Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare

Ayman Mir
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
ayman.mir@spit.ac.in

Sudhir N. Dhage
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
sudhir_dhage@spit.ac.in

***Abstract*—Healthcare domain is a very prominent research field with rapid technological advancement and increasing data day by day. In order to deal with large volume of healthcare data we need Big Data Analytics which is an emerging approach in Healthcare domain. Millions of patients seek treatments around the globe with various procedure. Analyzing the trends in treatment of patients for diagnosis of a particular disease will help in making informed and efficient decisions to improve the overall quality of healthcare. Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. This paper aims at building a classifier model using WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. The research hopes to recommend the best algorithm based on efficient performance result for the prediction of diabetes disease. Experimental results of each algorithm used on the dataset was evaluated. It is observed that Support Vector Machine performed best in prediction of the disease having maximum accuracy.**

***Keywords*—Healthcare, Big Data, Machine Learning, Disease Prediction, Naive Bayes, Support Vector Machine, Random Forest, Simple CART**

I. Introduction

Health is always a priority even before technology exists. Healthcare domain provides a lot of scope for research as it has tremendously evolved. There is a necessity of upgrading the existing Healthcare technology by embracing digitization of medical information, both in terms of patient provided data as well as medical results generated from advanced equipment. A common outcome of this information revolution is that we are faced with the daunting task of interpreting and understanding the huge data gathered. Since there is huge amount of data hence Big Data Analytics comes to rescue.

Big Data Analytics is the emerging approach which is dealing with many sectors thus applying it to the healthcare domain will result in improved healthcare services. While big data approach in healthcare is still in developing phase, it is clear that the designing of a healthcare platform for a better tomorrow will be of great help in enhancing the quality delivery of healthcare.

Machine learning is another emerging and trending approach which closely works to solve the real time problems. Currently in the health care domain various data mining methods are used to find interesting pattern of disease using statistical medical data with the help of machine learning algorithms. Machine learning approach can be applied for prediction of diseases and provide automated diagnosis under the validation of professional doctor.

The paper focuses on the prediction of the critical disease diabetes. Diabetes is ranked as the fifth deadliest disease world wide. The following figure Fig 1 represents the WHO report for Diabetes.

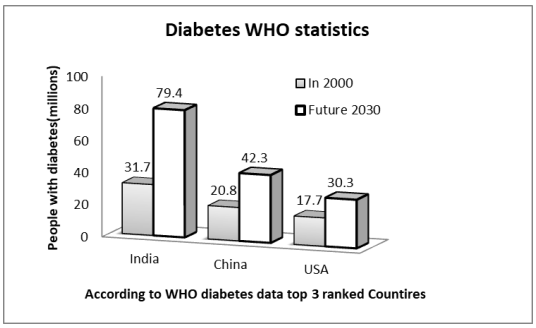

Fig. 1. WHO Diabetes statistics

According to the WHO report India is ranked No. 1 with 31.7 million no. of diabetic patient in 2000 and is likely to increase upto 79.4 million. Since there is a huge risky likelihood of increasing no. of diabetic patient where accurate diagnosis will be the need of the hour. Hence this statistic provides the motivation to carry out the research to find the best performing algorithm. In this paper we will make use of four supervised learning algorithm and perform comparison on their result to recommend the best approach to predict the diabetes which will be helpful for making informed decisions accurately.

The remainder of the paper is organized as follows :

Section II is Literature Survey describing the already existing work. Section III is about Methodology which highlights the dataset being worked on and the proposed methodology. Section IV describes the Experimental results that are obtained after building classifiers through WEKA. Section V is about Results and Discussions which discusses about the performance evaluation of all classifiers. Section VI is about Conclusion which concludes the overall results.

978-1-5386-5257-2/18/$31.00

II. Literature Survey

This section reviews the existing recent literature work and provides insights in understanding the challenges and tries to find the gaps in existing approaches.

Various computing techniques are applied in Healthcare domain. The focus of literature survey here is on the use of Big Data Analytics and Machine Learning in healthcare domain . In order to make a smart learning Healthcare system there are unresolved analytical data challenges[1]. Through Big data analytics the relationship between data patterns are understood and additional value from the huge healthcare data is uncovered [2]. There are various research trends and challenges throughout the data life cycle while implementing Big Data Analytics and is well described in [3][4][5]. Healthcare domain challenges are in improving research phases.The reality mining is a new approach i.e. using big data to study the patient's behavior through mobile phone sensors that helps to improve the healthcare quality[5]. Extracting useful information from Electronic Health Record is surveyed in [6]. An intelligent design using Big Data is proposed in [7] which is a web based application that provides efficient platform for simplification of complex assessment of health along with monitoring procedure.

