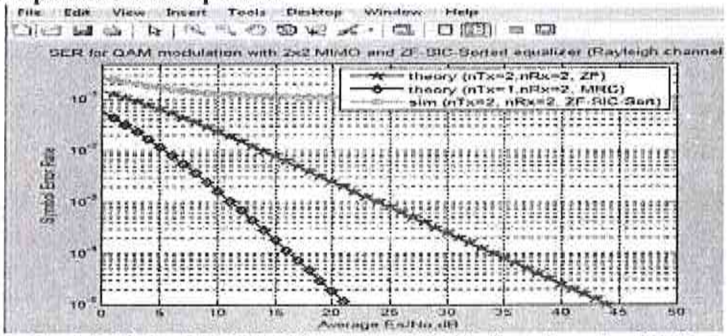


Fig.4 BER for ZF-SIC OPTIMAL ORDER Equalizer for QAM

Table 4
Theoretical and Simulated BER table for ZF-SIC with optimal order equalizer (QAM Modulation)

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC)	Simulated BER for nTx=2,nRx=2(ZF-SIC sort)
0	0.146447	0.058058	0.24786
10	0.023269	0.001599	0.08414
20	0.002481	1.84E-05	0.05823
30	0.00025	1.87E-07	0.05447
40	2.5E-05	1.87E-09	0.05456

Fig and table 4 shows the BER performance for QAM modulation for 2x2 MIMO system using ZF-SIC with optimal order equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. From the graph we can see that ZF-SIC with optimal order does not show the much improvement in SNR.

(e) BER For MMSE-SIC Optimal Order Equalizer

Table 5
Theoretical and Simulated BER table for MMSE with optimal order equalizer

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC	Simulated BER of MMSE-SIC for nTx=2,nRx=2
0	0.146447	0.058058	0.22533
10	0.023269	0.001599	0.08221
20	0.002481	1.84E-05	0.05679
30	0.00025	1.87E-07	0.0548
40	2.5E-05	1.87E-09	0.05359

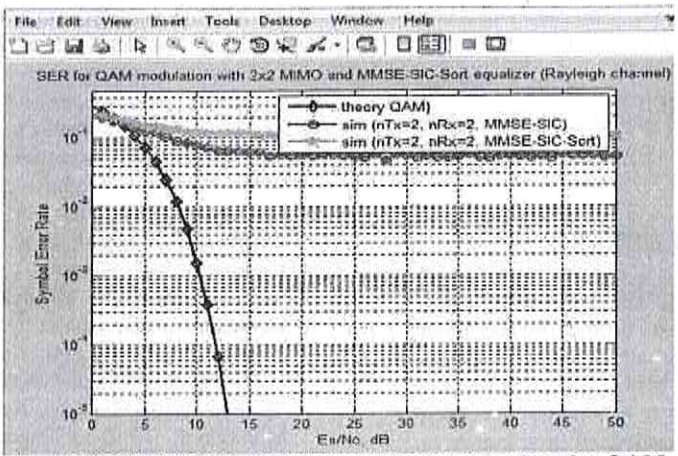


Fig. 5 BER for MMSE-SIC Optimal Order Equalizer for QAM

Fig and table 5 shows the BER performance for QAM modulation for 2x2 MIMO system using MMSE-SIC with optimal order equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. From the graph we can see that MMSE-SIC with optimal order does not show the improvement in SNR.

(f) BER for ML Equalizer

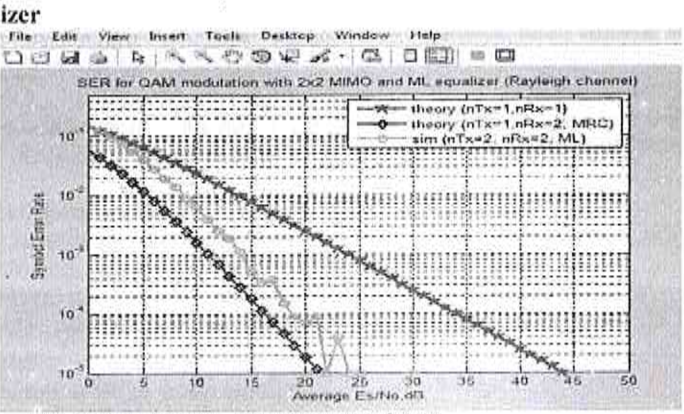


Fig. 6 BER for ML Equalizer for QAM

Table 6
Theoretical and Simulated BER table for ML equalizer (QAM Modulation)

Eb/No In dB	Theoretical BER for nTx=1,nRx=1	Theoretical BER for nTx=1,nRx=1(MRC)	Simulated BER for nTx=2,nRx=2(ML)
0	0.146447	0.058058	0.15812
10	0.023269	0.001599	0.0058
20	0.002481	1.84E-05	0.00012
30	0.00025	1.87E-07	0
40	2.5E-05	1.87E-09	0

Fig.6 and table 6 shows the BER performance for QAM modulation for 2x2 MIMO system using ML equalizer in Rayleigh channel. Black lines show the theoretically ideal value for BER. Green line shows the simulation result. From the graph we can see that performance of ML equalizer is similar to MMSE equalizer and it shows better improvement then ZF and ZF-SIC equalizer.

From comparing all the six equalizer for QAM modulation the result for ML and MMSE Equalizer is best among these above mentioned equalizer in terms of cancelling the interference to optimum level.

V CONCLUSION

To achieve higher data rate and least BER is the demand of wireless system design. Equalization techniques play very important role for designing such system. In this paper performance comparison of different key equalization techniques has been carried out under the fading and noisy environment to find out the appropriate equalizer for 2x2 MIMO systems. From the result obtained it is evident that zero forcing equalizer shows better performance if noise is zero and shows degradation under fading environment.

The performance of ZF-SIC, MMSE, and ZF-SIC with optimal ordering, MMSE-SIC with optimal ordering and ML equalizer are in increasing order. From the results it can be concluded that the ZF equalizer is best among these above mentioned equalizer in term of cancelling the interference to optimum level.

REFERENCES

[1] G.L. Stuber, J.R. Barry, S.W. Mclaughlin, Ye Li, M.A. Ingram And T.G. Pratt, "Broadband MIMO-OFDM Wireless Communications," Proceed-Ings Of The IEEE, Vol. 92, No. 2, Pp. 271-294, February. 2004.

[2] 'Fading' – Online Article In Wikipedia En.Wikipedia.Org/Wiki/Fading 'A Comparative Study Of Rayleigh Fading Wireless Channel Simulators' By VRS Ramaswamy Bit Error Rate Performance In OFDM System Using MMSE & MLSE Equalizer Over Rayleigh Fading Channel Through The BPSK, QPSK,4 QAM & 16 QAM Modulation Technique.

[3] D. W. Tufts, "Nyquist's Problem-The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation," Proc. IEEE, Vol. 53, Pp. 248-260, Mar. 1965. "Techniques For Adaptive Equalization Of Digital Communication Systems," Bell Syst. Tech. I., Vol. 45, Pp. 255- 286,Feb. 1966 "Adaptive Equalization" By Shahid U. H. Qureshi, Senior Member, IEEE Vol.4, No.4, October, 2010. "Zero-Forcing Equalization For Time-Varying Systems With Memory "By Cassio B. Ribeiro, Marcello L. R. De Campos, And Paulo S. R. Diniz. Zero-Forcing Frequency Domain Equalization For Dmt Systems With Insufficient Guard Interval By Tanja Karp, Martin J. Wolf , Steffen Trautmann , And Norbert J. Fliege. Digital Communications. New York: Mccraw-Hill, 1983.

[4] G. Leus, S. Zhou, and G. B. Giannakis, "Orthogonal Multiple Access Over Time- And Frequency-Selective Channels," IEEE Transactions on Information Theory, Vol. 49, No. 8, Pp. 1942–1950, 2003.

[5] Rashmi Kashyap, Jaspal Bagga, "Performance Evaluation Of Equalization Techniques Under Fading and Noisy Environment For Mimo Systems In Wireless Communication" IJRIT. Volume 2, Issue 6, June 2014, Pg: 478-487

[6] Anuj Kanchan, Shashank Dwivedi, "Comparison of BER Performance In OFDM Using Different Equalization Techniques", International Journal Of Engineering And Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume- 1,Issue-6, August 2012, Pp. No. - 139-143

Secure Modified QC- LDPC Code Based McEliece Public key Encryption Scheme

Renuka Sahu[1], B.P. Tripathi[2]
[1,2]Dept. of Mathematics, Govt. N. P.G College of Science, Raipur (C.G), India

ABSTRACT
In this paper, a modified method of QC-LDPC McEliece cryptosystem is presented. Authors of [3] suggested that the recovery of Q is possible by an adversary using power trace method. To overcome with this problem, we present a modification in the existing scheme so that QC-LDPC McEliece cryptosystem by using a random linear non-singular matrix. And also we have tried to hide the structure of the generator matrix. The modified scheme has the ability to detect all the errors and correct nearly up to 30% of the errors that occurs. Thus, it has better error correcting capacity than the existing schemes.

I INTRODUCTION

With the rapid development of technologies, our society is concerned completely upon the security of current public key infrastructures. The fundamental components for current public key infrastructure are based on public cryptographic strategies like RSA and DSA, etc. However, which was proved that these cryptographic strategies would be simply broken by the super powerful quantum computers. Thus, it becomes very important to develop some new public key cryptographic strategies in such a way that the new PKC method would be secure against quantum attack.
In 1978, Robert. J. McEliece presents a public key cryptosystem primarily based t-error correcting Goppa codes [10], which is known as McEliece cryptosystem. The general decoding problem of the linear block codes is NP-Complete, which was hard to handle. McEliece Cryptosystem has been considered as one of the candidate for the post-quantum cryptography. The originalversion of McEliece cryptographic scheme uses Goppa codes. As compared with the other cryptographic schemes like RSA, McEliece cryptographic schemes have high-speed encryption/ decryption algorithms. However, because of its large public key size and low code rate aren't in demand these days. To overcome these drawbacks of McEliece Cryptosystem, a number of variants are introduced by replacing the Goppa code with the other significant codes. For example, H. Niederreiter et al. [12] presented "GRS Codes" primarily based scheme. Then sub codes of H. Niederreiter were suggested by T. Berger and Loidreau [16]. V. M. Sidelnikov et al. [17] used "R. M. codes", H. Janwa and Morena [5] used "Algebraic Geometric codes", M. Baldi et al. [1] used "Low-Density parity check (LDPC) codes, and Misoczki et al. [11] used "Moderate Density Parity Check (MDPC) codes, and Londahl et al. [7] used "Convolutional Codes". Most of these cryptographic schemes have been broken through Moderate density parity check codes (MDPC) / Low density parity check codes (LDPC) based on McEliece cryptographic encryption schemes. Recently in [18] RLCE Scheme is presented using Hexi code was presented .The original McEliece Cryptographic encryption scheme was considered to be secure.

In [1], Baldi and Chiaraluc introduced "quasi-cyclic low density parity check code (QC- LDPC codes)" in the McEliececryptosystem which reduces its public key size. The Cryptosystem is now called as the QC-LDPC McEliece cryptosystem. In[13], Otmani et al. showed that the proposed system had several vulnerabilities. An amended version of the cryptographic techniques was introduced by Baldi et al. in [2], which was secure against the Otmani's et al. [13] attack. In 2016 [3], Fabsic et al. demonstrate that a threat is present in the QC- LDPC variant of the McEliece cryptosystem. In this paper, we present a modified method for constructing and encoding QC- LDPC Codes.

This paper is organized as follows: Section II gives a brief introduction of QC- LDPC McEliece cryptosystem. In sectionIII, we present the modified method of the QC-LDPC McEliece cryptosystem. In Section IV, we have shown that our new modified scheme would resists against attacks given by [3] by adopting a different form for its constituent matrices, without altering other parameters. In section V, the performances of the modified scheme are compared and finally conclude the paper.

II PRELIMINARIES

In this section, we recall the keywords concerning with the modified Encryption scheme.

(a) QC- LDPC Codes

QC-LDPC codes are called as "reputable structured" type Low density parity check (LDPC) codes. Quasi- Cyclic codes was first studied by Townsend and Welson, where a QC-codes is defined as linear block code with dimension "$k = p . k_0$" and length "$n = p . n_0$ "having the following properties:

(i) A series of " p " blocks of " n_0 " symbols will form each code word, each codeword is formed by k_0 information symbols defined by $r_0 = n_0 - k_0$ rebundancy symbols and

(ii) Another valid codeword is formed by each cyclic shift of codeword by n_0 symbols.

(b) Generator Matrix of a Quasi-Cyclic Code

A first form for the generator matrix of a quasi-cyclic code directly follows from the code definition, as shown by the following

Lemma 1.[1] The generator matrix G of a quasi-cyclic code has the form of a "blocks circulant" matrix, where each block $G_i$has size $k_0 \times n_0$.

$$\mathbf{G} = \begin{bmatrix} G_0 & G_1 & \cdots & G_{p-1} \\ G_{p-1} & G_0 & \cdots & G_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ G_1 & G_2 & \cdots & G_0 \end{bmatrix} \qquad \textbf{(1)}$$

(c) Parity-Check Matrix of a Quasi-Cyclic Code

Similarly to the generator matrix G, the following form holds for the parity check matrix H of a quasi-cyclic code.

$$\mathbf{H} = \begin{bmatrix} H_0 & H_1 & \cdots & H_{p-1} \\ H_{p-1} & H_0 & \cdots & H_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ H_1 & H_2 & \cdots & H_0 \end{bmatrix} \qquad \textbf{(2)}$$

Lemma 2[1]. The parity-check matrix H of a quasi-cyclic code has the form of a "blocks circulant" matrix, where each block H_i has size $r_0 \times n_0$

(d) Alternative "Circulants Block" form

Lemma 3 [1]. Given a matrix in the "blocks circulant" form (1) or (2), it can be put in an alternative "circulants block" form that, for the matrix H in (2), is:

$$\mathrm{H^c} = \begin{bmatrix} H^c_{00} & H^c_{01} & \cdots & H^c_{0(n_0-1)} \\ H^c_{10} & H^c_{11} & \cdots & H^c_{1(n_0-1)} \\ \vdots & \vdots & \ddots & \vdots \\ H^c_{(r_0-1)0} & H^c_{(r_0-1)1} & \cdots & H^c_{(r_0-1)(n_0-1)} \end{bmatrix} \qquad (3)$$

where each matrix $\mathrm{H^c_{ij}}$ is a "p × p" circulant matrix:

$$\mathbf{H^c_{ij}} = \begin{bmatrix} h_0^{ij} & h_1^{ij} & h_2^{ij} & \cdots & h_{(p-1)}^{ij} \\ h_{(p-1)}^{ij} & h_0^{ij} & h_1^{ij} & \cdots & h_{(p-2)}^{ij} \\ h_{(p-2)}^{ij} & h_{(p-1)}^{ij} & h_0^{ij} & \cdots & h_{(p-3)}^{ij} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_1^{ij} & h_2^{ij} & h_3^{ij} & \cdots & h_0^{ij} \end{bmatrix} \qquad \textbf{(4)}$$

III MCELIECE CRYPTOSYSTEM WITH QC-LDPC CODES

The use of LDPC codes in the McEliece cryptosystem should allow reducing the public key length, at least in principle, since such codes are defined by sparse parity-check matrices, whose storage size increases linearly in the code length.

(a) Key Setup

(i) "$(n-k) \times n$" parity check matrix **H** of an LDPC code correct less than "t " errors.

(ii) "$k \times k$ " invertible matrix S blocks of " $p \times p$ " circulants.

(iii) "$n \times n$ " a sparse invertible matrix **Q** blocks of " $p \times p$ " circulants.

S and **Q** are formed by blocks of "$p \times p$ " circulant matrices.

Public Key:$\mathbf{G}' = \mathbf{S}^{-1} \times \mathbf{G} \times \mathbf{Q}^{-1}$ **(4)**

Private Key: (**H, S, Q**)

(b) Encryption

For a row vector messageu $\in GF(q)^k$, Choose a random row vector error having length "n" and weight "t ".

Sender computes the ciphertext as

$$\mathbf{x} = \mathbf{u} \times \mathbf{G}' + \mathbf{e}$$
$$= \mathbf{c} + \mathbf{e}$$

(c) Decryption

For a received ciphertext "x", receiver computes

$$x' = x \times Q$$
$$= (u \times S^{-1} \times G) + (e \times Q) \qquad (5)$$

x' → Codeword vector of the QC-LDPC code choosen by receiver corresponds to the information vector $u' = u \times S^{-1}$.
$e \times Q$ → error vector ,
$t = t' m$ → maximum weight.
Receiver is able to correct all the errors with high probability, by means of LDPC decoding. Thus recovering u' and then u through a post-multiplication by S.

IV MODIFIED QC- LDPC CODE BASED MCELIECE CRYPTOSYSTEM

In this section, we have proposed a modified QC-LDPC code in order to hide the structure of the private key. To receive the message, the receiver randomly chooses a code in a family of $(n_{0,}\ d_v,\ p)$ QC-LDPC code based on Random Difference Families [ ].

(a) Key Setup

(i) Select"$(n-k) \times n$" parity check matrix H and produces a " $k_0 \times n_0$" generator matrix G in reduced echelon form. The matrix H is formed by a row $\{H_0, \dots, H_{n_0-1}\}$ of $n_0 = \frac{n}{n-k}$ binary circulant blocks with size "$p \times p$", where $p = n - k$. Generator marix $\mathcal{G}$ is formed by a "$k \times k$" identity matrix I with $k = k_0 . p$ and $k_0 = n_0 - 1$, followed by a column of k_0 binary circulant blocks with size p. If $H_{n_0-1}$is non-singular , then Generator matrix can be obtained as follows:

$$G = \begin{bmatrix} I & \begin{matrix} (H_{n_0-1}^{-1}.H_0)^T \\ (H_{n_0-1}^{-1}.H_1)^T \\ \vdots \\ (H_{n_0-1}^{-1}.H_{n_0-2})^T \end{matrix} \end{bmatrix}$$

(ii) Let $C_0, C_1, \dots, C_{n-1} \in GF(q)^{k_0 \times r}$ be "$k \times r$" matrices drawn at random and let $G_\emptyset = [G_0, C_0, G_{1,}C_1, \dots, G_{n-1}, C_{n-1}]$ be the $k_0 \times n_0(r+1)$ matrices obtained by inserting the random matrices C_i into G.

(iii) Let us choose uniformly random dense invertible $(r+1) \times (r+1)$ matrices $A_0, \dots, A_{n-1} \in GF(q)^{((r+1)\times(r+1))}$.

$$A = \begin{pmatrix} A_0 & & & \\ & A_0 & & \\ & & \ddots & \\ & & & A_{n-1} \end{pmatrix}$$

be an $n_0(r+1) \times n_0(r+1)$ invertible matrix.

(iv) Let "S" be a randomly selected dense "$k \times k$" binary non-singular matrix.

(v) Let "Q" be a "$n \times n$" sparse invertible matrix having fixed "m". ["S" and "Q"are formed by block of "$p \times p$" circulant matrices].

(vi) Public key is the $k_0 \times n_0\ (r+1)$ matrix.
$G^\emptyset = S^{-1} \times G_\emptyset \times A \times Q^{-1}$.

(vii) Private key $(S, G_\emptyset, A, Q)$.

(b) Encryption

Sender, who wants to send he encrypted message to recviver extracts $G^\emptyset$ from the public key and divides the message into k- bit blocks.If "Ψ" is one of these bocks, sender computes the encrypted, message as follows.

$$E_c = (\Psi \times G^\emptyset) + e$$

(c) Decryption

When receiver receives the encrypted message E_c, then receiver compute

$$\begin{aligned} E_c^\emptyset &= E_c Q A^{-1} \\ &= (\Psi G^\emptyset + e) Q A^{-1} \\ &= \Psi G^\emptyset Q A^{-1} + e Q A^{-1} \\ &= \Psi S^{-1} G_\emptyset A Q^{-1} Q A^{-1} + e Q A^{-1} \\ &= \Psi S^{-1} G_\emptyset + e Q A^{-1} \end{aligned}$$

Where

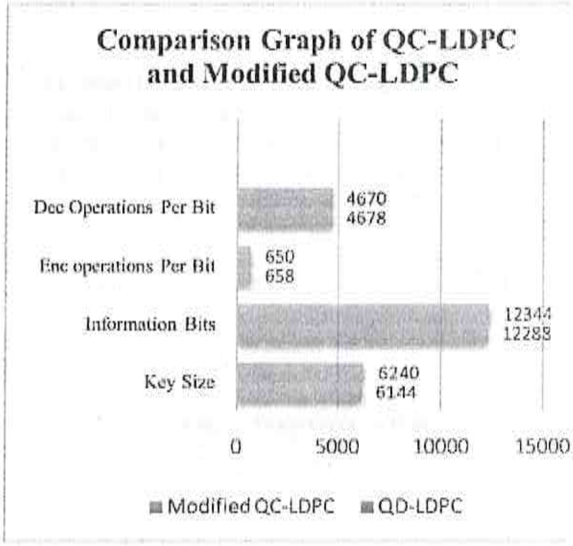
$$A^{-1} = \begin{pmatrix} A_0^{-1} & & & \\ & A_1^{-1} & & \\ & & \ddots & \\ & & & A_{n-1}^{-1} \end{pmatrix}$$

Vector $E_c^\emptyset$ is a codeword of the LDPC code choosen by receiver corresponding to the information vector $\Psi^\tau = \Psi . S^{-1}$ affected by an error vector e. Q, whose maximum weight is $t = t' m$.
By using efficient LDPC decoding algorithm, receiver is able to correct all the errors, thus recovering Ψ and the obtaining Ψ^τthrough multiplication by S.

V SECURITY ANALYSIS

The security of the modified encryption scheme is based on the fact that the adversary needs decoding in the code G, while G is not only different to the code G as in McEliece cryptosystem. But, after inserting C in G, it becomes more complicated to decode for an adversary. In [3], it is shown that the attacker can recover the matrix "Q" in the QC-LDPC McEliece Cryptosystem by used power trace method. However, only knowing the pattern of "Q" doesn't completely reveal the secret key. Since there is one another matrix "S" whose weight is approximately equal to "p/2". Also, in this paper we have suggested a random non-singular matrix "A" which will multiply with "Q" to make it more complex for an adversary to decode easily. If an adversary try to recover the matrix "Q" from the positive pattern in the power trace, then get the matrix multiplication of "QA^{-1}". Then decomposing the matrix "QA^{-1}" in Q and A^{-1}

is not feasible for an adversary. Thus, our modified QC-LDPC McEliece Cryptosystem will resist against the attack given in [3] without compromising with its performance with the existing scheme.



VI CONCLUSION

In this paper, a Modified version of QC-LDPC McEliece cryptosystem is proposed. This system belongs to the class of cryptosystems based on complete decoding task. We have shown that our modified scheme is more secure without compromising with the performance of the existing QC-LDPC McEliece Cryptographic scheme.

REFERENCES

[1] Baldi M., Chiaraluce F., (2007): Cryptanalysis of a new instance of McEliece cryptosystembased on QC-LDPC codes, in: Proceedings IEEE ISIT '07, Nice, France,,pp. 2591–2595.

[2] Baldi M., Bodrato M., Chiaraluce, F.(2008,): A new analysis of the McEliece cryptosystem based on QC-LDPC codes,in: 6th Internat. Conf. on Security and Cryptographyfor Networks—SCN '08 (R. Ostrovsky et al., eds.), Lecture Notes in Math., Vol. 5229, Springer-Verlag, Berlin, pp. 246–262.

[3] Fabsic T., Gallo O., Hromada V., (2016). Simple Power analysis attack on the QC-LDPC McEliece Cryptosystem, Slovak Academy of Sciences, Tatra Mt. Math. Publ. 67, pp-85-92.

[4] Heyse S., Moradi A., Paar C .(2010): Practical power analysis attacks on software implementationsof McEliece, in: Post-Quantum Cryptography (N. Sendrier, ed.), LectureNotes in Math., Vol. 6061, Springer-Verlag, Berlin, pp. 108–125.

[5] Janwa H,. and Moreno O.(1996), "McEliece public cryptosystem using algebraic-geometric codes." Des. Codes Cryptography, 8 pp. 293-307, (1996).

[6] Koochak Shooshtari M., Ahmadian-Attari M., Johansson T., Reza Aref M., (2016): Cryptanalysis of McEliece cryptosystem variants based on quasi-cyclic low-density parity check codes, IET Information Security **10, pp-**194–202.

[7] Londahl C., Johansson T., Shooshtari M. K., Ahmadian Attari M,. Aref M. R., (2016): Squaring attacks on McEliece public-key cryptosystems using quasi-cycliccodes of even dimension, Des. Codes Cryptogr. **80** pp. 359–377.

[8] Loidreau P.,(2000) "Strengthening McEliece Cryptosystem", International conference on the theory and application of cryptology and information security, Asiacrypt 2000, pp. 585-598.

[9] Mceliece R. J.,(1978): A public-key cryptosystem based on algebraic coding theory, Deep Space Network Progress Report **44, PP-**114–116.

[10] Misoczki R., Tillich J. P., Sendrier N., Barreto P. S. L. M,.: MDPC--McEliece: new McEliece variants from moderate density parity-check codes, in: IEEEInternat. Symp. on Information Theory—ISIT '13), Istanbul, pp. 2069–2073.

[11] Niederreiter H., (1986), Knapsack-type cryptosystems and algebraic coding theory. ProblemsControl Inform. Theory, 15(2):159–166

[12] Otmani A., Tillich J. P., Dallot L., (2010): Cryptanalysis of two McEliece cryptosystems based on quasi-cyclic codes, in: The 1st Internat. Conf. on Symbolic Computationand Cryptography—SCC '08, Beijing, China, 2008, Math. Comput. Sci., **3** no. 2,pp-129–140.

[13] Repka M., Zajac P., (2014): Overview of the McEliece cryptosystem and its security. Tatra Mt. Math. Publ. **60** pp. 57–83.

Studies on Recent Machine Learning Approaches to Explore Performance, Security Issues and New Dimensions to Deal with the Challenges

Reshamlal Pradhan[1], **S R Tandan**[2]

[1,2]Dept. of CSE, Dr. C.V.Raman University, Bilaspur (C.G.) India.

ABSTRACT

In the era of Computer technology, Machine learning is centre of attraction for the researchers of data mining. Organizational and individuals information in the Computers and we are going through serious issues of threats and intrusions. Malicious activities are increased in the computer and web usage. With rapid advancement in computer technology and networking services, huge amount of data has been generating, which is difficult to handle by traditional data processing applications. Datasets in the web are composed of structure and unstructured set of data. To deal with the unstructured set of data is a prime area of attention for the researchers. There is need of advancements in the data mining and data processing techniques, to deal with this massive amount of structure and unstructured datasets. Challenges are to improve accuracy and performance of data classification, regression and clustering, to analyze and update data storage, to maintain security and information privacy. Today machine-learning techniques, which are getting key attention, are Feature reduction, Decision tree techniques, Ensemble techniques, Neural Networks, Statistical techniques, Genetic algorithm, Fuzzy logic and big data analytics. In this paper, we are trying to gain attention on some of recent work done in these fields to explore data processing, data analysis and security challenges issues.

Keywords: Data Mining, IDS, Big data, Ensemble techniques.

I INTRODUCTION

Data mining has attracted more attention in recent years as scientific organizations and business organizations are dealing with very huge amount of data. Probably Big Data is the key attention of data mining in current era. Challenges in big data are extracting data, data storage and analysis, searching, querying and updating data, information privacy. Data mining is the process of discovering knowledge and interesting patterns from large amount of data. Today Intrusion detection system (IDS) is a necessary addition to the security infrastructure of most organizations. IDS collect and analyses information from different areas within a computer or network to detect possible security violations defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer or network.

Domain fields of data mining are business intelligence, scientific discovery, Web search and digital libraries etc. Big data has become crucial for numerous application domains as it deals with the large amount of unstructured data. Rapid growth of cloud services are the reason behind popularity of Big Data.

(a) **Process of KDD:** Data mining is often referred to as Knowledge discovery of data, which highlights the goal of mining process. To extract knowledge from data following steps are performed in KDD:
Step 1. Data preprocessing
Step 2. Data transformation
Step 3. Data mining
Step 4. Pattern evaluation and presentation.

(b) **Types of IDS**

Different types of intrusion detection system are:

(i) **NIDS and HIDS:** Host based Intrusion detection system (HIDS) monitors only individual workstation or system. HIDS are unaffected by switched network. HIDS can be thought of as an agent which monitors whether anyone or anywhere any unusual or subspecies activity is done. Network based intrusion detection system (NIDS) on the other hand monitors network traffic for particular network segment or device and analysis the network and application protocol activity to identify any sign of suspicious activity [1].

(ii) **Misuse and anomaly detection based IDS:** Misuse detection or signature based detection technique uses the previous data or pattern for detection: if previous pattern or signature is not available it cannot detect the new attack. Anomaly detection is adaptive in nature. They attempt to identify behaviors that do not conform to normal behavior [1, 2].

(c) **Data Mining Techniques:** There are varieties of data mining techniques. Using data processing techniques, it perceives and extrapolates knowledge that may scale back the probabilities of fraud detection [5]. These techniques are used for knowledge discovery and pattern recognization in order to detect intrusions and extract information.

(i) **Genetic algorithm:** Genetic Algorithm is an adaptive search technique initially introduced by Holland [7]. Genetic algorithm operates on a set of individuals called population, where each individual is an encoding of the problem input data and are called chromosomes. The search for best solution is guided by an objective function called fitness function. The selected solution of fitness function replace those of less function .as they are able to produce new solution that are more fitted in the environment. Fitness function controls the selection of best solution and provides criteria to evaluate the candidate individuals [8].

(ii) **Decision tree:** Decision trees are unit arborous structures that represent decision sets. These choices generate rules that are used to classify data [6]. A decision tree classifies a sample through a sequence of decisions, in which the current decision helps to make the subsequent decision. Such a sequence of decision is represented in a tree structure. The classification of a sample proceeds from the root node to a suitable end life node, where each end life node represents a classification category. The attributes of the samples are assigned to each node, and the value of each branch is corresponding to the attributes [9].Some of decision tree techniques are CHAID, CART, ID3 etc.

(iii) **Artificial Neural Network:** Artificial Neural Network is unit non-liner predictive models that learn through training. Though there are powerful predictive modeling techniques. The auditors simply use them is reviewing records to spot fraud and fraud-like actions. they are higher utilized in things wherever they will be used and reused, like reviewing MasterCard transactions each month to envision for anomalies [6].

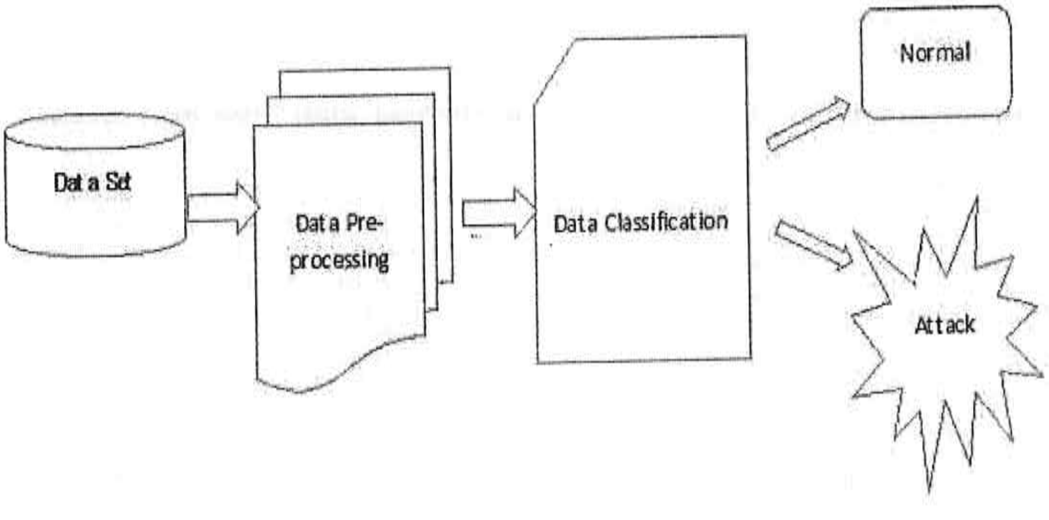


Fig. 1 Classification as IDS technique

(iv) **Statistical Techniques:** Statistical models [22,23] involving a latent structure often sup-port clustering, classification, and other data mining tasks. Because of their ability to deal with minimal information and noisy labels in a systematic fashion, statistical models of this sort have recently gained popularity. Statistical techniques are Bayesian net, support vector machine etc.

(v) **Ensemble Classifier:** The idea behind a Ensemble classifier[24,25] is to combine several machine learning techniques so that the system performance can be significantly improved. Ensemble model is combination of two or more IDS techniques; here combination of techniques is done to detect the attack, increase to avoid the drawbacks of individual models and to achieve high accuracy [26]. Ensemble techniques are Bagging, Boosting and Stacking.

II RELATED WORKS

Data mining is one of the key research area in the field of computational science. There are varieties of data mining techniques used for knowledge discovery. Mining techniques of key attention in current time are ensemble techniques. Genetic algorithm. Fuzzy logic and big data analytics. Data processing of Big data encounters many challenges. To explore data processing and security challenges, novel data processing and analysis techniques, some of recent works in these fields are:

(a) **Jemal H. Abawajy, Andrei KelREV, Morshed Chowdhury have worked on "Large iterative Multitier Ensemble Classifiers for security of Big data"[4].** The paper introduces and investigates large iterative multitier ensemble classifiers (LIME) for big data. They are generated automatically as a result of several iterations in applying ensemble meta classifiers. They incorporate diverse ensemble meta classifiers into several tiers simultaneously and integrate them into one iterative system. The four tier LIME classifiers based on Random forest achieved better performance compared with the base classifiers. The four tier LIME classifiers based on Multiboost at the fourth tier, decorate at third tier and bagging at second tier obtained the best outcome with AUC 0.998.

(b) **Peiying Tao , Zhe Sun, And Zhixin Sun have worked on "An Improved Intrusion Detection Algorithm Based on GA And SVM"[13].** The Author proposed an alarm intrusion detection algorithm (FWP-SVM-GA) based on the genetic algorithm (GA) and support vector machine (SVM) algorithm for use in human centred smart IDS. First, this paper makes effective use of the GA population search strategy and the capability of information exchange between individuals by optimizing the crossover probability and mutation probability of GA. The convergence of the algorithm is accelerated, and the training speed of the SVM is improved. A new fitness function is proposed that can decrease the SVM error rate and increase the true positive rate. Simulation and experimental results show that the improved intrusion detection technology based on the genetic algorithm (GA) and support vector machine (SVM) proposed in this paper increases the intrusion detection rate, accuracy rate and true positive rate; decreases the false positive rate; and reduces the SVM training time.

