AN ALGORITHM FOR PREDICTIVE DATA MINING APPROACH IN MEDICAL DIAGNOSIS

Shakuntala Jatav[1] and Vivek Sharma[2]

[1]M.Tech Scholar, Department of CSE, TIT College, Bhopal
[2]Professor, Department of CSE, TIT College, Bhopal

ABSTRACT

The Healthcare industry contains big and complex data that may be required in order to discover fascinating pattern of diseases & makes effective decisions with the help of different machine learning techniques. Advanced data mining techniques are used to discover knowledge in database and for medical research. This paper has analyzed prediction systems for Diabetes, Kidney and Liver disease using more number of input attributes. The data mining classification techniques, namely Support Vector Machine(SVM) and Random Forest (RF) are analyzed on Diabetes, Kidney and Liver disease database. The performance of these techniques is compared, based on precision, recall, accuracy, f_measure as well as time. As a result of study the proposed algorithm is designed using SVM and RF algorithm and the experimental result shows the accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively.

KEYWORDS

Data Mining, Clinical Decision Support System, Disease Prediction, Classification, SVM, RF.

1. INTRODUCTION

Computational health informatics is rising research topic that involving varied sciences like biomedical, medical, nursing, data technology, computer science, and statistics [1]. Data mining techniques are applied to predict the effectiveness of huge and complex clinical data in order to diagnose disease and extract information to suggest effective medical assistance [2]. In bioscience, doctor's facilities introduced different knowledge frameworks with plenty of data to manage medical insurance and patient information however unfortunately, knowledge don't seem to be mined to find hidden data for effective decision [2][3].

Clinical test outcomes are often created on the basis of doctor's perception and experience instead of on the knowledge enrich data masked within the database and generally this procedure prompts unintended predispositions, doctor's experience might not be capable to diagnose it accurately that affects the disease diagnosis system [2][3]. In aid sector, the term data mining will mean to research the clinical data to predict patient's health status. Therefore discovering fascinating pattern from healthcare information, different data mining techniques are applied with statistical analysis, machine learning and database technology.

Predictive systems are the systems that are wont to predict some outcome on the basis of some pattern recognition, as shown in figure 1. Disease detection is that the method by which patient's diagnosis is performed on the basis of symptoms analyzed which may causes difficulty while predicting disease affect [4]. As an example, fever itself could be a symptom of the many disorders that doesn't tell the healthcare professional what exactly the disease is. Because the results or

opinion vary from one physician to a different, there's a requirement to help a medical physician which will have similar opinion certainly symptoms and disorders [5]. This may be done by analyzing the information generated by medical data or medical records.