Currently in the health care domain by implementing machine learning algorithm and making use of statistical medical data interesting patterns in disease data are discovered [8][9] For improved diagnosis and prognosis machine learning can be used effectively is demonstrated with two case studies for disease diagnosis in [10][11]. The Expert decision support system using data mining techniques for disease classification is proposed in [12] where it focuses on diagnoses of diabetes disease and uses the Decision Tree and KNN Algorithm.The EM- psychiatry system is an ambient intelligent system intended for emergency psychiatry care and it makes use of machine learning technique that is maximum entropy Markov model (MEMM) [13]. Another system is iHANDS that is an intelligent health and decision support agent built with various artificial intelligence mechanisms that assists individuals in their healthcare decision-making process [14]. A similar machine learning approach is applied to another E-medication system for Sickle Cell Disease diagnosis where it tries bridge the gap between the doctor and patient through application on the smart phone [15]. Since the machine learning domain consists of various techniques the researches make comparisons so as to which gives efficient and faster results in prediction. A comparison performed using SVM and implemented multilayer perceptron neural networks for heart disease prediction where SVM gives higher accuracy [16] SVM has got preference over other machine learning algorithms due to its accuracy is better in comparison[16].

The following table TABLE I consists of various techniques that different researchers have used to make comparisons and their working dataset is mentioned along with outcome and the possible limitations are listed

Table I: Comparison of various Literature Works

Ref	Methodology	Dataset	Outcomes	Limitations
Bhargava et al. 2017 [17]	Simple CART algorithm in WEKA to predict heart attack	Real world Male Heart disease dataset. Instances Used =209	Accuracy of correctly classified instances is 79.9 %	Only one algorithm considered therefore couldn't state if it is best approach.
Dhomse, Mahale, 2016 [18]	SVM, Decision Tree and Naive Bayes applied with and without feature selection to predict heart disease	Heart disease dataset from Cleveland Clinic Foundation. Instances Used =303	After reducing dataset SVM outperforms Naive Bayes	Accuracy results not mentioned direct graph plotted.
Dhomse, Mahale, 2016 [18]	SVM and Naive Bayes applied with and without feature selection to predict diabetes disease using WEKA tool	Diabetic patients dataset is collected from hospital repository. Instances Used = 1865	Naive Bayes have better accuracy results and takes less time for building the training model than SVM	Classification Accuracy of Naive Bayes is 34.89% which is quite risky for prediction
Ramzan, 2016 [19]	Naive Bayes, J48 Decision Tree, Random Forest are used to compare classifiers to predict critical disease using WEKA tool	Disease classification dataset collected from Global Health Data Exchange. Instances Used = 9242	Random Forest turns out with an Accuracy of 99.83% beating both Naive Bayes and J48	Random Forest requires more time for building the training model
Naik, Samant, 2016 [20]	Decision Tree, K-Nearest Neighbours, Naive Bayes are correlately reviewed to predict liver disorder using WEKA, Orange, Tungara, KNIME and Rapid miner tool	Liver patient dataset collected from Indian Liver Patient Dataset. Instances Used = 583	As for tool's performance all three algorithms performed well using KNIME tool. As for algorithm then Decision Tree and KNN outperformed Naive Bayes using all the tools	Requires a powerful machine learning to analyze the outcome of model using all tools to improve classification accuracy.
Iyer et al., 2015 [21]	J48 Decision Tree and Naive Bayes approach for diagnosis of diabetes	Pima Indians Diabetes Database. Instances Used = 768	Naive Bayes gives least error rate and thus outperforms J48 decision Algorithm	Comparison of only 2 algorithm is not sufficient to build best diagnosis model.

III. METHODOLOGY

This section includes the methodology describing the approach that is used to carry out the research in order to perform comparative analysis

A. WEKA Tool Description

The WEKA tool is briefly described below :

WEKA [Waikata Enviroment for Knowledge Analysis]

- It is a very popular machine learning and data mining toolkit for conducting data driven researches.
- Developed in New Zealand at the University of Waikato
- The collection of machine learning and data mining algorithms present are written in Java
- The version of WEKA used for experimentation in this paper is WEKA Version 3.82

The research made use of WEKA tool as it helps in performance evaluation and performing comparison of various machine learning techniques conveniently on real time data.