(c) **Abdel-Rahman Hedar, Mohamed A. Omer, Ahmad F. Al-Sadek And Adel A. Sewisy, proposed "Hybrid Evolutionary Algorithms For Data Classification in Intrusion Detection System"[03].** This research work is based on classification attack for Intrusion detection system. AGAAR is used to reduce features. A Classifier model built by GPLS is used to classify attacks in the NSL-KDD. The classifier trained with the full feature of 20%-NSL-KDD. The classifier was trained with 19.51% of the dataset features. The classifier accuracy improves the result after reducing dimensionality of the dataset. This reduction makes a significant improvement in term of memory and CPU time. This shows that AGAAR can remove the relevant features in intrusion detection. The experiment shows that the GPLS using reduced features is more accurate than that uses all of the features. These classifiers (AGAAR-GPLS) compared with others methods, show better results than many methods. The classification rate increased from 75.98% to 81.44% after reducing the features. In few cases, the results were close for few methods. The reduction of the dataset leads to minimizing the modeling time and computational costs.

(d) **Mohammad Saniee Abadeh, Hamid Mohamadi and Jafar Habibi, proposed "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks"[29].** In this paper, the use of different GFS (genetic fuzzy system) approaches is investigated to develop an intrusion detection system capable of detecting intrusive behaviors in a computer network. The characteristic features of the proposed GFSs can be summarized as follows:

(i) As intrusion detection is a high-dimensional classification problem one of the important properties of the proposed GFSs in this paper is that the class labels of all of the rules in the population and in each rule set (in Pittsburgh approach) are the same. This feature allows the algorithm to focus on learning of each class independently. Therefore the genetic fuzzy rule generation algorithm is repeated for each of the classes in the classification problem.

(ii) An initialization procedure is used to generate fuzzy if-then rules directly from the training data set. These rules enable the algorithm to focus on finding fuzzy rules, which are related to a special class. A same procedure is used to reinitialize the population at the end of each generation of the Michigan based GFS.

(iii) The genetic operators (i.e., crossover, and mutation) of the Michigan approach based GFS guaranteed to generate valid individuals. To achieve this, after performing the operator, consequent class of the generated individual is determined. If this class is the same as the parent class then the generated individual is accepted, otherwise the operator is repeated.

(e) **Bolón-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, proposed "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset"[30].** This paper proposes a method based on the combination of discretization, filtering and classification methods that maintains the performance results of the classifiers but using a reduced set of features. Specifically, it has been applied over the KDD Cup 99 dataset, a benchmark in the intrusion detection field. The proposed method, is based on classifiers like naive Bayes or C4.5, is applicable to large databases, since this machine learning algorithms have the advantages of being faster and more computational efficient than other classifiers used by other authors of the literature, such as SVM's, multilayer perceptions or functional networks. The comparative study denotes that the proposed method achieved a better performance than the other authors' results, specifically in those classes more difficult to detect.

(f) **Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni, worked on "Experiments on Detection of Denial of Service Attacks using Ensemble of Classifiers"[31].**The classifiers used are Naive Bayesian(NB), Bayesian Network(BN), Sequential Minimal Optimization(SMO), J48(C4.5) and Reduced Error Pruning Tree(REPTree). Result shows that Ensemble of Reduced Error Pruning Tree, Bayesian Network and J48 classifiers with no data pre-processing provided significant accuracy increment with minimal resource requirement. It is also proven that instead of developing a new classifier, one can achieve extremely high accuracy by using Ensemble of these multi-category classifiers.

(g) **Fatma Gumus, C. Okan Sakar, Zeki Erdem and Olcay Kursun, proposed "Online Naive Bayes Classification for Network Intrusion Detection" [32].** The proposed online naive Bayes classification method is first tested on the well-known Fisher-iris dataset which is available at UCI machine learning repository. The dataset has 150 samples, 3 classes, and each instance is represented with 4 features. The dimensionality of the data is reduced to two using the principal component analysis (PCA) to enable visualization on two axes. This method is based on the idea that weighting the most recent samples more allows the classifier to adapt to the recent attacks which may develop over time. The proposed method is time efficient compared to k-NN and more accurate than the linear perceptron.

(h) **Winston et. al. proposed "A Novel Technique to Detect DDoS and Sniffers in Smart Grid"[12].** In this paper, an IDS technique called double layer protection method is used for detection and isolation of sniffers, in first layer detection MD-5 technique is used and in the second layer detection PMD technique is used. DDoS attacks are detected using TTL analysis technique. Tools used for this detection purpose are NS-2 and CISCO. MD-5 safeguards the data integrity by encryption and decryption technique. PMD helps to find the source of the sniffing packets. NS-2 and network analyzer are the tools used for result comparison. Using these techniques and some features it gives the high efficiency.

(i) **Adhikari et. al. proposed Developing a "Hybrid Intrusion Detection System Using Data Mining for Power Systems"[14].** The IDS was trained an evaluated for a three-bus two-line transmission system which implements a two zone distance protection scheme. Twenty five scenarios consisting of stocktickerSLG faults. control actions, and cyber-attacks were implemented on a hardware-in-the-loop test bed. Scenarios were run in a loop 10000 times with randomized system parameters to create a dataset for IDS training and evaluation. The IDS correctly classified 90.4% of tested scenario instances. Evaluation also included a tenfold cross validation to evaluate the detection accuracy of zero-day attack scenarios. The average detection accuracy for zero-day attack scenarios was 73.43%. The common paths mining-based IDS out performs traditional machine learning algorithms and is better suited for the high volume of data present in power system.

(j) **Jaiyen et. al. proposed "Intrusion Detection Model Based on Ensemble Learning for U2R and R2L Attacks"[18].** In this paper, an algorithm is used for increasing the accuracy, and decreases the false alarm rate of U2R and R2L attacks by using Correlation-based feature selection and multiple weak classifiers such as Naïve Bayes, Decision Tree, MLP, k-NN and SVM based on Adaboost algorithm.

(k) **Lei xu, Chunxiao Jiang, Jian wang, Jian yuan and Yong ren have an article on "Information security in Big data: Privacy and Data Mining"[8].** In this paper they have highlights an emerging research topic in data mining known as privacy preserving data mining (PPDM). It focused on how to reduce the privacy risk brought by data mining operations. It reviews the privacy issues related to data mining by using a user role based methodology. Particularly four different types of users are identified which are involved in data mining applications, namely data provider, data collector, data miner, and decision maker. Each user role has its own privacy concerns. Hence the privacy preserving approach adopted by one user role different from those adopted by others.

(l) **Yenduri et. al. proposed "Analyzing Intrusion Detection System: An Ensemble based Stacking Approach"[19].** In this paper, an intrusion detection system uses stacking classifier which has detected different types of intrusions, for achieving good accuracy, precision, recall and ROC values. KDD cup99 dataset and weka data mining tool is used. Accuracy rate is 82.7206%.

(m) **Kulkarni et. al. worked on "Experiments on Detection of Denial of service Attacks using Ensemble of Classifiers"[20].** In this paper J48, Naive Bayesian classifiers and KDD 99 as a dataset are used. Ensemble of Naive Bayesian, J48 and Sequential Minimal Optimization classifiers, when combined with Numeric to Binary Data Pre-processing method provides maximum accuracy of 99.89785%. The same accuracy is also provided by the ensemble of Bayesian network, J48 and Sequential Minimal Optimization.

(n) **Pradhan et. al. worked "Performance Assessment of Robust Ensemble Model for Intrusion Detection using Decision Tree Techniques"[21].** In this paper J48, robust forest as a data mining technique and ensemble classifiers (bagging, boosting, and stacking) are used. Dataset which is used here is NSL-KDD and weka tool. The experiment results shows that Bagging classifiers provides highest accuracy 98.71%, stacking provides accuracy of 98.66% and boosting provides accuracy of 98.60% which is better than the accuracy of individual classifier J48 and Random forest. Not only in accuracy, in precision, have recall and f-measure also had ensemble techniques provided better results than individual classifiers.

(o) **Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni, worked on "Experiments on Detection of Denial of Service Attacks using Naïve Bayesian Classifier"[27].** This paper evaluates variation in performance of Naive Bayesian classifier for intrusion detection when used in combination with different data pre-processing and feature selection methods. Naive Bayesian classifier performed significantly better when combined with Numeric to Binary data pre-preprocessing. It can be also observed that, instead of going for an improved version of Naive Bayesian classifier or completely different set of multi-classifiers, one can achieve better performance using Naive Bayesian classifier along with Numeric to Binary data pre-processing. Experimental results prove that accuracy of Naive Bayesian classifier is improved and performs better than other classifiers when used in combination with Feature Selection and data pre-processing methods.

III PROBLEM IDENTIFICATION AND FINDINGS

Above literature review explores data processing and security challenges, novel data processing and analysis techniques to deal with security challenges. With the study of these research works the advantage and disadvantage of these literatures are summaries. It is observed that hybrid models, accuracy percentage is high from the base classifier. The author [12] uses MD-5, PMD and TTL techniques for detection of DDoS and Sniffers in smart grid. MD-5 safeguards the data integrity by using cryptography techniques. Detection of sniffers is done using very less bandwidth. Author [4] has uses large iterative multitier ensemble classifier for security of big data. With random forest, adaboost and bagging in Meta classifier in multitier, it results in high accuracy. Author [13] stated with genetic algorithm convergence of the algorithm is accelerated, and the training speed of the SVM is improved. A new fitness function decreases the SVM error rate and increases the true positive rate. The hybrid of cryptographic methods reduces the overhead. EAACK mechanism is used for this. The hybrid of GPLS and AGAAR has the classification rate 81.44%. If both of the models are used as a single classifier the rate will be low. AGAAR is used to reduce the features from the dataset; it removes the relevant features in intrusion detection. The hybrid of hardware and software based model which is used for grid, is mainly apply on high volume of data's. Author [18] uses Adaboost algorithm to create the ensemble of Decision Tree, Naïve Bayes, SVM, and MLP classifiers for detecting U2R and R2L attacks which are difficult to detect. The hybrid of Naive Bayes and MLP produces the highest sensitivity. Decision tree results the least performance. Author [19] uses stacking based ensemble classifier technique which provides the efficient result, accuracy rate is high and dataset which is used here is easily available. The multi-layer hybrid technique uses PCA for feature selection and comparison purpose. Classifiers uses here are fast and highly independent. The hybrid of J48 and Naive Bayesian network with no data pre- processing provides accuracy and less resource requirement. Author [29] investigated the use of different GFS (genetic fuzzy system) approaches to develop an intrusion detection system capable of detecting intrusive behaviors in a computer network.

IV CONCLUSION

In this paper, a literature study of recent work in the field of data mining is presented. In which different hybrid classifiers are analyzed based on their performance and result. It explored data processing and security challenges, novel data processing and analysis techniques to deal with security challenges. We see some of the single classifiers with their drawbacks and strengths. There is a need of hybrid system for better result. It is not necessary that every single model has some drawback but it is observed that hybrid models give better results and performance. Many combinations of machine learning techniques are tested on hybrid models and still there are provisions for different combination of machine learning techniques which can be tested on ensemble (hybrid) models for better result. It is well known that Challenges in big data are extracting data, data storage and analysis, searching, querying and updating data, information privacy. Various dimensions to deal with these issues are ensemble techniques. Genetic algorithm, Fuzzy logic and big data analytics, which are playing key role in the security concern of Big data and data mining.

REFERENCES

[1] Farid Lawan Bello,Kiran Ravulakollu, Amrita (2015) "Analysis and Evaluation of Hybrid Intrusion Detection System models", in international conference on co.nputers, communication and systems.

[2] James P Anderson "Computer security threat monitoring and surveillance" in technical report, James P Anderson company, fort woshington, Pennsylvania.1980.

[3] Abdel-Rahman Hedar, Mohamed A. Omer and Ahmed F. Al-Sadek, Adel A. Sewisy,(2015) "hybrid evolution algorithm for data classification in IDS". in IEEE SNPD 2015, June 1-3 2015, Takamatsu, Japan.

[4] Jemal H. Abawajy, Andrei KelREV, Morshed Chowdhury (2014)" Large iterative Multitier Ensemble Classifiers for security of Big data", in IEEE Transactions on EMERGING TOPICS IN COMPUTING, Volume 2, No. 3, September 2014.

[5] Lei Li,De-Zhang Yang, Fang-Cheng Shen (2010)"A Novel Rule Based Intrusion detection system Using Data Mining" in the Proc . Of 3rd IEEE International conference on computer science and information technology, pp. 169-172,2010.

[6] S.Revathi, Dr. A .Malathi (2013)"A detailed analysis on NSL-KDD Dataset using various machine learning techniques for intrusion detection", in International Journal of Engineering &Technology (IJERT),ISSN: 2278-0181.vol. 2 issue 12. December-2013.

[7] DE Goldberg and H Holland, Genetic algorithms and machine learning, Machine learning, Vol. 3(2), 95-99,1988.

[8] Lei xu, Chunxiao Jiang, Jian wang, Jian yuan , Yong ren (2014)"Information security in Big data: Privacy and Data Mining" in IEEE access, The journal for rapid open access publishing, Volume 2, 2014.

[9] Chih-Fong Tsai a, Yu-Feng Hsu b, Chia-Ying Lin c, Wei-Yang Lin d,(2009) "Intrusion detection by machine learning: A review" in Expert Systems with Applications, Elsevier, 36 (2009) 11994–12000, 0957-4174/ 2009.

[10] Amrita anand, Brajesh Patel (2012)"An Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering Different Protocol" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8, August 2012.

[11] Sanjiban Sekhar Roy, P Venkata Krishna, Sumanth Yenduri, (2014)"Analyzing Intrusion Detection System: An Ensemble based Stacking Approach", 978-1-4799-1812-6/14 ©2014 IEEE.

[12] S.Shitharth, Dr.D.Prince Winston,(2016) "A Novel IDS Technique to Detect DDoS and Sniffers in Smart Grid" in 2016 world conference on futuristic trends in research and innovation for social welfare.

[13] Peiying Tao , Zhe Sun, And Zhixin Sun (2018) "An Improved Intrusion Detection Algorithm Based On GA and SVM" in Special section on human-centered smart systems and technologies, IEEE ACCESS, volume 6, 2018.

[14] Shengyi Pan, Thomas Morris, Uttam Adhikari (2015) "Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems" in IEEE TRANSACTIONS ON SMART GRID, VOL. 6, NO. 6, NOVEMBER 2015.

[15] M. Revathi, T. Ramesh, "Network intrusion detection system using reduced dimensionality" in Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, Vol. 2 No. 1 pp. 61 -67,96.

[16] Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda, Zhiyuan Tan (2016)"Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm" in IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 10, OCTOBER 2016.

[17] W. Feng, Q. Zhang, G. Hu, J. X. Huang, "Mining network data forIntrusion detection through combining SVMs with Ant colony networks" in Elsevier, Future Generation Computer Systems 37(2014) 127 – 140.

[18] Ploypharn Sornsuwit, SaichonJaiyen, (2015)"Intrusion Detection Model Based on Ensemble Learning for U2R and R2L Attacks" in 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE),chiangmai , Thailand.

[19] Sanjiban Sekhar Roy, P Venkata Krishna, Sumanth Yenduri (2014)"Analyzing Intrusion Detection System: An Ensemble based Stacking Approach" in 978-1-4799-1812-6/14 ©2014 IEEE.

[20] Mr. Vijay D.Katkar, Mr.Siddhant Vijay Kulkarni,(2013) "Experiments on Detection of Denial of Service Attacks using Ensemble Classifier" in 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 978-1-4673-6126-2/13/ 2013 IEEE.

[21] Reshamlal Pradhan, Deepak Kumar Xaxa,(2014)" Performance Assessment of Robust Ensemble Model for Intrusion Detection using Decision Tree Techniques", in International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 3, Issue 3 May 2014.

[22] Arun K. Pujari, (2001), Data mining techniques, 4th edition, Universities Press (India) Private Limited.

[23] Jiawei Han, Micheline Kamber, (2006), "Data mining concepts and tech-niques", Second edition, San Francisco, Margan Kaufmann Publishers, USA.

[24] Manish Kumar Nagle, Dr. Setu Kumar Chaturvedi (2013)"Feature Extraction Based Classification Technique for Intrusion Detection System" in International Journal of Engineering Research and Development (August 2013).

[25] Mrutyunjaya Panda, (2011) "A hybrid intelligent approach for network intrusion detection", Procedia Engineering.

[26] Anup Ashok Patil, ShitalMali (2016)"Hybrid Cryptography Mechanism for SecuringSelf-Organized Wireless Networks" in 2016 3rd International Conference on Advanced Computing and Communication Systems (*ICACCS* -2016), Jan. 22 – 23, 2016, Coimbatore, INDIA.

[27] Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni, (2013)"Experiments on Detection of Denial of Service Attacks using Naïve Bayesian Classifier" in International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), IEEE(2013).

[28] Reshamlal Pradhan, Deepak Kumar Xaxa,(2014) "Robust Ensemble Model for Intrusion Detection using Data Mining Techniques" in International Journal of Scientific & Engineering Research, Volume 5, Issue 4, April-2014 781 ISSN 2229-5518.

[29] Mohammad Saniee Abadeh, Hamid Mohamadi, Jafar Habibi (2011)"Design and analysis of genetic fuzzy systems for intrusion detection in computer networks" in Expert Systems with Applications 38 (2011) 7067–7075, Elsevier.

[30] V. Bolón-Canedo, N.Sánchez-Maroño, A. Alonso-Betanzos, (2011) "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset" in Expert Systems with Applications 38 (2011) 5947–5957, Elsevier.

[31] Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni, (2013)" Experiments on Detection of Denial of Service Attacks using Ensemble of Classifiers" in International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), IEEE(2013).

[32] Fatma Gumus, C. Okan Sakar, Zeki Erdem, Olcay Kursun (2014)"Online Naive Bayes Classification for Network Intrusion Detection" [32] in 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014).

[33] Amira Sayed A. Aziz, Aboul Ella Hassanien, Sanaa El-Ola Hanafy, M.F.Tolba (2013)" Multi-layer hybrid machine learning techniques for anomalies detection and classification approach", 978-1-4799-2439-4/13/ ©2013 IEEE.

A Survey on Features and Techniques of Blog Spammer Identification

Rupali Dohare[1], Dr. Pratima Gautam[2]

[1]Research Scholar, Rabindranath Tagore University, Bhopal (M.P.) India.

[2]Dean of CS/IT, Rabindranath Tagore University, Bhopal (M.P.) India.

ABSTRACT

Blogging becomes a popular way for a Web user to publish information on the Web. Bloggers write blog posts, share their likes and dislikes, voice their opinions. Activities happened in Blogosphere affect the external world. This attract many promoters hire some bloggers who post to increase heights of those brands or products. So spamming is a major problem in internet-based things as well as in social media. Different techniques have been proposed for spam filtering have been exposed across various platforms with varies degree of measures. This survey focused on some of the present strategies used for filtering social spam. Starting with different types of social spam, the paper has discussed about recent developments in the field of elimination of social spam. This paper gives a concise study of methods proposed by different researchers. Here various features of spammer profile identification were also done with a comprehensive and comparative understanding of existing literature.

***Index Terms*—** Content filtering, Fake Profile, Online Social Networks, Spam Detection.

I INTRODUCTION

During the recent few years social media has evolved many folds and has become much more interactive and integral part of our lives. The interaction channels in the social media have changed from traditional media like newspapers and television to mobile phones, social media websites, micro blogging sites etc. It has changed the way people communicate with each other on the personal as well as on public from as described in [1]. There are varieties of social media sites that offer diverse functionality, some are for common people like Facebook, which started as an experimental social network in the Harvard University by some students, while others like LinkedIn is a network formed by professionals from every field. Many sites are exclusively for sharing videos and pictures media like YouTube, Instagram, Flickr etc. while others focused on blogs where people from varied domains express and share their views. There are even social tagging and news sites like Reddit, Delicious etc. which allow the user to rank the websites on the basis of quality of content and usefulness of the sites. Most recent trend of micro – blogging let people update the real – time status of their daily routine or happenings via app like Twitter which has more than 200 million users exchanging more than 400 million tweets per day [2] where the length of tweets is limited to 140 characters.

According to Teen, Social Media and Technology Overview 2015 [3], —More than 24% of the teen are constantly online and 71% of them use more than one social networking sitell. This ease of sending and receiving data over Internet has resulted in some notorious people sending unwanted messages to large number of recipients over the network trying to take advantage by getting access to their privacy. Initial spread of spams started with email spam. According to M3AAWG report, the abusive email content amounts to 87.1% - 90.2% of the total email content during 2012 – 2014 [4] which has increased the financial burden by increasing the storage requirement and technological requirement for spam detection. Slowly spams started spreading in every digital media like from mobile network through mobile phone, social networking sites, blogs, review sites etc.

The rest of this paper is organized as follows: in the second section, the requirement of blog spamming was discussed. Third section list various techniques adopt by spammer to promote their target website, brand, product, etc. While fourth section provide related work of the current approaches applied by different researchers to identify blog spammers. Research problem is pointed out, and then the proposed problem is formalized in detail. The conclusion of the whole paper is made in the last section.

II REQUIREMENT OF BLOG SPAMMING

Due to machine-generated nature and its focus on search engines manipulation, spam shows abnormal properties such as high level of duplicate content and links; rapid changes of content; and the language models built for spam pages deviate significantly from the models built for the normal Web.

(a) Spam pages deviate from power law distributions based on numerous web graph statistics such as Page Rank or number of in-links.

(b) Spammers mostly target popular queries and queries with high advertising value.

(c) Spammers build their link farms with the aim to boost ranking as high as possible, and therefore link farms have specific topologies that can be theoretically analyzed on optimality.

(d) According to experiments, the principle of approximate isolation of good pages takes place: good pages mostly link to good pages, while bad pages link either to good pages or a few selected spam target pages. It has also been observed that connected pages have some level of semantic similarity – topical locality of the Web, and therefore label smoothing using the Web graph is a useful strategy.

(e) Numerous algorithms use the idea of trust and distrust propagation using various similarity measures, propagation strategies and seed selection heuristics.

(f) Due to abundance of "neponistic" links, that negatively affect the performance of a link mining algorithm, there is a popular idea of links removal and down weighting. Moreover, the major support is caused by the k-hop neighborhood and hence it makes sense to analyze local sub graphs rather than the entire Web graph.

(g) Because one spammer can have a lot of pages under one website and use them all to boost ranking of some target pages, it makes sense to analyze host graph or even perform clustering and consider clusters as a logical unit of link support.

(h) In addition to traditional page content and links, there are a lot of other sources of information such as user behaviour or HTTP requests. We hope that more will be developed in the near future. Clever feature engineering is especially important for web spam detection.

(i) Despite the fact that new and sophisticated features can boost the state-of-the-art further, proper selection and training of a machine learning models is also of high importance.

III TECHNIQUES OF SPAMMING

Term spamming techniques can be grouped based on the text field in which the spamming occurs [5]. Therefore, we distinguish:

(a) **Body spam** In this case, the spam terms are included in the document body. This spamming technique is among the simplest and most popular ones, and it is almost as old as search engines themselves [6].

(b) **Title spam** Today's search engines usually give a higher weight to terms that appear in the title of a document. Hence, it makes sense to include the spam terms in the document title [7].

(c) **Meta tag spam** The HTML meta tags that appear in the document header have always been the target of spamming. Because of the heavy spamming, search engines currently give low priority to these tags, or even ignore them completely. Here is a simple example of a spammed keywords meta tag:

(d) **Anchor text spam** Just as with the document title, search engines assign higher weight to anchor text terms, as they are supposed to offer a summary of the pointed document. Therefore, spam terms are sometimes included in the anchor text of the HTML hyperlinks to a page. Please note that this spamming technique is different from the previous ones, in the sense that the spam terms are added not to a target page itself, but the other pages that point to the target. As anchor text gets indexed for both pages, spamming it has impact on the ranking of both the source and target pages [7].

(e) **URL spam** Some search engines also break down the URL of a page into a set of terms that are used to determine the relevance of the page [8]. To exploit this, spammers sometimes create long URLs that include sequences of spam terms. For instance, one could encounter spam URLs.

Some spammers even go to the extent of setting up a DNS server that resolves any host name within a domain. Often, spamming techniques are combined. For instance, anchor text and URL spam is often encountered together with link spam. Another way of grouping term spamming techniques is based on the type of terms that are added to the text fields. Correspondingly:

(i) Repetition of one or a few specific terms. This way, spammers achieve an increased relevance for a document with respect to a small number of query terms.

(ii) Dumping of a large number of unrelated terms, often even entire dictionaries. This way, spammers make a certain page relevant to many different queries. Dumping is effective against queries that include relatively rare, obscure terms: for such queries, it is probable that only a couple of pages are relevant, so even a spam page with a low relevance/importance would appear among the top results.

(iii) Weaving of spam terms into copied contents. Sometimes spammers duplicate text corpora (e.g., news articles) available on the Web and insert spam terms into them at random positions. This technique is effective if the topic of the original real text was so rare that only a small number of relevant pages exist. Weaving is also used for dilution, i.e., to conceal some repeated spam terms.

IV RELATED WORK

Muhammad U. S. Khan et. al. [8] proposes a framework that separates the spammers and unsolicited bloggers from the genuine experts of a specific domain. The proposed approach employs modified Hyperlink Induced Topic Search (HITS) to separate the unsolicited bloggers from the experts on Twitter on the basis of tweets. The approach considers domain specific keywords in the tweets and several tweet characteristics to identify the unsolicited bloggers.

Y. Chen et. al. [9] utilize graph-based detection due to less security guarantee in feature-based detection. Assuming that fake profiles can establish limited number of intruded (attack) edges, the sub graph formulated by the set of all real accounts is sparsely connected to false account, that is, the cut over intruded edges is sparse. This method makes prediction and find out such sparse cut with formal guarantees. For example, Tuenti deploy SybilRank to rank accounts according to their perceived likelihood of being false, based on structural properties of its social graph and based on their formulation.

H. Gao et. al. [10] utlize graph-based detection provides comfortable security guarantees, real-world social graphs do not conform to the main assumption on which it depends. In particular, various surveys conform that intruders can interstices OSNs on a large scale by deceiving users into befriending their fake profile.

Taghi Javdani et. al. [11]apply the hybrid graph analysis method and behavior analysis, is to increase the diagnostic accuracy and detection rate with the help of appropriate classification algorithms and the most effective features. So, two scenarios were used to achieve higher accuracy level and lower false positive. The first scenario was based on using the entire data to build and evaluate the model. The results showed that despite the high precision of this approach, due to the high levels of false positive, this approach is not appropriate. In the second scenario, the ratio of the normal users to spammers was considered equal to 2 to 1 which led to satisfactory results. After reviewing the confusion matrix and false positives in different algorithms, the Logistic algorithm was chosen as an appropriate algorithm which meets the objective of this study.

Hailu Xu et al.[12]. Studied a methodology to detect spam across online social networks. This methodology focuses on combining spam in one soial network to another social network. They had used 1937 spam tweets and 10942 ham tweets and 1338 spam posts and 9285 ham posts. In TSD, out of 1937 spam tweets, 75.6% spam tweets contained in URL links, 24.4% spam tweets contained in words. From 10942 ham tweets, 62.9% tweets are in URL links and words, remaining 37.1% consist of only words.For the spam posts of FSD, 32.8% spam posts consists of URL links and words, 67.2% of spam posts consist of words. For ham posts 95.1% consist of URL links and 4.9% only consist of words. They had used top 20 word features from Twitter spam data and Face book spam data. They had split the TSD and FSD into training and test data sets .The training and test data sets of TSD, FSD are used to train and test various classifiers like Random forest, logistic, random tree, Bayes Net, Naïve bayes.

M. Okazaki et al. [13].presented an initial study to quantify and characterize spam campaigns launched using accounts on Face book. They studied a large anonym zed dataset of 187 million asynchronous wall messages between Face book users, and used a set of automated techniques to detect and characterize coordinated spam campaigns. Authors detected roughly 200,000 malicious wall posts with embedded URLs, originating from more than 57,000 user accounts.

Fire et al. [14].developed the Social Privacy Safeguard (SPS) software, which is a set of applications for Face book that aim to improve user account privacy policies. The application examines a user's friends list in order to determine accounts that have a risk to the user's privacy. Such accounts could then be protected by users from accessing their profile information. Using these set of data from the SPS developed over Face book, the authors could test several machine learning classifiers to detect fake profiles, some algorithms are been used: Naïve Bayes, Rotation Forest and Random Forest are been used for fake profile detection.

Pern Hui et al. [15] Third-party applications capture the attractiveness of web and platforms providing mobile application. Many of these platforms accept a decentralized control strategy, relying on explicit user consent for yielding permissions that the apps demand. Users have to rely principally on community ratings as the signals to classify the potentially unsafe and inappropriate apps even though community ratings classically reflect opinions regarding supposed functionality or performance rather than concerning risks. To study the advantages of user-consent permission systems through a large data collection of Face book apps, Chrome extensions and Android apps. The study confirms that the current forms of community ratings used in app markets today are not reliable for indicating privacy risks an app creates. It is found with some evidences, indicating attempts to mislead or entice users for granting permissions: free applications and applications with mature content request; "look alike" applications which have similar names as that of popular applications also request more permissions than is typical. Authors find that across all three platforms popular applications request more permissions than average.

J. Kim et al. [16] Twitter can suffer from malicious tweets containing suspicious URLs for spam, phishing, and malware distribution. Attackers have limited resources and thus have to reuse them; a portion of their redirect chains will be shared. We focus on these shared resources to detect suspicious URLs. We have collected a large number of tweets from the Twitter public timeline and trained a statistical classifier with features derived from correlated URLs and tweet context information. Our classifier has high accuracy and low false- positive and false negative rates.

Malik Mateen et al.[17] studied an approach for spam detection in Twitter network. To detect spam in Twitter dataset used different kind of features like user based features, content based features and graph based features. User based features are based on users relationships and properties of user accounts. The spammers have to reach large number of profiles to spread misinformation. Different user account related features are Number of followers, Number of following, age of account, FF ratio and reputation. Content based features are related to tweets posted by user. Different features are total number of tweets, hash tag ratio, URL's ratio, mentions ratio, tweet frequency and spam words. Graph based features are used to identify spammer behaviour. Different features are in/out degree and between's. In the proposed methodology used Twitter dataset consist of 10,256 users and 467480 tweets. To develop a spam detection model used J48, decorate and Naive ayes classifiers. These three classifiers are individually trained on various dataset features and classify the dataset as spam or ham dataset. Out of these three classifiers J48 classifier highest accuracy to classify the data as spam or non spam.

Content based features are best suitable for classifying the dataset. To classify the dataset with highest accuracy combine the content, user based and graph based features. The combined feature set is given as input to the three classifiers. But decorate and J48 classifiers have given highest accuracy up to 97.6%.

Fire et al. [18] developed the Social Privacy Safeguard (SPS) software, which is a set of applications for Facebook that aim to improve user account privacy policies. The application examines a user's friends list in order to determine accounts that have a risk to the user's privacy. Such accounts could then be protected by users from accessing their profile information. Using these set of data from the SPS developed over Facebook, the authors could test several machine learning classifiers to detect fake profiles, some algorithms are been used:Naïve Bayes, Rotation Forest and Random Forest are been used for fake profile detection.

V CONCLUSION

With the rapid growth of social networks, people tend to misuse them for unethical and illegal conducts, fraud and phishing. Creation of a fake profile becomes such adversary effect which is difficult to identify without appropriate research. So this paper have summarize current solutions that have been practically developed and theorized to solve this issue of spam detection issue and spam identification of fake profiles. Here it was obtained that spammers develop high social networking sites than create fake profile on that and start there blogging for target product. It was obtained that most of work use clustering techniques for segregating spammer from real users by reading their behavior on sites. In future it is desired to develop the highly accurate algorithm which not only detects the spam but spammer profile as well.

REFERENCES

[1] Van Dijck, José. The culture of connectivity: A critical history of social media. Oxford University Press, 2013.

[2] Boyd, D., & Ellison, N. (2008). Social network sites: Definition, history, and scholarship. Journal of Computer Mediated Communication, 13(1), 210—23.

[3] Lenhart, Amanda. "Teens, social media & technology overview 2015." Pew Research Center 9 (2015).

[4] Gauri Jain* , 2Manisha, 3Basant Agarwal. "An Overview of RNN and CNN Techniques for Spam Detection in Social Media". Volume 6, Issue 10, October 2016 ISSN: 2277 128X.

[5] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In Proceedig of the 16th European Conference on Machine Learning, ECML'05, 2005.

[6] C. D. Manning, P. Raghavan, and H. Schtze. Introduction to Information Retrieval. Cambriuge University Press, New York, NY, 2008.

[7] O. A. Mcbryan. GENVL and WWWW: Tools for taming the web. In Proceedings of the First World Wide Web Conference, WWW'94, Geneva, Switzerland, May 1994.

[8] Muhammad U. S. Khan, Mazhar Ali, Assad Abbas,Samee U. Khan, *and* Albert Y. Zomaya. "Segregating Spammers and Unsolicited Bloggers from Genuine Experts on Twitter". IEEE Computer Society, 2017.

[9] 28. H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. N. Choudhary. Towards Online Spam Filtering in Social Networks. In NDSS, 2012.

[10] 29. H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and Characterizing Social Spam Campaigns. In Internet Measurement Conference, pp 35–47. ACM, 2010.

[11] Mona Najafi Sarpiri, Taghi Javdani Gandomani, Mahsa Teymourzadeh, Akram Motamedi. "A Hybrid Method for Spammer Detection in Social Networks by Analyzing Graph and User Behavior". Journal of computers, Volume 13, Number 7, July 2018.

[12] Hailu Xu,Weiqing sun,Ahmad javaid: Efficient spam detection across online social networks,IEEE-2015.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo: "Realtime event detection by social sensors", In Proceedings of the 19thinternational conference on World wide web ACM, 2010.

[14] M. Fire, D. Kagan, A. Elyashar, Y. Elovici, Friend or foe? fake profile identification in online social networks. Social Network Analysis and Mining 4 (1) 1–23, 2014.

[15] Chia, Pern Hui, Yusuke Yamamoto, and N. Asokan. "Is this app safe? a large scale study on application permissions and risk signals." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[16] S. Lee and J. Kim, WarningBird: Detecting suspicious URLs in Twitter stream,‖ in Proc. NDSS, 2012.

HSES Knowledge Portal: Study of Feedbacks of Users

Santosh Kumar Miri[1], **Dr. Neelam Sahu**[2]
[1]Research Scholar, IT & CA, Dr. C. V. Raman University, Bilaspur (C.G.) India.
[2]Associate Professor, Dept. of IT, Dr. C. V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

HSES Knowledge portal is a web portal in which syllabus, eBooks, question papers and video lectures for class XI & XII are ported. The students, teachers and other people are benefited from study material of the portal. IK, CLSP, TM, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL and ASM are fields of student table of HSES database. Similarly, IK, CLSP, LM, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL and ASM are fields of teacher fields of HSES database. IK, CLSP, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL, ASM, Help and Parent are fields of others table of HSES database. LM, TM, Help and Parent are a unique fields of the student, teacher and others tables. Feedback can give by the registered members of the portal. Data is collected during the filling of feedback forms. Different counters are developed for parameter wise counting of feedback. We have arbitrary took a standard that any grade above 80 percent is marked as very good, between 60 to 80 as fairly good and below 60 as average. The appreciable grade is found from the teachers, students or others for all feedback parameters. Grading of feedback parameters are done based on yes answers of feedback questionnaires.

