Comparative Study - Prediction of Diabetes and Heart Disease using Data Mining Approaches

Amee Upadhyay
Department of Computer Science / IT
SVM Institute of Technology
Bharuch, India

Vaishali R. Patel
Department of Computer Science / IT
SVM Institute of Technology
Bharuch, India

Abstract— Data mining is the process of selecting, discovering, and modeling large amounts of data to discover unknown patterns or relationships useful to the data analyst. Now-a-days in health care industry; data mining plays a vital role in prediction of diseases from clinical databases. It has huge amount of information about patients and their diseases. Medical data mining has emerged perfect with potential for exploring hidden patterns. For fast and improved clinical decision making for protective and suggestive medicine, these patterns can be utilized. For discovering hidden patterns and structures in data, data mining can work better. Here the study and comparison is done of different data mining techniques used for prediction of diabetes and heart disease from clinical dataset with different accuracy. Identifying hidden pattern for any disease aims to avoid patients from undertaking blood tests, checking diastolic and reduced the costs of treatment.

Keywords—Diabetes, Heart Disease, Health care, Data mining techniques

I. INTRODUCTION

Data Mining is one of the most innovative areas of computer science that uses various statistical techniques. The methodology lies in the ability to find patterns and relationships. For predicting the various diseases from the medical field the data mining techniques are useful. There are different types of diseases predicting in data mining namely diabetes, heart diseases, lung cancer and breast cancer.

According to World Health Organization (WHO) tells that 37crores of people live with diabetes worldwide and it doubles before the year 2030. Because of diabetes 48 lakhs of people were died in the year 2012. From lower and middle class families 80% of people were died. 5 crores and above are affected by diabetes and this becomes 7 crores by some years in India. India holds number 2 place worldwide. The data mining techniques includes different works for variety of diseases like Diabetes, Cancer and Heart diseases to discover. Heart disease is the most important reason of mortality in the UK, USA, Canada, and England [1].

The Fig.1 shows the overall estimated figures of prediabetes, diabetes and total cases in India which clearly indicates the seriousness of the issue concerned.

![Fig. 1. Overall estimate of pre-diabetic, diabetic and total prevalence in India](image)

This paper organized as follows: In Section II defines the problem statement for predict diseases. Section III, defines causes of Diabetes and Heart Disease. Section IV, defines related survey work on prediction of diabetes & heart disease. In Section V, we have defined prediction accuracy by different DM techniques. Finally, Section VI gives conclusion and future work.

II. PROBLEM STATEMENT

Clinical database are widely used today with many clinical specialists and new database are rapidly emerging. Clinical database are a wide concept extending over many different clinical tasks.

- Due to uncertainty and huge volume in clinical data, information is imperfect, fragmentary, not fully reliable, unclear, ambiguous or deficient in some way [2-4].
• Diabetic’s complication goes undiagnosed due to lack of standardized diagnostic criteria and no clear explanation or Information [3].
• Current diabetes diagnosis methods and procedures are very expensive and time consuming [6].
• Difficulties in identification of types due to mistaken entries in the clinical’s records and lack of information or lack of understanding about symptoms and risk factors from diabetic patients by the healthcare professional [5].

III. DIABETES AND HEART DISEASE

A. Diabetes

Diabetes usually referred to a group of metabolic disorder in which the person has high glucose level in the blood either because of insulin production inadequate or because the body’s cell do not respond properly to insulin. The uses of glucose level increase in the blood leading to symptoms such as heavy thirsty, frequent urination, unexplained weight loss etc. Not only insulin regulates the glucose in the blood but it is also responsible for lipid metabolism. Insulin is offered as medicine only when the above criteria are broken. Physicians become familiar with other aspects of managing the patients with diabetes, and also the importance of controlling the postprandial glucose, diabetes self-management training etc. [7]. Diabetes causes serious complications such as heart disease, blindness, stroke, kidney failure and cancer. Some cancer types such as pancreatic cancer, liver cancer, and breast cancer are common in diabetic patients. The diagnosis of diabetes is one of the important classification problems.

Diabetes can be classified into three types [8].

1. Type-1 Diabetes

This occurs when body failure to produce insulin completely. It requires injecting insulin or wearing an insulin pump. It affects mostly children usually thin. But it may strike at any age. This type is called as “Insulin Dependent Diabetes Mellitus” or “Adult onset diabetes”.