B. Diabetes Disease Dataset

Here is the description of the dataset that has been used as an input to classifiers implemented using various algorithms. The name of the dataset that has been considered is **Pima Indians Diabetes Database** which is collected from National Institute of Diabetes and Digestive and Kidney Diseases. The total No. of Instances are 768 and the size is 37 KB. The total no. of attributes are 9 including the target class attribute. The name of two target classes are tested_positve and tested_negative. The no. of instances for tested_positive are 268 and the no. of instances for tested_negative are 500. The data pre-processing is automatically performed by WEKA tool.

The following table TABLE II describes the 9 attribute of the diabetes dataset briefly

Table II: Diabetes Disease Dataset

Sr No.	Attribute Used	Attribute Type	Attribute Description
1	preg	Numeric	No. of times pregnant
2	plas	Numeric	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	pres	Numeric	Diastolic blood pressure (mm Hg)
4	skin	Numeric	Triceps skin fold thickness (mm)
5	insu	Numeric	2-Hour serum insulin (mu U/ml)
6	mass	Numeric	Body mass index (weight in kg / (height in square m)
7	pedi	Numeric	Diabetes pedigree function
8	age	Numeric	Age (years)
9	Class	Nominal	Class variable (tested_positive or tested_negative)

C. Flow Chart of Proposed Methodology

Here is the brief description about the flow of proposed methodology.

The Proposed Classifier model specifically considered diabetes disease and takes input of the dataset for diabetes. The input dataset is processed using four machine learning algorithms that are Naive Bayes, SVM, Random Forest, Simple CART and for each algorithm respective classifier model is trained and tested and the results are gathered. Based on the experimental results the best performing algorithm can be determined which will help in accurate prediction of the disease.

The following figure Fig 2. depicts the approach that has been applied to perform the comparative analysis in order to recommend the best algorithm for building classification model in order to predict the diabetes disease.

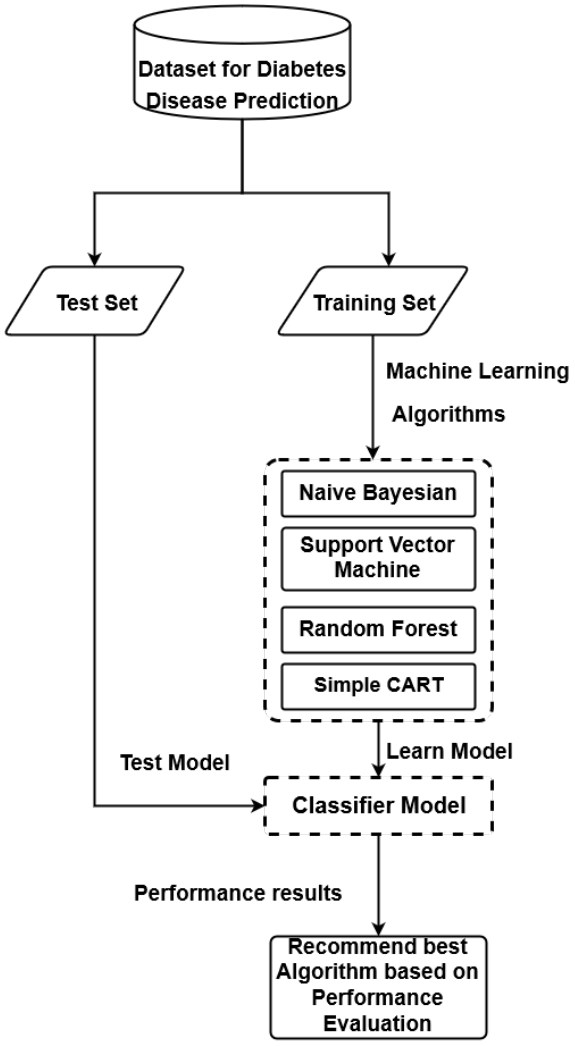

Fig. 2. Proposed Methodology Flowchart

The following describes the steps involved in the procedure of the Fig 2. Proposed Classifier Methodology

Stepwise Procedure of Proposed Methodology

- **Step 1 :** - Preprocess the input dataset for diabetes disease in WEKA tool

- **Step 2 :** - Perform percentage split of 70% to divide dataset as Training set and Test set
- **Step 3 :** - Select the machine learning algorithm i.e. Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm.
- **Step 4 :** - Build the classifier model for the mentioned machine learning algorithm based on training set.
- **Step 5 :** - Test the Classifier model for the mentioned machine learning algorithm based on test set
- **Step 6 :** - Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
- **Step 7 :** - After analyzing based on various measure conclude the best performing algorithm.