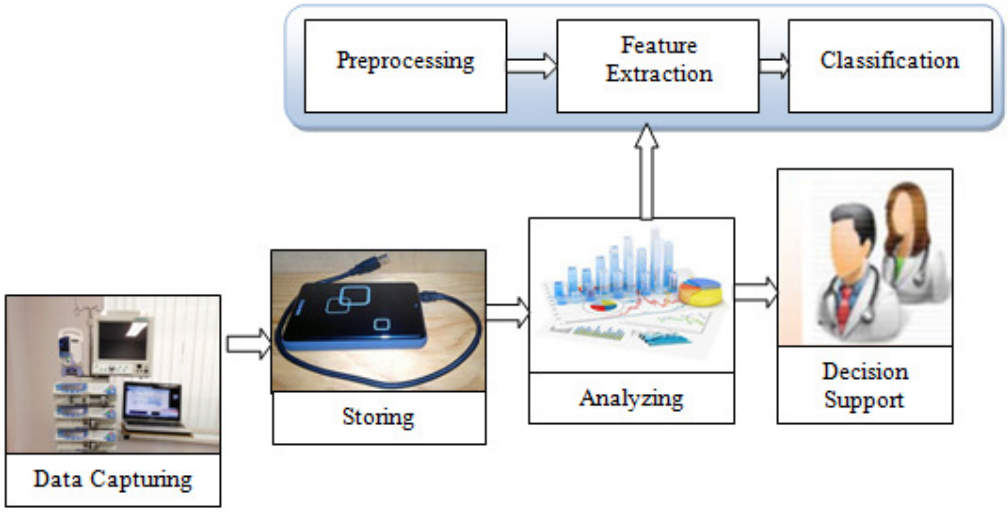

Figure 1: A Typical Health Informatics Processing

As a result, the new information is often compared with previous records and optimistic diagnosis is often done. Predictive medical diagnosis could be a net application which is able to predict a selected disorder on the basis of symptoms and supply diagnosis for same disorder which is able to be detected by rule. Healthcare professionals use their previous data and insights to reach a particular decision regarding any disease or disorder [6]. Within the similar manner, this paper proposes different classification techniques for diagnosis by using generic disease datasets. This paper brings into limelight all the benefits and drawbacks of using the various data mining techniques for the prediction of diseases. It conjointly accounts for the prediction rate for various techniques therefore, bringing out the comparison between each of them [7]-[10].

Data mining has been successfully used in knowledge discovery for predictive purposes to make more active and accurate decision [11]. The main focus of the paper is on classification as well as clustering techniques. In clustering process such as K-means, EM, Fuzzy c-means, etc, data is partitioned into sets of clusters or sub-classes [12].Machine learning techniques such as KNN, SVM, Naïve bayes etc, can be used to classify different objects on the basis of a training set of data whose outcome value is known.

Clustering Techniques

The clustering process divides the data into cluster groups or subclasses. We used four clustering algorithms, namely K-Means, EM, PAM, Fuzzy C-Means [12]. The K-Means classification algorithm works by partitioning n observations in k-subclasses defined by centroids, where k is chosen before the algorithm begins. K-Means and EM are two iterative algorithms. EM (expectation-maximization) is a statistical model that depends on the unobserved latent variables to estimate the maximum likelihood parameters. Partitioning around medoids (MAP) is similar to K-means that partitioning is based on the K-medoids method, which divides data into a number of disjoint clusters [12]. In fuzzy clustering, data elements can belong to multiple clusters. This is also called soft clustering.

Classification Techniques

Machine learning based classification techniques can be used to classify various objects based on a series of training data whose result value is known. In this study four classification algorithms are used: KNN, SVM, Naive Bayes and C5.0. In the nearest neighbor KN, the object is classified by the majority of its neighbors, with the object being assigned to the most commonly used class among its nearest neighbors. In SVM (Support Vector Machines), data is first converted into a set of points and then classified into classes that can be separated linearly. The Naive Bayes model calculates the probability of a set of data that can belong to a class using the Bayes rule. The C5.0 algorithm is a decision tree that recursively separates observations in branches to form a tree to improve prediction accuracy. It is an improved version of the C4.5 and ID3 algorithms [10]. It also provides the powerful gain method to increase the accuracy of this classification algorithm [11].

2. RELATED WORK

In [1] neural networks, decision tree and naïve bayes machine learning approach is used to diagnose heart disease. For optimized feature selection genetic algorithm is used and obtained an accuracy of 100%, 99.62 and 90.74 respectively.

In [2], performed a prediction of heart disease detection using neural network with genetic algorithm based feature extraction. Back propagation based neural network weight optimization by Genetic algorithm is designed and obtained the 89% accuracy of prediction of heart disease.

In [3], author developed a heart disease prediction system using an approach of ANN with LVQ and achieved accuracy of about 80%, sensitivity of about 85% sensitivity as well as specificity of about 70%.

In [4] proposed a heart disease diagnosis is proposed using lazy data mining approach with data reduction strategies i.e. principal component analysis is used to get category association rules. The result analysis shows that J4.8 has 10.26 enhancement as well as 8.6% enhancement over naïve bayes.

In [5] presented a heart disease prediction system using data mining approach with two additional features i.e. obesity and smoking to boost the prediction rate. Neural networks, Decision trees and Naive Bayes was used in for predicting heart disease with an accuracy of 99.25%, 94.44% and 96.66% respectively.