2. Type-2 Diabetes

This occurs when the body cannot effectively use the produced insulin. It requires diet, exercise and blood sugar level is lowered using drugs. It occurs above age 40. Mostly 90% of people are living with type-2 diabetes. This type is called as “Non-Insulin Dependent Diabetes Mellitus” or “Adult onset diabetes”.

3. Gestational Diabetes

The third type occurs when pregnant woman’s receptivity to insulin. 4% of all pregnant women are affected with this type. It can be controlled with insulin and diet. But 50-70% it may affect again. This type is called as “Gestational diabetes”.

B. Heart Disease

Heart is essential part of our body. If heart is not working correctly, it will affect the other body parts. Heart disease effects on the operation of heart. Various factors increase the risk of Heart disease. The World Health Organization (WHO) has estimated that worldwide 12 million deaths occur every year due to the Heart diseases. Individual kills in each 32 seconds because of Heart disease. Over 80% of deaths in world are because of Heart disease [9].

The risk factors which cause heart disease are [9]:

1. Congenital of heart disease: The greater risk for Heart Disease may occur due to genetic reason. It may pass from one to another generation.

2. Smoking: Smoking is major cause of heart related disease. About 40% of people die from smoking tobacco due to heart and blood vessel diseases. It can reduce rapidly after only one year of not smoking.

3. Cholesterol: High level of lipids (fats) in the blood is cause for heart diseases. Cholesterol found among the lipids in the bloodstream and even in all the body’s cells. These fats combined with the cholesterol and increase the risk of heart diseases.

4. High blood pressure: High blood pressure injured the walls of our blood vessels. It also increases the risk of having heart attack or stroke and of developing heart failure and kidney failure.

5. Obesity: Heart problem may higher if the obese puts anybody at higher risk. Obesity term is used for the health condition of anyone significantly above their ideal healthy weight.

6. Lack of physical exercise: The risk for diabetes and high blood pressure may increase by the coronary artery disease (CAD). And it can increase by the lack of physical exercise.

IV. RELATED WORK

Large number of work has been done to find out efficient methods of medical diagnosis for various diseases. It is an attempt to predict efficiently
diagnosis of Diabetes with reduced number of attributes which are based on symptoms that occurs early stages of Diabetes Mellitus.

A. Prediction of Diabetes and Cancer using ANFIS & KNN
To get the better result in insulin variation and maintain the level of glucose of the body, Neuro Fuzzy System is used. It uses If-then inference rules. Adaptive group based k nearest neighbor (AGKNN) algorithm is used to train the neural network [10].

B. Diabetes & Heart disease prediction using Neural network, Fuzzy logic & K-means clustering

C. Predict Heart Disease using K-means clustering, MAFIA & C4.5 classification
The database is preprocessed to make the mining process more efficient. To get the relevant data in database, K-means algorithm is used to cluster the data. Maximal Frequent Itemset Algorithm (MAFIA) is used for mining maximal frequent patterns in database. The frequent patterns are classified using C4.5 algorithm as training algorithm to predict the heart disease [12].

D. Diabetes prediction using Decision tree & Rule based classifier
Rule set Classifier represents the high level of abstracted discovered knowledge. It is very helpful for decision making by predict the goal attribute. Decision tree classification can be used to mine Diabetes data. It has each branch node represents a choice between a number of alternatives and a decision is represented by each leaf node [13].

E. Predict Diabetes using Neural Network
In this approach, Neural Network consists of 28 nodes out of which 13 are input nodes, 13 are hidden nodes and one is output node, inputs are provided to the input nodes having some weights associated with it. It has values from 0 to 1.0 and a bias value is also attached with all the input value. Output will be in binary. If the value is 0, it means the person is not affected from Diabetes Mellitus and if it is 1 it reveals that the person is suffering from Diabetes Mellitus [14].

F. Diabetes prediction using Fuzzy inference system
This fuzzy classifier approach is used for risk classification of CAHD (coronary artery heart disease) in diabetes patients. Attribute ranking technique measures an attribute by calculating the variation of output caused by the changed value of sample attribute. These attributes are classified in different range with fuzzy logic. The fuzzifier converts the crisp value into degree of member ship then if-then inference rules are applied and again they are converted back to crisp value by defuzzification [15].