Keywords: Feedback parameters, Counters of feedback parameters, Grading of Feedback parameters

I INTRODUCTION

The HSES Knowledge portal is a web portal in which syllabus, eBooks, question papers and video lectures for class XI & XII are ported. The students, teachers and other people are benefited from study material of the portal. IK, CLSP, TM, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL and ASM are fields of student table of HSES database. Similarly, IK, CLSP, LM, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL and ASM are fields of teacher fields of HSES database. IK, CLSP, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL, ASM, Help and Parent are fields of others table of HSES database. LM, TM, Help and Parent are a unique fields of the student, teacher and others tables. IK, CLSP, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL and ASM are common feedback parameters. Feedback can give by the registered members of the portal. Data is collected during the filling of feedback forms. Different counters are developed for parameter wise counting of feedback. We have arbitrary took a standard that any grade above 80 percent is marked as very good, between 60 to 80 as fairly good and below 60 as average. The appreciable grade is found from the teachers, students or others for all feedback parameters. Grading of feedback parameters are done based on yes answers of feedback questionnaires. Data related to 'No' answer of feedback is negotiable for the grading system. The graphical representation of collected data for feedback parameters are shown in figures. The HSES database of HSES Knowledge Portal is created in the MySQL database management system under XAMPP Control Panel. Feedbacks are given by registered members of the portal. There are separate registration form is available for each category of the user.

II FEEDBACK PARAMETERS

IK, CLSP, LM, TM, QP, SY, Time, Region, HY, PExam, SMSearch, Euse, AMCL, SMDL, ASM, Help and Parent are feedback parameters. LM is feedback field of student table only. TM is feedback field of teacher table only. Parent and Help are feedback fields of others table only.

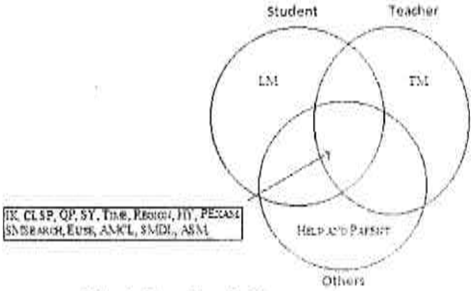

Student
Teacher
LM
TM
IK, CLSP, QP, SY, Time, Region, HY, PExam SMSearch, Euse, AMCL, SMDL, ASM
Help and Parent
Others

Fig.1 Feedback Parameters

III COUNTERS OF FEEDBACK PARAMETERS

Feedbacks of feedback parameters of student, teacher or others tables are counted by separate counters which are IK counter, CLSP counter, LM counter, TM counter, QP counter, SY counter, Time counter, Region counter, HY counter, PExam counter, SMSearch counter, Euse counter, AMCL counter, SMDL counter, ASM counter Parent counter and Help counter. This is done by given steps:

(i) Open HSES database
(ii) Select Student/Teacher/Others
(iii) Select Feedback field
(iv) For I=1 to N
(v) count=count+1
(vi) End of loop
(vii) Close HSES database

IV OBSERVATION & INTERPRETATION

In the portal www.hseshiksha.in, user's feedbacks are collected into Teachers, Students and Others. In the www.hseshiksha.in, feedbacks from 93 teachers are collected between September 2018 and January 2019. Feedbacks are collected online with 14 parameters of questionnaire. The 14 parameters of questionnaire are IK, CLSP, TM, QP, SY, Time, Region, HM, PExam, SMSearch, Euse, AMCL, SMDL and ASM.

Feedbacks from the teachers are shown in the Table 1. 79 teachers have knowledge of Internet (IK) which is 84.95% and 14 teachers have not IK which is 15.05%. 81 teachers can operate the computer, laptop or smart phone (CLSP) which is 87.10% and 12 teachers can't operate the CLSP which is 12.90%. It says that a very good number of teachers are registered member of the portal who has IK and can operate the CLSP. 82 teachers are said "Yes" about the availability of teaching materials which is 88.17% and 11 teachers are said "No" for the availability of teaching materials which is 11.83%. 81 teachers are said "Yes" about the availability of question papers (QP), syllabus (SY) and any time which is 87.10% each and 12 teachers are said "No" for the same which is 12.90% each. 75 teachers are said "Yes" about the availability of any region (Region) which is 80.65% and 18 teachers are said "No" for the same which is 19.35%.79 teachers are said "Yes" about the comfort-ability of Hindi medium users (HM) which is 84.95% and 14 teachers are said "No" for the same which is 15.05%. 81 teachers are said "Yes" for the preparation of examination which is 87.10% and 12 teachers are said "No" for the same which is 12.90%. 81 teachers are said "Yes" for the searching of study materials (SMSearch) which is 87.10% and 12 teachers are said "No" for the same which is 12.90%. 81 teachers are said "Yes" for the easier use of portal (Euse) which is 87.10% and 12 teachers are said "No" for the same which is 12.90%. 81 teachers are said "Yes" for the accessing of portal through the mobile, computer or laptop (AMCL) which is 87.10% and 12 teachers are said "No" for the same which is 12.90%.

77 teachers are said "Yes" for the downloading of study material (SMDL) which is 82.80% and 16 teachers are said "No" for the same which is 17.20%. 81 teachers are said "Yes" for the accessing of study material (ASM) which is 87.10% and 12 teachers are said "No" for the same which is 12.90%. 11.83% to 19.35% teachers have given the answer with "No" option. Similarly, 80.65% to 88.17% teachers have given the answer with "Yes" option. We have arbitrary took a standard that any grade above 80 percent is marked as very good, between 60 to 80 as fairly good and below 60 as average.

Table 1
Observation of Feedback from Teachers

Feedback Particular	Feedback from Teachers				
	In Figure		In Percentage		
	No	Yes	No	Yes	Grade
IK	14	79	15.05	84.95	Very Good
CLSP	12	81	12.90	87.10	Very Good
TM	11	82	11.83	88.17	Very Good
QP	12	81	12.90	87.10	Very Good
SY	12	81	12.90	87.10	Very Good
Time	12	81	12.90	87.10	Very Good
Region	18	75	19.35	80.65	Very Good
HM	14	79	15.05	84.95	Very Good
Pexam	12	81	12.90	87.10	Very Good
SMScarch	12	81	12.90	87.10	Very Good
Euse	12	81	12.90	87.10	Very Good
AMCL	12	81	12.90	87.10	Very Good
SMDL	16	77	17.20	82.80	Very Good
ASM	12	81	12.90	87.10	Very Good

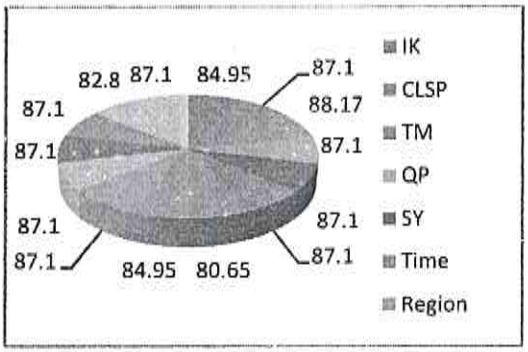


Fig. 2 Showing of Feedback from Teachers

In the www.hseshiksha.in, feedbacks from 301 students are collected between September 2018 and January 2019. Feedbacks are collected online with 14 parameters of questionnaire. Users can give answer of question. The 14 parameters of questionnaire are IK, CLSP, LM, QP, SY, Time, Region, HM, Pexam, SMSearch, Euse, AMCL, SMDL and ASM. Feedbacks from the students are shown in the Table-FS. 292 students have knowledge of Internet (IK) which is 97.01% and 9 teachers have not IK which is the availability of question papers (QP) which is 94.02% and 18 students are said "No" for the same 2.99%. 291 students can operate the computer, laptop or smart phone (CLSP) which is 96.68% and 10 students can't operate the CLSP which is 3.32%. It says that a very good number of students are registered member of the portal who has IK and can operate the CLSP.

288 students are said "Yes" about the availability of learning materials (LM) which is 95.68% and 13 students are said "No" for the availability of LM which is 4.32%. 283 students are said "Yes" about which is 5.98%. 288 students are said "Yes" about the availability of syllabus (SY) which is 95.68% and

13 students are said "No" for the same which is 4.32%. 284 students are said "Yes" about the availability of any time (Time) which is 94.35% and 17 students are said "No" for the same which is 5.65%. 279 students are said "Yes" about the availability of any region (Region) which is 92.69% and 22 students are said "No" for the same which is 7.31%. 273 students are said "Yes" about the comfort-ability of Hindi medium users (HM) which is 90.70% and 28 students are said "No" for the same which is 9.03%.

Table 2
Feedback from Students

Feedback Particular	Feedback from Students				
	In Figure		In Percentage		
	No	Yes	No	Yes	Grade
IK	9	292	2.99	97.01	Very Good
CLSP	10	291	3.32	96.68	Very Good
LM	13	288	4.32	95.68	Very Good
QP	18	283	5.98	94.02	Very Good
SY	13	288	4.32	95.68	Very Good
Time	17	284	5.65	94.35	Very Good
Region	22	279	7.31	92.69	Very Good
HM	28	273	9.30	90.70	Very Good
Pexam	38	263	12.62	87.38	Very Good
SMSearch	21	280	6.98	93.02	Very Good
Euse	13	288	4.32	95.68	Very Good
AMCL	18	283	5.98	94.02	Very Good
SMDL	19	282	6.31	93.69	Very Good
ASM	16	285	5.32	94.68	Very Good

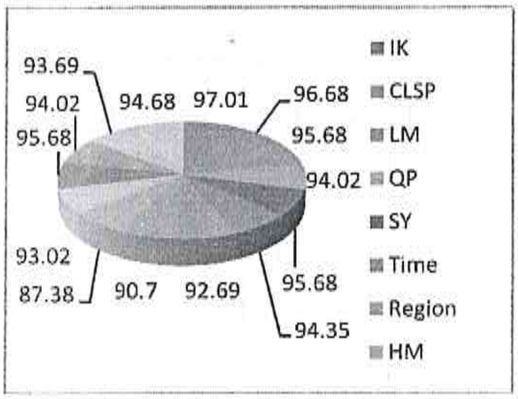


Fig. 3 Showing of Feedback from Students

263 students are said "Yes" for the preparation of examination which is 87.38% and 38 students are said "No" for the same which is 12.62%. 280 students are said "Yes" for the searching of study materials (SMSearch) which is 93.02% and 21 students are said "No" for the same which is 6.98%. 288 students are said "Yes" for the easier use of portal (Euse) which is 95.68% and 13 students are said "No" for the same which is 4.32%. 283 students are said "Yes" for the accessing of portal through the mobile, computer or laptop (AMCL) which is 94.02% and 18 students are said "No" for the same which is 5.98%. 282 students are said "Yes" for the downloading of study material (SMDL) which is 93.69% and 19 students are said "No" for the same which is 6.31%. 285 students are said "Yes" for the accessing of study material (ASM) which is 94.68% and 16 students are said "No" for the same which is 5.32%. 2.99% to 12.62% students have given the answer with "No" option. Similarly, 87.38% to 97.01% students have given the answer with "Yes" option. We have arbitrary took a standard that any grade above 80 percent is marked as very good, between 60 to 80 as fairly good and below 60 as average.

In the www.hseshiksha.in, feedbacks from 110 Others are collected between September 2018 and January 2019. Feedbacks are collected online with 15 parameters of questionnaire. Other category of users can give answers of questions related to these 15 parameters. The 15 parameters of questionnaire are IK, CLSP, QP, SY, Time, Region, HM, PExam, SMSearch, Euse, AMCL, SMDL, ASM, Parent and Help. Parent and Help are new parameters. Feedbacks from the others except teachers and students are shown in the Table-FO. 91 others have knowledge of Internet (IK) which is 82.73% and 19 others have not IK which is 17.27%. 97 others can operate the computer, laptop or smart phone (CLSP) which is 88.18% and 13 others can't operate the CLSP which is 11.82%. It says that a very good number of others are registered member of the portal who has IK and can operate the CLSP. 104 others are said "Yes" about the availability of question papers (QP) which is 94.55% and 6 others are said "No" for the same which is 5.45%. 102 others are said "Yes" about the availability of syllabus (SY) which is 92.73% and 8 others are said "No" for the same which is 7.27%. 104 others are said "Yes" about the availability of any time (Time) which is 94.55% and 6 others are said "No" for the same which is 5.45%. 89 others are said "Yes" about the availability of any

97 others are said "Yes" for the downloading of study material (SMDL) which is 88.18% and 13 others are said "No" for the same which is 11.82%. 103 others are said "Yes" for the accessing of study material (ASM) which is 93.64% and 7 others are said "No" for the same which is 6.36%. 101 others are said region (Region) which is 80.91% and 21 others are said "No" for the same which is 19.09%. 83 others are said "Yes" about the comfort-ability of Hindi medium users (HM) which is 75.45% and 27 others are said "No" for the same which is 24.55%. 83 others are said "Yes" for the preparation of examination (Pexam) which is 75.45% and 27 others are said "No" for the same which is 24.55%. 102 others are said "Yes" for the searching of study materials (SMSearch) which is 92.73% and 8 others are said "No" for the same which is 7.27%. 104 others are said "Yes" for the easier use of portal (Euse) which is 94.55% and 6 others are said "No" for the same which is 5.45%. 100 others are said "Yes" for the accessing of portal through the mobile, computer or laptop (AMCL) which is 90.91% and 10 others are said "No" for the same which is 9.09%.

"Yes" for accessing the parent (Parent) which is 91.82% and 9 others are said "No" for the same which is 8.18%. 98 others are said "Yes" for child helping by parent (Help) which is 89.09% and 12 others are said "No" for the same which is 10.91%.

Table 3
Feedback of Others

Feedback	**Feedback from Others**				
	In Figure		**In Percentage**		**Grade**
Particular	No	Yes	No	Yes	
IK	19	91	17.27	82.73	Very Good
CLSP	13	97	11.82	88.18	Very Good
QP	6	104	5.45	94.55	Very Good
SY	8	102	7.27	92.73	Very Good
Time	6	104	5.45	94.55	Very Good
Region	21	89	19.09	80.91	Very Good
HM	27	83	24.55	75.45	Good
Pexam	27	83	24.55	75.45	Good
SMSearch	8	102	7.27	92.73	Very Good
Euse	6	104	5.45	94.55	Very Good
AMCL	10	100	9.09	90.91	Very Good
SMDL	13	97	11.82	88.18	Very Good
ASM	7	103	6.36	93.64	Very Good
Parent	9	101	8.18	91.82	Very Good
Help	12	98	10.91	89.09	Very Good

We have arbitrary took a standard that any grade above 80 percent is marked as very good, between 60 to 80 as fairly good and below 60 as average.

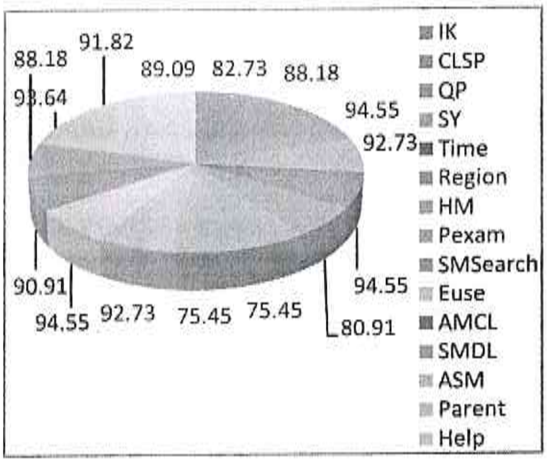


Fig. 4 Showing of Feedback from others

V RESULT

The grade of HM and PExam are found to be "Good" where other parameters are in the Category "Very Good".

VI CONCLUSION

KP-HSES counts feedback collected for teachers, students and other users. Based on counting of feedback parameters, values of counters are found appreciable. Overall hses portal is beneficial for students, teachers and others.

REFERENCES

[1] Subramanian D.Venkata, Geetha Angelina(2012). —Evaluation Strategy for Ranking and Rating of Knowledge Sharing Portal Usability— IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online): 1694-0814, pp. 395-400

[2] Suradi Nur Razia Mohd, Subramaniam Hema, Hassan Marina, and Omar, Siti Fatimah(2010) — Development of Knowledge Portal using Open Source Tools: A Case Study of FIIT, UNISEL—International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering Vol:4, No:2, pp. 94-97

[3] Irina Kondratova and Ilia Goldfarb(2010) —knowledge portal as a new paradigm for scientific publishing and collaboration— ITcon Vol. 9, pp. 161-174

[4] Alhawary A. Faleh, Irtaimeh J. Hani, Hamdan Bany Khaled(2011), —Building a Knowledge Repository: Linking Jordanian Universities E-library in an Integrated Database System— International Journal of Business and Management, Vol. 6, No. 4, pp:129-133

UMK$_{Gm}$ TP: User Friendly Multi Group Key Transfer Protocol with Circulant Matrices

Shruti Nathani[1], B.P. Tripathi[2], S.K. Bhatt[3]
[1,2,3]Dept. of Mathematics, Govt. N.P.G. College of Science, Raipur (C.G.) India.

ABSTRACT

Most existing traditional group key distribution protocols are largely designed for a single group. They establish a single key for a single group. Many group oriented applications require multi-group key establishments at time. In which user may join multiple groups simultaneously. Recently, in 2018, C.F. Hsu et al. gave new type of user oriented multi-group key establishments using secret sharing (UMKESS). As many other group Key establishments schemes this protocol (UMKESS) is also polynomial based in which to distribute and recover the secret group key, the key generation centre(KGC) and each group member has to solve t-degree interpolating polynomial. Inspire from Hsu et al.'s UMKESS, in this paper, we present a new design of user friendly group key distribution protocol using secret sharing with circulant matrices. Because of using circulant matrices as a tool, our proposed protocol $UMK_{G_m}TP$ is become more efficient, secure and robust. Also, all the required security features of group communications are handle in $UMK_{G_m}TP$.

Key words: multi-group key establishment, secret sharing scheme, circulant matrices, key transfer protocol.

I INTRODUCTION

The traditional one to one communication has been expanded into one-to-many and many-to-many communication. This type of communications involving multiple users($n \geq 2$) are called group communication [11]. For a secure group communication a group key is needed to be shared among all the group members. That is, before exchanging communication messages a key establishment protocol must be used to construct the session keys for legitimate participants in the communication [19]. This session a key is then uses by the group users to communicate their secrets, to encrypt and decrypt sensitive information and to authenticate messages in the group.

The group key establishment protocols are often classified into two types:[2]

(a) Centralized, also called distributive group key establishment protocols, where a server is responsible for generate a group key and distribute the group key to all the group members. This type of protocols is also called GKT/GKD protocol.

(b) Distributed, also called, contributory group key establishment, in which there is no server, is required and group key is generated by the contribution of all the group members. This type is also known as group key agreement (GKA) protocol.

In the past few years a large amount of research work on group key transfer protocol has been published in the literatures. The most widely used group key transfer protocols are based on secret sharing scheme(SSS), which was first introduced by both Blakley[7] and Shamir[1], independently in 1979. Then the first group key transfer protocol using secret sharing scheme (SSS) is proposed in 1989 by Laih et al.[5]. Later, there are several other group key transfer protocols [8,9,10] following the same concept of using SSS was proposed.

In 2010, Harn et al.[10] proposed, a first authenticated GKT protocol based on SSS. The confidentiality and authentication of this novel GKT protocol is information theoretically secure. But, in this protocol, to distribute and recover the secret group key, KGC and each group member has to compute a t-degree interpolating polynomial. At the same time, many research articles [11,12,13,16,17] based on Harn et al.'s[10] authenticated protocol using SSS with the computation of a t-degree interpolating polynomial has been proposed.

To overcome, this drawback, in 2016, Hsu et al. [2] gave an efficient GKT protocol. In their scheme the information related to group keys was hidden by vandermonde matrix and to distribute the group key efficiently they employed linear secret sharing scheme on vandermonde matrix, which reduces the computation load of each group member.

Recently in 2018, S. Nathani et al.[14] also gave an authenticated and secure GKT protocol based on secret sharing scheme with circulant matrices. But all this above cited conventional GKT protocols can establish a single group key at a time, that is, establish a single group key for a single group.

With the rapid development of group oriented services such as business conferencing system, wireless body area network, programmable routey communications and file sharing tools etc, require more and more multi-group communications in which users may join multiple groups simultaneously.

Recently, a new type of user oriented multi-group key establishments using secret sharing (UMKESS) is proposed by C.F. Hsu et al.[3] in 2018. This multi-group key establishment scheme is also polynomial based. That means, again to distribute and recover the secret group key, KGC and each group member has to solve t degree interpolating polynomial.

Therefore, inspire from C.F. Hsu et al.'s [3], UMKESS protocol, we extend our conventional GKT protocol [14] into multi-group key transfer protocol on SSS with circulant matrices. In this paper, we propose a new design of user friendly multi-group key distribution protocol using SS with circulant matrices.

Some unique features of our protocol are summarized below:

- A circulant matrices based key distribution protocol for multi-group communications is proposed.
- We use circulant matrix as a tool and present an efficient computation of group keys. Since information related to group keys is a hidden using circulant matrix. Thus, each participating group member and KGC has to calculate only first row of the matrix. This gives us much less computational complexity.
- Each user keeps only one share with KGC at the time of registration and the share can be used to recover multiple group keys.
- In the whole proposed scheme, the group key is authenticated by each user of distinct groups and KGC. Also, authentication has been done by only one message in each group.
- The KGC can manage user joining or leaving dynamically. There has no rekeying overhead.
- All the required security features are handling in our proposed multi-group key transfer protocol.

II PRELIMINARIES

(a) **Secret Sharing:** In a secret sharing scheme, a secret S is divided into n shares and shared among a set of n shareholders by a mutually trusted dealer in such a way that authorized subset of shareholders can reconstruct the secret but unauthorized subset of share holders cannot determine the secret. If any unauthorized subset of shareholders cannot obtain any information about the secret, then the scheme is called perfect.[2]

(b) **Circulant Matrix:[4]**A Circulant matrix is a square matrix where, given the first row, the successive rows are obtained by cyclically right shifting the present row by one element. Thus the i^{th}row of a circulant matrix of size $(n \times n)$ is obtained by cyclically right shifting the $(i-1)^{th})$ row by one position, for $i = 2$ to n , given the first row. Let the first row be the row vector .$[c(1), c(2), \ldots\ldots, c(n-1), c(n)]$. Then the circulant matrix C is obtained as

$$C = \begin{bmatrix} c(1) & c(2) & \cdots & c(n) \\ c(n) & c(1) & \cdots & c(n-1) \\ \vdots & \vdots & \cdots & \vdots \\ c(2) & c(3) & \cdots & c(1) \end{bmatrix}$$

The most important property of circulant matrices is they are multiplicatively commutative.

(c) **SSS based on Circulant matrix for multi-group communications:** Suppose a group of n participants $\{U_1, U_2, U_3, \cdots, U_n\}$ want to communicate in a secure multi-group communication with their long term secrets $\{x_1, x_2, \ldots\ldots, x_n\}$ shared with only KGC. Also for multi-groups communication we have to take a batch of group $\{G_1, G_2, \ldots\ldots, G_m\}$ and a mutually trusted KGC. Actually this scheme consists of two algorithms [14].

(d) **Secret generation algorithm**: To form Circulant matrix for each user $U_j (1 \leq i \leq n)$ in each particular group $G_i (1 \leq i \leq m)$ KGC first picks the shared secret x_j of each user U_j and make circulant matrix $[C_{ij}]$ as below :

$$[C_{ji}] = \begin{bmatrix} c(1) & c(2) & \cdots & c(n) \\ c(n) & c(1) & \cdots & c(n-1) \\ \vdots & \vdots & \cdots & \vdots \\ c(2) & c(3) & \cdots & c(1) \end{bmatrix} = \mathrm{Circ}(x_j^1, x_j^2, \ldots\ldots\ldots\ldots x_j^m)$$

where $1 \leq j \leq n$

and m denotes the number of group users in each particular group G_i and then calculate the secrets of S_{ji} of each user $U_j (1 \leq j \leq n)$ by computing

$$S_{ji} = [C_{ji}] + \mathrm{Circ}(r_{1i}, r_{2i}, \ldots\ldots, r_{ji})$$

for $1 \leq j \leq n, 1 \leq i \leq m$

Thus, this algorithm outputs with a list of secret shares $S_{ji} (1 \leq j \leq n, 1 \leq i \leq m\ )$.

(e) **Secret Reconstruction Algorithm:** This algorithm takes all the shares $S_{ji}(1 \leq j \leq n, 1 \leq i \leq m)$ each participating member U_j has long term private key x_j
and public vector $\underset{r\ ji}{\rightarrow} = (r_{1i}, r_{2i}, \dots \dots, r_{ni})$ as inputs and outputs the secret

$$S = s_1 + s_2 + \cdots + s_n$$

by computing each product

$$\boldsymbol{S_{ji} = Circ(x_j^1, x_j^2, \dots \dots, x_j^m). Circ(r_{1i}, r_{2i}, \dots, r_{ji})}$$

$\boldsymbol{(for\ 1 \leq j \leq n, 1 \leq i \leq m).}$

III PROPOSED PROTOCOL

We suppose that there are n users $\{U_1, U_2, \dots \dots, U_n\}$ participated in multi-group communications. Each user is required to register itself at KGC and KGC keeps tracking all the registered group member which includes removing any unsubscribed group participants or adding new member. To achieve secure multi-group communications, KGC has to selects multi-group session keys for all the running groups simultaneously and securely distributes these keys to all the valid registered members of particular groups. Therefore, the only valid members who belong to that particular group can easily derive this group's session key.

The proposed group key transfer protocol for multi-group communications consist of three phases: Initialization, user registration, multi group key distribution and establishment. Here we assume that there are n users $\{U_1, U_2, \dots \dots, U_n\}$ participated in multi-group communications denoted by $\{G_1, G_2, \dots \dots, G_m\}$.

(a) **Initialization:** The KGC selects a safe large prime p, and a secure one way hash function $h(.)$ whose domain is GF(p). The KGC publishes p and $h(.)$.

(b) **User Registration**: Each user is required to register at the KGC for subscribing the key distribution service. The KGC keeps tracking all the registered users or adding new users. During the registration each user $U_j(1 \leq j \leq n)$ shares his/her long term secret $x_j \in K$, $(1 \leq j \leq n)$ with KGC in a secure manner.

(c) **Multi-group key generation, distribution and establishment**: Suppose a group of n members $\{U_1, U_2, \dots \dots, U_n\}$ want to communicate in a secure multi-group communication with their long term secrets $\{x_1, x_2, \dots \dots, x_n\}$ shared with only trusted party KGC secretly. Here we also assume a batch of groups $\{G_1, G_2, \dots \dots, G_m\}$ which are handle by KGC simultaneously. The process of multigroup key generation, distribution and establishment contain five steps:

(i) **Step 1:** The initiator sends a key generation request to KGC for multiple groups with a list of groups $\{G_1, G_2, \dots \dots, G_m\}$ and each group is represented as $G_i = \{U_1, U_2, \dots \dots, U_j\}$, $1 \leq i \leq m$ *where* $j \in \{1, 2, \cdots, n\}$.

(ii) **Step 2:** KGC finally broadcast the list of all groups $\{G_1, G_2, \dots \dots, G_m\}$ to all members as a response.

(iii) **Step 3:** For each group member $U_j, 1 \leq j \leq n$, he/she decides to join more than one groups $G_i(1 \leq i \leq m)$ simultaneously. Then each group user sends their random value r_{ji}, (for $1 \leq j \leq n$, $1 \leq i \leq m$) for each group G_i in which they want to join.

(iv) **Step 4:** Now KGC received all the random values send by all the group participants $U_j, (1 \leq j \leq n)$. Then KGC broadcast the actual list of participants of each particular group according their random values sent by each group user. This list of number of participants in each particular group helps the group participants to make circulant matrices.

(v) **Step 5:** Now KGC randomly selects the group keys $K_{G_i}(1 \leq i \leq m)$ for all the groups $G_i(1 \leq i \leq m)$. Then KGC compute the secrets $S_j(1 \leq j \leq m)$ of each user U_j in each particular group $G_i(1 \leq i \leq m)$ by computing the product

[Circulant matrices of shared secrets of each user $\boldsymbol{U_j}$ in the group $\boldsymbol{G_i}$]* [Circulant matrix of random values $\boldsymbol{r_{ji}}$ of each user $\boldsymbol{U_j}$ in the group $\boldsymbol{G_i}$] =$\boldsymbol{s_{ji}}$.
$(1 \leq i \leq m,\ 1 \leq j \leq n)$
$[C_{ji}] * Circ(r_{1i}, r_{2i}, \dots \dots, r_{ji}) = s_{ji}$
Here, m denotes the number of members in the group G_i. After this computation of secret of each user U_j in particular groups, KGC also computes some additional values $u_{ji} = S_i - s_{ji}$, where
$S_i = Circ(K^1{}_{G_i}, K^2{}_{G_i}, \dots \dots, K^j{}_{G_i})$,
for $1 \leq j \leq n, 1 \leq i \leq m$ and

$$Auth_i \quad = h(K_{G_i}, U_1, U_2, \dots, U_j, r_{1i}, r_{2i}, \dots, r_{ji}, u_{1i}, u_{2i}, \dots, u_{ji})$$

for , $1 \leq j \leq n, 1 \leq i \leq m$.
At last, finally KGC broadcast $(Auth_i, (u_{ji})_{G_i}$ for $1 \leq i \leq m, 1 \leq j \leq n$.

Here, i represents number of groups and j represents number of participants in each group G_i.

(vi) **Step: 6** Now each participating group member $U_j, 1 \leq j \leq n$, knowing their corresponding public value u_{ji}, in each particular group $G_i, (1 \leq i \leq m)$, is able to compute the product

$$[C_{ij}] * Circ(r_{1i}, r_{2i}, \dots \dots, r_{ji}) = s_{ji}$$

and recover the group key K_{G_i} by computing,

$$S_i = (u_{ji} + s_{ji})$$

Which is of the form

$$S_i = Circ(K_{G_i}^1, K_{G_i}^2, \ldots\ldots K_{G_i}^j)$$

(for , $1 \le j \le n$, $1 \le i \le m$)
Afterwards, each u_{ji}, ($for\ 1 \le j \le n,\ 1 \le i \le m$) authenticates their corresponding groups $G_i$by computing
$Auth_i^* = \quad h(K_{G_i,}U_1, U_2, \ldots., U_j, r_{1i}, r_{2i}, \ldots, r_{ji}, u_{1i}, u_{2i}, \ldots u_{ji})$
for $1 \le j \le n,\ 1 \le i \le m$
and then checks this value by

$$Auth_i = Auth_i^*.$$

If this result is correct then each participant $U_j(1 \le j \le n)$, in the group $G_i(1 \le i \le m)$ authenticates the group key K_{G_i} is sent from KGC.

IV AN EXAMPLE

In our example we assume a group of 7 members $\{U_1, U_2, U_3, U_4, U_5, U_6, U_7\}$ want to generate a secure group communications in multiple groups simultaneously.

(a) User Registration: During registration each user $U_j, 1 \le j \le 7$, shares his/her long term secrets $x_i \in K$ with KGC. Suppose U_1 Shares $x_1 = 2$, U_2 Shares $x_2 = 1$, U_3 Shares $x_3 = 4$, U_4 Shares $x_4 = 3$, U_5 Shares $x_5 = 10$, U_6 Shares $x_6 = 5$, U_7 Shares $x_7 = 7$ in a secure manner. KGC publishes $h(\cdot)$.

(b) Group Key Generation and Distribution:

In our example we assume a batch of groups$\{G_1, G_2, G_3\}$, in which there 7 group members want to join simultaneously.

Step 1: Suppose $U_2(initiator)$ sends a key generation request to KGC with a list of groups $\{G_1, G_2, G_3\}$.
Step 2: KGC broadcast the list of groups $\{G_1, G_2, G_3\}$ to all members as a response.
Step 3: Here each group member $U_j, (1 \le j \le 7)$, he/she decides to join more than one groups G_i, ($1 \le i \le 3$). Then each group participants sends their radom values r_i, for each group G_i in which they want to join.

Suppose , U_1 sends $r_{11} = 2$, $r_{13} = 1$, U_2 sends $r_{21} = 1$, $r_{22} = 8$, U_3 sends $r_{32} = 2$, U_4 sends $r_{41} = 10$, $r_{43} = 3$, U_5 sends $r_{51} = 11$, $r_{52} = 6$, U_6 sends $r_{61} = 4$, $r_{62} = 2$, U_7 sends $r_{73} = 9$ to KGC.