A web based application has introduced in [6] using Naïve Bayesian algorithm which took symptoms from user and gave the diagnosis result to the user or patient. In [7] Association rule mining technique was used for diagnosis of diabetes. The authors concluded that the data mining techniques when used appropriately increases the computation and also the classification performance. These rules have the potential to improve the expert system and to make better clinical decision making. For predicting diabetes disease on weka tool, author in research work [8] had presented a comparison between Naïve bayes algorithm and decision tree algorithm and achieved system accuracy of about 79.56% and 76.96% respectively.

In [9] Decision Tree, Naive Bayes, and NBTree algorithms is used for liver disease detection with 10 features. The result analysis with respect to accuracy NBTree algorithm has the highest accuracy whereas with respect to computational time Naive Bayes algorithm performs better. In [11] author performed a comparative analysis on clustering and classification algorithms. The result analysis shows that the classification is better than clustering algorithms with an accuracy of about 81%.

In [12] author proposed classification with clustering technique i.e. KNN with FCM clustering and F-KNN with FCM clustering. It is clear that the Fuzzy KNN with Fuzzy c-means model produced the better result than the KNN with Fuzzy c-means model on both PIMA and Liver-disorder datasets. It is also clear that the use of Fuzzy c-means clustering algorithm for preprocessing of datasets improved the result in terms of classification accuracy and speed by reducing the number of tuples from the original datasets. From experiment, it is been found that KNN with Fuzzy c-means have accuracy of 97.02 and Fuzzy KNN with Fuzzy c-means have accuracy of 99.25 on PIMA dataset whereas on Liver disorder KNN with Fuzzy c-means have accuracy of 96.13 and Fuzzy KNN with Fuzzy c-means with accuracy of 98.95.

In [13] performed the chronic disease prediction by using data mining approach such as Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) for the diagnosis of diabetes and heart disease. The result analysis shows that SVM gives highest accuracy of 95.556% in case of heart disease and Naïve bayes gives accuracy of 73.588% in case of diabetes. Table I gives the comparative analysis of different existing techniques for heart, liver and diabetes diseases.

Table 1: Comparative Analysis of Different Techniques in terms of Accuracy

Name of Author	**Technique**	**Accuracy**
Bhatla et al.	Neural Network, Naïve Bayes and Decision Tree for heart disease detection.	Neural networks = 100 % Decision tree= 99.62 % Naïve Bayes 90.74 %
Amin et al.	Optimized Neural Network for heart disease detection.	Training data was = 89% Validation data = 96.2%.
Chen et al.	ANN based heart disease detection.	ANN = 80%.
Dangare et al.	Decision trees, Neural networks and Naive Bayes for heart disease detection.	Neural Networks = 99.25% Naive Bayes = 94.44 % Decision Tree =96.66 %
Iyer et al.	Decision tree and Naïve bayes algorithm for diabetes detection.	Decision Tree =76.96% Naïve Bayes= 79.56%
Uma Ojha and Savita Goel	Decision Tree, SVM, FCM	Decision tree (C5.0) =81% SVM =81% FCM = 37%
Chetty et al.	KNN and F-KNN for diabetes and liver disease detection.	KNN = 97.02% F-KNN = 99.25% for diabetes data. KNN = 96.13% F-KNN =98.95% for liver disease data.
Kumari Deepika and Dr. S. Seema	Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) for diabetes and heart disease detection.	SVM = 95.556% (heart disease) Naïve Bayes = 73.588% (diabetes).

3. METHODOLOGY

3.1 Proposed Methodology

One of the interesting and important subjects among researchers in the field of medical and computer science is diagnosing illness by considering the features that have the most impact on recognitions. The subject discusses a new concept which is called Medical Data Mining (MDM). Indeed, data mining methods use different ways such as classification and clustering to classify diseases and their symptoms which are helpful for diagnosing.

A disease diagnosis system is created in order to predict different diseases such as diabetes, kidney disease as well as liver disease, etc. System's workflow is discussed below:

Step 1: Through the proposed application user (doctor, patient, physician etc.) can input the attribute values of disease and send it to the decision support system for analysis.

Step 2: At decision support system, dataset of different diseases are loaded and apply data mining algorithms to train dataset. Requested user inputs are collected and processed on server to predict the diagnosis result.