The proposed classifier model has been built using WEKA tool and based on successful execution of each step we can evaluate the experimental results.

IV. Experimental Results

This section describes the experimental results that are obtained after training Naive Bayes, Support Vector Machine, Random Forest and Simple CART classifiers on the diabetes patient dataset. The purpose of these experimental results are for performance evaluation of all four classifier and to recommend the best algorithm suited for prediction.

A. Confusion Matrix

In machine learning, a Confusion Matrix is used to analyze the performance of the classification algorithm. The Confusion matrix is a tabular structure where the rows represents Actual class and columns represents Predicted class.

CONFUSION MATRIX STRUCTURE			
Total no. of instances		**Predicted Class**	
		No *a=tested_negative*	**Yes** *b=tested_positive*
Actual Class	**No** *a=tested_negative*	TrueNegative	FalsePositive
	Yes *b=tested_positive*	FalseNegative	TruePositive

Fig. 3. General Confusion Matrix Structure

Certain terminology as that appear in the general Confusion Matrix Structure are described below. These terminology will be further used for Performance Evaluation of each classifier.

- **Actual Class** : Class label representing the Actual class before building the classifier
- **Predicted Class** : Class label representing the Predicted Class after building the classifier
- **TruePositive** : No. of instances predicted positive and are actually positive
- **FalsePositive** : No. of instances predicted negative and are actually negative
- **TrueNegative** : No. of instances predicted positive but are actually negative
- **FalseNegative** : No. of instances predicted negative but are actually positive
- **Total no. of Instances** : The sum of all the instances that have been classified by the classifier.

The Confusion Matrix thus obtained after building the classifier using Naive Bayes, Support Vector Machine, Random Forest, Simple CART machine learning algorithm in WEKA tool are shown below:

```
      Naïve Bayes
=== Confusion Matrix ===

   a   b   <-- classified as
 133  25 |   a = tested_negative
  28  44 |   b = tested_positive
```

Fig. 4. Naive Bayes Confusion Matrix

As per Fig 4. According to Naive Bayes Confusion Matrix the values of TrueNegative=133, FalseNegative=28, FalsePositive=25, TruePositive=44

```
   Support Vector Machine
=== Confusion Matrix ===

   a   b   <-- classified as
 143  15 |   a = tested_negative
  33  39 |   b = tested_positive
```

Fig. 5. Support Vector Confusion Matrix

As per Fig 5. According to Support Vector Machine Confusion Matrix the values of TrueNegative=143, FalseNegative=33, FalsePositive=15, TruePositive=39.

```
      Random Forest
=== Confusion Matrix ===

   a   b   <-- classified as
 138  20 |   a = tested_negative
  34  38 |   b = tested_positive
```

Fig. 6. Random Forest Confusion Matrix

As per Fig 6. According to Random Forest Confusion Matrix the values of TrueNegative=138, FalseNegative=34, FalsePositive=20, TruePositive=38.

```
      Simple CART
=== Confusion Matrix ===

   a   b   <-- classified as
 133  25 |   a = tested_negative
  29  43 |   b = tested_positive
```

Fig. 7. Simple CART Confusion Matrix

As per Fig 7. According to Simple CART Confusion Matrix the values of TrueNegative=133, FalseNegative=29, FalsePositive=25, TruePositive=43.

B. Classification Accuracy

The classification accuracy is one of the performance evaluation measure. Accuracy represents how well the classifier performs prediction of the instances based on the training data.

- **Accuracy** : It is the ratio of the no. of true predicted instance both positive and negative to the total no. of instances.

$$\textbf{Accuracy} = \frac{TruePositive + TrueNegative}{Total\ no.\ of\ instances}$$

The following table Table III represents the experimental classification accuracy results of Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. The table displays the Training time, testing time and Accuracy value of each algorithm.