Step 4: Now KGC received all the random keys send by the 7 users $\{U_1, U_2, U_3, U_4, U_5, U_6, U_7\}$.
Then, KGC broadcast the actual list of participants $U_j(1 \le j \le 7)$ of each particular group $G_i(1 \le i \le 5)$. That means KGC broadcast

$$(\{U_1, U_2, U_4, U_5, U_6,\} \in G_1,$$

$\{U_2, U_3, U_5\} \in G_{2,}$ $\{U_1, U_4, U_7\} \in G_3$) list of all group members publicly.
Step 5: Now KGC randomly selects the 3 group keys $K_1 = 100$, $K_2 = 200$, $K_3 = 50$, to all the 3 groups $\{G_1, G_2, G_3\}$.
Now KGC compute the secrets s_j of each user U_j of each particular groups $G_i(1 \le j \le 7,\ 1 \le i \le 3)$.
For this KGC, first has to make the circulant matrices of each participating group user $U_j(1 \le j \le 7)$ in each particular group $G_i(1 \le i \le 3)$, with the help of their corresponding shared secret values.

$$x_1 = 2, x_2 = 1, x_3 = 4, x_4 = 3, x_5 = 10, x_6 = 5, x_7 = 7$$

That means, for G_1, $\{U_1, U_2, U_4, U_5, U_6,\}$,

$$C_{11} = Circ(2^1, 2^2, 2^3, 2^4, 2^5) = Circ(2,4,8,16,32)$$

$$C_{21} = Circ(1^1, 1^2, 1^3, 1^4, 1^5) = Circ(1,1,1,1,1)$$

$$C_{41} = Circ(3^1, 3^2, 3^3, 3^4, 3^5) = Circ(3,9,27,81,243)$$

$$C_{51} = Circ(10^1, 10^2, 10^3, 10^4, 10^5) = Circ(10,100,1000,10000,100000)$$

$$C_{61} = Circ(5^1, 5^2, 5^3, 5^4, 5^5) = Circ(5,25,125,625,3125)$$

Then, $s_{11} = \quad [C_{11}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$

$$= Circ(2,4,8,16,32) * Circ(2,1,10,11,4)$$

$$= Circ(300,538,446,230,212).$$

$$s_{21} = [C_{21}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$

$$= Circ(1,1,1,1,1) * Circ(2,1,10,11,4)$$

$$= Circ(28,28,28,28,28).$$

$$s_{41} = [C_{41}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$

$$= Circ(3,9,27,81,243) * Circ(2,1,10,11,4)$$

$= Circ(1392,3450,3090,1284,948)$.

$$s_{51} = [C_{51}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$

$$= Circ(10,100,1000,10000,100000) \quad * Circ(2,1,10,11,4)$$

$= Circ(211420,1114210,1142200,$

$$422110,221140)$$
$$s_{61} = [C_{61}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(5,25,125,625,3725) \quad * Circ(2,1,10,11,4)$$
$$= Circ(11460,44680,43800,$$

16580,9620)

For group G_2, $\{U_2, U_3, U_5\}$,

$$C_{22} = Circ(1^1, 1^2, 1^3) = Circ(1,1,1).$$

$$C_{32} = Circ(4^1, 4^2, 4^3) = \qquad Circ(4,16,64).$$
$$C_{52} = Circ(10^1, 10^2, 10^3) = \qquad Circ(10,100,1000).$$

Then,

$$s_{22} = [C_{22}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(1,1,1) * Circ(8,7,6)$$
$$= Circ(21,21,21).$$
$$s_{32} = [C_{32}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(4,16,64) * Circ(8,7,6)$$
$$= Circ(576,540,648).$$
$$s_{52} = [C_{52}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(10,100,1000) * Circ(8,7,6)$$
$$= Circ(7680,6870,8760).$$

For group G_3, $\{U_1, U_4, U_7\}$,

$$C_{13} = Circ(2^1, 2^2, 2^3) = Circ(2,4,8).$$

$$C_{43} = Circ(3^1, 3^2, 3^3) = \qquad Circ(3,9,27).$$
$$C_{52} = Circ(7^1, 7^2, 7^3) = \qquad Circ(7,49,343).$$

Then,

$$s_{13} = [C_{13}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(2,4,8) * Circ(1,3,9)$$

$$= Circ(62,82,38).$$
$$s_{43} = [C_{43}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(3,9,27) * Circ(1,3,9)$$

$$= Circ(165,261,81).$$

$$s_{73} = [C_{73}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(7,49,343) * Circ(1,3,9)$$
$$s_{73} = Circ(1477,3157,553).$$

Now, KGC computes the five additional values for group G_1,

$$u_{11} = S - s_{11}$$
$$u_{11} = Circ(100^1, 100^2, 100^3, 100^4, 100^5) \ - Circ(300,538,446,230,212).$$
$$= Circ(-200,9462,999554,99999770,$$

9999999788).

$$u_{21} = S - s_{21}$$
$$u_{21} = Circ(100^1, 100^2, 100^3, 100^4, 100^5) \ - Circ(28,28,28,28,28).$$
$$= Circ(72,9972,999972,99999972,$$

9999999972).

$$u_{41} = S - s_{41}$$
$$u_{41} = Circ(100^1, 100^2, 100^3, 100^4, 100^5) \ - Circ(1392,3450,3090,1284,948).$$
$$= Circ(-1292,6550,996910,99998716$$

,9999999052).

$$u_{51} = S - s_{51}$$
$$u_{51} = Circ(100^1, 100^2, 100^3, 100^4, 100^5) \ - Circ\begin{pmatrix}211420, 1114210,1142200, \\ 422110, 221140\end{pmatrix}.$$

$$= Circ(-211320, -1104210, -142200,$$

99577890,9999778890).

$$u_{61} = S - s_{61}$$
$$u_{61} = Circ(100^1, 100^2, 100^3, 100^4, 100^5) \ - Circ(11460,44680,43800,16580,9620).$$
$$= Circ(-11360, -34680,956200,99983420,$$

9999990380).

and the value of

$Auth_1 = h(K_{G_1} = 100,\ \{U_1, U_2, U_4, U_5, U_6\},\ r_{11}, r_{21}, r_{41},\ r_{51}, r_{61,} u_{11}, u_{21}, u_{41}, u_{51}, u_{61})$.

KGC computes three additional values for group G_2,

$$u_{22} = S - s_{22}$$
$$u_{22} = Circ(200^1, 200^2, 200^3) \ - Circ(21,21,21)$$
$$= Circ(179,39979,7999979).$$

$$u_{32} = S - s_{32}$$
$$u_{32} = Circ(200^1, 200^2, 200^3) \ - Circ(576,540,648)$$
$$= Circ(-376,39460,7999352).$$

$$u_{52} = S - s_{52}$$
$$u_{52} = Circ(200^1, 200^2, 200^3) \quad - Circ(7680,6870,8760)$$
$$= Circ(-7480,33130,7991240).$$

and the value of

$Auth_2 = h\left(K_{G_2} = 200,\ \{U_2, U_3, U_5\},\ r_{22}, r_{32}, r_{52}, u_{22}, u_{32}, u_{52}\right)$.

Also, KGC has to compute 3 additional values for group $G_3 \in \{U_1, U_4, U_7\}$.

$$u_{13} = S - s_{13}$$
$$u_{13} = Circ(50^1, 50^2, 50^3) \ - Circ(62,82,38)$$
$$= Circ(-12,2418,124962).$$
$$u_{43} = S - s_{43}$$
$$u_{43} = Circ(50^1, 50^2, 50^3) \ - Circ(165,261,81)$$

$= Circ(-115,2239,124919)$.

$$u_{73} = S - s_{73}$$
$$u_{73} = Circ(50^1, 50^2, 50^3) \ - Circ(1477,3157,553)$$
$$= Circ(-1427, -657,124447).$$

and the value of

$$Auth_3 = h\left(K_{G_3} = 50, \qquad \{U_1, U_4, U_7\}. \qquad r_{13}, r_{43}, r_{73}, u_{13}, u_{43}, u_{73}\right).$$

Thus, KGC finally broadcast,

$$\{Auth_1, Auth_2, Auth_3, \{u_{11}, u_{21}, u_{41}, u_{51,} u_{61}\}_{G_1},$$

$\{u_{22}, u_{32}, u_{52,}\}_{G_2}, \{u_{13}, u_{43}, u_{73}\}_{G_3}\}$.

Step 6: At last to compute the common group key, each participating group members of group,

$$G_1 \in \{U_1, U_2, U_4, U_5, U_6\}, \qquad G_2 \in \{U_2, U_3, U_5\},$$
$$,\ G_3 \in \{U_1, U_4, U_7\},$$

has to solve the equation

$$S = (u_{ji} + s_{ji})$$
$$\text{where, } S = Circ(K_i^1, K_i^2, \dots, K_i^j)$$

here, j denotes the number of participants in the group i .

Therefore, for group G_1,

User U_1 , computes

$$s_{11} = [C_{11}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(2,4,8,16,32) * Circ(2,1,10,11,4)$$
$$= Circ(300,538,446,230,212).$$

So. $S = u_{11} + s_{11}$

S= $Circ(-200,9462,999554,$

99999770,9999999788)+ $\quad Circ(300,538,446,230,212)$

S=Circ(100,10000,1000000,100000000,10000000000)

S=Circ(100,100^2, 100^3, 100^4, 100^5)

Thus, $G_{K_1} = 100$.

$$s_{21} = [C_{21}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(1,1,1,1,1) * \ Circ(2,1,10,11,4)$$
$$= Circ(28,28,28,28,28).$$

So, $S = u_{21} + s_{21}$

S= $Circ(72,9972,999972,$
$99999972, 9999999972) + \quad Circ(28,28,28,28,28)$
=Circ(100,10000,1000000, 100000000,10000000000)
S = Circ($100, 100^2, 100^3, 100^4, 100^5$)
Thus, $G_{K_1} = 100$.

$$s_{41} = [C_{41}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(3,9,27,81,243) * Circ(2,1,10,11,4)$$
$= Circ(1392,3450,3090,1284,948).$
So, $S = u_{41} + s_{41}$
S = $Circ(-1292,6550,996910,$
$99998716,9999999052)+ \quad Circ(1392,3450,3090,1284,948)$
S=Circ(100,10000,1000000, 100000000,10000000000)
S =Circ($100, 100^2, 100^3, 100^4, 100^5$)
Thus, $G_{K_1} = 100$.

$$s_{51} = [C_{51}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(10,100,1000,10000,100000) \quad * Circ(2,1,10,11,4)$$
$= Circ(211420,1114210,1142200422110,221140).$

So, $S = u_{51} + s_{51}$
S = $Circ(-211320, -1104210,$
$-142200,99577890,9999778860)+ \ Circ(211420,1114210,1142200,422110,221140).$
S=Circ(100,10000,1000000, 100000000,10000000000)
S=Circ($100, 100^2, 100^3, 100^4, 100^5$).
Thus, $G_{K_1} = 100$.

$$s_{61} = [C_{61}] * Circ(r_{11}, r_{21}, r_{41}, r_{51}, r_{61})$$
$$= Circ(5,25,125,625,3725) \quad * Circ(2,1,10,11,4)$$
$= Circ(11460,44680,43800, 16580,9620).$

So, $S = u_{61} + s_{61}$
S = $Circ(-11360, -34680,956200,$
$99983420,9999990380) +$
$Circ(11460,44680, 43800,$
$16580,9620)$
S=Circ(100,10000,1000000,100000000,10000000000)
S=Circ($100, 100^2, 100^3, 100^4, 100^5$)
Thus, $G_{K_1} = 100$.

Hence, all the group users of group G_1 gets the group key $K_{G_1} = 100$.

For. group $G_2 \in \{U_2, U_3, U_5\}$,
User U_2 computes,

$$s_{22} = [C_{22}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(1,1,1) * Circ(8,7,6)$$
$$= Circ(21,21,21).$$

So, $S = u_{22} + s_{22}$
S = $Circ(179,39979,7999979) + \ Circ(21,21,21)$
S=Circ(200,40000,8000000)
S=Circ($100, 200^2, 200^3$)
Thus, $G_{K_2} = 200$.
User U_3 computes,

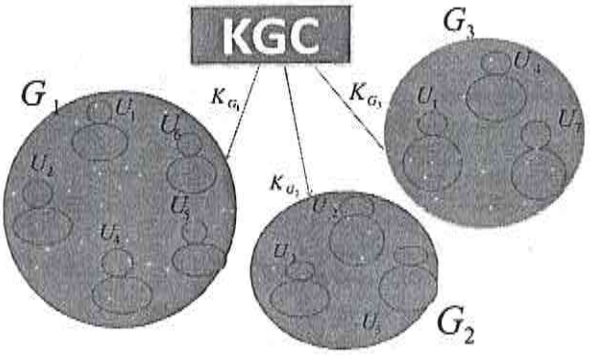
$$s_{32} = [C_{32}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(4,16,64) * Circ(8,7,6)$$
$$= Circ(576,540,648).$$

So, $S = u_{32} + s_{32}$
S= $Circ(-376,39460,7999352)$+$Circ(576,540,648)$
S=Circ(200,40000,8000000)
S=Circ(200,200^2, 200^3).
Thus, $G_{K_2} = 200$.
User U_5 computes,

$$s_{52} = [C_{52}] * Circ(r_{22}, r_{32}, r_{52})$$
$$= Circ(10,100,1000) * Circ(8,7,6)$$
$$= Circ(7680,6870,8760).$$

So, $S = u_{52} + s_{52}$
S= $Circ(-7480,33130,7991240)$ + $Circ(7680,6870,8760)$.

S = Circ(200,40000,8000000)
S = Circ(200,200^2, 200^3).
Thus, $G_{K_2} = 200$.
Hence, all the group users of group G_2 gets the group key $K_{G_2} = 200$.



For, group $G_3 \in \{U_1, U_4, U_7\}$,
User U_1 computes,

$$s_{13} = [C_{13}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(2,4,8) * Circ(1,3,9)$$
$$= Circ(62,82,38).$$

So, $S = u_{13} + s_{13}$
S = $Circ(-12,2418,124962)$+ $Circ(62,82,38)$.
S=Circ(50,2500,125000)
S=Circ(50,50^2, 50^3)
Thus, $G_{K_3} = 50$.
User U_4 computes,

$$s_{43} = [C_{43}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(3,9,27) * Circ(1,3,9)$$
$$= Circ(165,261,81).$$

So, $S = u_{43} + s_{43}$
S = $Circ(-115,2239,124919)$ + $Circ(165,261,81)$.
S=Circ(50,2500,125000)
S=Circ(50,50^2, 50^3)
Thus, $G_{K_3} = 50$.

$$s_{73} = [C_{73}] * Circ(r_{13}, r_{43}, r_{73})$$
$$= Circ(7,49,343) * Circ(1,3,9)$$
$$s_{73} = Circ(1477,3157,553).$$

User U_7 computes,
So, $S = u_{73} + s_{73}$

S= $Circ(-1427,-657,124447)+$ $Circ(1477,3157,553)$.
S=Circ(50,2500,125000)
S=Circ($50,50^2,50^3$)
Thus, $G_{K_3}=50$.
Hence, all the group users of group G_3 gets the group key $K_{G_3}=50$.

V SECURITY ANALYSIS

Theorem: The proposed protocol possesses key freshness, key confidentiality and key authentication.

Proof: Key Freshness: In our proposed protocol for each new communication session m new group keys $\{G_{K_1},G_{K_2},\dots\dots\dots,G_{K_m}\}$ associated with $\{G_1,G_2,\dots\dots,G_m\}$ are randomly selected by KGC for each multi-group key service request. Also, to compute the group key $K_{G_i}(1\le i\le m)$ each group user $U_j(1\le j\le n)$ has to calculate

$S=(u_{ji}+s_{ji})$, where

$$s_{ji}=[C_{ij}]*Circ(r_{1i},r_{2i},\dots\dots,r_{ji})$$

$$s_{ji}=(x_j^1,x_j^2,\dots\dots,x_j^m)*Circ(r_{1i},r_{2i},\dots\dots,r_{ji})$$

Which is a function of shared secrets of each user U_j and random challenges(public values) $r_{ji}(1\le j\le n,\ 1\le i\le m)$ selected by each group member $U_j(1\le j\le n)$ for each new communication service request. Thus, it is obvious that the group key K_{G_i} will be fresh that is new and different for each new communication session.

$$\boldsymbol{S_i=(u_{ji}+s_{ji})(=Circ[K_{G_i}^1,K_{G_i}^2,\dots\dots,K_{G_i}^t])}$$

Key Confidentiality: Key secrecy is provided due to the security feature of SSS based on circulant matrices for multiple groups. To handle multiple groups at a time KGC has to select multiple group keys $\{K_{G_1},K_{G_2},\dots\dots\dots,K_{G_m}\}$, the respective group members have calculate

Where, u_{ji} are the public values sent by KGC and

$$\boldsymbol{s_{ji}=[C_{ji}]*Circ(r_{1i},r_{2i},\dots\dots,r_{ji})}$$
$$\boldsymbol{s_{ji}=(x_j^1,x_j^2,\dots\dots,x_j^t)*Circ(r_{1i},r_{2i},\dots\dots,r_{ji})}$$

Where t denotes the number of members in the group G_i. This shared secret value s_{ji} assured that only authorized group member is able to recover the group key K_{G_i} which is of the form

$$\boldsymbol{S_i=Circ(K_{G_i}^1,K_{G_i}^2,\dots\dots,K_{G_i}^t)}$$

where t represent the number of members in the group G_i.
Hence, key confidentiality is surely achieved in our proposed scheme.

Key Authentication: In key distributing phase, the KGC also compute $Auth_i$ for all the multiple groups G_i simultanously. Also, each user U_j authenticates their corresponding groups G_i by computing

$$\mathrm{Auth^{*}i}=h(K_{G_i},U_1,U_2,\dots.,U_j,r_{1i},r_{2i},\dots,r_{ji},u_{1i},u_{2i},\dots.,u_{ji})$$

for , $1\le j\le n, 1\le i\le m$.
and then check this hash value by $\mathrm{Auth_i}=\mathrm{Auth_i^*}$.
Also this key authentication is done only by one message for each group G_i.

Theorem(Insider attack): The proposed protocol **UMK$_{G_m}$TP** is secure against insider attack.
Proof: At the time of registration, each participating group member U_j shared his/her long term secret key x_j only with KGC (a trusted authority). For each new communication session a new group key K_{G_i} is selected by KGC and makes some values $u_{ji}=(S-s_{ji})(1\le i\le m,\ 1\le j\le n)$ publicly known. Then each authorized group member knows their shared secret x_j with KGC and public values u_{ji} is able to compute the group key K_{G_i} which is of the form

$$S=Circ(K_G^1,K_G^2,\cdots,K_G^t).$$

Since , $S=u_i+s_i$,
where ,

$$\mathbf{s_{ji}=(x_j^1,x_j^2,\dots\dots,x_j^t)*Circ(r_{1i},r_{2i},\dots\dots,r_{ji})}$$

Therefore, the secret $x_j \in K$ of each group member shared with KGC remains unknown to outsiders and also each authorized group member is able to recover the group key but not able to obtain other member's long term secret x_j. Thus, our proposed protocol resist against insider attack.

Theorem (Forward and Backward Secrecy): The proposed protocol $\mathbf{UMK_{G_m}TP}$ provide backward and forward secrecy, that is newly joined members cannot recover the old group keys and those old members who left the group cannot access the current group key.

Proof: In our proposed $\mathbf{UMK_{G_m}TP}$ protocol, for every multi-group session, if new members join in or old members left from groups, the KGC needs to distribute new group keys to all existing group members. In each group the group key K_{G_i} is derived from the current group members long term secrets $x_j's$ and fresh random challengesr_{ji}. Also, our whole computation is totally depends on the number of members in the current group. Thus, the newly joined members can recover the current group key but cannot recover the previous group keys and those old members who left the group cannot recover the current group key. Thus, our protocol achieves both forward and backward secrecy of group communication.

VI CONCLUSION

We defined a new type of, circulant matrices based key transfer protocol for multi-group communications. Because of using circulant matrices as a tool, our proposed multi-group key transfer protocol takes much less time than other existing multi-group key transfer protocols. Also all the required security attributes are addressed in detail and the confidentiality of our proposed protocol is unconditionally secure.

REFERENCES

[1] A. Shamir, "How to share a secret ", Commun. ACM vol. 22, no. 11. pp. 612-613, Nov. (1979).

[2] C.F. Hsu, L. Harn, Y. Mu, M. Zhang, X. Zhu, "Computation efficient key establishment in wireless group communications ", wireless network , vol. 23, PP. 289-297, (2016).

[3] C.F. Hsu, L. Harn, B. Zeng, "UMKESS: user oriented multigroup key establishments using secret sharing", wireless networks ,(2018).

[4] C. Rajarama, J. N. Sugatoor, T. Y. Swamy, " Diffie-Hellman type key exchange, ElGamal like ecryption/decryption and proxy re-encryption using circulant matrices ", International Journal of Network Security, vol. 20, no. 4, PP. 617-624, July (2018).

[5] C.S. Laih and J. Y. Lee, "A new threshold scheme and its applications in designing the conference key distribution cryptosystem", Inf. Process. Lett., vol. 32, no. 3, PP. 95-99, (1989).

[6] C.Y. Lee, Z.H. Wang, L.Harn , C.C. Chang, "Secure key transfer protocol based on secret sharing for group communications", IEICE Trans. Inf. & Syst. , vol. E94-D, no. 11, (2011).

[7] G. R. Blakely, "Safegaurding cryptographic keys" , in proc. AFIPS 1979, National Computer Conference, PP. 313-317. AFIPS, (1979).

[8] G. Saze, "Generation of key predistribution schemes using secret sharing schemes ", Discrete Applied Mathematics , vol. 128, PP. 239-249, (2003).

[9] L. Ch, J. Pieprzyk, "Conference key agreement from secret sharing ", Proc. Fourth Australasian Conf. Information Security and Privacy(ACISP'99), PP. 64-76, (1999).

[10] L. Harn, C. Lin, "Authenticated group key transfer protocol based on secret sharing", IEEE Trans. Comuter , vol. 59, no. 6, PP. 842-846, (2010).

[11] L. Harn, G. Gong, "Conference key establishment protocol using a multivariate polynomial and its applications", Security and Communication Networks, vol. 8, no. 9, PP. 1794-1800,(2014).

[12] L. Harn, C. Lin, "Efficient group Diffie-Hellman key agreement protocols", Comput. Elect. Eng., (2014).

[13] R. F. Olimid, "Cryptanalysis of a password based group key exchange protocol using secret sharing", Appl. Math. Inf. Sci., vol. 7, no. 4, PP. 1585-1590,(2013).

[14] S. Nathani, B.P. Tripathi, " An authenticated and secure group key transfer protocol with circulant matrices", Journal of Computer and Mathematical Sciences , vol. 9, no.12, PP. 2086-2095, (2018).

Prediction of Lifestyle Diseases Such as Diabetes Using Supervised Machine Learning Approach

Animesh Tayal[1], Amruta K. Chimote[2], S.R. Tandan[3]
[1,2]Research Scholar, Dr. C.V. Raman University, Bilapsur (C.G.) India.
[3]Associate Prof., Dept of CSE, Dr. C.V. Raman University, Bilapsur (C.G.) India.

ABSTRACT

***Lifestyle diseases** are defined as diseases linked with the way people live their life. This is commonly caused by alcohol, drug and smoking abuse as well as lack of physical activity and unhealthy eating. Diseases that impact on our lifestyle are heart disease, stroke, obesity and type II diabetes. Habits that detract people from activity and push them towards a sedentary routine can cause a number of health issues that can lead to chronic non-communicable diseases that can have near life-threatening consequences. We are preparing model for prediction of these diseases which will help user to change daily habits and get healthy lifestyle. We are generating primary data for Proposed system as per suggested by medical practitioner, our model will analyzed the data and predict whether person is prone to these diseases or not using supervised machine learning technique. Supervised machine learning help to classify the available data and finds the relation between them which is necessary to predict future consequences.*

Keywords: Lifestyle diseases, Supervised Classification, Machine learning.

I INTRODUCTION

Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not return correctly to insulin, or by both reasons. The day is now to prevent and diagnose diabetes in the early stages.

According to the WHO (world health organization) report in Nov 14, 2016 in the world diabetes day "*Eye on diabetes*" reported 422 million adults are with diabetes, 1.6 million deaths, as the report indicates it is not difficult to guess how much diabetes is very serious and chronic.

Diabetes diseases damage different parts of the human body from those parts some of them are: eyes, kidney, heart, and nerves. *William's text book of endocrinology was* predictable that in 2013 more than 382 million populations in the world or all over the world were with diabetes or had diabetes. There are so many people's are died every year by diabetes disease (DD) both in poor and rich countries in the world.

According to the centers for disease control and prevention (CDCP) they give information for the duration of 9 ensuing years that is between 2001 and 2009 type II diabetes increased 23% in the United States (US). There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person death.

Diabetes. Most in the current time diabetes is grouped into two types of diabetes, type I and Type II diabetes. Type I diabetes this type of diabetes in heath language or in doctors' language this type of diabetes also called Insulin dependent diabetes illness. Here the human body does not produce enough insulin. 10 % of diabetes caused by this type of diabetes.

Type II diabetes this type of diabetes. According to CDA (Canadian Diabetes Association) during 10 years, between 2010 and 2020, expected to increase from 2.5 million to 3.7 million. Therefore, as the above mentioned Diabetes diseases needs early prevention and diagnosis to safe human life from early death .By considering how much this disseises is very series and leading one in the world. Moloud [2] Algorithms which are used in machine learning have various powers in both classification and predicting.

This study follows different machine learning algorithms to predict diabetes disease at an early stage. Such as, Logical regression. SVM to predict this chronic disease at an early stage for safe human life.

II RELATED WORK

(a) Describe and explain different classification Algorithms using different parameters such as Glucose, Blood Pressure, Skin Thickness, insulin, BMI. Diabetes Pedigree, and age. The researches were not included pregnancy parameter to predict diabetes disease (DD). In this research, the researchers were using only small sample data for prediction of Diabetes. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally. The researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for prediction of Diabetes.

(b) Machine learning algorithms are very important to predict different medical data sets including diabetes diseases dataset(DDD).in this study they use support vector machines(SVM) ,Logistic Regression ,and Naïve Bayes using 10 fold cross validation to predict different/varies medical datasets including diabetes dataset(DD) .the researchers' was compare the accuracy and the performance of the algorithm based on their result and the researchers conclude that SVM(support Vector Machine) algorithm provides best accuracy than the other algorithm which are mentioned on the above . The researchers were use those machine learning algorithm on a small sample of data in this study factors for accuracy were identified such factors are Data origin, Kind, and dimensionality.

(c) CART (classification and Regression Tree) was used for generating fuzzy rule. Clustering algorithm also was used (principal component Analysis (PCA) and Expectation maximization (EM) for pre-processing and noise removing before applying the rule. Different medical dataset (MD) was used such as breast cancer, Heart, and Diabetes Develop decision support for different diseases including diabetes. The result was CART (Classification and Regression tree) with noise removal can provide effective and better in health/diseases prediction and it is possible to safe human life from early death.

(d) This study was the new approach that used KNN algorithm by removing the outlier/OOB(out of bag) using DISKR(decrease the size of the training set for K-nearest neighbor .and also in this study the storage space was minimized. There for .the space complexity is become less and efficient .after removing a parameters or instances which have less effect or factor the researchers got better accuracy.

(e) Feature selection is one of the most important steps to increase the accuracy. Hoeffding Tree(HT) ,multi-layer perceptron (MP),Jrip,BayeNet,RF(random forest).and Decision Tree machine learning Algorithms were used for prediction .From different feature selection algorithm in this study they were use best first and greedy stepwise feature selection algorithm for feature selection purpose . The researchers conclude that Hoeffding Tree (HT) provides high accuracy.

(f) In this study the researchers concentrate on different datasets including Diabetes Dataset (DD).The researcher were investigate and construct the models that are universally good and capability for varies/different medical datasets (MDs).the classification algorithm did not evaluate using Cross validation evaluation method .ANN,KNN, Navie Bayes.J48.ZeroR.Cv Parameter selection, filtered classifier .and simple cart were some of the algorithm used in this study. From those algorithms Naïve Bayes provide better accuracy in diabetes dataset (DD) in this study. The two algorithms KNN and ANN provide high accuracy in other datasets on this study.

(g) By using CPCSSN(Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death .on this study Bagging ,Adaboost, and decision tree(J48) were used to predict the diabetes and the researcher was compare the result of those methods and concluded that Adaboost method was provide effective and better accuracy than the other methods in weka data mining tools

(h) Classification problems were identified in this study, one of the most problem in classification is data reduction .it has a vital role in prediction accuracy .to get better and efficient accuracy the data should be reduced as the researchers studied here. On this study PCA (principal component Analysis) for data pre-processing including data reduction for better accuracy. For prediction modified decision tree (DT) and Fuzzy were used for prediction purpose .finally it was concluded as to get better result the dataset should be reduced.

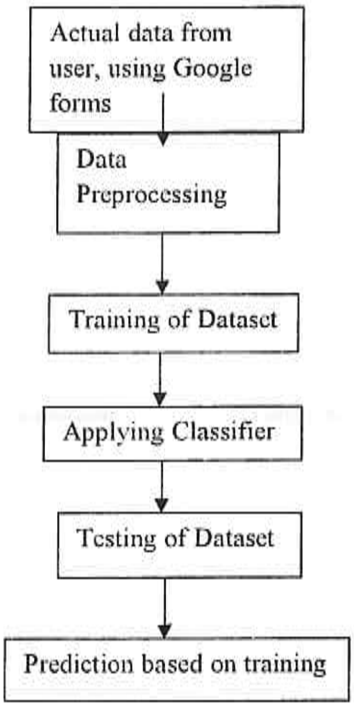
(i) In this study the performance of machine learning techniques were compared and measured based on their accuracy. The accuracy of the technique is varying from before pre-processing and after pre-processing as they identified on this study. This indicates the in the prediction of diseases the pre-processing of data set has its own impact on on the performance and accuracy of the prediction. Decision tree techniques provide better accuracy in this study before pre-processing to predict diabetes diseases. Random forest and support vector machine provide better prediction after pre-processing in this study using diabetes data set.

(j) K-means and Genetic algorithm used in this study for Dimension reduction in order to get better performance. The integration of support vector machine for prediction technique was used and provides better accuracy in small sample diabetes data set by selecting only five factors or parameters. 10 cross validation on this study used as evaluation method. finally reduced data set provide better performance than large dataset.

(k) In this study the researchers were use different data mining techniques to predict the diabetic diseases using real world data sets by collecting information by distributed questioner .in this study SPSS and weka tools were used for data analysis and prediction respectively .in this study the researchers compare three techniques ANN, Logistic regression. and j48 .finally it was concluded as j48 machine learning technique provide efficient and better accuracy.

(l) Oracle Data miner and Oracle Database 10g used for Analysis and storage respectively .the parameters or factors were identified in this study .the target variables were identified based on their percentage .this study concentrated on the treatment of the patient .the patient divided into two categories old and young based on their age and predict their treatment for both young and old diet control indicates high percentage on this study. The treatment predictive percentage done by support vector machine.

III PROPOSED METHODOLOGY



IV EXPECTED OUTCOME

User has to enter his information regarding height, weight, age, sex, type of exercise he/she performs and system will be able to predict whether that person can have diabetes in future or not? This will help user to take action at right time to prevent future mishap.

REFERENCES

[1] Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.

[2] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.

[3] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.

[4] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Computer Science, 112(C), 2519-2528.

[5] Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. Expert Systems with Applications, 72, 335-343.

[6] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). Risk prediction of type II diabetes based on random forest model. In Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on (pp. 382-386). IEEE.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.

[8] Komi. M., Li. J., Zhai. Y., & Zhang. X. (2017, June). Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.

[9] Nilashi. M., bin Ibrahim. O., Ahmadi, H., & Shahmoradi, L. (2017). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering.

[10] Balpande, V. R., & Wajgi, R. D. (2017, February). Prediction and severity estimation of diabetes using data mining technique. In Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on (pp. 576-580). IEEE.

[11] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017, February). An expert clinical decision support system to predict disease using classification techniques. In Electrical, Computer and Communication Engineering (ECCE), International Conference on(pp. 396-400). IEEE.

[12] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: a medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.

[13] Mekruksavanich, S. (2016, August). Medical expert system based ontology for diabetes disease diagnosis. In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on (pp. 383-389). IEEE.

[14] Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584-1589). IEEE.

[15] Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on (pp. 347-352). IEEE.

[16] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.

[17] Kamadi, V. V., Allam, A. R., & Thummala, S. M. (2016). A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. Applied Soft Computing, 49, 137-145.

[18] Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).

[19] Ramzan, M. (2016, August). Comparing and evaluating the performance of WEKA classifiers on critical diseases. In Information Processing (IICIP), 2016 1st India International Conference on (pp. 1-4). IEEE.

[20] Negi, A., & Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on (pp. 237-241). IEEE.

[21] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, 76-83.

[22] Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. International journal of advanced research in computer and communication engineering, 4, 196-1.

[23] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in (pp. 122-127). IEEE.

[24] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In Next Generation Computing Technologies (NGCT), 2015 1st International Conference on (pp. 673-676). IEEE.

[25] Pavate, A., & Ansari, N. (2015, September). Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques. In Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on(pp. 371-375). IEEE.

[26] Nam, J. H., Kim, J., & Choi, H. G. (2015). Developing statistical diagnosis model by discovering principal parameters for Type 2 diabetes mellitus: a case for Korea. Public Health Prev. Med, 1(3), 86-93.

[27] Lukmanto, R. B., & Irwansyah, E. (2015). The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. Procedia Computer Science, 59, 312-319.

[28] Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S. J., & Yu, K. S. (2015). An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Systems with Applications, 42(9), 4265-4273.

[29] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.

[30] kumar Dewangan, A., & Agrawal, P. (2015). Classification of Diabetes Mellitus Using Machine Learning Techniques. International Journal of Engineering and Applied Sciences, 2(5), 145-148.

[31] Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.

[32] Mounika, M., Suganya, S. D., Vijayashanthi, B., & Anand, S. K. (2015). Predictive analysis of diabetic treatment using classification algorithm. IJCSIT, 6, 2502-2505.

[33] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.

[34] Wang, K. J., Adrian, A. M., Chen, K. H., & Wang, K. M. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. Journal of biomedical informatics, 54, 220-229.

[35] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An Efficient Rule-Based Classification of Diabetes Using ID3, C4. 5, & CART Ensembles. In Frontiers of Information Technology (FIT), 2014 12th International Conference on (pp. 226-231). IEEE.

[36] Lee, B. J., Ku, B., Nam, J., Pham, D. D., & Kim, J. Y. (2014). Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE journal of biomedical and health informatics, 18(2), 555-561.

[37] Sankaranarayanan, S. (2014, March). Diabetic prognosis through Data Mining Methods and Techniques. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 162-166). IEEE.

[38] Varma, K. V., Rao, A. A., Lakshmi, T. S. M., & Rao, P. N. (2014). A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering, 40(5), 1758-1765.

[39] Li, L. (2014, November). Diagnosis of Diabetes Using a Weight-Adjusted Voting Approach. In Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on (pp. 320-324). IEEE.

[40] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.

[41] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.

[42] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[43] Guo, Y., Bai, G., & Hu, Y. (2012, December). Using bayes network for prediction of type-2 diabetes. In Internet Technology And Secured Transactions, 2012 International Conference for (pp. 471-472). IEEE.

[44] Yıldırım, E. G., Karahoca, A., & Uçar, T. (2011). Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, 1374-1380.

[45] Al Jarullah, A. A. (2011, April). Decision tree discovery for the diagnosis of type II diabetes. In Innovations in Information Technology (IIT), 2011 International Conference on (pp. 303-307). IEEE.

Deep Learning in Medicine

Tarun Jaiswal[1], Ragini Shukla[2]
[1]M.Phil. Dept. of Computer Science, Dr. C. V. Raman University, Bilaspur (C.G.) India.
[2] Asst. Prof., Dept of CS & IT, Dr. C. V. Raman University, Bilaspur (C.G.) India.

ABSTRACT
Spurred by advances in processing power, memory, storage, and an unprecedented wealth of data, computers are being asked to tackle increasingly complex learning tasks, often with astonishing success. Computers have now mastered a popular variant of poker, learned the laws of physics from experimental data, and become experts in video games – tasks which would have been deemed impossible not too long ago. In parallel, the number of companies centered on applying complex data analysis to varying industries has exploded, and it is thus unsurprising that some analytic companies are turning attention to problems in healthcare. The purpose of this review is to explore what problems in medicine might benefit from such learning approaches and use examples from the literature to introduce basic concepts in machine learning. It is important to note that seemingly large enough medical data sets and adequate learning algorithms have been available for many decades – and yet, although there are thousands of papers applying machine learning algorithms to medical data, very few have contributed meaningfully to clinical care. This lack of impact stands in stark contrast to the enormous relevance of machine learning to many other industries. Thus part of my effort will be to identify what obstacles there may be to changing the practice of medicine through statistical learning approaches, and discuss how these might be overcome.

Keywords: Computers; statistics; risk factor; prognosis; machine learning.

I INTRODUCTION

Today we are engaged in frequent enduring studies of the healthcare effects of numerous ingredients, the ultimate consequences of rival approaches of treatment, and decease irrefutable development of diseases. Huge databases on noteworthy inhabitants, focused on brain cancer, cardiovascular disease, arthritis, cancer and other major medicinal problems, are now being collected and used to clarify the true occurrence of diseases, to identify demographic influences and to measure salutary efficacy of drugs and procedures [6-8].