G. Diabetes prediction using FP growth &Association rule mining
Association rule mining techniques are used to identify relationships among a frequent itemsets in database. In frequent pattern growth, FP-tree is structured and then directly frequent itemsets extracted. FP-tree is implemented for storing compressed and crucial information about frequent pattern [16].

H. Diabetes prediction using Decision tree & Android application
This mobile app, MobDBTest, an important tool that can help in predicting the probabilities of diabetes and also provides knowledge and suggestion about this disease. Machine learning techniques to classify diabetes levels as low, medium and high. Decision Tree classifier is used to design the machinery for the mobile application for diabetes prediction [17].

I. Diabetes prediction using K-means clustering & SVM
Dimensionality reduction is placed by the K-means and Genetic algorithm. Here K-Means clustering is used for removing the noisy, outliers and inconsistent data. And the reduced data is used for selecting the best features with Genetic algorithm. SVM classifier classifies the dataset to achieve better accuracy [18].

J. Prediction of Heart disease using Neural network & Genetic algorithm
There are many of risk factors like sex, age, family history blood pressure, Smoking Habit, alcohol consumption, physical inactivity, diabetes, blood cholesterol, poor diet which causes heart disease. Genetic algorithm is specialized for the global search. A new hybrid model is to optimize the connection weights of ANN to improve the performance of the Artificial Neural Network [19].
K. Heart disease prediction using Decision tree classification

Decision Tree is a classifier method of classifier which results in a value of target variable with using some defined rules [20].

L. Heart disease prediction using Decision tree, Naïve bayes & KNN

The decision tree structure includes root, branch and leaf node. Each root node denotes a test on an attribute, each branch denotes the outcome of test and each leaf node holds the class label. Naïve Bayes classifier is based on probabilities. These probabilities are used to regulate the most likely next event for the given instance from all the training data. KNN remembers all the instances in dataset. To encounter new instance, it uses previous instances as a model and compares it with the new instance [21].

M. Heart disease prediction using Fuzzy KNN classifier

Fuzziness in the measured data is to remove uncertainty of unstructured data. A membership function is designed with the measured value to remove uncertainty and fuzzified data is used to predict the heart disease patients [22].

N. Diabetes prediction using clustering & classification

The data preprocessing transforms the data in suitable form to execute the subsequent stages. The class label assignment phase assigns the each patient record as high risk, medium risk or low risk. The clustering technique is to group the input patient data into mentioned clusters. For building the classifier, divide the dataset into two groups as training set and test set. Then, using test dataset, the classifier is validated for prediction accuracy of the new patient data to classify it in one of the risk category of diabetes [23].

O. Prediction of Diabetes using ANFIS

The textural features of the DM fundus image are extracted. Normalization is a process where the features are scaled within an equal range. Using the process of fuzzification, crisp data are converted to fuzzy set. The if–then Inference rules used fuzzy logic and it formulates the mapping function from a given input to an output. Defuzzification is the process of producing fuzzy set to crisp result values [24].