Step 3: For analyzing healthcare data, major steps of data mining approaches like preprocess data, replace missing values, feature selection, machine learning and make decision are applied on train dataset. On the decision support system end different classification algorithms would be executed on train dataset and ready to classify the test dataset.

In the proposed algorithm Support Vector Machine and Random Forest is used to give clustering level for different subspaces. The voting model will ensemble all these results and output the final classification result. Finally, the predicted results will be compared with true labels of the testing phase.

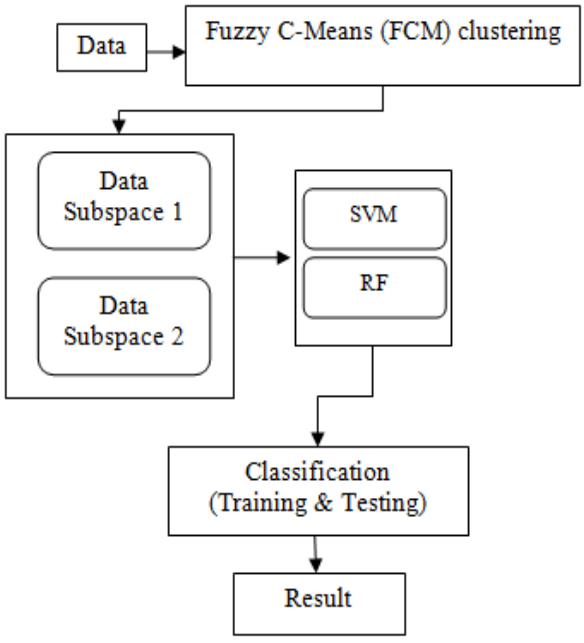

Figure 2: Proposed Model

Support Vector Machine

Support vector machine is a machine learning approach that can be used as classifier as well as for regression. SVM classifies the data into different classes by finding hyperplane (line) which separates training data into classes. SVM does not overfit the data and gives best classification performance in terms of precision and accuracy.

SVM does not make any strong assumptions on data. It shows more efficiency for correct classification of the future data. SVM is classified into 2 categories i.e. Linear and non-Linear. In linear approach, training data is separated by line i.e. hyperplane.

Random Forest

Random Forest algorithm is capable of performing each classification and regression tasks. the fundamental principle of RF is that a group of weak learner's come together to create a robust learner. Random forest rule uses bagging approach to form the bunch of decision trees with random set of the information. The model is trained few times on random sample of the dataset to attain best prediction performance from the RF rule. In this ensemble technique of learning, the output of all decision trees within the RF is combined to form a final prediction. The final prediction of the RF rule is derived after polling the results of every decision tree.

Suppose there are N cases within the training set. Then these N samples are taken randomly however with replacement. These samples are training set for growth of tree. If $m < M$ is specific. The simplest split of this m is employed to separate the node. The value of m is constant whereas growing the forest.

3.2 Proposed Algorithm

Input: D {Clininal data};

Output: Label {Disease Label};

Patient's Label{Normal, Disease}

Step1: Pre-processing and data cleansing

Step2: For each instance in D, do

Find feature vector (V)

Step 3: For each V do

Data clustering using FCM and split data in two halves and classify data using SVM and RF algorithm

Step 4: Determine the total class label

Find

True_positive (TP)

True_negative (TN)

False_positive (FP)

False_negative (FN)

Step 5: Find Performance Parameters

Step 6: Predict Disease Class as

if (class=1) Patient=Normal State

else_if(class =0) Patient=Disease State

end for

3.3 Dataset Description

Pima Indians Diabetes Database

This study used data sets from the Pima Indians Diabetes Database of National Institute of Diabetes [14]. This dataset consists of 768 samples with 8 numerical valued attribute where 500 are tested negative and 268 are tested positive instances.

Chronic Kidney Disease Dataset

This study used data sets from the university of California Irvine (UCI) repository. The data set contains 400 patients, where 250 patients were positively affected by kidney disease and as many as 150 patients do not suffer from kidney disease.

Liver Disorders Data Set

This study used data sets from the university of California Irvine (UCI) repository. This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

3.4 Performance Measures

In this study, we used three performance measures: Precision, Accuracy, Recall, F_measure and Total execution time.