An enlightening review of the history of AI and the bouts between its supporters and challengers may be found in the recently published Machines Who Think [9]. According to Szolovits, P. [10], Medication is a field in which such help is judgmentally needed. Our cumulative expectations of the highest quality health care and the speedy evolution of ever more detailed remedial knowledge leave the physician without adequate time to devote to each case and besieged to keep up with the newest expansions in his field. For lack of time, most medicinal decisions must be based on fast decisions of the case relying on the physician's single-handed remembrance. Only in infrequent circumstances can a nonfiction search or other extended examination be undertaken to assure the doctor and the patient, that the modern knowledge is transported to accept on any particular case. Sustained training and recertification events encourage the surgeon to keep more of the pertinent evidence continuously in mind, but important confines of human recollection and remembrance attached with the growth of information assure that most of what is known cannot be known by most entities. It is the chance for new computer based tools: to help organize, store, and retrieve suitable medical knowledge needed by the consultant in dealing with each difficult case, and to recommend appropriate diagnostic, prognostic and therapeutic decisions and decision making techniques.

Artificial Intelligence is the study of thoughts which allow computers to do the things that make people seem intelligent ... The central goals of Artificial Intelligence are to make computers more useful and to understand the principles which make intelligence possible [11].

"Machine Learning is the discipline of getting computers to learn and act like humans do, and improve their learning over time in self-governing fashion, by feeding those data and information in the form of observations and real-world communications."

In a 1970 review article, Schwartz speaks of -the possibility that the computer as an intelligent tool can redesign the present system of health care, basically alter the role of the doctor, and deeply change the nature of medical manpower employment and medical education--in short, the possibility that the healthcare system by the year 2000 will be basically different from what it is today [12].

According to the knowledge data discovery (KDD0 workshop of machine learning [3] Over the current ages, the decreasing cost of data acquisition and ready availability of data sources such as Smart card Based Health records, claims, administrative data and patient Based health data , as well as shapeless data, have controlled to an increased focus on data-driven and ML methods for medicinal and healthcare area. From the systems natural science point of view, large multimodal data typically including omics, clinical extents, and imaging data are now readily obtainable. Appreciated information for obtaining machine-like insight into the disease is also currently available in shapeless formats for example in the scientific works. The loading, incorporation, and examination of these data current noteworthy challenges for translational medication research and impact on the effective mistreatment of the data. Additionally, intellectual analysis of observational data from Smart card based record and patient based data sources and integration of insights generated from the same to the system natural science sphere can greatly improving longsuffering human involvement, consequence, and refining the complete health of the populace while reducing per capita cost of care. However, the black-box landscape, characteristic in some of the best performing ML methods, has widened the hole between how human and machines think and often unsuccessful to provide clarifications to make understandings tortious. In the novel era with users of "right aimed at explanation", this is detrimental to the acceptance in repetition. To drive the usage of such rich yet assorted datasets into actionable insights, we aim to bring together a wide array of investors, including doctors, biomedical and data science experts, and industry solution subject matter professionals. We will pursue to start deliberations in the area of precision medicine as well as the importance of interpretability of ML models towards the increased practical use of ML in drug and healthcare.

Artificial Intelligence is transforming the world of drug. AI can help doctors make quicker, more accurate diagnoses. It can forecast the risk of a disease in time to prevent it. It can help researchers understand how genetic disparities lead to disease. Although AI has been around for eras, new advances have ignited a prosperous in deep learning. The AI technique powers driver less cars, super-human image recognition, and life-changing, even life-saving, improvements in medicine.

Deep learning helps researchers examine medical data to treat diseases. It improves doctors' ability to analyze medical images. It's proceeding the future of personalized medicine. It even helps the blind "see."[5]

"Deep learning is transforming a wide range of scientific arenas," said Jensen Huang, NVIDIA CEO and co-founder. "There could be no more important application of this new capability than improving patient care." Three trends drive the deep learning revolution: further powerful GPUs, sophisticated neural network algorithms modeled on the human brain, and access to the explosion of data from the internet[5].

II REVIEW OF MACHINE LEARNING IN MEDICINE

Imagine physically as a young graduate student in Stanford's Artificial Intelligence lab, building a system to diagnose a common infectious disease. After years of sweat and toil, the day comes for the test: a head-to-head comparison with five of the top human experts in infectious disease. Over the first expert, system squeezes a narrow victory, winning by just 4%. It beats the second, third, and fourth doctors handily. Against the fifth, it wins by an astounding 52% [13].

Would you believe such a system exists already? Would :ou believe it existed in 1979? This was the MYCIN development, and in vindictiveness of the brilliant investigation outcomes, it certainly not through its way into scientific practice [14].

ML is routinely used in biological expansion and for evaluating and interpreting data in genomics, transcriptomics and proteomics pathways (e.g. [16, 17]), whereas in clinical laboratory medicine, it has been applied to classical biomarker testing of biological resources. Numerous "expert systems (ES)" have been newly labeled, original and commercialized for scientific workshop drives. Designed to appraise specific data in hematology, urinalysis or clinical chemistry, they are conventionally based on a predefined decision tree (DT) encompassing logic rules and checks to exclude diagnostic hypotheses or define them or propose additional examination to complete the diagnosis and support decision making (SDM). By contrast, ML is a totally different approach, where "rules" and everything are learned by the machine, or we can say machine intelligence. More often than not, speaking of obvious rules is unsuitable, as the forecast is somehow hidden in the model's non-linear restrictions that bend the decision boundaries around the data, literally. More specific queries using the search terms laboratory medicine, laboratory tests and machine learning in either the title or abstract identified 34 papers in Scopus and only three in PubMed, one of which was a research journal article [18]. We suppose that ML methods will become more broadly used in the analysis of research laboratory restrictions, and especially for data that can be easily grouped and compared across different groups. The application of ML in laboratory medicine should be supported as a means to enhance research laboratory association and expand the core skills set of research laboratory specialists, within a broader process of change and origination (e.g. [15, 19]). We entitlement this for a quantity of reasons. First, research laboratory are a foremost part of today's healthcare organizations. However, despite high throughput with low turnaround times, the capacity to screen data for results of special interest has decreased and few tests are directly diagnostic [20]. Second, technical improvements have enabled the integration of ES capabilities and software presentations, including automatic analyzers and modules of research laboratory information systems [2'1]. Since this kind of support is usually based on dichotomous thresholds or rigid mutual exclusion of data, it can be difficult if not impossible to obtain precise or personalized results [22], suggesting an obvious edge for development.

Third, because patients can now directly access their research laboratory test results of their investigative benefactor, there is an increasing demand for meaningful, possibly personalized reference limits and the need to interpret precision asterisks [20], the conventional signs indicating abnormal or borderline values. Finally, with the convergence of smartphones and innovative biosensors based on microfluidics and microelectronics, the vision of the lab-on-a-chip (LOC) and related models for laboratory medicine has opened opportunities, "in which a smartphone-enabled portable laboratory is brought to the patient instead of the patient being brought to the laboratory" [23]. In this context, apomediation refers to progressive disintermediation whereby traditional intermediaries, such as healthcare professionals who give "relevant" information to their patients, are functionally replaced by apomediaries, i.e. network/group/collaborative filtering processes [24]. ML systems can be seen as new and "smarter" apomediaries that act as gap fillers that analyze the increasing amount of diagnostic data a patient can access without mediation by a general practitioner or laboratory specialist, and then refer the patient to a specialist only in case of likely positive or anomalous results. This can be done by factoring together the diverse phenotypic attributes of a patient (i.e. in addition to body mass index [BMI], age, gender and ethnicity) or, better yet, of the patient's history of past basal values associated with a healthy condition. In this case, the very notion of reference limits would change, and ML, by leveraging and improving other statistical approaches, could help limit the misinterpretation of values outside of reference limits or of apparently normal data but also diagnostic for some conditions (e.g. [25]). Furthermore, some envision an ML-based clinical decision support that, by predicting correlated test results and enhancing the diagnostic value of multianalyte sets of test results, could help to reduce redundant laboratory testing [25] and, hence, lower healthcare costs, which are estimated to total $5 billion yearly in the United States alone [27]. Finally, the growing number of available and affordable types of diagnostic tests, different measurement methods and patient phenotypes (e.g. ethnic subtypes) has produced an unprecedented complexity of data interpretation and integration that calls for novel management technologies. In the following section, we report on research into the potential of ML models to address these challenges in laboratory medicine.

Hoong [4] summarized the potential of AI techniques in medicine as follows:

(a) Provides a research laboratory for the checkup, association, demonstration and classification of medical acquaintance.

(b) Harvests new tools to support medical decision-making, training and research.

(c) Participates happenings in medical, computer, cognitive and other disciplines.

(d) Propositions a content-rich discipline for forthcoming scientific medical area.

Many intelligent system have been developed for the purpose of enhancing health-care and provide a better health care facilities, reduce cost and etc. As express by many studies [28-33], intelligent system was developed to assist users and provide early diagnosis and prediction to prevent serious illness. Even though the system is equipped with "human" knowledge, the system will never replace human expertise as human are required to frequently monitor and update the system' s knowledge. Therefore, the role of medical specialist and doctors are important to ensure system validity.

Early studies in intelligent medical system such as MYCIN, CASNET, PIP and Internist-I have shown to out performs manual practice of diagnosis in several disease domain [34]. MYCIN was developed in the early 1970s to diagnose certain antimicrobial infections and recommends drug treatment. It has several facilities such as explanation facilities, knowledge acquisition facilities, teaching facilities and system building facilities. CASNET (Causal ASsociationalNETworks) was developed in early 1960s is a general tool for building expert system for the diagnosis and treatment of diseases[1-2].

CASNET major application was the diagnosis and recommendation of treatment for glaucoma. PIP an abbreviation for Present Illness Program was developed in 1970s to simulates the behaviour of an expert nephrologist in taking the history of the present illness of a patient with underlying renal disease. The work on Internist-I in early 1982s was concentrated on the investigation of heuristic methods for imposing differential diagnostic task structures on clinical decision making. It was applied in diagnoses of internal medicine.

In 1990s, the study in intelligent system was enhanced to utilize the system based on current needs. In several studies two or more techniques were combined and utilized the function of the system to ensure system performance. ICHT (An Intelligent Referral System for Primary Child Health Care) developed to reduce children mortality especially in rural areas [35]. The system success in catering common paediatric complaints, taking into consideration the important risk factors such as weight monitoring, immunization, development milestones and nutrition. ICHT utilized expert system in the process of taking the history data from patients. Other expert system have been developed such as HERMES (HEpathology Rule-based Medical Expert System) an expert system for prognosis of chronic liver diseases [36], Neo-Data expert system for clinical trails [37], SETH an expert system for the management on acute drug poisoning [38], PROVANES a hybrid expert system for critical patients in Anesthesiology [39] and ISS (Interactive STD Station) for diagnosis of sexually transmitted diseases [40].

Experienced Based Medical Diagnostics System an interactive medical diagnostic system is accessible through the Internet [29]. Case Based Reasoning (CBR) was employed to utilize the specific knowledge of previously experienced and concrete problem or cases. The system can be used by patients to diagnose them without having to make frequent visit to doctors and as well as medical practitioner to extend their knowledge in domain cases (breast cancer).

Data mining is an AI technique for discovery of knowledge in large databases, could be used to collect hidden information for medical purposes [41-43]. It could also be combined with neural network for classification of fuzzy pattern of HIV and AIDS using unsupervised learning [41]. Patients status life or dead was classified as training and testing pattern. Data mining was also used to generate a scatter diagram and a model of rules statement to enhance current rule base system [42]. Neves et al [43] developed information system that supports knowledge discovery and mining in medical imaging.

Fuzzy logic is another branch of artificial intelligence techniques. It deals with uncertainty in knowledge that simulates human reasoning in incomplete or fuzzy data. Meng [44] applied fuzzy relational inference in medical diagnosis. It was used within the medical knowledge based system, which is referred to as Clinaid. It deals with diagnostic activity, treatment recommendations and patient's administration.

Neural Network (NN) is one of the powerful AI techniques that has the capability to learn a set of data and constructs weight matrixes to represent the learning patterns. NN is a network of many simple processors or units [45]. It simulates the function of human brain to perform tasks as human does. As an example, a study on approximation and classification in medicine with incremental neural network shows superior generalization performance compared with other classification models [46]. NN has been employed in various medical applications such as coronary artery [47], Myocardial Infarction [48], cancer [49-50], pneumonia [51] and brain disorders [52]. In Karkanis et al [50] NN was implemented as a hybrid with textual description method to detect abnormalities within the same images with high accuracy.

Partridge et al [53] listed several potential of NN over conventional computation and manual analysis:

(i) Implementation using data instead of possibly ill-defined rules.

(ii) Noise and novel situations are handled automatically via data generalization.

(iii) Predictability of future indicator values based on past data and trend recognition.

(iv) Automated real-time analysis and diagnosis.

(v) Enables rapid identification and classification of input data.

(vi) Eliminates error associated with human fatigue and habituation.

III CENTRALIZED DATABASES AND WWW

For system using AI techniques, when the number of patients is high the system will produce more accurate results compared to the system with less number of patients. The patient's records are valuable information for the knowledge-based system. The current patients data would enhance and strengthen the validity of the system reasoning [29].

Current enhancements in information technology such as development of information superhighway inevitably encourage many organizations including government to develop electronic medical information and make it available on the Internet. The patients can use the information and monitor their risk level from their home or office without having to consult the physician [29]. The Internet supports two-ways communications between users around the world at minimum cost. In medical, communication is very important as new information or new discovery is the key for the future survival for example [54]. In addition, communications helps doctors sharing their knowledge or expertise [55].

IV WEB-BASED MEDICAL DIAGNOSIS AND PREDICTION

The model for Web-Based medical diagnosis and prediction consists of four components, they are databases, prediction module, diagnosis module and user interface. The databases consist of patient's database and patients-disease databas^. Patients database will be used to store patient's information such as name, addresses, and others particulars details. Patients-disease database stored all the information about patients and their illness. The information stored in the database includes types of diseases, the treatments and other details about the test and administering therapy. Patients information are separated in a different database to enhance the patients records storage, so that other departments could use the records when the patients are referred to them. This method could prevent other departments or unauthorized users from accessing the information about patient's diseases and provide a centralized information access for the patient's records.

Prediction module and diagnosis module are two of the main features in Web-Based Medical Diagnosis and Prediction. Prediction module utilizes neural networks techniques to predict patients illness or conditions based on the previous similar cases.

V APPLICATION OF MACHINE LEARNING IN MEDICINE

ML has made great advances in pharma and biotech efficiency. The following table 1 shows application of machine learning in medicine.

Table 1

Diagnose diseases	Correctly diagnosing diseases takes years of medical training. Even then, diagnostics is often an arduous, time-consuming process. In many fields, the demand for experts far exceeds the available supply. This puts doctors under strain and often delays life-saving patient diagnostics. ML – particularly Deep Learning algorithms – have recently made huge advances in automatically diagnosing diseases, making diagnostics cheaper and more accessible.
Develop drugs faster	Developing drugs is a notoriously expensive process. Many of the analytical processes involved in drug development can be made more efficient with Machine Learning. This has the potential to shave off years of work and hundreds of millions in investments.
Personalize treatment	Different patients respond to drugs and treatment schedules differently. So personalized treatment has enormous potential to increase patients' lifespans. But it's very hard to identify which factors should affect the choice of treatment. ML can automate this complicated statistical work – and help discover which characteristics indicate that a patient will have a particular response to a particular treatment. So the algorithm can predict a patient's probable response to a particular treatment. The system learns this by cross-referencing similar patients and comparing their treatments and outcomes. The resulting outcome predictions make it much easier for doctors to design the right treatment plan.
Improve gene editing	Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), specifically the CRISPR-Cas9 system for gene editing, is a big leap forward in our ability to edit DNA cost effectively – and precisely, like a surgeon. This technique relies on short guide RNAs (sgRNA) to target and edit a specific location on the DNA. But the guide RNA can fit multiple DNA locations – and that can lead to unintended side effects (off-target effects). The careful selection of guide RNA with the least dangerous side effects is a major bottleneck in the application of the CRISPR system. Machine Learning models have been proven to produce the best results when it comes to predicting the degree of both guide-target interactions and off-target effects for a given sgRNA. This can significantly speed up the development of guide RNA for every region of human DNA.

VI CONCLUSION

AI is already helping us more efficiently diagnose diseases, develop drugs, personalize treatments, and smooth oversee genetic factor. But this is **just the commencement**. The more we digitize and unify our medicinal data, the more we can use AI to help us find appreciated patterns – patterns we can use to make precise, cost-effective judgements in complex analytical processes. ML has the probable to meaningfully aid medical rehearsal. The future for medicine will be better and better. The use of computer and communication tools can change the medical practice into a better implementation. Consolidation in health-care provider will happen by focusing on cost and later on quality of services. Advancement in technology will form a platform for development a better design of telemedicine application. Telephone line and Internet will be the most important tools in medical applications. The main features in medical diagnosis and prediction using artificial intelligence techniques will make the consultation to be more interactive.

REFERENCES

[1] Harvey. J. 1986 Expert Systems: An Introduction. Electrical Communication. 60(2): 100-108.

[2] Kulikowski, C. A. and Weiss, S. M. "Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects." Chapter 2 in Szolovits, P. (Ed.) Artificial Intelligence in Medicine. Westview Press, Boulder, Colorado. 1982.

[3] 2018 KDD Workshop on Machine Learning for Medicine and Healthcare, Location: London, United Kingdom, Workshop Date: August 20, 2018

[4] Hoong, N. K. (1988). Medical Information Science - Framework and Potential. International Seminar and Exhibition Computerization for Development-the Research Challenge, UniversitiPertanian Malaysia: Kuala Lumpur, pp. 191 - 198.

[5] Deep Learning,Advances In Medicine,Blog, Https://Www.Nvidia.Com/Object/Deep-Learning-In-Medicine.Html

[6] Mabry, J. C., Thompson. H. K., Hopwood, M.D., and Baker, W. R., "A Prototype Data Management and Analysis System--CLINFO: System description and user experience," MEDINFO 77, North-Holland, Amsterdam, (1977), 71-75,

[7] Rosati, R. D., McNeer, J. F., and Stead, E. A., pages 1017-1024. "A New Information System for Medical Practice," Archives of Internal Medicine 135, (1975).

[8] Weyl. S., Fries, J., Wiederhold, G., and Germano, F., "A modular self-describing clinical databank system," Comp. Biomed. Res. 8, (1975), 279-293,

[9] McCorduck, P., Computers Who Think, W. H. Freeman and Co., (1980).

[10] Szolovits, P, "Artificial Intelligence and Medicine." Chapter 1 in Szolovits, P. (Ed.) Artificial Intelligence in Medicine. Westview Press, Boulder, Colorado. 1982.

[11] Winston, P. W., Artificial Intelligence. Addison-Wesley, Reading, Mass., (1977).

[12] Schwartz. W. B., "Medicine and the Computer: The Promise and Problems of Change," New Engl. J. Med. 283, (1970), 1257-1264.

[13] Brandon Ballinger, Co-Founder @Cardiogram,Blog, Three Challenges for Artificial Intelligence in Medicine,Sep. 20,2016.

[14] An interesting snapshot of history can be found in the 1982 book Artificial Intelligence in Medicine, long out of print.

[15] Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216.

[16] Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, et al. Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification. ClinChem Lab Med 2010;48:1319–26.

[17] Madabhushi A, Doyle S, Lee G, Basavanhally A, Monaco J, Masters S, et al. Integrated diagnostics: a conceptual framework with examples. ClinChem Lab Med 2010;48:989–98.

[18] Demirci F, Akan P, Kume T, Sisman AR, Erbayraktar Z, Sevinc S. Artificial neural network approach in laboratory test reporting. Am J ClinPathol 2016;146:227–37.

[19] Forsting M. Machine learning will change medicine. J Nucl Med 2017;58:357–8.

[20] Horowitz GL. The power of asterisks. ClinChem 2015;61: 1009–11.

[21] Connelly DP. Embedding expert systems in laboratory information systems. Am J ClinPathol 1990;94(4 Suppl 1):S7–14.

[22] Lippi G, Bassi A, Bovo C. The future of laboratory medicine in the era of precision medicine. J Lab Precis Med 2016;1:7.

[23] Komatireddy R, Topol EJ. Medicine unplugged: the future of laboratory medicine. ClinChem 2012;58:1644–7.

[24] Eysenbach G. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res 2008;10:e22.

[25] Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. J Biomed Inform 2016;59:276–84,

[26] Lindbury BA, Richardson AM, Badrick T. Assessment of machinelearning techniques on large pathology sets to address assay redundancy in routine liver function test profiles. Diagnosis 2015;2:41–51.

[27] Jha AK, Chan DC, Ridgway AB, Franz C, Bates DW. Improving safety and eliminating redundant tests: cutting costs in U.S. hospitals. Health Aff 2009;28:1475–84.

[28] Mahabala, H. N., Chandrasekhara, M. K., Baskar, S., Ramesh, S., and Somasundaram, M. S. (1992). ICHT: An Intelligent Referral System for Primary Child Health Care.

[29] Manickam, S., and Abidi, S. S. R. (1999). Experienced Based Medical Diagnostics System Over The World Wide Web (WWW), Proceedings of The First National Conference on Artificial Intelligence Application In Industry, Kuala Lumpur, pp. 47 - 56.

[30] Alexopoulos, E., Dounias, G. D., andVemmos, K. (1999). Medical Diagnosis of Stroke Using Inductive Machine Learning. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 20-23.

[31] Zelic, I., Lavrac, N., Najdenov, P., Rener-Primec, Z. (1999). Impact of machine Learning of the Diagnosis and Prognosis of First Cerebral Paroxysm. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 24-26.

[32] Ruseckaite, R. (1999). Computer Interactive System for Ascertainment of Visual Perception Disorders. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 27-29.

[33] Bourlas, P., Giakoumakis, E., and Papakonstantinou, G. (1999). A Knowledge Acquisition and management System for ECG Diagnosis. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 27-29.

[34] Shortliffe, E. H. (1987). Computer Programs to Support Clinical Decision Making. Journal of the American Medical Association, Vol. 258, No. 1.

[35] Mahabala, H. N., Chandrasekhara, M. K., Baskar, S., Ramesh, S., and Somasundaram, M. S. (1992). ICHT: An Intelligent Referral System for Primary Child Health Care. Proceedings SEARCC' 92: XI Conference of the South East Asia Regional Computer Confederation. Kuala Lumpur.

[36] Bonfa, I., Maioli, C., Sarti, F., Milandri, G. L., and Monte, P. R. D. (1993). HERMES: An Expert System for the Prognosis of Hepatic Diseases. Technical Report UBLCS-93-19. Universiti of Bologna.

[37] Theodorou, T., and Ketikidis, P. (1995). Neo-Dat An Expert System to Support the Designers of Clinical Trails. 5 th Hellenic Conference on Informatics.

[38] Droy, J. M., Darmoni, S. J., Massari, P., Blanc, T., Moritz, F., and Leroy, J. (1993). SETH: An Expert System for the Management on Acute Drug Poisoning. http://www.churousen.fr/dsii/publi/seth.htm

[39] Passold, Fs., Ojeda, R. G., and Mur, J. (1996). Hybrid Expert System in Anesthesiology for Critical Patients. In Proceedings of the 8 th IEEE Mediterranean Electrotecnical Conference - MELECON' 96 (ITALIA). Vol. III, pp. 1486-1489.

[40] Walker, N. J., and Kwon, O. (1997). ISS: An Expert System for the Diagnosis of Sexually Transmitted Diseases. 11 thAnnual Midwest Computer Conference (MCC' 97) March 21, Springfield, Illinois.

[41] SitiNurul Huda Sheikh Abdulah and MiswanSurip (1999). SatuMetodologiPerlombongan Data UntukPesakit AIDS. Proceedings of the First National Conference on Artificial Intelligence Application in Industry. Kuala Lumpur, pp. 57-71.

[42] Siti FatimahMdSaad and RogayahGhazali (1999). Data Mining for Medical Database.Proceedings of the First National Conference on Artificial Intelligence Application in Industry. Kuala Lumpur, pp. 72-79.

[43] Neves, J., Alves, V., Nelas, L., Romeu, A., and Basto, S. (1999). An Information System That Supports Knowledge Discovery and Data Mining in Medical Imaging. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 37-42.

[44] Meng, Y. K. (1996). Interval-Based Reasoning in Medical Diagnosis. Proceedings of National Conference on Research and Development in Computer Science and Its Applications (REDECS'96), UniversitiPertanian Malaysia: Kuala Lumpur, pp. 220 -224.

[45] Sarle, W. S. (1999). Neural Network FAQ, part 1 of 7: Introduction. Periodic posting to the Usenet Newsgroup comp.ai.neural-nets, ftp://ftp.sas.com/pub/neurl/FAQ.html

[46] Jankowski, N. (1999). Approximation and Classification in Medicine with IncNetNeuralNetworks. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 53-58.

[47] Lippmann, R. P., Kulkolich, L., Shahian, D. (1995). Predicting the Risk of Complications in Coronary Artery Bypass Operations Using Neural Networks. Advances in Neural Information Processing Systems 7, The MIT Press, Cambridge, pp. 1053-1062.

[48] Heden, B., Ohlsson, M., Rittner, R., Pahlm, O., Haisty, W. K., Peterson, C., and Edenbrandt, L. (1996). Agreement Between Artificial Neural Networks and Human Expert for the Electrocardiographic Diagnosis of Healed Myocardial Infarction. Journal of the American College of Cardiology, Vol. 28, pp. 1012-10s16.

[49] Street, W. N., Mangasarian, O. L., andWolberg, W. H. (1996). Individual and Collective Prognostic Prediction. Thirteenth International Conference on Machine Learning.

[50] Karkanis, S. A., Magoulas, G. D., Grigoriadou, M., and Schurr, M. (1999). Detecting Abnormalities inColonoscopic Images by Textual Description and Neural Networks. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 59-62.

[51] Caruana, R., Baluja, S., and Mitchell, T. (1996). Using the Future to "Sort Out" the Present: Rankrop and Multitask Learning for Medical Risk Evaluation. Advances in Neural Information Processing Systems 8, The MIT Press, Cambridge, pp. 959-965.

[52] Pranckeviciene, E. (1999). Finding Similarities Between An Activity of the Different EEG' s by means of a Single layer Perceptron. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 49-52.

[53] Partridge, D., Abidi, S. S. R., and Goh, A. (1996). Neural Network Applications in Medicine. Proceedings of National Conference on Research and Development in Computer Science and Its Applications (REDECS'96), UniversitiPertanian Malaysia: Kuala Lumpur, pp. 20 - 23.

[54] Shortliffe, E. H., Fagan, L. M. and Yu, V. L. (2000). The Infectious Diseases Physician and the Internet. In Mandell, G.L., Bennett, J.E. and Dolin, R. (Eds.), Mandell, Douglas, and Bennett' s Principles and Practice of Infectious Diseases. Churchill Livingstone, Inc., Pennsylvania, pp. 3258-3263.

[55] Detmer, W. M. andShortliffe, E. H. (1997). Using the Internet to Improve Knowledge Diffusion in Medicine. Communications of the Associations of Computing Machinery, Vol. 40, No. 8, pp. 101 - 108.

Faculty Performance Appraisal

Preeti Jain[1], Dr. Umesh kumar Pandey[2]
[1]Research Scholar, Dept. Computes Science, MATS University, Raipur (C.G.) India.
[2]Associate Professor, MATS University, Raipur (C.G.) India.

ABSTRACT

Appraisal as a lively process produces data, which acts as a performance indicator for an individual and subsequently impacts on the decision making of the stakeholder's as well as the individual. The idea proposed in this paper is to perform an analysis considering number of parameter s for the derivation of performance prediction indicator's needed for faculty performance appraisal, monitoring and evaluation. The aim is to predict the quality, productivity and potential of faculty across various disciplines which will enable higher level authorities to take decisions and understand certain patterns of faculty motivation, satisfaction, growth and decline. The analysis depends on many factors, encompassing student's feedback, organizational feedback, institutional support in terms of finance, administration, research activity etc. The data mining methodology used for extracting useful patterns from the institutional database is able to extract certain unidentified trends in faculty performance when assessed across several parameters.

Keywords-Data-Mining, Performance, Analysis

I INTRODUCTION

The applications of Data Mining in the field of higher education can truly be supported with the findings that typical type of data mining questions used in the business world has counterpart questions relevant to higher education [2]. The need in higher education is to mine faculty and students data from various stakeholders' perspective [7]. The methodology adapted to design the system comprises of Phase-I - Finding the key parameters needed for the assessment and evaluation of the faculties [10]. Phase-II – Finding the most appropriate data mining techniques needed to evaluate the performances with substantial accuracy and to derive the indicators, which help in revising the policies of the institute and the intellectual stature of the faculties.

II PHASE I - PARAMETER IDENTIFICATION

The proposed model as shown in Figure – 1 portrays the framework for faculty performance evaluation system. Figure 2 lists the model depicting seventy seven parameters which have been identified for assessing faculty performance. A database consisting of [50 (faculties) * 77(parameters)] was subjected to data mining algorithms for analysis. The faculties were from Information Technology stream from one Institute. Figure – 1 FPMES -Framework

III TRADITIONALAPPROACH

The Faculty Performance if done using the traditional approach as shown in Figure 3B does not identify the hidden patterns in their performances and is not of muchuse to the management as no clear differentiation emerges in the analysis. The traditional approach uses cumulative values of all parameters taken into consideration. This necessitates using data mining concepts for the performance evaluation so that hidden trends and patterns in faculty performance can be unearthed and can be a benefactor for the management in restoring potential faculties, encouraging faculty growth, honoring and awarding faculties.

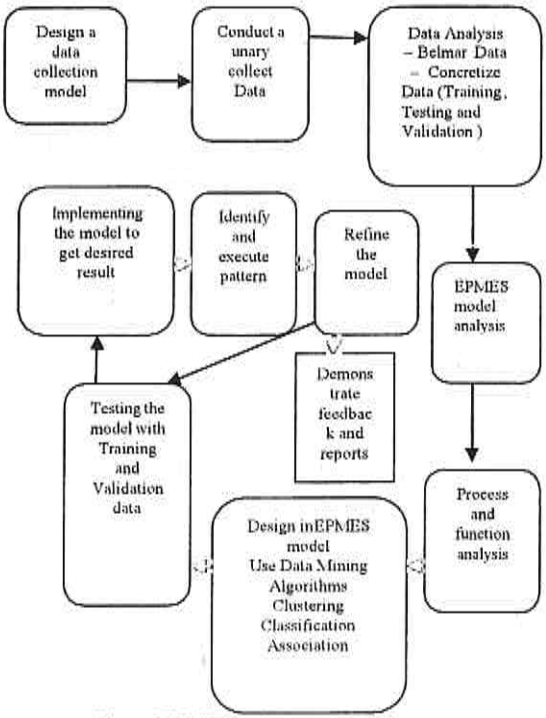

Design a data collection model
Conduct a unary collect Data
Data Analysis – Belmar Data – Concretize Data (Training, Testing and Validation)
Implementing the model to get desired result
Identify and execute pattern
Refine the model
EPMES model analysis
Demons trate feedbac k and reports
Testing the model with Training and Validation data
Design inEPMES model Use Data Mining Algorithms Clustering Classification Association
Process and function analysis

Fig. 1 FPMES – Framework

Evaluation by Management Parameters	Faculty Profile	Educational qualification Industry Experience	
	Teaching	Content Knowledge of the subject Technical Know – how Programming skills	
		Course Design	Appropriate Syllabus Formulation
			Continuous course content improvement
		Instructional Delivery	Syllabus Specific Instructional Delivery
			Continuous improvement in instructional delivery with improvement in course content Follow case based approach
		Instructional Relationships	Support of departmental instructional efforts
			Support from students
		Course Management	
		Class control	
		Guidance to students	Class Advisor
			Living Advisor
			Club Advising
			Summer and Winter coaching
			Student exchange program
			One to one monitoring
		Course organization	
		Years of teaching	
		Thesis advising work	
		Teaching workload	
		Student achievement based on performance exams and projects Project supervision of graduate and postgraduate level	
	Professional Development	Commitment to Pupils and Pupil Learning	The teacher demonstrate commitment to the well-being and development of all pupils The teacher is dedicated in his or her effort to teach and support pupil learning and achievement The teacher trends all pupil equality and with respect The teacher provide an environment for learning that encourage pupils to be problem solving , decision makers , lifelong learners and contributing members of a changing society
		Organizing Professional Learning	The teacher engage in organizing professional learning and applies it to improve his or her teaching practices
			Seeks input from colleagues , consultants or other appropriate support staff and effectively applies it to enhance teaching practices
			Identifies areas for professional growth . attend workshops, appropriate seminar to respond to change in education/policies and practices effectively applies information to enhance teaching practices
			Participates willingly and effectively in professional learning , study groups and in service program to enhance skill development or broaden knowledge

Fig.-2Snap-shot of Performance Parameters

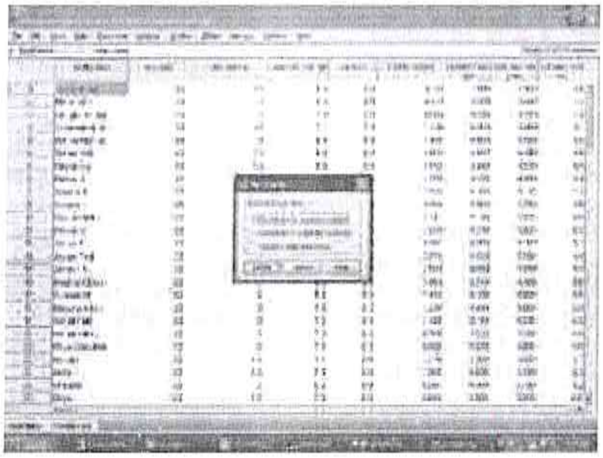


Fig. 3A-Statistics for Pie Chart

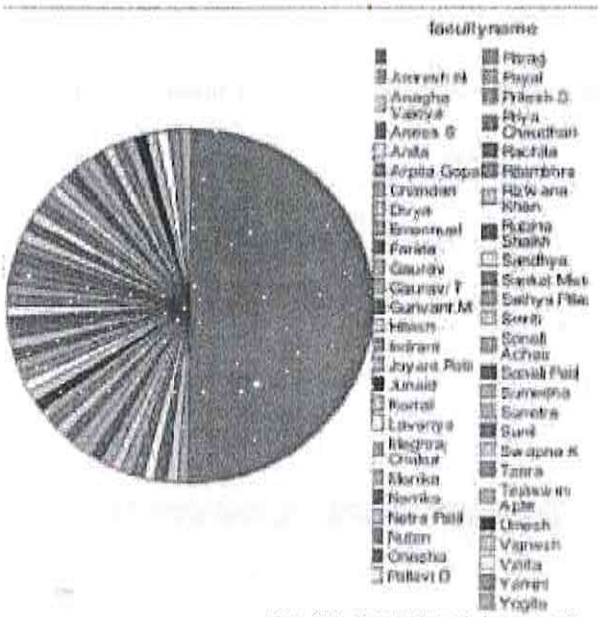


Fig. 3B Traditional Approach

IV PHASE II -OUR APPROACH

To evaluate Faculty Performance using Data Mining Techniques we used PASW Statistics 17.0 to classify the data [12]. The statistical file was subjected to classification using K means algorithm to generate the clusters and the number of cases identified in both the clusters is shown by the results in Table 1.