Table III: Experimental Classification Accuracy Results

Algorithm	Training Time	Testing Time	Accuracy Value
Naive Bayes	0.03 sec	0.02 sec	0.77
Support Vector Machine	0.14 sec	0.03 sec	0.7913
Random Forest	0.67 sec	0.06 sec	0.765
Simple CART	1.38 sec	0.02 sec	0.765

The following figure Fig 8. shows the Classificaion Accuracy value plot of all four classifiers

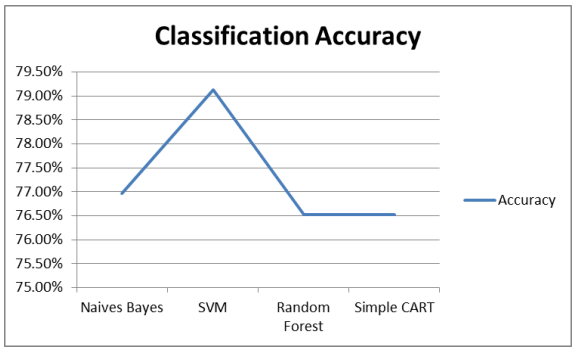

Fig. 8. Classification Accuracy Values

C. Accuracy Measure Values

Following are the Classifier Accuracy Measure Values description:

- **TP - Rate** : It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances

$$TP{-}Rate = \frac{TruePositive}{TruePositive + FalseNegative}$$

- **FP - Rate** : It is the ratio of the no. of predicted negative instances to the actual total no. of negative instances

$$FP{-}Rate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

- **Precision** : It is the ratio of no. of predicted positives instances to the total of all predicted positive instances.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

- **Recall** : It is the ratio of the no. of predicted positive instances to the actual total no. of positive instances

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

- **F- Measure** : Used to represent overall performance. It is weighted harmonic mean of the precision and recall

$$F{-}Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The following table Table IV represents Accuracy Measure Value of all four classifiers that are obtained after building all four classifiers on diabetes dataset in WEKA

Table IV: Classification Major Accuracy Measure Values

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure
Naive Bayes	0.770	0.317	0.767	0.770	0.768
Support Vector Machine	0.791	0.345	0.784	0.791	0.782
Random Forest	0.765	0.326	0.756	0.765	0.758
Simple CART	0.765	0.364	0.762	0.763	0.446

The following figure Fig. 9. shows the Accuracy Measure Value of all four classifiers based on the values obtained after experimentation which is represented in Table IV

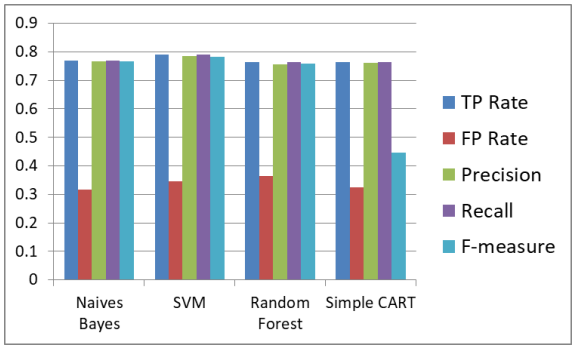

Fig. 9. Classifier Accuracy Measure Results

V. Results and Discussions

This section discusses the overall experimental results thus obtained through WEKA tool.

According to the Classification Accuracy figure Fig 8. The Accuracy of SVM is the highest which is 0.7913. The Accuracy of Naive Bayes is 0.77 better than Random Forest and Simple CART. The accuracy of Random Forest and Simple CART is almost equal with value 0.765. The training time of Naive Bayes is less than SVM. The training time of Simple CART is the highest. Overall according to classification Accuracy SVM outperformed all other classifier.

According to the Classification Accuracy Measure depicted in figure Fig 9. the Support Vector Machine has the highest F-measure value of 0.782 and the least is of Simple CART which is 0.446. The Precision value of Support Vector Machine is highest with a value of 0.784 and the precision value of Randome Forest is the least with a value of 0.756

VI. Conclusion

In this research work, four classifiers based on machine learning algorithm which are Naive Bayes, Support Vector Machine, Random Forest and Simple CART have been used for experimentation on WEKA tool to predict Diabetes disease. The four classifiers thus build have been compared based on training time, testing time and accuracy value. Another performance evaluation method was classifier accuracy measure which included TP-rate, FP-rate, precision, recall, F-Measure. The overall performance of Support Vector machine to predict the diabetes disease is better than Naive Bayes, Random Forest and Simple Cart. Hence the effectiveness of the proposed model is clearly depicted throughout the experimental results mentioned.