Accuracy is termed as ratio of the number of correctly classified instances to the total number of instances.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision is the ration of actually true predicted instances out of the total true instances.

Precision = TP/(TP+FP)

Recall is the ratio of actual true instances out of all the items which are true.

Recall = TP/(TP+FN)

F-measure is the harmonic mean of both precision and recall.

F_Measure = 2*(Precision*Recall)/(Precision + Recall)

Where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively.

4. RESULT ANALYSIS

Below Table 2 and 3 as well as Figure 3 shows the comparative analysis of proposed algorithm with some existing algorithms.

Table 2: Result Analysis of Diabetes Disease Detection

Diabetes Disease Detection			
Recall	Precision	Accuracy	F_measure
1	0.9821	0.9935	0.991

Table 3: Comparative Result Analysis of Diabetes Disease Detection

Accuracy Measurement	
Existing Work [14]	90.43%
Proposed Work	99.35%

Figure 3: Comparative Chart of Diabetes Disease Detection

Below Table 4, 5 and 6 shows the parameter values for different diseases such as diabetes disease, kidney disease as well as liver disease. Similarly figure 4 and 5 shows the corresponding parametric chart of different disease detection using proposed algorithm.

Table 4: Result Analysis of Kidney Disease Detection

Kidney Disease Detection			
Recall	Precision	Accuracy	F_measure
1	0.9875	0.9937	0.9937

Table 5: Result Analysis of Liver Disease Detection

Liver Disease Detection			
Recall	Precision	Accuracy	F_measure
0.9667	1	0.9914	0.9831