The pattern recognized was that cluster 1 contains all unique values of faculty performance and cluster 2 contains performance values which are common or occur more than once in the faculty performance database. In Table 2 the distance between the two clusters is moderate as observed from the result and the pattern which is identified indicates that for segment of faculties, performance differs significantly if assessed across 77 parameters which is not the case if few performance parameters are taken into consideration. The statistical file was then subjected to rigorous analysis using Classification and Regression Tree Algorithm (C&R Tree Figure 4) which is a tree based classification and prediction method that uses recursive partitioning to split the training records into segments with similar output field values. Figure 5 shows the interactive tree formation after the C&R tree algorithm was executed on that data set. The interactive tree helps classify tuples as per the parameters taken into consideration. .The Tree Growing Process of C&R tree is as follows: The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest"[1]. In this algorithm, only univariate splits are considered. That is, each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. If X is a nominal categorical variable of I categories, there are 2I-1 possible splits for this predictor. If X is an ordinal categorical or continuous variable with K different values, there are K - 1 different split on X. A tree is grown starting from the root node by repeatedly using the following steps on each node.

(a) **Step–1:** Find each predictor's best split. For each continuous and ordinal predictor, sort its values from the smallest to the largest. For the sorted predictor, go through each value from top to examine each candidate split point (call it v, if x ?v, the case goes to the left child node, otherwise, goes to the right.) to determine the best. The best split point is the one that maximize the splitting criterion the most when the node is split according to it. For each nominal predictor, examine each possible subset of categories (call it A, if xε?A, the case goes to the left child node, otherwise, goes to the right.) to find the best split.

(b) **Step-2:** Find the node's best split. Among the best splits found in step 1, choose the one that maximizes the splitting criterion.

(c) **Step-3:** Split the node using its best split found in step 2 if the stopping rules are not satisfied. The tree has been generated using the expert model with specific stopping criterion. The Gains chart in Figure 6 and 7 shows the performance chart which categorizes the performance depending on the flag associated with the variable faculty performance.

Table 1
Number of cases in each cluster

Cluster	1	3.000
	2	47.000
Valid		50.000
Missing		0.000

Table 2
Distances between Final Cluster Centers

Cluster	1	2
1		22.232
2	22.232	

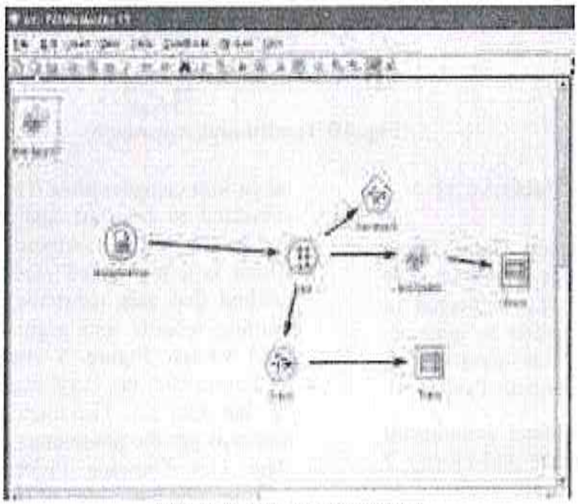


Fig. 4 C & R Tree-Model

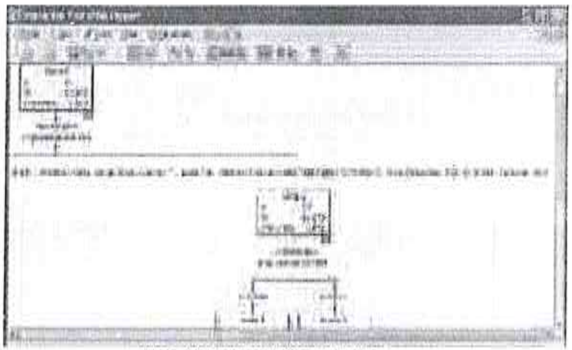


Fig. 5 Interactive Tree

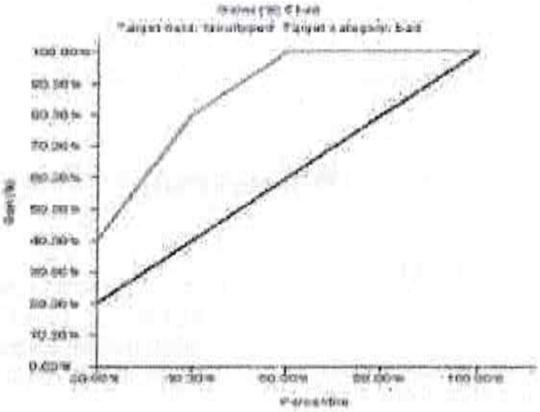


Fig.6 Target Category –Bad

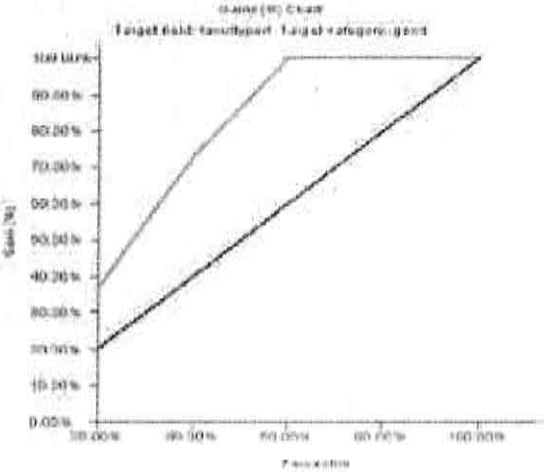


Fig.7 Target Category-Good

Figure 8 shows the Gains chart depicting mean faculty performance measured against the target field faculty performance. Using this classification model it was easy to analyze the known outcomes like a faculty with experience performed better than a novice though while assessing individual cases like faculty acceptance to changes in education policies it was found that newly joined faculties easily accepted the changes while the experienced faculties resisted to the same. Other data mining models like the Segmentation model can also predict the unknown outcomes and patterns of faculty performance.

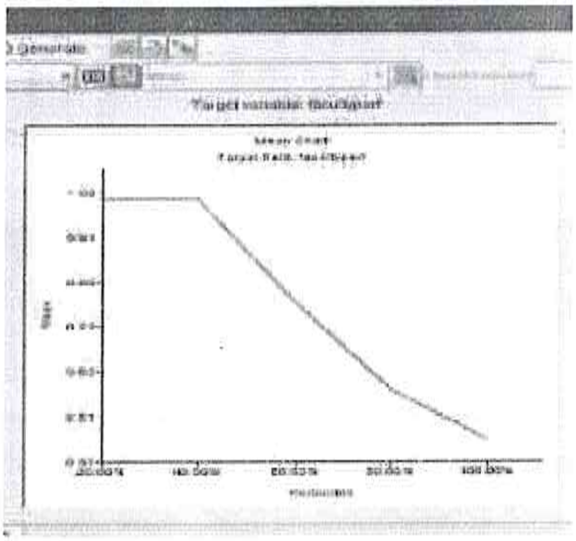


Fig. 8 Mean Chart

V CONCLUSION AND FUTURE WORK

The proposed technique justifies the use of Data Mining to provide effective monitoring tools for faculty performance with considerable accuracy using derived variables which are fine tuned to improve prediction quality. In future we can take into consideration varied segments of faculties across various disciplines and try to find unidentified pattern in their performances using Data Mining models which can help predict unknown outcomes. The reports which will be generated in future will serve mainly to compare changes over time in performances as may be affected by the different predictors that are available plus other well chosen variables.

REFERENCES

[1] Breiman. L., Friedman. J.H., Olshen. R., and Stone. C.J., 1984. Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books &Software. Pacific California.

[2] A.K. Jain and R. C. Dubes. [1988]. Algorithms for Clustering Data. Prentice Hall.

[3] RAgrawal.RSrikant FastAlgorithms for MiningAssociation rules in Large Databases (1994) by Proceedings of the VLDB

[4] Ganti. V., Gehrke, J. and Ramakrishnan, R. 1999a. CACTUSClustering Categorical Data Using Summaries. In Proceedings of the 5thACMSIGKDD, 73-83, San Diego, CA.

[5] GUHA, S., RASTOGI, R., and SHIM, K. 1999. ROCK: A robust clustering algorithm for categorical attributes. In Proceedings of the 15th ICDE, 512-521, Sydney, Australia.

[6] Zaki, M.J. Scalable algorithms for association mining Knowledge and Data Engineering. IEEE Transactions on Volume 12, Issue 3, May/Jun 2000 Page(s):372 390 Digital Object Identifier 10.1109/69.846291

[7] Chiu, T., Fang, D., Chen, J., and Wang, Y. 2001. A Robust and scalable clustering algorithm for mixed type attributes in large database environments. In Proceedings of the 7th ACM SIGKDD. 263-268, San Francisco, CA.

[8] Luan J. [2002] "Data Mining and Knowledge Management in higher Education" Presentation atAIR Forum,Toronto, Canada.

[9] Fathi Elloumi, Ph.D., David Annand, [2002] Integrating Faculty Research Performance Evaluation and the Balanced Scorecard in AUStrategic Planning:ACollaborative Model.

[10] Raoul A. Arreola, Michael Theall, and Lawrence M. Aleamoni [2003] Beyond Scholarship: Recognizing the Multiple Roles of the Professoriate." Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL. April 21-25, 2003).

[11] M.R.K. Krishna Rao. [2004] Faculty and Student Motivation: KFUPMFaculty Perspectives

User Identification across Multiple Social Networks

Anjli Barman[1], Rohit Miri[2], S.R. Tandan[3]
[1,2,3]Dept. of CSE, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

Multinle social networks have become a way to connect with people they know and share their opinions on news and events with friends, colleagues and relatives etc. With the source of picture and video media, multiple social networks are a great way of passing time. Some of the most popular social networks include Facebook, Twitter, Instagram, Whatsapp, YouTube, Snapchat etc. There are number of social network networks to connect more number of people around the world. All social network networks different from each other based on various components such as Graphical User Interface, functionality, features etc. Many users have virtual identities on multiple social network networks. It is common that people are users of more than one social network and also their friends may be registered on multiple social network networks. User may login or sign in to multiple social network sat different timing, so user may not find his friends online when he logins to the particular social networking website. To overcome this issue their proposed system will bring together their online friends on multiple social networks into a single integrated environment. This would enable the user to keep up-to-date with their virtual contacts more easily, as they'll as to provide improved facility to search for people across multiple social networks.

Keywords: Online Social Networks, Identity search, Identity resolution, Privacy, Digital activities, User Identification, Cross-Media Analysis, Social media net work, Friend relationship, Anonymous identical users.

I INTRODUCTION

In this article, they propose a method to identify users based on profile matching. To match profile they evaluate the importance of fields in the web profile and develop a profile comparison tool. By using this profile comparison tool user can easily find out other friends who are available on multiple social networks. This system is a web application where user will register himself and will login to the system using his user id and password. User can view his friends who are online on multiple social networks in a single integrated environment. User can search for friends who are on other social networking networks using profile comparison tool. This system will help many people to connect with each other. The effectiveness and efficiency of profile comparison tool is that it identifies and finds duplicate users on different social networks.

In this article, we propose a method to identify users based on profile matching. To match profile we evaluate the importance of fields in the web profile and develop a profile comparison tool. By using this profile comparison tool user can easily find out other friends who are available on multiple social networks. This system is a web application where user will register himself and will login to the system using his user id and password. User can view his friends who are online on multiple social networks in a single integrated environment. User can search for friends who are on other social networking networks using profile comparison tool. This system will help many people to connect with each other. The effectiveness and efficiency of profile comparison tool is that it identifies and finds duplicate users on different social networks.

(a) **Features**

(i) User can see his friends who are online on other social networking networks in a single integrated environment.

(ii) This system allows finding user who is registered on multiple **social networks**.

(iii) This system uses profile comparison tool to find out user's friends who are available on multiple **social networks**.

(iv) This system will evaluate the importance of fields in the web profile and develop a profile comparison tool. These important fields in the web profile will be used to search duplicate users on multiple social networks.

(v) This system helps many people to connect with each other.

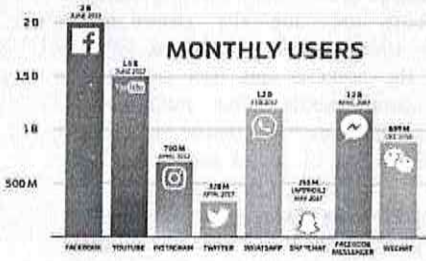


Fig. 1: Users activities in multiple social networks

II STUDY AREA

Many people use more than one social network. They create accounts for sharing both private and public information. This information is digital footprints that can be used to identify the owners. To identify the account identities, it is necessary to gather user information regarding their online behaviours. In this paper, the architecture for matching accounts across multiple social networking networks was proposed. It is designed for extensibility and configurability so that, given an account of a user on a social networking networks, it can be used to find other accounts belonging to the same user on any social networking networks. The system collects account information, such as username, friends, and interests from accounts on multiple social networking networks. User may login to different social networking networks at different timing, so user may not find his friends online when he logins to the particular social networking website.

III METHODOLOGY

To approach the earlier mentioned problems, one has to first identify users across multiple social networks. Our methodology for identifying users across multiple social networks is based on unique behavioural patterns that individuals exhibit on social media. Our methodology has direct roots in behavioural theories in sociology and psychology. These behaviours are due the environment, personality, or even human limitations of the individuals in the content and link individuals generate on social media. Our methodology performs feature discovery to capture traces that these behaviours leave in social media for user identification. Before introducing our methodology, we discuss the types of information that can help us identify users across networks. Network structure and friendship information is known to carry information that could prove useful in many tasks, such as link and attribute prediction, spam detection, Behavioural analysis and group behaviour. Recent studies have indicated that link-based Methods outperform many other techniques on various tasks.

(a) **Definitions**

(i) **User identity:** A user identity of a user on a social network is composed of three dimensions of attributes Profile, Content and Network. Profile is describes such as username, name, age, location, etc. Content is describes the content creates or is shared by user such as text, time of post, etc., and Network is describes the network, which user creates to connect to this users such as number of friends. A real-world user is denoted by **I** and his user identity on a social network $\mathbf{SN_A}$ is denoted by $\mathbf{I_A}$.

(ii) **Problem Definition:** Given an identity I_A of user I on social network SN_A, and his correct identity I_B on social network SN_B:

$$\mathbf{I_A \rightarrow \{I_B\}}$$

The process of user identification in social networks follows two sub-processes **user identity search** and **user identity matching**. User identity search process is a set of user identities on SN_B, which are similar to given identity I_A and belong to user I. User identity matching process then calculates the similarity score between I_A and every user identity returned by user identity search process, on certain metrics. User identities are then ranked on the basis of similarity score, and the user identity with highest match-score is returned as I_B.

User Identity Search: For a user I, given his identity I_A on social network SN_A and a search parameter S, and a set of identities I_{Bj} on social network SN_B such that

$$\mathbf{S\ (I_A) \cong S\ (I_{Bj})}.$$

$$\mathbf{\{I_A, S\} \rightarrow \{I_{B1}, \ldots.., I_{Bj} \ldots\ldots, I_{BN}\}}$$

Any search method takes a source and a set of search parameters as input and retrieves a setof user items which hold similar values for the search parameters. For a user identity search algorithm, source can be given identity I_A and search parameters can be I_A attributes defined on her three identity dimensions namely profile, content, and network. Identity Search by profile implies searching for user identities on SN_B by profile attributes as search parameters extracted from I_A. The user identities I_{Bj} returned are similar to I_A in terms of profile attributes as username, name, gender, school, education, etc. Identity Search by content implies searching for user identities on SN_B with content attributes of I_A as search parameters. The user identities I_{Bj} returned are similar to I_A in terms of content creation, URLs posted, platform used for content creation, timestamp, etc. Identity Search by network implies searching for user identities on SN_B by network attributes of I_A as search parameters. The user identities I_{Bj} are similar to I_A in terms of friends, network in-degree, network out-degree, etc.

(iii) **User Identity Matching:** Given a user identity I_A of user Ion social network SN_A, a set of user identities:

$$\mathbf{Q = \{I_{B1}, \ldots, I_{Bj}, \ldots, I_{BN}\}}$$

On social network SN_B and a match function M, locate an identity pair (I_A; I_{Bj}) such that

$$\mathbf{M\ (I_A, I_{Bj}) = Max\ \{M\ (I_A, I_{B1}), \ldots, M\ (I_A, I_{BN})\}}$$

I_{Bj} with highest match score is inferred as I_B.

$$\mathbf{\{I_A, Q, M\} \rightarrow \{I_A, I_{Bj}\} \rightarrow I_B}$$

IV RESULTS AND DISCUSSION

Nowadays, more and more people have their virtual identities on the multiple social networks. It is common that people are users of more than one social network and also their friends may be registered on multiple web networks. A facility to aggregate our online friends into a single integrated environment would enable the user to keep up-to-date with their virtual contacts more easily, as well as to provide improved facility to search for people across multiple social networks.

In this article, we propose a method to identify users based on profile matching. To match profile we evaluate the importance of fields in the web profile and develop a profile comparison tool. By using this profile comparison tool user can easily find out other friends who are available on multiple social networks. This system is a web application where user will register himself and will login to the system using his user id and password. User can view his friends who are online on multiple social networks in a single integrated environment. User can search for friends who are on other social networking networks using profile comparison tool. This system will help many people to connect with each other. The effectiveness and efficiency of profile comparison tool is that it identifies and finds duplicate users on different social networks.

V CONCLUSION

This system will be useful for user who use social networking networks and likes to use multiple social networks. In this article, we have provided empirical evidence on the existence of a mapping between identities of individuals across the social media networks and studied the possibility of identifying users across multiple social networks. Both link and content information were used to identify individuals. In the link section, we found that when an individual is present on both networks, there are not relationships between the numbers of friends that the individual has on each network.

It was also shown that when the same individual had some friends shared across the two networks, no correlations were observed regarding what percentage of friends on each network was shared? Furthermore, we found that the target-node is not very likely to be connected to the crossed-over friends of the base node, and even in the cases that it is found to be connected, it is challenging to identify it among all connected nodes. These findings and evaluation results of the proposed method show that counter-intuitively, link information is not sufficient to identify individuals across social media networks. However, content information and in particular usernames can be used quite successfully to identify corresponding usernames on various networks. We demonstrated a content-based methodology for connecting individuals across social media networks.

REFERENCES

[1] Paridhi Jain, Ponnurangam Kumaraguru, Anupam Joshi Indraprastha Institute of Information Technology (Iiit-Delhi), India, Identifying Users Across Multiple Online Social Networks.

[2] Anshu Malhotra, Ponnurangam Kumaraguruy Indraprastha Institute of Information Technology, New Delhi, India, Studying User Footprints in Different Online Social Networks, 29 Jan 2013.

[3] N. Chandramouli, Velpula Mounica Assistant Professor, M.Tech, Department Of Cse, Vaageswari College of Engineering, Karimnagar, Telangana, India, Identifying Users across Multiple Online Social Networks, International Journal of Scientific Research in Computer Science, Engineering and Information Technology Nov-Dec2017 Ijsrcseit.

[4] B. Guruprasath, B. Dinesh, R. Kalyana Raman, S. Mohamed Riswan Assistant Professor, Ug Students,Department Of Information Technology A.V.C College Of Engineering, Tamilnadu, India, Unique User Identification Across Multiple Social Networks, International Journal For Research In Applied Science & EngineeringTechnology (Ijraset), Volume 5 Issue Iii, March2017.

[5] Paridhi Jain, Ponnurangam Kumaraguru Indraprastha Institute of Information Technology (Iiit-Delhi),India, Finding Nemo: Searching and Resolving Identities of Users across Online Social Networks, 26 Dec 2012.

[6] Reza Zafarani And Huan Liu Department Of Computer Science And Engineering,Arizona State University, Connecting Corresponding Identities Across Communities, Proceedings Of The Third International Icwsm Conference (2009).

[7] Reza Zafarani, Syracuse University Lei Tang, Clari Huan Liu, Arizona State University, User Identification Across Social Media, Article 16 (October 2015).

[8] Reza Zafarani and Huan Liu Computer Science and Engineering Arizona State University, Connecting Users across Social Media Sites: A Behavioral-Modeling Approach

Inversion of Integral Equation Involving Polynomial Suggested by Hermite Polynomial

Priyank Jain[1], **Dr. Archana Lala**[2], **Dr. Chitra Singh**[3]
[1]Ph.D. Scholar, RNT University, Bhopal (M.P.) India.
[2]Dept. of Mathematics, SRGI, Jhansi (U.P.) India.
[3]Dept. of Mathematics, RNT University, Bhopal (M.P.) India.

ABSTRACT

The purpose of this paper is to drive a solution of certain integral equation whose kernel involves Generalized Hermite Polynomial. We believe that our result is unify in nature and many results can be obtained by considering suitable parameters involved in Generalized Hermite Polynomial. For the purpose of illustration we mentioned a special case briefly by choosing suitable parameters involved in Generalized Hermite Polynomial.

Keyword: Generalized Hermite Polynomial, Mellin Transform, Convolution Theorem, Fox-H function.

I INTRODUCTION

Many boundary value problems reduced to the problem of solving integral equations whose kernel involves many well known classical polynomials like those of Hermite, Laguerre, Bessal, Legendre, Jacobi etc. During the recent past attempts have been made to generalize and unify these classical polynomials with the help of Rodrigue's formulae. To mention Goued Hopper [8] gave a generalization of Hermite polynomials by formulae.

$$H_n^r\left(x,a,p\right) = (-1)^n x^{-a} e^{px^r} D^n\left[x^a e^{-px^r}\right]$$

(1.1)

and we have used

$$H_n^2\left(x,0,1\right) = (-1)^n e^{x^2} D^n\left[e^{-x^2}\right]$$

(1.2)

where $D = \frac{d}{dx}$ and $r, a,$ and p are parameters, for suitable value of $r, a,$ and p (1.1) reduced to modified Hermite, modified Laguerre and modified Bessel polynomials. In view of these generalizations it is worth considering integral equations involving $H_n^2\left(x,0,1\right)$ as kernel and such we prove the following theorem.

II THEOREM

If f is an unknown function satisfying the integral equation.

$$g\left(x\right) = \int_0^\infty k\left(x \mid y\right) f(y)\frac{dy}{y}, x > 0$$

(2.1)

Where $k(x) = e^{-x^2}.H_n^2\left(x,0,1\right)$

and g is a prescribed function then f is given by

For $r = 2, p = 1$

$$f(x) = (-1)^n x^n$$

$$\times \int_0^\infty 2H_{3,2}^{2,0}\left[\frac{x}{y}\middle|\begin{matrix}(0,1)(-n,1/2)(-n,1)\\(-2n,1)(-n,1)\end{matrix}\right]$$

$$\times\left[\left(\frac{d}{dy}\right)^n\left\{g\left(y\right)\right\}\right].\frac{dy}{y}$$

III SOLUTION

To Prove the Theorem we make use of Mellin Transform and discuss case $r > 0$, $p > 0$.

$(r = 2, p = 1\ in\ our\ case)$

By the convolution theorem for Mellin Transform (2.1) reduces to

$$k^*(s)f^*(s) = g^*(s)$$

(3.1)

Where $k^*(s), f^*(s), g^*(s)$ are respective Mellin Transform of $k\left(x\right), f\left(x\right), g\left(x\right)$ and by Sneddon

$$f^*(s) = \int_0^\infty f(x)x^{s-1}dx = M\left[f(x),s\right] \qquad (3.2)$$

When $r > 0$ and $p > 0$, Applying Mellin Transform of equation (3.1) and use the result of Erdelyi $k^*(s) = (-1)^n M\left[D^n\left(e^{-x^2}\right);s\right]$ [9], we get

$$= \frac{\Gamma s\,\Gamma\frac{(s-n)}{2}}{2\,\Gamma s-n} \qquad (3.3)$$

Where

$\mathrm{Re}(s) > n,\ where\ \mathrm{Re}(a) = 0$

$\mathrm{Re}(s) > n - \mathrm{Re}(a);\ where\ \mathrm{Re}(a) \le 0$

We write equation (3.1) in the form

$$f^{*}(s) = \frac{g^{*}(s)}{k^{*}(s)}$$

replacing s by $s - n + a$ where $a = 0$

$$f^{*}(s-n) = (-1)^{n} L^{*}(s)\left[(-1)^{n} \frac{\sqrt{s}}{\sqrt{s-n}}.g^{*}(s-n)\right]$$

(3.4)

Where

$$L^{*}(s) = \frac{\sqrt{s-n}}{\sqrt{s}.k^{*}(s-n)}$$

(3.5)

Then from (3.3) and (3.5)

$$L^{*}(s) = \frac{2\sqrt{s-2n}\sqrt{s-n}}{\sqrt{s}\sqrt{s-n}\sqrt{\frac{s-2n}{2}}} \qquad (3.6)$$

By use of definition of H function. We get the inverse transform $L(x)$ of $L^{*}(s)$ as

$$L(x) = 2H_{3,2}^{2,0}\left[x\,\middle|\,\begin{matrix}(0,1)(-n,1/2)(-n,1)\\(-2n.1)(-n,1)\end{matrix}\right] \qquad (3.7)$$

Where $H_{m,n}^{p,q}$ are Fox's H functions defined by [5].

And now taking Mellin Transform on both sides of (3.4), using convolution theorem and result of Mellin Transform. We get

$$M^{-1}\left[f^{*}(s-n)\right]$$

$$= (-1)^{n}\int_{0}^{\infty} L\left(\frac{x}{y}\right)\left[M^{-1}\left\{(-1)^{n}\frac{\sqrt{s}}{\sqrt{s-n}}g^{*}(s-n)\right\}\right]\frac{dy}{y}$$

$$x^{-n}.f(x) = (-1)^{n}\int_{0}^{\infty} L\left(\frac{x}{y}\right)\left[\left(\frac{d}{dy}\right)^{n}\{g(y)\}\right]\frac{dy}{y}$$

$$f(x) = (-1)^{n} x^{n}\int_{0}^{\infty} L\left(\frac{x}{y}\right)\left[\left(\frac{d}{dy}\right)^{n}\{g(y)\}\right]\frac{dy}{y}$$

Hence using (3.7)

$$f(x) = (-1)^{n} x^{n}$$

$$\times\int_{0}^{\infty} 2H_{3,2}^{2,0}\left[\frac{x}{y}\,\middle|\,\begin{matrix}(0,1)(-n,1/2)(-n,1)\\(-2n,1)(-n,1)\end{matrix}\right]$$

$$\times\left[\left(\frac{d}{dy}\right)^{n}\{g(y)\}\right].\frac{dy}{y} \qquad (3.8)$$

Thus we have prove the following theorem – If f is unknown function satisfying (2.1), where g is some known function then f is given by (3.8) according $r > 0$.

REFERENCES

[1] Goyal S P and Salim T O. (1998), A class of convolution integral equations involving a generalized polynomial set, Proc. Indian Acad. Sci. (Math. Sci.): Vol 108(1), PP.55-62

[2] Srivastava R. (1994), The inversion of an integral equation involving a general class of polynomials, J. Math, Anal. Appl: Vol 186, PP.11-20

[3] Lala A and Shrivastava P N (1990), Inversion of an integral involving a generalized function, Bull. Calcutta Math. Soc.Vol 82, PP.115-118

[4] Lala A and Shrivastava P N (1990), Inversion of an integral involving a generalized Hermite polynomial, Indian J. Pure App/. Math. Vol 21, PP.163-166

[5] Srivastava H M, Gupta K C and Goyal S P (1982), The H-functions of One and Two Variables with Appilcations, (New Delhi: South Asian Publ.),

[6] I N Sneden (1974), The use of Integral Transforms. Tata McGraw Hill, New Delhi.

[7] Srivastava H M and Singhal J P (1971), A class of polynomials defined by generalized Rodrigues formula, Ann. Mat. Pura Appl., Vol 90, PP.75-85

[8] Gould H Wand Hopper A T (1962), Operational formulas connected with two generalizations of Hermite polynomials, Duke Math. J. Vol 29, PP.51-63

[9] Erdelyi A, Magnus W, Oberhettinger F and Triconi F G (1954), Tables of Integral Transforms (New York: McGraw-Hili) Vol. I

Solution of Dual Integral Equations by Reducing It into an Integral Equation

Anil Tiwari[1], Dr. Archana Lala[2], Dr. Chitra Singh[3]
[1]Ph.D. Scholar, RNT University, Bhopal (M.P.) India.
[2]Dept. of Mathematics, SRGI, Jhansi (M.P.) India.
[3]Dept. of Mathematics, RNT University, Bhopal (M.P.) India.

ABSTRACT

The aim of this paper is to solve a dual integral equation by changing it into an integral equation by use of mellin transform whose kernel involves Generalized Hermite Polynomial with suitable parameter. We believe that there are some more possible way to reduce such dual integral equations using different transform like those of Henkel, Fourier etc. For the sake of example we choose a dual integral equation of certain type and obtained an integrai equation by use of fractional operator and mellin transform.

Keywords: Generalized Hermite Polynomial; Mellin Transform; Fractional operators; Fox-H function.

I INTRODUCTION

Dual integral equations are often encountered in different branches of mathematical physics. In the solution of certain mixed boundary value problem of mathematical physics, it is worth converting the dual integral equation into an integral equation. In the present paper, we try to solve the certain type of dual integral equations, whose kernel involves Generalized Hermite Polynomial, by converting them into integral equations. Many attempts have already been made in the past to solve such problems. The following integral representation is basic tool for our illustration.

$$\int_0^\infty k_1(x,u)\,A(u)\,du=\lambda(x);\ 0\le x\le 1$$

(1.1) $\int_0^\infty k_2(x,u)\,A(u)\,du=\omega(x);\ x>1$

(1.2)

k$_1$ & k$_2$ are kernels defind over x-u plane.

$$H_n^r(x,a,\ p)=(-1)^r x^{-a}e^{px^r}\ D^n[x^a e^{px^r}]\ D=\frac{d}{dx}$$

a, r, p parameter.

II THEOREM

If f is unknown function satisfying the dual integral equation.

$$\int_0^\infty (x\,|\,y)^{a_1}\ e^{-(x|y)^r}\ H_n^r(x\,|\,y,a_1,1)\ f(y)\frac{dy}{y}=h(x);\ 0\le x<1$$

(2.1)

$$\int_0^\infty (x\,|\,y)^{a_2}\ e^{-(x|y)^r}\ H_n^r(x\,|\,y,a_2,1)\ f(y)\frac{dy}{y}=g(x);1\le x<\infty$$

(2.2) When h and g are prescribed function and a_1, a_2 *and* r are parameters, then f is giving by

$$f(x)=\frac{1}{r}\int_0^\infty L(x\,|\,y)\ t(y)\frac{dy}{y}$$

Where

$$L(x)=H_{2,1}^{1,0}\left(x\left|\begin{matrix}(n,1)\\(1,1)\left(\left(1-\frac{1}{r}(a_1-n)\right),\frac{1}{r}\right)\end{matrix}\right.\right)$$

and

$$t(x)=h(x), 0\le x<1$$

$$t(x)=\frac{r\,x^{-n+a_1}}{\Gamma\left(\frac{1}{r}(a_2-a_1)\right)}$$

$$\times\int_0^\infty (v^r-x^r)^{\left(\frac{1}{r}(a_2-a_1)-1\right)}v^{n-a_2+r-1}\,g(v)\ dv;1\le x<\infty$$

III MATHEMATICAL PRELIMINARY

To prove the theorems we shall use Mellin transformer and fractional integral operator.

$$f^*(s)=M[f(x);s]=\int_0^\infty f(x)x^{s-1}\ dx \qquad (3.1)$$

When $s=\sigma+i\tau$ is a complex variable.

The inverse melling transform f(x) of f*(s) is given by

$$M^{-1}[f^*(s)]=f(x)=\frac{1}{2\pi i}\int_{\sigma+i\infty}^{\sigma-i\infty} f^*(s)x^{-s}\ ds \qquad (3.2)$$

By convolution theorem for mellin transform

$$M\left[\int_0^\infty k(x\,|\,y)\,f(y)\frac{dy}{y};s\right]=k^*(s)\,f^*(s)$$

$$\int_0^\infty k(x\,|\,y)\,f(y)\frac{dy}{y}=M^{-1}\left[k^*(s)\,f^*(s);s\right]$$

$$=\frac{1}{2\pi i}\int_L k^*(s)\,f^*(s)\,x^{-s}\,ds$$

(3.3) When L is suitable contour.