References

[1] Rahul C. Basole, Mark L. Braunstein, And Jimeng Sun, "Data and Analytics Challenges for a Learning Healthcare System", ACM Journal of Data and Information Quality, Vol. 6, No. 2–3, Article 10, Publication date: July 2015

[2] J.Archenaa and E.A. Mary Anita, "A Survey Of Big Data Analytics in Healthcare and Government", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), 2015

[3] Agusti Solanas, Fran Casino, Edgar Batista and Robert Rallo, "Trends and Challenges in Smart Healthcare Research: A Journey from Data to Wisdom", IEEE 2017

[4] Fuad Rahman, "Application of Big-Data in Healthcare Analytics – Prospects and Challenges", IEEE 2017

[5] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime and Thomas Noel, "Big Data in healthcare: Challenges and Opportunities", IEEE 2015

[6] Pranjul Yadav, Michael Steinbach, Vipin Kumar and Gyorgy Simon, "Mining Electronic Health Records (EHRs): A Survey", ACM Computing Surveys, Vol. 50, No. 6, Article 85. Publication date: January 2018.

[7] Weider D. Yu, Jaspal Singh Gill, Maulin Dalal, Piyush Jha and Sajan Shah, "Big Data Approach in Healthcare used for Intelligent Design", 2016 IEEE International Conference on Big Data (Big Data)

[8] Rohan Bhardwaj, Ankita R. Nambiar and Debojyoti Dutta, "A Study of Machine Learning in Healthcare", 2017 IEEE 41st Annual Computer Software and Applications Conference

[9] Athmaja S., Hanumanthappa M. and Vasantha Kavitha, "A Survey of Machine Learning Algorithms for Big Data Analytics", 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[10] Oleg Roderick, Nicholas Marko, David Sanchez and Arun Aryasomajula, "Data Analysis And Machine Learning Effort In Healthcare: Organization, Limitations, And Development Of An Approach", Internet of Things and Data Analytics Handbook, First Edition. 2017

[11] Niharika G. Maity, Dr. Sreerupa Das, "Machine Learning for Improved Diagnosis and Prognosis in Healthcare", IEEE 2017

[12] Emrana Kabir Hashi, Md. Shahid Uz Zaman , Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, IEEE

[13] Md. Golam Rabiul Alam, Rim Haw, Sung Soo Kim, Md. Abul Kalam Azad, Sarder Fakhrul Abedin, Choong Seon Hong, "EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 6, DECEMBER 2016

[14] Brett Hannan, Xiaoqin Zhang, Kristen Sethares, "iHANDs: Intelligent Health Advising and Decision-Support Agent", 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

[15] Dhafar Hamed Abd,Jwan K. Alwan, Mohamed Ibrahim, Mohammab B Naeem, "The Utilisation of Machine Learning Approaches for Medical Data Classification and Personal Care System Mangement for Sickle Cell Disease", Annual Conference on New Trends in Information & Communications Technology Applications(NTICT'2017)7 9 March 2017, IEEE 2017

[16] Parisa Naraei, Abdolreza Abhari, Alireza Sadeghian, "Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data", FTC 2016 - Future Technologies Conference 2016, 6-7 December 2016 — San Francisco, United States, IEEE 2016

[17] Dr.Neeraj Bhargava, Sonia Dayma, Abishek Kumar, Pramod Singh, "An Approach for Classification using Simple CART Algorithm in Weka", 2017 11 th International Conference on Intelligent Systems and Control (ISCO)" , IEEE 2017

[18] Dhomse Kanchan B., Mr. Mahale Kishor M, "Study of Machine Learning Algorithms for Special Disease Prediction using Principal Component Analysis", 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE 2016

[19] Munaza Ramzan, "Comparing and Evaluating the Performance of WEKA Classifiers on Critical Diseases" , 2016 1st India International Conference on Information Processing (IICIP), IEEE 2016

[20] Amrita Naik, Lilavati Samant, "Correlation review of classification algorithm using data mining tool: WEKA , Rapidminer , Tanagra ,Orange and Knime", International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85 (2016) 662 – 668

[21] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015