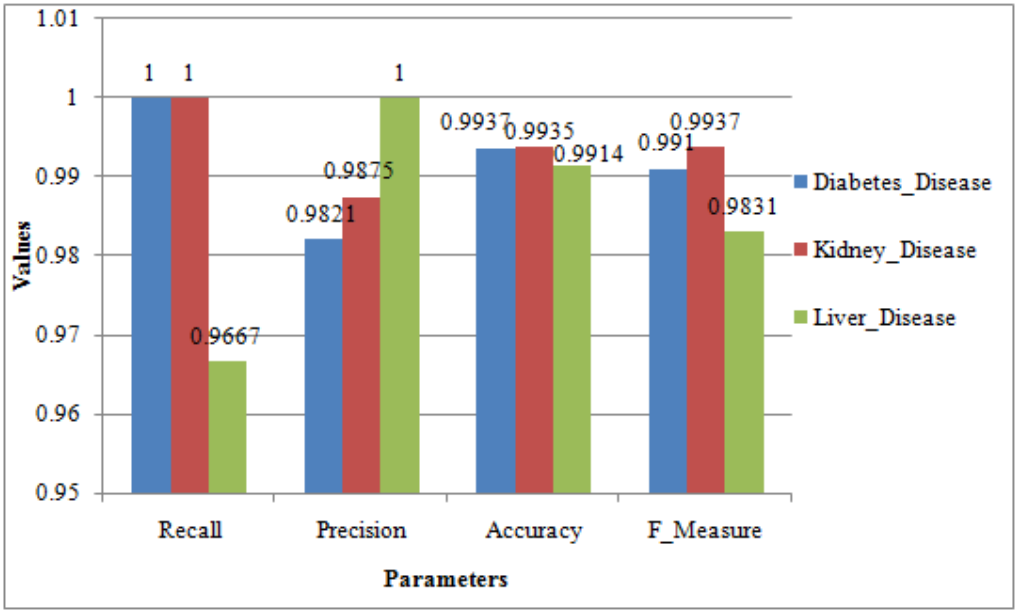

Figure 4: Parameter Comparison Chart of Different Disease Detection

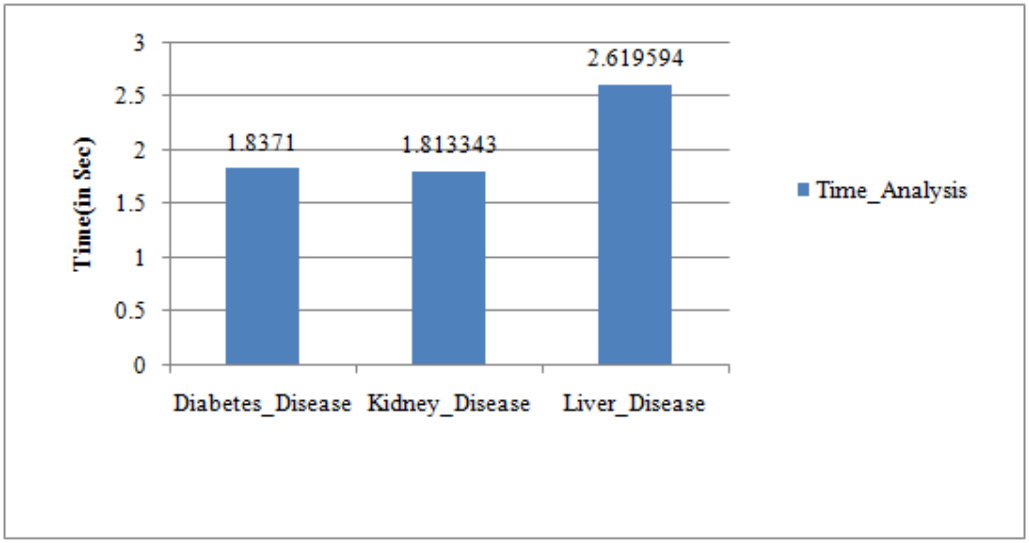

Figure 5: Time Comparison Chart of Different Disease Detection

5. CONCLUSION

This research paper is mainly focused to predict disease possibility using data mining or machine learning approach in order to enhance the accuracy or precision of the disease detection expert system. This paper also shows the related work study of different approaches such as neural network, naïve bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques. As a result of study the proposed algorithm is designed using SVM and RF algorithm and the experimental result shows the accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively. In future using data mining approach a new optimized intelligent system can be designed which can give accurate and efficient result.

REFERENCES

[1] Nidhi Bhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques",IJERT,Vol 1, Issue 8, 2012.

[2] Syed Umar Amin, Kavita Agarwal, Rizwan Beg, "Genetic Neural Network based Data Mining in Prediction of Heart Disease using Risk Factors", IEEE, 2013.

[3] A H Chen, S Y Huang, P S Hong, C H Cheng, E J Lin, "HDPS: Heart Disease Prediction System", IEEE, 2011.

[4] M. Akhil Jabbar, B. L Deekshatulu, Priti Chandra, "Heart Disease Prediction using Lazy Associative Classification", IEEE, 2013.

[5] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Volume 47– No.10, June 2012.

[6] P. Bhandari, S. Yadav, S. Mote, D.Rankhambe, "Predictive System for Medical Diagnosis with Expertise Analysis", IJESC, Vol. 6, pp. 4652-4656, 2016.

[7] Nishara Banu, Gomathy, "Disease Forecasting System using Data Mining Methods", IEEE Transaction on Intelligent Computing Applications, 2014.

[8] A. Iyer, S. Jeyalatha and R. Sumbaly, "Diagnosis of Diabetes using Classification Mining Techniques", IJDKP, Vol. 5, pp. 1-14, 2015.

[9] Sadiyah Noor Novita Alfisahrin and Teddy Mantoro, "Data Mining Techniques for Optimatization of Liver Disease Classification", International Conference on Advanced Computer Science Applications and Technologies, IEEE, pp. 379-384, 2013.

[10] A. Naik and L. Samant, "Correlation Review of Classification Algorithm using Data Mining Tool: WEKA, Rapidminer , Tanagra ,Orange and Knime", ELSEVIER, Vol. 85, pp. 662-668, 2016.

[11] Uma Ojha and Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques", International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.

[12] Naganna Chetty, Kunwar Singh Vaisla, Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach", International Conference on Advances in Computing and Communication Engineering, IEEE, pp. 568-572, 2015.

[13] Kumari Deepika and Dr. S. Seema, "Predictive Analytics to Prevent and Control Chronic Diseases", International Conference on Applied and Theoretical Computing and Communication Technology, IEEE, pp. 381-386, 2016.

[14] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", IEEE, 2017.