Fractional integral operator

$$\tau(\alpha;\beta;r;w(x))=\frac{r\,x^{-r\alpha+r-\beta-1}}{\Gamma(\alpha)}\int_0^{x}(x^r-v^r)^{\alpha-1}v^\beta w(v)\,dv$$

(3.4)

$$R(\alpha;\beta;r;w(x))=\frac{r\,x^\beta}{\Gamma(\alpha)}\int_x^\infty (v^r-x^r)^{\alpha-1}v^{-\beta-r\alpha+r-1}\,w(v)\,dv$$

(3.5)

IV SOLUTION

Now taking

$$k_i(x) = x^{a_i}\ e^{-x^r} H_n^r(x; a_i; 1),\ i = 1,2$$

Then from Erdeeyi [10] We get

$$k_i^*(s) = \frac{\Gamma s\ \Gamma\frac{1}{r}(s-n+a_i)}{r\ \Gamma s-n},\ i = 1,2$$

(4.1)

Hence by use at (3.3),(2.1) & (2.2) can be written as

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma s\ \Gamma\left(\frac{1}{r}(s-n+a_i\right)}{\Gamma s-n} f^*(s)x^{-s}\,ds$$

$$= h(x);\ 0 \le x < 1$$

(4.2)

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_2\right)}{\Gamma s-n} f^*(s)x^{-s}\,ds$$

$$= g(x);\ 1 \le x < \infty$$

(4.3) Now operating a (4.2) by the operator (3.5) we get

$$\frac{rx^\beta}{\alpha}\cdot\frac{1}{2r\pi i}\int_L \frac{\Gamma s\ \Gamma\left(\frac{1}{r}(s-n+a_2)\right)}{\Gamma(s-n)} f^*(s)x^{-s}\,ds$$

$$\int_x^\infty (v^r - x^r)^{\alpha-1} v^{-\beta-r\alpha+r-1} g(v)dv$$

$$= \frac{rx^\beta}{\Gamma\alpha}\int_x^\infty (v^r - x^r)^{\alpha-1} v^{-\beta-r\alpha+r-1} g(v)dv$$

Now putting $v^r = \frac{x^r}{t}$ and simplifying we get

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_2\right)}{\Gamma s-n}\cdot\frac{x^{-s}}{\Gamma\alpha}$$

$$\int_0^1 (1-t)^{\alpha-1}.t^{\left(\frac{\beta+s}{r}-1\right)}\,dt\ f^*(s)ds$$

$$= \frac{rx^\beta}{\Gamma\alpha}\int_x^\infty (v^r - x^r)^{\alpha-1} v^{-\beta-r\alpha+r-1} g(v)dv$$

(4.4)

$$\Rightarrow \frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_2\right)}{\Gamma s-n}\cdot\frac{x^{-s}}{\Gamma\alpha}$$

$$\times\frac{\Gamma\left(\frac{1}{r}(\beta+S)\right)\Gamma\alpha}{\Gamma\left(\alpha+\frac{1}{r}\beta+\frac{1}{r}s\right)} f^*(s)ds$$

$$= \frac{rx^\beta}{\Gamma\alpha}\int_x^\infty (v^r - x^r)^{\alpha-1} v^{-\beta-r\alpha+r-1} g(v)dv$$

$$\Rightarrow \frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_2\right)}{\Gamma s-n} x^{-s}$$

$$\times\frac{\Gamma\frac{1}{r}(\beta+s)}{.\Gamma\left(\alpha+\frac{1}{r}\beta+\frac{1}{r}s\right)} \times f^*(s)\,ds$$

$$= \frac{rx^\beta}{\Gamma\alpha}\int_x^\infty (v^r - x^r)^{\alpha-1} v^{-\beta-r\alpha+r-1} g(v)dv$$

In equation (4.4), we put $\beta = -n + a_1$ and $\alpha = \frac{1}{r}(a_2 - a_1)$, so that (4.4) Changes to

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_2\right)}{\Gamma s-n} x^{-s}$$

$$\times\frac{\Gamma\frac{1}{r}(s-n+a_1)}{\Gamma\frac{a_2}{r}-\frac{a_1}{r}+\frac{1}{r}(-n+a_1)+\frac{1}{r}s} \times f^*(s)\,ds$$

$$= \frac{r\,x^{-n+a_1}}{\Gamma\left(\frac{1}{r}(a_2-a_1)\right)}$$

$$\times\int_x^\infty (v^r - x^r)^{\left(\frac{1}{r}(a_2-a_1)-1\right)} v^{-n-a_2+r-1} g(v)dv\ 1 \le x \le \infty$$

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\ \Gamma\left(\frac{1}{r}(s-n+a_1\right)}{\Gamma(s-n)} x^{-s}\ f^*(s)ds$$

$$= \frac{r\,x^{-n+a_1}}{\Gamma\left(\frac{1}{r}(a_2-a_1)\right)}\times$$

$$\int_x^\infty (v^r - x^r)^{\left(\frac{1}{r}(a_2-a_1)-1\right)} v^{-n-a_2+r-1} g(v)dv\ 1 \le x \le \infty$$

(4.5)

Now we write

$t(x) = h(x),\quad 0 \le x < 1$

and $t(x) = \dfrac{r\,x^{-n-a_1}}{\Gamma\left(\frac{1}{r}(a_2-a_1)\right)}$

$$\times\int_x^\infty (v^r - x^r)^{\left(\frac{1}{r}(a_2-a_1)-1\right)} v^{n-a_2+r-1} g(v)\ dv.1 \le x < \infty$$

(4.6) Now from (4.2), (4.5), (4.6) we get

$$\frac{1}{2r\pi i}\int_L \frac{\Gamma(s)\,\Gamma\left(\frac{1}{r}(s-n+a_1\right)}{\Gamma(s-n)}\,x^{-s}\;f^*(s)ds=t(x)$$

(4.7)

Again using (3.3), (4.1) & (4.6) becomes

$$\int_0^\infty k_1(x\mid y)f(y)\frac{dy}{y}=t(x);\;0\le x<\infty$$

(4.8)

When $k_1(x)=x^{a_1}e^{-x^r}H_x^r(x;\,a_1;1)$

Thus pair at dual integral equation (1.1) &(1.2) we have been reduced to single integral equation (4.8). Hence by mellin transform (4.8) can be written as –

$$k_1^*(s)f^*(s)=T^*(s)$$

(4.9)

$$\text{Where } k_1^*(s)=\frac{\Gamma s\,\Gamma\left(\frac{1}{r}(s-n+a_1\right)}{\Gamma(s-n)}$$

and T*(s) is the mellin transform of t(x).

Now

$$F^*(s)=L^*(s)\;T^*(s)$$

(4.10)

Where

$$L^*(s)=\frac{1}{k^*(s)}$$

$$=\frac{\Gamma s-n}{\Gamma(s)\,\Gamma\frac{1}{r}(s-n+a_1)}$$

By use of definition of H – function, we get the inverse transform L(x) at L*(S) as

$$L(x)=H_{2,1}^{1,0}\left(x\left|\begin{matrix}(n,1)\\(1,1)\left(\left(1-\frac{1}{r}(a_1-n)\right),\frac{1}{r}\right)\end{matrix}\right.\right)$$

(4.11)

Taking inverse mellin transform of (4.10)

$$f(x)=\int_0^\infty L(x\mid y)\,t(y)\frac{dy}{y}$$

Hence using (4.11) we get

$$f(x)=\frac{1}{r}\int_0^\infty H_{2,1}^{1,0}\left(\frac{x}{y}\left|\begin{matrix}(n,1)\\(1,1)\left(\left(1-\frac{1}{r}(a_1-n)\right),\frac{1}{r}\right)\end{matrix}\right.\right)t(y)\frac{dy}{y}$$

When t (y) is given by (4.6).

Hence proved the theorem.

REFERENCES

[1] Ahdiaghdam S., Ivaz K. and Shahmorad S. (2016), "Approximate solution of dual integral equations" Bull. Iranian Math. Soc. Vol. 42, No. 5, pp. 1077–1086

[2] Chakrabarti A. and Martha S. C. (2012), "Methods of solution of singular integral equations", Math. Sci. 29 pages.

[3] Hoshan N. A. (2009), "Exact solution of certain dual Integral equations involing heat conduction equation", Far East Journal of Applied Mathematics 35(1) 81-88.

[4] Manam S. R.(2007), "On the solution of dual integral equations", Appl. Math. Lett. 20, no. 4, 391–395.

[5] Chakrabarti A. and Berghe G. V.(2004), "Approximate solution of singular integral equations", Appl. Math. Lett. 17, no. 5, 533–559.

[6] Jerry A. J. (1999), "Introduction to Integral Equations with Applications, Second Edition", WileyInterscience, New York,

[7] Nasim C. and Aggarwala B. D.(1984), "On some dual integral equations", Indian J. Pure Appl. Math.15, no. 3, 323–340.

[8] Gould H Wand Hopper A T (1962), "Operational formulas connected with two generalizations of Hermite polynomials", Duke Math. J. Vol 29, PP.51-63.

[9] I. N. Sneddon(1960), "The elementary solution of dual integral equations", Proc. Glasgow Math. Assoc. 4, 108–110,

[10] Erdelyi A, Magnus W, Oberhettinger F and Triconi F G (1954), "Tables of Integral Transforms" (New York: McGraw-Hili) Vol. I.

[11] Jain P., Lala A., Singh C.(2019), "Inversion of Integral Equation Associated with Leguerre Polynomial Obtained from Hermite Polynomial", Int. J. of Engg. Research & Tech. Vol. 8 issue 4, 83-84

An Analytical Study on Earnings Growth and Price-Earnings Ratio Expansion of Indian Listed Mid-Cap Companies and Its Relation with Their Share Price Performance

A K Asnani[1], Dr. Deepti Maheshwari[2], Dr. Sangeeta Jauhari[3]
[1]Research Scholar, Rabindranath Tagore University, Bhopal (M.P.) India.
[2]Dean, Faculty of Commerce, Rabindranath Tagore University, Bhopal (M.P.) India.
[3]Head, Dept. of Management, Rabindranath Tagore University, Bhopal (M.P.) India.

ABSTRACT
Investors are always on the lookout for the stocks which could deliver multiple returns in the long run. Apart from the huge returns in terms of stock price there is additional benefit in terms of lower tax outgo as long term capital gains tax is always lower than short term gains tax. In addition investors have to monitor only few companies in terms of performance and need not change the stocks frequently and look for short term gains. With few potential wealth multipliers in portfolio investor has to only monitor the continuation of has good future prospects. The present study is an attempt to establish relation of the dependent variable (share price) with independent variable (growth in earnings per share for minimum three years), so as to predict the share price in relation to growth in earnings. Earnings Per Share (EPS) is the most important parameter because it directly tells us as to how much we pay in terms of share price and how much company earns in terms of Net Profit per year. This paper attempts to provide empirical evidence on how growth in EPS and PE ratio over a minimum three year period affect the share price movement.

Keywords: Earnings Per Share, PE ratio, CAGR, Multibagger

I INTRODUCTION

There is a strong theory that stock price movements are random in nature. If the stock prices do not follow the trend ofrandom walk, then one possibility is that stock prices followed mean-reversion process. In that case the share price movement's shouldbe predictable from the changes in firm fundamental values.

Reasonably good research is available on finding relation between one year EPS and share price but too little for extended times. It needs more study. Consistent growth in bottom-line should be the ideal case for multiple returns from share price. India being an emerging economy, the stock market is on the radar of domestic as well as foreign investors who poured in staggering 18 lac crores since the year 1991 when huge reforms were carried out.

Merely the thought of mega returns has always generated considerable interest among investorsbe it retail or an institution domestic as well as foreign.Often markets are filled with rumors and share price exhibits volatility. Unfortunately not much research has been carried out in this field and more study is required.

II LITERATURE REVIEW

The Dividend yield (dividend to share price) (Fama and French, 1988) and earnings-to-price ratio (Campell and Shiller, 1988) contributed significantly to the explanation of long-term stock price variation.

Kent Hickman and Glenn H. Petry,(1990) concluded that the predictions of the academically popular dividend discount models are far inferior to the court and price/earnings regression models.

Ansotegui and Esteban (2002) established a long term relationship between the Spanish stock market and itsfundamentals.

In the year 2008,Chang, Hsu-Ling, Yahn-Shir Chen, Chi-Wei Su, and Ya-Wen Chang from Taiwan found that a relationship does exist between EPS and share price but very feeble. Study was performed on Taiwan Panel Data. The empirical result indicated that the relationship existed between stock prices and EPS in the long-run. Furthermore, it was found that for the firm with a high level of growth rate, EPS has less power in explaining the stock prices however, for the firm with a low level of growth rate; EPS has a strong impact in stock prices.

III OBJECTIVES

(a) Study the share price performance of the companiesin terms of EPS growth and PE expansion for minimum three years in a row.

(b) To analyse the bearing of EPS growth and PE expansion growth on share price movement.

IV RESEARCH DESIGN

(a) Sample Design

Introduction: There are total 5137listed entities on BSE. Of these 4713 stocks are listed for equity capital. Out of the 6565 companies are suspended for trading of their shares. Stocks listed on BSE are categorized into various groups like A-Group, B-Group, Z-Group and T-Group. Window for our analysis has been confined to only mid cap shares, as this category comprises of companies which have potential to deliver EPS with consist growth.These company stocks were midcap at the time of start of reporting consistently rising EPS. Many of these stocks are still midcap and few turned into large caps. Also, we included different market sectors so as to cover the whole economy of the country, like NBFC (Non-Banking Finance Company), Consumer goods, Textile, Automobile, Information technology, Leather, FMCG (Fast Moving Consumer Goods), Packaging, Pharmaceuticals, Finance, Chemical, Tiles etc. After screening the stocks we could finalise on 37 stocks. From this population, the stocks have been selected based on the availability of the data.for the purpose of the present study sample of 13 companies has been taken into consideration.

There are two stock exchanges actively being operated at national level – Bombay Stock Exchange (BSE) and National Stock Exchange (NSE), both are located at Mumbai. Almost entire company stocks listed on NSE is also listed on BSE. Thus, BSE stocks represent the whole universe/population. All the stocks listed on BSEare the Universe or Population for the subject study.

Parameters to be studied:

EPS (Earnings per Share) – High Earnings per Share means high level of net profits delivered by the company. Since investors invest for faster rate of returns, higher EPS is always rewarded with high price by investors.
PE ratio – PE ratio is a ratio of current ruling share price divided by the EPS (Earnings Per Share) for the latest completed financial year. In a simple way it indicates the number of year it would take to recover the price paid assuming that the EPS does not grows.

PE ratio being price paid for given earnings indicates that higher the ratio higher the confidence of investors in future prospects of the company. Thus it is partly derived from past performance which has already been captured in EPS growth. Consistently healthy EPS growth makes the investors confident of its future prospects.

In this study data has been considered for minimum three years. Three years duration has been considered to rule out the inconsistent players.

Technique used for studying the three years data: CAGR (Compounded Annual Growth Rate)
Mathematically, it can be represented as
CAGR = Compounded Annual Growth Rate = (Final amount / Initial amount)$^{\wedge(1/n)}$
Where, n = Number of years
This has been derived from Compound Interest formula
i.e. A = P (1 + r / 100)$^{\wedge n}$
Where,
A = Final Amount
P = Starting or Principal Amount
R = Rate of Interest
N = Number of periods (mostly years)

V ANALYSIS & INTERPRETATION

CAGR (Compound Annual Growth Rate) is very a useful measure of measuring growth over multiple time periods generally years. It can be considered as the growth rate that multiplies the initial investment amount to the final investment amount assuming that the investment has been compounding over the time period under consideration.

In other words CAGR is that geometric progression ratio that provides a fixed rate of return over the time period.

Analysis of EPS and PE ratio CAGR for minimum three years after posting a consistently rising EPS. All the data has been adjusted wherein the company if issued Bonus shares or stock split.

Sr. No	Company	Period	No. of years	EPS CAGR (%)	PE ratio CAGR (%)	Share Price CAGR (%)
1	**Amara Raja Batteries**	FY13 - 16	3	19.52%	23.92%	48.11%
2	**Aurobindo Pharma**	FY13 - 18	5	29.41%	10.65%	43.21%
3	**Bajaj Finance**	FY10 - 19	9	42.85%	13.10%	61.56%
4	**Control Print**	FY12 - 15	3	27.43%	40.34%	78.83%
5	**Garware Technical Fibres**	FY12 - 19	7	27.04%	24.48%	58.13%
6	**Divi's Laboratories**	FY11 - 16	5	20.46%	2.78%	23.82%
8	**Greenply Industries**	FY11 - 18	7	25.48%	4.55%	31.19%
9	**Indusind Bank**	FY12 - 18	6	23.22%	10.71%	36.41%
10	**Kajaria Ceramics**	FY11 - 17	6	26.60%	24.84%	58.04%
11	**P I Industries**	FY13 - 17	4	44.83%	10.69%	60.32%
12	**Tata Elxsi**	FY14 - 19	5	31.02%	-2.28%	28.03%
13	**VIP Industries**	FY15 - 19	4	27.94%	14.04%	45.91%
	Average		**5.33**	**28.90%**	**14.82%**	**47.80%**

(a) Interpretation

(i) Average EPS CAGR is placed at 28.90%.

(ii) Average PE ratio CAGR is placed at 14.82%.

(iii) Except for Tata Elxsi all the companies have reported positive PE ratio CAGR. This company is quite dependent on group company Tata Motors which is not doing well from pretty long time.

(iv) Excluding Tata Elxsi average PE ratio CAGR would have been 16.37%

(v) Average share price CAGR is 47.80%

(vi) In case of Amara Raja Batteries and Control Print, PE ratio has much higher contribution than EPS growth which was later corrected when the investor hopes shattered and share price declined due to lower EPS figures in subsequent years.

(vii) In case of Bajaj Finance, Divi's Lab and Greenply Industries investors PE expansion is low as they already are ruling at high PE ratio.

(viii) Higher the duration of consistent growth, higher the share price gains. This is amply indicated by Bajaj Finance, Garware Technical Fibres, Kajaria Ceramics and Aurobindo Pharma.

(ix) Average CAGR share price increase is very high at 47.80% that means doubling of money in one year and 10 months only. As much as 2/3rd of this rise is explained by the financial performance and rest through expansion in PE ratio

VI FINDINGS/ CONCLUSION

Based on the tabulated data there is conclusive evidence that the share price rises rapidly when most of the below mentioned criteria is met.

(a) Minimum three years EPS growth is positive and upward of 25% yearly growth

(b) EPS growth adds double the contribution of PE ratio in explaining the multiple returns in terms of share price.

(c) PE does expand rapidly when the company starts delivering fast growth in Earnings Per Share. But investors are not willing to expand the PE further once the earnings growth stabilizes.

(d) It is not the financial performance alone which explains the rise in share price. Expansion in PE plays a very significant role.

(e) Wide variation in expansion in PE indicates that past good performance explains only partly in investor confidence (PE ratio). Other factors can be management, sector prospects etc.

(f) Dividend can have crucial contribution in PE expansion. Needs further study.

(g) The parameters for most of the companies are near to average which indicates that there is no sector specific bias for investor confidence.

REFERENCES

[1] Chang, Hsu-Ling, Yahn-Shir Chen, Chi-Wei Su, and Ya-Wen Chang, (2008) "The Relationship between Stock Price and EPS: Evidence Based on Taiwan Panel Data." Economics Bulletin, Vol. 3, No. 30 pp. 1-12

[2] Samir K. Barua, V. Raghunathan, Jayanth R. Varma IIM, Ahmedabad, (Year 1994) *Research on Indian Capital Market: A Review*. The article appeared in 'Vikalpa', the journal of the IIM, Ahmedabad, in which the paper was first published (January-March 1994, 19(1). 15-31),

[3] Ansotegui, C. and M. V. Esteban (2002), Cointegration for market forecast in the Spanish stock market, Applied Economics, 34, 843-857.

[4] Beaver, M. McAnally, and C. Stinion (1997), The information content of earnings and prices: A simultaneous equations approach, Journal of Accounting and Economics, 23, 53-81.

[5] Kent Hickman and Glenn H. Petry, Financial Management, Vol. 19, No. 2 (Summer, 1990), pp. 76-87. A Comparison of Stock Price Predictions Using Court Accepted Formulas, Dividend Discount, and P/E Models

[6] Websites of the companies under consideration.

[7] Campbell, J. Y. and R. Shiller (1988), Stock prices, earnings and expected dividends,Journal of Finance, 43, 661-676.

[8] Respective company Annual Reports

[9] Bombay Stock Exchange website www.bseindia.com

[10] National Stock Exchange website www.nseindia.com

[11] Press release by the companies under consideration

Handling Negation for Sentiment Analysis: A Case Study Using Dependency Parse Tree on Amazon Reviews of Kindle

Smita Suresh Daniel[1], Ani Thomas[2], Neelam Sahu[3]
[1,3]Dept. of Computer Application and IT, Dr. C.V. Raman University, Bilaspur (C.G.) India.
[2]Dept. of Computer Applications, Bhilai Institute of Technology, Durg (C.G.) India.

ABSTRACT

The process of sentiment analysis is a task of detecting, extracting and classifying sentiments expressed in texts. It includes the understanding of the meaning of words within the text through natural language processing rules using dependency based parse trees, using grammatical relations among words to model a sentence, and hence to determine words that are affected by negation. This paper presents a framework for identifying. Calculating and representing the presence of negation in textual data using dependency parsers. It includes a list of rules for negative polarity identification and calculation. These negation rules are designed to improve sentiment analysis. This paper is a demonstration of an approach for identifying the scope of negation in a review and its calculation for the Amazon product- Kindle dataset.

Keywords: Negation Identification, Negation Calculation, Sentiment Analysis, Dependency Parsing

I INTRODUCTION

The aim of sentiment analysis is to find out the positive and negative feelings written in a text, but it contain negations that are very frequently used in text that completely change the polarity of words. Negation identification and detecting its scope within a sentence (text) are necessary in finding out the sentiments from a piece of text. Proper addressing of negation identification is an important aspect of sentiment analysis.. Negation identification is a difficult task and its complexity increases, since negation words such as not, nor etc., (syntactic negation) are not the only criterion for negation calculation. The linguistic patterns - prefixes (e.g.,n-, dis-, etc.) or suffixes (e.g., -less) also introduce the context of negation in textual data . Similarly, word intensifiers and diminishers (contextual valence shifter) also change the polarity of sentiments. These valence shifters do not only flip the polarity but also increase or decrease the degree to which a sentimental term is positive or negative [2].

On the other hand, negation does not mean to handle only 'not'. There are words and clauses like; no, not, n't, no way, without, nowhere, never, no longer, by no means, no more, by no means, at no time, etc. [2] which also inverts the meaning of a sentence.. This paper is an effort towards finding a method to handle the syntactic negation for sentiment analysis by not only using the sentiment and its intensity for words but also using the dependencies of these words and their relation within the sentences and sentence structure.

The negation in a text is assessed with the help of diminishes,. intensifiers and negation terms during the process of sentiment analysis.

The paper is structured as follows: Section II presents the related work in the area of sentiment analysis. Section III. Describes the proposed framework for sentiment analysis, and the existing resources used to generate dependencies. Section IV presents an application for negation handling in sentiment analysis. It also explains the basic techniques used in this framework for handling negation. Section V involves an analysis of the technique.

II RELATED WORK

Most researchers in the field of opinion mining have used the lexicons and lists of words, with word as basic unit of expression of emotions in any language. Lexicon based negation i.e., negation introduced by suffix and/or prefix is easily handled with the help of a good lexical resource, i.e., dictionary, ontology, database etc.

However, more emphasis on opinion analysis should be on how these words are joined and correlated with other words to give specific meanings in any language. This interrelationship of words makes up sentences, which is why it is important to emphasis on finding the scope of negation, diminshers or intensifiers. Syntactic and semantic differences make it difficult to interpret the intensity of polarity. When calculating the value of intensity of any sentence, there are always modifiers, which not only change the polarity of other words in the sentence but also affect its intensity. It is also difficult to identify which part of a clause a negation is changing in a sentence. The different methods used for negation identification and how they affect sentiment analysis of text are discussed below.

(a) **Bag of Words-** Bag of words (BOW) is a technique where each word in a document is represented by a separate variable numeric value or weight. It is the most widely used technique for sentiment analysis where negation in a sentence reverses the meanings of the sentence. Words like "not", "never", "no", etc., serve to reverse sentence meaning. Limitation of this method is that it is based on the list of words, and lists in any language can are very vast.

(b) **Contextual Valence Shifter-**Contextual Valence Shifters or modifiers are words, which changes, boosts, enhances and diminishes the meaning. Many researchers have shifted their research on sentiment analysis from BOW to Parts of Speech (POS) especially Verbs, Adjectives and Adverbs. Polarity is associated with every word. However, lots of modifiers are still needed to change or modify the valance associated with words. Negatives, intensifiers or diminishers are examples of contextual shifter. For example :

Negatives: The battery is good versus the battery is not good.

Intensifier: The charger is working well versus the charger is working very well.
Diminishers: It starts versus it hardly starts.

There is a need for relationship finder to define the scope of negation terms .Researchers have tried to define the scope by defining lists of verbs, adjectives and adverbs and defining their relationships for sentiment analysis [6]. Lists of positive and negative terms and a set of lists for modifiers define the scope of these modifiers as n- terms before and after positive or negative terms, although this n remained a constant.

(c) **Semantic Relations-** Semantic relations refer to the relationship between concepts or meanings for example antonym, synonym, homonym etc. It is evident from existing research that semantic relationship is also used for negation identification. It is clear that atomic words, which can provide a misleading polarity for sentences as words can be modified (weakened, strengthened, or reversed) based on lexicons. The use of linguistic structure of sentence for sentiment analysis was proposed in [9], where the polarity of a sentence is dependent upon the polarities of its parts: noun phrases (NP), verb phrases (VP) and individual parts of speech. Negation is handled by defining different intensities of negation words. In other words, the negation of words can change the polarity of an entire sentence or only parts of it [7].

Dependency Analysis on the semantic verb frames of each sentence, and apply a set of rules to each dependency relation to calculate the contextual valence of the whole sentence. A two phase process was proposed in [3] as another way of compositional semantics. In the first phrase they identify the polarity of words where all the words are classified on the basis of the level of their strength in terms of the scope in the sentence and in the second phase inference rules are used, which identify the polarity modification feature.

(d) **Relations and Dependency Based-** The grammatical relationships between the words within a sentence and syntactic dependencies help in extraction of textual relations. For context aware approach for sentiment analysis where the sentiment is evaluated towards a target entity or an event. The scope of words is defined by the clauses or phrases (noun phrase, verb phrase) in the sentence. The sentiment of sentences is understood by the heuristic rules defined to join the clauses. Simple tree based rules can identifying the dependent terms and later use some parts of speech based tools to understand the sentimental behavior of negation [2].

(e) **Analysis of Negation-** For the sentence level sentiment analysis clauses and phrases are required to be understood. They are further divided into sentences and into different types of sentences (simple, complex and compound). The sentence is made more complicated by adding declarative, interrogative, exclamatory and imperative sentences. In order to further complicate the problem as the comparison, contradiction, negation and irony and sarcasm might also be introduced in the sentences. Negation needs to identify its scope. Negation can be local (e.g., not good), or it can involve far distance dependencies (e.g., does not look very good) or the negation of the subject (e.g., no one thinks that it's good). It can also change its roles i.e., instead of negating and it can also intensify (e.g., not only good but amazing). In order to find out the scope of the negation, the sequence of words in the sentence should also be identified. On the whole, it is not only the negation of a word but the negation of the sentence.

The expression of negation within a sentence can be a verb, adverb, suffix or prefix. It might also occur more than once in a sentence and rather than cancelling each other it can give negative meaning, the following section explains the proposed framework.

III FRAMEWORK FOR SENTIMENT ANALYSIS

This section introduces a framework for sentiment analysis and explains how it is handling negation identification, scope of negation and calculation of sentiment on sentence level. The framework uses a combined approach to lexical and syntactic resources for sentiment analysis.Our system use the dependency parsing by the Stanford Parser, described in[9].The output consists in a sentiment score in [1, 0 , -1] for positive, neutral and negative reviews. In order to enhance the accuracy of sentiment evaluation, the text analysis process includes several pre- processing phases like tokenization, Parts of Speech tagging, lemmatization and reduction.

Its main components are briefly described in Sections (a) through (c).

The framework is presented in Figure 1 as shown below.

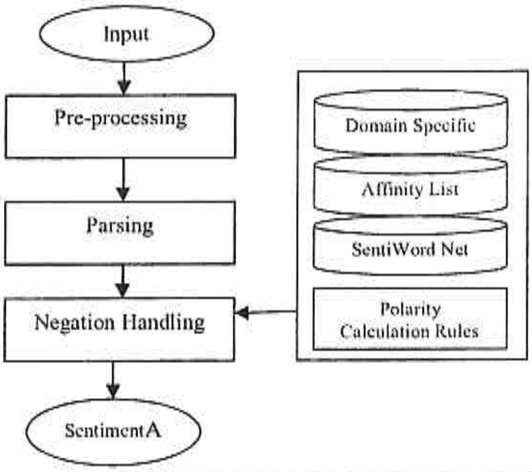


Fig.1 Framework for Sentiment Analysis

Its main components are briefly described in Sections (a) through (c).

(a) **Pre-processor-** The pre-processing phase of the system takes text as input and arranges all the data in required format. It splits data into sentences. Filter all sentences not containing negation triggers and extracts sentences containing negation triggers into a file and provides them to the syntactic parser.

(b) **Syntactic Parser-** The Parts of Speech tagger (POS) tags each word in the sentence. The name entities and phrases involved in a sentence are also identified in syntactic parsing. The Stanford dependency Parser identifies how different words are interacting within a sentence and identifies the syntactic dependencies/relationship within a sentence. The syntactic parser parses each sentence iteratively with all the identified information classifying the sentence as a question, an assertion or a comparison, using the rule of sentence type identification. It is parsed and stored in a file in CoNLLx format which is sent to the negation algorithm.

(c) **Negation Algorithm-** Here we exploit the dependency based parse tree to define a Negation algorithm. An approach used here is not only to invert the polarity of the term followed by the negation word but when a negation occurs in a sentence it is necessary to detect its scope that is the number of following words affected by it, which states that there is no fixed negation window. It depends on the structure of the sentence .The algorithm uses Stanford dependency parser to analyse the grammatical structure of the sentence. A dependency based parse tree is built for each sentence in the review having negation triggers .It explores it to find negation word using Depth First Search rule. If a negation word is found in a tree node, the algorithm inverts the polarity of the following child node and its sub trees. The sentiment score of the negation word is set to 0 as it does not have any sentiment values by itself.

The Negation Algorithm

Let S be an extracted Sentence from a split review having negation triggers.

Let T be the parse tree of Converted to CoNLL format.

Get_Neg_Scope(T) :Set rules to determine the scope of negation using maximum spanning tree and return range for the scope in a sentence.

If POS of trigger is RB/ DT/ JJ/ CC then tree from its immediate governor is used;

If POS of trigger is VB/IN/NN then tree from itself is used;

Span towards right or span left relatively to the position of trigger word till SUB arc, span nothing if there's no SUB arc or right part. (This rule is only applied to root node)

Let Polarity Score () calculate the sentiment score of each term within the scope of negation.

Invert the sentiment score if negation is true

IV USAGE OF FRAMEWORK FOR NEGATION

All the sentences having negation terms and clauses are forwarded from the pre-processor to the syntactic parser with other sentences. However, syntactic parser identifies the negation and POS that are involved in negation with the help of Parser during the syntactic parsing phase and takes care of the negation word and the scope associated with it. . In the negation identification process, the kind of negation i.e.," no one is interested in this new feature", where 'no' is used to determine the interest of one, is also identified. This process also takes care of the negation in conjunction sentences. The Dependency parsers identify the scope of the negation, if it is a single word or a phrase or a clause within a sentence.

These pharses and clause which is within this scope is extracted for polarity calculation by the sentiment analyzer. The polarity is inverted as per table no 1 . The following section explains how sentiment analyzer uses the extracted part of the sentence from the negation algorithm for polarity calculation.

(a) **Sentiment Analyser-** The sentiment analyser uses resources like Sentiwordnet [10] to extract the sentiment oriented words by using dependency relationships within the text. It identifies the semantics involved in a sentence, their meaning, polarity of each term and their intensity. It identifies the positive, negative and neutral words and assigns a value to each term and hence their intensities can also be calculated.

(b) **Polarity Calculator-** The words in a sentence, their meanings, alternative words, polarity of each word and intensity associated with each word are basic elements used by sentiment analyzer for sentiment identification. The polarity of sentence is usually based on the meaning of words. However, the negation changes the meaning of the words and polarity of the sentence. In order to calculate the polarity of a sentence, some rules are defined in Table 1.. Most negation words are classified as adverbs, suffix, prefix or verbs.

The scope of negation will be identified by the dependency tree, which indicates how negation is interacting with other words in the sentence. This dependency parser will identify the scope of the negation - whether it is a single word or a phrase / clause within a sentence. The negation is handled in each phrase accordingly. The intensity of polarity will not exceed (+/-) 1, where + is for positive and – is for negative polarity. The polarity of a sentence is calculated using sentiwordnet.:

The Resulting Intensity = Total positives –total negatives

The positive/negative value of words in the Equation 1 is extracted from the SentiWordNet in order to calculate the polarity of a sentence. The extracted value from the SentiWordNet is reversed during this process if negation is 'True' as presented in Table 1.

Algorithm to Calculate Polarity

```
Function Calculate_Polarity  Returns Polarity
{
 polarity = 0
For Each  Extraction_Of_Sentence
{
get SentiWordNet value of all Adjectives and adverbs
in the phrase
If (Sentence is Marked NEGATION by Syntax
Parser) {
Reverse the SentiWordNet values of related
Adverbs /Adjectives }
            } }
Return polarity
}
```

Table 1
Rules specifying negation

Negation Word/phrase/ Clause	Associated Word/Phrase /Clause	Negation	Result
Negative	Positive	True	Positive
Negative	Positive	False	Negative
Negative	Negative	True	Negative
Negative	Negative	False	Positive

V ANALYSIS

The sentence polarity is calculated on the basis of the parts of a sentence. A sentence may contain either simple POS (Verb, Adverb, Adjectives, etc.) or complex parts of speech (Noun Phrase [Pronoun, Noun] or Verb Phrase [Verb, Noun Phrase], relations of possession, determiner, etc.). The following hierarchy is an example of POS in a complete sentence.

```
(Sentence
(Noun Phrase (Pronoun. Noun))
(Adverbial Phrase (Adverb))
(Verb Phrase (Verb)
 (Sentence
  (Verb Phrase (Verb)
   (Noun Phrase (Noun))) ) ) )
```

Sentiment polarity identification and calculation is a nested process. This process calculates the sentiment of the most inner most level first and then it calculates along with the next higher level, If there is a negation word the polarity will be calculated accordingly. If a negation word is found in a tree node, the algorithm inverts the polarity of its sub trees as they belong to the same clause.

For Example "Charger is never successful at charging." .

The dependency tree structure is as follows.

```
        (Sentence
            (Noun Phrase (charger))
```

```
(Verb Phrase (is)
  (Adverbial Phrase (never))
  (Adjectival Phrase (successful)
    (Prepositional Phrase (at)
      (Noun Phrase (charging))))))
```

Stanford Parser output:-

```
(ROOT
 (S
  (NP (DT
The) (NN
charger))
(VP (VBZ is)
 (ADVP (RB never))
 (ADJP (JJ successful)
  (PP (IN at)
   (S
    (VP (VBG charg'ng))))))
(. .)))
```

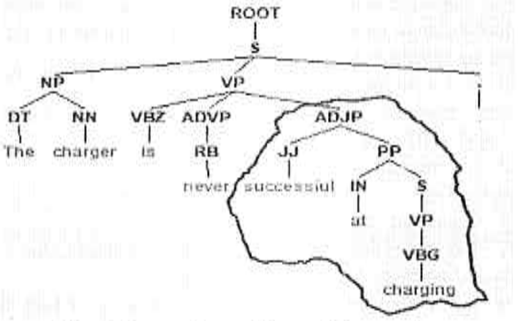


Fig. 2 Dependency Tree of the sentence

The resultant sentence structure is as shown below:

The charger is <SCOPE><NEG>never</NEG> successful at charging .</SCOPE>

Negation'never' is for the scope ('successful at charging') which is positive. This negation of positive phrase is a simple negation, which is presented in Figure no 2.

The confusion matrix for our experiment is shown in Table no 2. There is a 6% increase in accuracy after implementation of the above negation algorithm for our model for predicting sentiments for amazon reviews for the product kindle.

Table 2
Confusion Matrix for the Dataset used

	Predicted Negative	Predicted Neutral	Predicted Positive
Actual Negative	297	0	62
Actual Neutral	0	65	0
Actual Positive	27	0	349

We can calculate the Error estimation for n as Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - t_i)^2 = 0.08$$

where N is the number of review ie 600, x_i is the estimated sentiment score of the review i and t_i is the real score derived from human evaluation .

VI CONCLUSION

Negation has received little attention and its implication on the semantic understanding of sentences. This paper presents an approach for negation identification and calculation using a developed framework for sentiment analysis. These negation rules are designed in order to improve the sentiment text analysis. While, there are still a number of cnallenges to be addressed in the field of negation in sentiment analysis, the developed rules for negation calculation has good improvements in classification accuracy and helps to find the correct polarity.

REFERENCES

[1] Horn, L. R. and Kato, Y. (2000) Introduction: Negation and Polarity. at the Millennium.

[2] Jia , L., Yu, C. and Meng, W. (2009) The effect of negation on sentiment analysis and retrieval effectiveness. In: The 18th ACM conference on Information and knowledge management. Hong Kong, China. ACM, 1827-1830

[3] Choi, Y. and Cardie, C. (2008) Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: The conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii. Association for Computational Linguistics, pp. 793-801.

[4] Kennedy, A. and Inkpen, D. (2005) Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. In: FINEXIN 2005. Ottawa.

[5] Kennedy, A. and Inkpen, D. (2006) Sentiment Classification of Movie Reviews Using Contextual Valence Shifters Intelligence. , Vol. 22 2, pp. 110-125

[6] Subrahmanian, V. S. and Reforgiato, D. (2008) AVA: Adjective- Verb-Adverb Combinations for sentiment Analysis. IEEE Intelligent Systems, Vol. 23, 4, pp. 43-50.

[7] Wiegand, M., et al. (2010) A survey on the role of negation in sentiment analysis. In: The Workshop on Negation and Speculation In Natural Language Processing. Uppsala, Sweden. Association for Computational Linguistics, 60-68

[8] Amna Asmi, Tanko Ishaya , IMMM 2012 . The second International Conference on Advances in Information Mining and Management .

[9] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430

[10] Miller , G.A(1995) WordNet, a lexical database for English Communications of ACM , Vol. 38,11, pp. 39-41

A Survey Based on 3D Fingerprint Matching Technique -Minutiae Cylindrical Code (MCC)

Sunil M. Wanjari[1], S.R.Tandan[2]
[1]St. Vincent Pallotti College of Engineering and Technology, Nagpur (M.S.) India.
[2]Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT

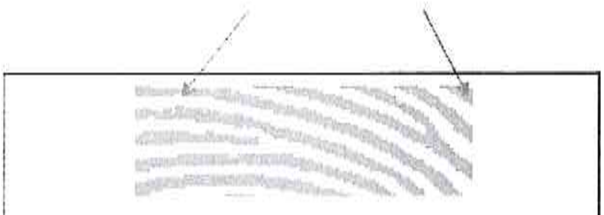
Fingerprints are the most common authenticbiometrics for personal identification, especially for forensic security. A 2D minutiae matching is widely used for fingerprint recognition and can be classified as ridge ending and bifurcation. This paper is survey of 3D Minutiae Matching Technique referred as Minutiae Cylindrical code (MCC). This technique is based on 3D data structures called as Cylinders. In this digital era lots of physical data has been transferred. Based on this rationale, this paper this paper aims to improve the fingerprint matching performance. In the current state of art liner solution, by using Minutiae cylindrical Code Technique. False acceptance rate (FAR), False rejection rate(FRR), Execution Time, Matching Time, Enrollment Time is being going to improved.

Keywords—Fingerprint Matching; MCC; FAR; FRR; Matching Time; Execution Time; Enrolment Time.

I INTRODUCTION

Fingerprint recognition is an intriguing pattern recognition problem that has been studied for more than 40 years. Although very effective solutions are currently available, fingerprint recognition cannot be considered a fully solved problem, and the design of accurate, interoperable, and computationally light algorithms is still an open issue.

Fingerprints are the patterns formed on the epidermis of the fingertip. The fingerprints are of three types: arch, loop and whorl. The fingerprint is composed of ridges and valleys. The interleaved pattern of ridges and valleys are the most evident structural characteristic of a fingerprint. There are two main fingerprint -a) Global Ridge Pattern b) Local Ridge Detail.



(a) Ridge-ending (b) Bifurcation

Fig. 1 Fingerprint Ridge ending and Bifurcation

A fingerprint is a smoothly owing pattern of alternating ridges and valleys. The ridges do not flow continuously but rather display various types of imperfections known as minutiae (minor details in fingerprints). At the time of enrollment in a fingerprint system, important minutiae information (typically, positions of ridge endings and bifurcations, and the associated orientations) is extracted and stored in the database in the form of a template. Fingerprint matching is accomplished by comparing the minutiae distribution of two fingerprints via sophisticated point pattern matching techniques. Minutiae have been studied extensively in the forensic literature specifically in the context of fingerprint individuality models.

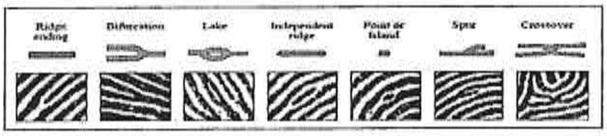


Fig. 2 Seven most common Types of minutiae

The advancement of technology has given contributions to the rapid growth of the use of digital data. In this digital era, lots of physical data have been transformed into the digital ones. One example of the use of digital data is the digital biometric fingerprint data on the Electronic Identity Card (KTP-el).To identify a person, fingerprint matching can be used. There are 3 approaches in fingerprint matching:

(i) Correlation-based-matching
(ii) Minutiae-based-matching
(iii) Ridge feature-based matching.

One of popular technique is minutia-based fingerprint matching.

II FINGERPRINT MACTCHING

There are 2 approaches in fingerprint matching: correlation based matching, minutiae-based matching, and ridge feature based matching. In this we focus on Minutia Cylinder-Code.

(a) Minutia-Based Matching - A minutia is either a ridge bifurcation or a ridge ending. Ridge bifurcation is a point where a ridge splits into two ridges; meanwhile ridge ending is a point where a ridge meets a dead-end. A minutia is represented by its position, angle, and type. In general, there are two algorithms in minutiae-based matching.

(i) Nearest neighbor Fixed Radius

In nearest neighbor-based algorithm, the neighbourhood of a given minutia is defined as K nearest minutiae. Thus, the number of neighbors in this algorithm is fixed so fingerprint matching can be performed fast efficient. Disadvantage of this algorithm is it is intolerant to missing and spurious minutiae.

In fixed radius-based algorithm, the neighbourhood of a given minutia is defined as all minutiae that its distance is within a circle radius R. The number of neighbours in this algorithm can vary, depends on the density of a minutia. Thus, fingerprint matching with this algorithm is harder than the former one. However, this algorithm is more tolerant to missing and spurious minutiae.

(b) Minutia Cylinder-Code

Minutia Cylinder-Code (MCC) is one of the best performing algorithms in minutia-based fingerprint matching. It combines the advantages of both nearest neighbour-based and fixed radius-based algorithms without having their drawbacks. It has an efficient performance as nearest neighbour-based algorithms and high tolerance to minutiae deformations as fixed radius-based algorithms.

MCC aims to achieve high accuracy while maintaining interoperability with other algorithms by using standard features in minutiae. It uses the position and direction of the minutiae but not the type and quality. It is due to the type is not a robust feature and the quality is not semantically clear in the standards.

The local minutiae representation introduced in this paper is based on 3D data structures (called cylinders), built from invariant distances and angles in a neighborhood of each minutia. Four global-scoring techniques are then proposed to combine local similarities into a unique global score denoting the overall similarity between two fingerprints. The main advantages of the new method, called Minutia Cylinder-Code (MCC), are:

MCC is a fixed-radius approach and therefore it tolerates missing and spurious minutiae better than nearest neighbor-based approaches.

Unlike traditional fixed-radius techniques, MCC relies on a fixed-length invariant coding for each minutia and this makes the computation of local structure similarities very simple.

Border problems are gracefully managed without extra burden in the coding and matching stages.

Local distortion and small feature extraction errors are tolerated thanks to the adoption of smoothed functions (i.e., error tolerant) in the coding stage.

MCC effectively deals with noisy fingerprint regions where minutiae extraction algorithms tend to place numerous spurious minutiae (close to each other); this is made possible by the saturation effect produced by a limiting function.

The bit-oriented coding (one of the possible implementations of MCC) makes cylinder matching extremely simple and fast, reducing it to a sequence of bit-wise operations (e.g., AND, XOR) that can be efficiently implemented even on very simple CPUs.

III LITERATURE SURVEY

In Minutiae based 2-D approach the ridge features called minutiae are extracted and stored in a template for matching. It is invariant to translation, rotation and scale changes. It is however error prone in low quality images. The minutiae based approach is applied. Usually before minutiae extraction, image preprocessing is performed. Before applying minutiae-based approach we should do the preprocessing and extraction stage. Fingerprint enhancements techniques are used to reduce the noise and improve the clarity of ridges against valleys.

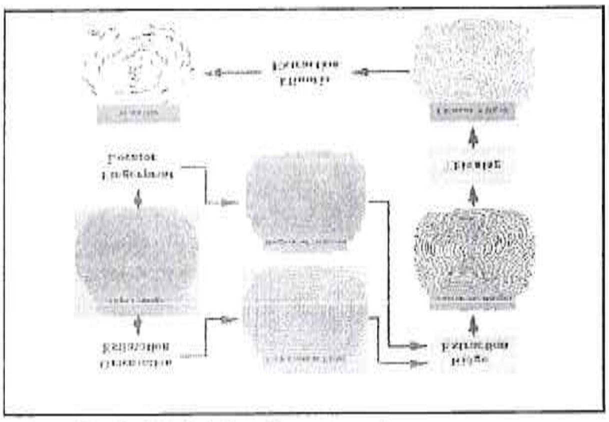


Fig.3 Typical 2-D Minutiae extraction process

In Minutiae Cylinder code (MCC) based 3-D Approach MCC representation associates a local structure to each minutia. This structure encodes spatial and directional relationships between the minutia and its (fixed-radius) neighborhood and can be conveniently represented as a cylinder whose base and height are related to the spatial and directional information, respectively.

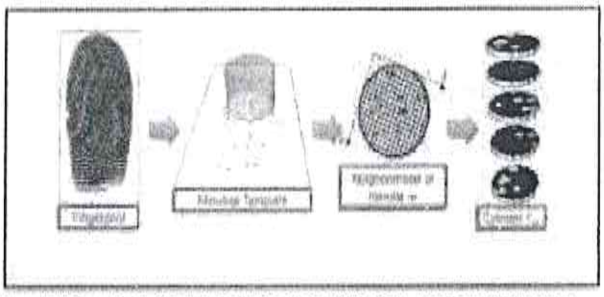


Fig. 4 Minutiae Cylindrical code (MCC) Representation.

Dr. Anil K. Jain et al. (2003) thoroughly explains allthe aspects of Fingerprints and Fingerprint Recognition in their book "Handbook of Fingerprint recognition".

David Maltoni (2005) presented Fingerprint matching techniques in his paper " A tutorial on fingerprint Recognition". This tutorial introduces fingerprint recognition systems and their main components: sensing, feature extraction and matching. The basic technologies are surveyed and some state-of-the-art algorithms are discussed.

Raffaele Cappelli et al. (2010) introduce the MinutiaCylinder-Code (MCC): a novel representation based on 3D data structures (called cylinders), built from minutiae distances and angles. The cylinders can be created starting from a subset of the mandatory features (minutiae position and direction) defined by standards like ISO/IEC 19794-2 (2005). They advice that some simple but very effective metrics can be defined to compute local similarities and to consolidate them into a global score. Extensive experiments over FVC2006 databases prove the superiority of MCC with respect to three well-known techniques and demonstrate the feasibility of obtaining a very effective (and interoperable) fingerprint recognition implementation for light architectures.

Matteo Ferrara et al. (2011) proposes a new hash-based indexing method to speed up fingerprint identification in large databases. A Locality-Sensitive Hashing (LSH) scheme has been designed relying on Minutiae Cylinder-Code (MCC), which proved to be very effective in mapping a minutiae-based representation (position/angle only) into a set of fixed-length transformation-invariant binary vectors. A novel search algorithm has been designed thanks to the derivation of a numerical approximation for the similarity between MCC vectors. Extensive experimentations have been carried out to compare the proposed approach against 15 existing methods over all the benchmarks typically used for fingerprint indexing. In spite of the smaller set of features used (top performing methods usually combine more features), the new approach outperforms existing ones in almost all of the cases.

David Maltoni et al. (2012) propose a two factorprotection scheme that makes P-MCC templates revocable to avoid that MCC templates can disclose sensitive information about position and angle of minutiae, a protected MCC representation was recently introduced (called P-MCC).Inspite of a satisfactory level of accuracy and reversibility, P-MCC templates cannot be revoked. [5]

M. Hamed Izadi et al. (2012) propose an alternativemethod to estimate the cylinder quality measures directly from fingerprint quality maps, in particular ridge clarity maps, by taking into account the number of involving minutiae as well. Integration of MCC with the proposed cylinder quality measures was evaluated through experiments on the latent fingerprint database NIST SD27. These experiments show clear improvements in the identification performance of latent fingerprints of ugly quality.

Matteo Ferrara et al. (2012) propose a novelprotection technique for Minutia Cylinder-Code (MCC), which is a well-known local minutiae representation. A sophisticate algorithm is designed to reverse MCC (i.e., recovering original minutiae positions and angles). Systematic experimentations show that the new approach compares favorably with state-of-the-art methods in terms of accuracy and, at the same time, provides a good protection of minutiae information and is robust against masquerade attacks.

A. Pasha Hosseinbor et al. (2017) propose a minutia-based fingerprint matching algorithm that employs iterative global alignment on two minutia sets. The matcher considers all possible minutia pairings and iteratively aligns the two sets until the number of minutia pairs does not exceed the maximum number of allowable one to-one pairings. The optimal alignment parameters are derived analytically via linear least squares.

WajihUllah Baig, et al. (2018) present a modificationto the underlying information of the MCC descriptor and show that using different features, the accuracy of matching is highly affected by such changes. MCC originally being a minutia only descriptor is transformed into a texture descriptor. The transformation is from minutiae angular information to orientation, frequency and energy information using Short Time Fourier Transform (STFT) analysis. The minutia cylinder codes are converted to minutiae texture cylinder codes (MTCC). Based on a fixed set of parameters, the proposed changes to MCC show improved performance on FVC 2002 and 2004 data sets and surpass the traditional MCC performance.

This Paper proposes a 3D fingerprint identification pre-treatment algorithm in Matlab. Based on Matlab, this article provides an algorithm implementation, and an improved method. The results of each fingerprint picture processing module, mainly including image segmentation which could be separated, obtained a fingerprint image from a background area. Image filtering, removing burr, cavity management and binarization processing (with the thought of self-adapted local threshold binariztion) which make the fingerprint image clearer, eliminate unnecessary noises and are beneficial to further identification. To thin the image, we first, adopt the quick thinning algorithm to handle the preliminary thinning other languages, including C, C++, C#, Java, Fortran and Python. The data doesn't have to be in structured form or uniform because each instance of data is taken care by separate process on a different node.

The streakline after thinning has a certain width, and secondly, the advanced one-pass thinning algorithm (OPTA) is adopted for use the fingerprint image that after preliminary thinning: this makes all areas, except the bifurcation point, remain a single-pixel wide, correcting the streakline that has been thinned by advanced OPTA. Then we get a clear fingerprint picture, extract the fingerprint feature point (spurious minutiae) from this picture; this feature point contains a lot of false features that take a lot of time and influences the matching precision. In this paper, the author adopt eliminating the false features by edge and distance, lessening the false feature points by approximately a third, and then next extract reliable information of the feature points and store in the book building template. When matching a fingerprint, we get clear fingerprint image using the same method, and build a contrast template; at last, we compare the contrast template with book building template and then get ideal results. Based on Matlab, with this method we are unable to do the simulation step-by step with the fingerprint identification pre-treatment algorithm, but also see the result of image processing algorithm intuitively, which cooperates with the algorithm research effectively. Experimental results show that with this algorithm, which is on the basis of Matlab, the processing result is more ideal, and this method is not only simple and quick, but also has a high precision, and satisfy the identification applicability.

IV CONCLUSION

This survey paper gives the detail survey of the work carried out in 3D fingerprint recognition in biometric security or personal identification. MCC relies on a robust discretization of the neighborhood of each minutia into a 3D cell-based structure named cylinder. Simple but effective techniques for the computation and consolidation of cylinder similarities are provided to determine the global similarity between two fingerprints.

It is found after analysis that there is need of some constructive, robust secured method of fingerprint recognition in adverse situation where we may have partial images or environmentally affected images which we would be trying in future course of my dissertation work.

REFERENCES

[1] Baig Wajih Ullah, Umar Munir, Waqas Ellahi, Adeel Ejaz, Kashif Sardar, (2018)," Minutia Texture Cylin der Codes for fingerprint Matching", arXiv:1807.02251v1 [cs.CV] 6 Jul 2018,pp 1-12.

[2] Biometric System Laboratory, DISI University of Bologna. ITALY, Web site : http://bias.csr.unibo.it

[3] Cappelli R., Ferrara M. and Maltoni D. (2011) , "Fingerprint Indexing Based on Minutia Cylinder-Cod e", IEEE Transactions on Pattern Analysis Machine Intelligence, vol.33,No.,5,pp 1051-1057 , May 2011, pp 1051-1057

[4] Dario Maio, Cappelli R., Maltoni Davide, Wayman James L., and Jain Anil K., (2004) "FVC 2004: Third Fingerprint Verification Competition", Internationa l Conference on Biometric Authentication, ICBA 2004, pp 1-7.

[5] Ferrara M. and Maltoni D. (2010), " Minutia Cylinder - Code: a new representation and matching technique for fingerprint recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no.12, pp.2128-2141, December 2010, pp 2128-2141.

[6] Ferrara Matteo, Maltoni, Davide Cappelli ,Raffaele

[7] (2012),"Noninvertible Minutia Cylinder-Code Representation",IEEE Transactions on Information Forensics and Security, Vol 7, No. 6, December 2012,pp 1727-1737.

Cybercrime: A Major Problem

Uvika Kujur[1], Swati Bareth[2], Jayshree bajaj[3]
[1,2,3]Dept. of Comp. Sc. & App., Loyola College, Kunkuri (C.G.) India.

ABSTRACT

Nowadays, Cybercrime has caused a lots of deface to an individual, organization and even the government sectors. Many cybercrime detection and classification methods have overcome with various levels of success in order to prevent, protect and authorized data from Cyber-attacks. Cybercrime is that activity done by human being knowingly or unknowingly to ruin organizations network, stealing important data and documents, hacking bank accounts details, transferring money to their own and so own.This paper describes about the common areas where cybercrime usually occurs and also classify various types of cybercrimes and laws to prevent and protect from cyber-attacks. This paper also deals about the main causes of cybercrimes occurred.

Keywords: cybercrime, cyber-attacks, causes, protect, cyber laws

I INTRODUCTION

Cybercrime is the latest and the most complicated problem in cyber world. The term crime is denoted an unlawful act which is punishable by a state (Ramdinmawaii et al. 2014). Crime is also called as an offense or a Criminal offense. Cyber-criminal use internet and computer technology to hack user's personal Computers, Smartphone data, personal details fromsocial media, national secrets etc. In general. We can define computer as the machine that oar stores and manipulate or process information or instruction, instructed by the users.

The term Cybercrime can be defined as an act of committed or omitted in violation of a law forbidding or commanding it and for which punishment is imposed upon conviction (Saini H. et. al., 2012). The term "Cyber law" doesn't have a fixed definition, but we can defined it as the law that governs the Cyberspace. Cybercrimes, Digital and electronic signatures data protections etc, are comprehended by the cyber law (Saini H. et. al., (2012).The UN'S general assembly recommended the first IT act of India which was based on the "united nations Model law on Electronic commerce "Model(Saini H. et. al., 2012). Cyber law is generic term which refers to all the legal and regulatory aspects of internet. It is a constantly evolving process, if the internet grows, numerous legal issues also arises. Cybercrime may be used of an instrument for an illegal ends of activity such as online gambling, financial crimes, cyber stalking, email spoofing, sales of illegal articles, forgery, committing fraud, violating privacy etc.

II CYBERCRIME

(a) History of Cybercrime

The first Cybercrime was recorded in the year 1820. The ancient type of computer has been in Japan, China, and India. Since 3500 B.C. The era of modern computer began with the analytical engine of Charles Babbage.

US $ 10 million were fraudulently transferred out of the bank and into a bank account in Switzerland. A Russian hacker group led by Vladimir Kevin, a renowned hacker, perpetrated the attack. The group compromised the bank's security systems. Vladimir was allegedly using his office computer at AO Saturn, a computer firm in St. Petersburg Russia, to break into Citibank computers. He was finally arrested on Heathrow airport on his way to Switzerland (Choudhury R. R. et. al., 2013).

The Cybercrime is enlarged from Morris worm to the ransom ware. Several countries like America, China, Germany, Britain including India are working to stop such Cybercrimes and attacks, but these attacks are frequently changing and influencing our nations. Here are lists of some types of attacks given below and they are as:

Table 1

(b) Evolution of Cybercrime

Years	Types of Attacks
1971	A phone phreak
1995	Macro-viruses
1997	Cybercrimes and viruses initiated, that includes Morris code worm and other.
1999	Melissa viruses
2002	Shadow crew's website
2004	Malicious code, Trojans, Advanced worm etc.
2007	Identifying thief, Phishing etc.
2010	DNS Attack, Rise of Botnets, SQL attacks etc.
2013	Social Engineering, DOS Attacks, BotNets, Malicious Emails, Ransomware attack etc.
Present	Banking Malware, Keylogger, Bitcoin wallet, Phone hijacking, Anroid hack, Cyber warfare etc.

(c) Types of Cybercrimes

There are many types of cybercrimes and they have been discussed given below

(i) Email spoofing
(ii) Salami attack
(iii) Worm/virus attacks
(iv) Web jacking
(v) Phishing
(vi) Forgery
(vii) Online Gambling

(i) **Email Spoofing:** E-mail spoofing basically means sending an email from a source while it appears to have been sent from another source. These tactics are used in phishing and spam campaigns mostly people think that the email has been sent by any legal source and they used to open that email. The mail goal ofemail spoofing is to get recipients to open and possibly even respond to a solicitation. Financial crimes are commonly committed through E-mail spoofing.

(ii) **Salami attacks:** A Salami attack is also known as salami slicing. It is often used to carry out illegal activities. Attackers uses customer online database information like bank details, etc. Attackers reduces very few amounts from every account over a period of time and the customer remains unaware of thisslicing and hence no complain is filled.

(iii) **Virus/worm attacks:** Viruses are programs that attach themselves to a computer or a file and then circulate themselves to other files and to other computers on a network. They usually affect the data on a computer, either by altering or deleting it. Worms, unlike viruses do not need the host to attach themselves to (Dashora K., 2011). They merely make functional copies of themselves and do this repeatedly till they eat up all the available space on a computer's memory. The world's most famous worm was the internet worm let loose on the internet by Robert Morris sometime in 1988.

(iv) **Web Jacking:** This term Web jacking is derived from the term hi jacking. In these kinds of offensethe attacker creates a fake websites and when the victims opens the link a new page appears with the message and they need to clicks the link that looks real he will redirected to a fake page(Saini H. et. al., 2012). Hence these types attacks are done to get approach or to get access and control the cite of another. The attacker may also change the information of the victim's webpage.

(v) **Phishing:** In Phishing, the attacker's tries to gain information such as login information or passwords, details of credit card, account's information by simulate as a reputable individual or entity in several communication channels or in emails. Phishing e-mails are likely to contain hyperlinks to the sites containing malwares.

(vi) **Forgery:** It means making of false documents, signature, currency, revenue stamp etc.

(vii) **Online Gambling:** It is offered by thousands of websites that have servers hosted abroad. Theses websites are the one of the most important sites for money launderers (Ramdinmawaii E. et. al., 2014).

III CAUSES OF CYBERCRIME

(i) **Loss of evidence** – Loss of evidence is a very general and common problem as all the data is frequently destroyed. For the collection of data outside the territorial extent also paralyzes the system of Cybercrime Investigation.

(ii) **Easy to Access** – The problem encountered in guarding a computer system from unauthorized access is that there is every possibilities of breach not due to human error but due tocomplex technology. (Dashora K., 2011). By secretly implanted logic bomb, key loggers that can steal access code, advanced voice recorders, retina imagers etc. that can fool biometric systems and bypass firewalls can be utilized to get pass many a security system.

(iii) **Capacity to store data in comparatively small space** – The computer has a unique characteristic of storing data in a very small space. This allows for much easier access or removal of information through either physical or virtual media (Choudhury R. R. et. al., 2013).

(iv) **Negligence** – Negligence is one of the characteristics in human conduct so there may be a possibly that secure and protecting the computer system we may make negligence which provides a Cyber-criminals the access and control over the computer system.

(v) **Complex** – The computers works operating system and this operating system are programmed of millions of codes. The human mind is imperfect so, they can do many mistakes or errors at several stages.

(a) **Laws of Cybercrime** - All laws aren't the same in many countries especially when it comes to Cybercrimes. For different countries have specific laws governing problems such as Cybercrimes. For example, insome countries such as India accepted The Information Act which was passed and enforces in 2000 on Electronic Commerce by the United Nations Commission on Trade Law. However, the act states that it will legalize e-commerce and supplementary modify the Indian Penal Code 1860, the act 1872, the Banker's Book Evidence Act 1891 and the Reserve Bank of India Act 1934. The Information Technology Act deals with the various Cybercrimes. From this Act, The important sections are: Section 43,65,66,67. Section 43 which explain and enforces the unlawful access, transferring virus outbreaks causes harm for DOA. Section 67 of information Technology Act, 2000 deals with obscenity and pornographic content on internet.

(b) **Cyber Laws in India:** Cyber Crimes, in India are registered under three main heads, The IT Act, The IPC (Indian Penal Code) and SLL (State Level Legislations) (https://www.jagranjosh.com/general knowledge/what-is-cyber-crime-and-how-it-is-increasing-day-by-day-1479450153-1).

(c) **Cases of Cyber Laws under IT Act:**

(i) Tampering with computer source documents – Sec. 65

(ii) Hacking with computer systems, Data alteration – Sec.66

(iii) Publishing obscene information – Sec. 67

(iv) Breach of Confidentiality and Privacy – Sec. 72

(v) Publishing false digital signature certificates – Sec.73

(d) **Cases of Cyber Laws under IPC and special Laws:**

(i) Sending threatening messages by email – Sec. 404 IPC

(ii) Email abuse – Sec.500 IPC

(iii) Web-jacking – Sec.383 IPC

(iv) Forgery of Electronic records – Sec 463 IPC

(v) Email spoofing – Sec.463 IPC

(vi) Bogus websites, Cyber Frauds – Sec 420 IPC

(e) **Cyber Crime under special cells:**

(i) Online sale of Arms Act

(ii) Online sale of Drugs under Narcotic Drugs and Psychotropic Substances Act.

- **Section 65**-Tempering with the computers source documents. Whoever intentionally or knowingly destroy, conceal or change any computer's source code that is used for a computer, computer program and computer system or computer network (Sarmah A. et. al., 2017).**Punishment:** Any person who involves in such crimes could be sentenced upto 3 years imprisonment or with a fine of Rs. 2 lakhs or with both.
- **Section 66**- Hacking with Computer system, data alteration etc. Whoever with the purpose or intention to cause any loss, damage or to destroy, delete or to alter any information that resides in a public or any person's computer. Diminish its utility, values or affects it injuriously by any means, commits hacking(https://cybercrimelawyer.wordpress.com/category/information-technology-act-section-

65/https:/cybercrimelawyer.wordpress.com/category/information-technology-act-section-65/). **Punishment:** Any person who involves in such crimes could be sentenced upto 3 years imprisonment, or with a fine that may extend upto 2 lakhs rupees, or both.

- **Section 66C-** Identity theft using of one's digital or electronic signature or one's password or any other unique identification of any person is a crime. **Punishment:** any person who involves in such crimes could be sentenced either with a description for a term which may extend upto 3 years of imprisonment along with a fine that may extend upto rupee 1 lakh.

Here are the some lists of Cybercrimes and Cyber Laws under the following section:

Table 2

Cyber Attacks	Laws
Un-authorized access to protected system.	Section 70
Penalty for misrepresentation.	Section 71
Breach of confidentiality and privacy.	Section 72
Publishing false digital signature certificates.	Section 73
Publication for fraudulent purpose.	Section 74
Act to apply for contravention or offence that is committed outside India.	Section 75
Compensation, confiscation or penalties for not to interfere with other punishment.	Section 77
Compounding of Offences.	Section 77A
Offences by Companies.	Section 85
Sending threatening messages by e-mail.	Section 503 IPC
Sending defamatory messages by e-mail.	Section 499 IPC
Bogus websites, Cyber Frauds.	Section 420 IPC
E-mail Spoofing.	Section 463 IPC
E-mail Abuse.	Section 500 IPC
Criminal intimidation by anonymous communications.	Section 507 IPC
Online sale of Drugs.	NDPS Act
Online sale of Arms.	Arm Act

IV PREVENTION OF CYBERCRIME

(i) To prevent cyber stalking avoid disclosing any information pertaining to one self.

(ii) Never send your credit card number to any site that is not secured, to guard against frauds.

(iii) Always avoid sending any photograph online particularly to strangers and chat friends as there have been incidents of misuse of photographs.

(iv) It is better to use a security program that gives control over the cookies and send information back to the sites as leaving the cookies unguarded might prove fatal.

(v) Use of firewalls may be beneficial.

(vi) Always keep back up data's so that one may not suffer data loss in case of virus contamination.
(vii) Use strong biometrics as a Password/locker.
(viii) Secure your mobile device.
(ix) Avoid suspicious E-mail.
(x) Protect your identity online.
(xi) Call the right person for help.
(xii) Check your accounts and your credit reports regularly.

V CONCLUSION

From this research paper it has been found that there are several ways and means through which an individual can enact crimes are an offense and are punishable by law (Ramdinmawaii E. et. al., 2014). In this paper we have discussed about the types of cybercrimes, laws of cybercrimes in India, the causes of cybercrimes and how to prevent or avoid cybercrimes. The solution to such crimes can't be simply based on the technology. These technologies can just be one such weapon to track and put a break to such activities to some extent. The way to overcome these crimes can broadly be classified into three categories: Cyber Laws (referred as Cyber laws), Education and policy making. All the above ways to handle cybercrimes either are having very less significant work or having in many of the countries. This lack of work requires to improve the existing work or to set new paradigms for controlling the cyber-attacks.

REFERENCES

[1] Choudhury R. R. et. al., (2013), Cyber Crimes-challenges and Solutions, International Journal of Computer Science and Information Technology, Volumes: 04, PP. 729-732.

[2] Dashora K., (2011), Cyber Crime in the Cociety: Problems and Preventions, Journal of Alteranative Perspectives in the Social Sciences, Volume: 03, Issue: 01, PP. 240-259.

[3] Ramdinmawaii E. et. al., (2014), A Study on Cyber-crime and Cyber Criminals: A Global problem, International Journal of Web Technology, Volume: 03, PP. 172-179.

[4] Saini H. et. al., (2012), Cyber-Crimes and their Impacts: A Review, International Journal of Engineering Research and Applications, Volume: 02, Issues: 02 PP. 202-209.

[5] Sarmah A. et. al., (2017), A study on Cyber-crime and Cyber Law's of India, International Research Journal of Engineering and Technology, Volume: 04 Issue: 06 PP. 1633-1640.

[6] https://www.jagranjosh.com/general-knowledge/what-is-cyber-crime-and-how-it-is-increasing-day-by-day-1479450153-1

[7] https://cybercrimelawyer.wordpress.com/category/information-technology-act-section-65/https:/cybercrimelawyer.wordpress.com/category/information-technology-act-section-65/

Proceedings
National Conference
On
Advances in Computer Science and Information Technology (NCACSIT)

22nd & 23rd February 2019

Chief Patron & Patron

Chief Patron
Shri Santosh Choubey
(Hon'ble Chancellor)
Dr. C. V. Raman University

Patron
Dr. R. P. Dubey
(Hon'ble Vice Chancellor)
Dr. C. V. Raman University

National Advisory Committee

Prof. V.K. Verma
Hon'ble Chancellor, CVRU,
Vaishali, Bihar

Prof A.K. Gwal
Hon'ble Vice Chancellor,
RNTU, Bhopal

Prof. R.N. Yadava
Hon'ble Adviser, Aisect
Group of Universities, Bhopal

Prof. Amitabh Saxena
Hon'ble Vice Chancellor,
CVRU, Khandwa

Prof. S.K. Shrivastava
Hon'ble Vice Chancellor,
Aisect University, Hazaribagh

Prof. D.S. Pandey
Hon'ble Vice Chancellor,
CVRU, Patna

Dr. Vijay Singh
Registrar, RNTU, Bhopal

Dr. Sitesh Sinha
Registrar, CVRU, Patna

Dr. Munish Govind
Registrar, Aisect University,
Hazaribagh

Dr. Ravi Chaturvedi
Registrar, CVRU, Khandwa

Local Advisory Committee

Dr. P. K. Naik
Pro-Vice Chancellor,
CVRU, Bilaspur

Shri Gaurav Shukla
Registrar,
CVRU, Bilaspur

Dr. Manish Upadhyaya
Principal,
CVRU, Bilaspur

Organizing Secretary
Dr. S. K. Tiwari
Dean of IT, CVRU

Local Organizing Committee

Dr. Ragini Shukla
Dr. Neelam Sahu
Mr. Sunil Sharma
Mr. Vineet Awasthi
Ms. Shagufta Farzana
Mr. Ayush Agrawal
Mr. Vikas Pandey
Mr. Suraj Keshri
Mr. Anurag Rao
Mr. Vinay Soni
Ms. Subhangi Pathak
Dr. Vaibhav Sharma
Mr. Amit Dewangan
Mr. Praveen Choksey
Mrs. Yukti Kesherwani
Mr. Rahul Pandey
Mr. Ayaz Ahmed
Mr. Somesh Mishra
Mr. Bisahu Sahu
Mr. Vivekanand Verma
Mr. Jitendra Gupta
Mr. Amit Kashyap

Keynote Speakers

Dr. Konda Srinivas
CMR Technical Campus
Hyderabad (T.G.), India

Dr. A.K. Saxena
Guru Ghasidas Vishwavidyalaya
Bilaspur (C.G.), India

Dr. Radha Krishna Rambola
SVKM'S NMIMS University (MH), India

Dr. H. S. Hota
Atal Bihari Vajpai University
Bilaspur (C.G.), India

Dr. R.R. Janghel
National Institute of Technology
Raipur (C.G.), India

Mr. Santosh Soni
Guru Ghasidas Vishwavidyalaya
Bilaspur (C.G.), India

Conveners

Dr. Rohit Miri
HoD, Dept. of CSE

Dr. Abhinav Shukla
HoD, Dept. of IT

Co-Conveners

Dr. S. R. Tandan
Associate Professor, Dept. of CSE

Dr. Akhilcsh Kumar Shrivas
Assistant Professor, Dept. of IT

Science Technology & Management Journal of RNTU

In This Special Issue

Vol VIII/Issue XVI Special Issue April- 2019 ISSN: 2278-4187