**TABLE I.

<table>
<thead>
<tr>
<th>Publication / Years</th>
<th>Title</th>
<th>Technique</th>
<th>Disease</th>
<th>Positive Aspect</th>
<th>Negative Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE - 2014</td>
<td>A New Approach for Diagnosis of Diabetes and Prediction of Cancer using ANFIS [10]</td>
<td>ANFIS(Adaptive Neuro fuzzy Inference System), KNN</td>
<td>Diabetes</td>
<td>Accuracy is better and complexity is reduced. It reduces the cost for different medical tests and helps the patients to take precautionary measures in advance.</td>
<td>ANN can give better result than KNN.</td>
</tr>
<tr>
<td>IEEE - 2014</td>
<td>Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques [11]</td>
<td>Neural Network, Fuzzy Logic, Hybrid Genetic Algorithm, K-means and Fuzzy C-means Clustering</td>
<td>Diabetes</td>
<td>These techniques reduce the cost for various medical tests and facilitate patients to require preventive measures well beforehand.</td>
<td>Neural network is capable to tolerate noise data but sometimes noise data will mislead result when health conditions of patients become more complex.</td>
</tr>
<tr>
<td>IEEE - 2014</td>
<td>Disease Forecasting System Using Data Mining Methods [12]</td>
<td>K-means clustering, MAFIA (Maximal Frequent Itemset Algorithm), C4.5 classifier Algorithm</td>
<td>Diabetes</td>
<td>The objective has better accuracy, high precision and recall metrics.</td>
<td>These techniques give output only in if-then rules.</td>
</tr>
<tr>
<td>IEEE - 2014</td>
<td>A Predictive Approach for Diabetes Mellitus</td>
<td>Rule based Classifier, Decision Tree</td>
<td>Diabetes</td>
<td>These techniques help to improve the quality of health care for diabetes</td>
<td>Decision tree gives the less accurate result with compare to other</td>
</tr>
<tr>
<td>Conference</td>
<td>Title</td>
<td>Dataset</td>
<td>Disease</td>
<td>Algorithm/Technique</td>
<td>Result/Note</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>IEEE - 2012</td>
<td>A Data Mining Approach for the Diagnosis of Diabetes Mellitus</td>
<td>Patients</td>
<td>Diabetes</td>
<td>Neural Network with back propagation</td>
<td>Artificial neural network gives the better result than the back propagation.</td>
</tr>
<tr>
<td>IEEE - 2014</td>
<td>A Fuzzy logic system with attribute ranking technique for risk-level classification of CAHD in female Diabetic patients</td>
<td>Patients</td>
<td>Diabetes</td>
<td>Fuzzy classifier, Fuzzy Inference System</td>
<td>This fuzzy logic approach is classified more accurate CAHD risk in type-II diabetes patients.</td>
</tr>
<tr>
<td>IEEE - 2014</td>
<td>Diabetic prognosis through Data Mining Methods and Techniques</td>
<td>Patients</td>
<td>Diabetes</td>
<td>Association rule mining, Apriori, FP-growth</td>
<td>Apriori can’t avoid costly candidate generation.</td>
</tr>
<tr>
<td>IEEE - 2015</td>
<td>MobDDBTest: A machine learning based system for predicting diabetes risk using mobile devices</td>
<td>Patients</td>
<td>Diabetes</td>
<td>Decision Tree, Machine learning algorithm, Android Application</td>
<td>Decision tree classifier is only based on the sensitivity and specificity results not for accuracy.</td>
</tr>
<tr>
<td>ELSEVIER – 2015</td>
<td>Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis</td>
<td>Patients</td>
<td>Diabetes</td>
<td>K-means clustering, Genetic algorithm, SVM</td>
<td>Instead of using missing values, box plot for outlier detection, algorithms for feature selection, standard deviation can used and to experiment classifiers from statistical, neural, fuzzy to enhance the accuracy.</td>
</tr>
<tr>
<td>IEEE - 2013</td>
<td>Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors</td>
<td>Patients</td>
<td>Heart Disease</td>
<td>Neural Network, Genetic Algorithm</td>
<td>Initialization of the NN weights is a blind process and backpropagation algorithm is very slow in convergence.</td>
</tr>
<tr>
<td>SPRINGER – 2015</td>
<td>Prediction of Occurrence of Heart Disease and Its Dependability on RCT Using Data Mining Techniques</td>
<td>Patients</td>
<td>Heart Disease</td>
<td>Decision Tree classification</td>
<td>This work does not cover the technique based on generalization of fuzzy logic using Neutrosophic logics.</td>
</tr>
</tbody>
</table>
Prediction of Heart Disease Using Classification Based Data Mining Techniques [21]  
Decision Tree, Naïve Bayes, K Nearest Neighbour  
Heart Disease  
Naïve Bayes can handle both discrete and numeric attribute values.  
KNN is slow when the dataset is large because it evaluates all instances.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy KNN classifier</td>
<td>Preprocessing Clustering Classification</td>
<td>Build a risk prediction model for type-2 diabetes disease.</td>
<td>Adaptive Neuro-Fuzzy Inference classifier</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>Diabetes</td>
<td></td>
<td>Diabetes</td>
</tr>
<tr>
<td>Fuzzy KNN classifier is more accurate as compared with KNN classifier.</td>
<td>Technique of new indicators do not identify more sensitive interrelationship.</td>
<td>Produce more effective adaptive rules and classification system.</td>
<td>ANFIS system creates corresponding rules for coarsely distributed attributes. It can also use for finely distributed attributes.</td>
</tr>
</tbody>
</table>

V. DISCUSSION

From the above related work table I, different techniques define various advantages as well as disadvantages. They can predict diabetes and heart disease with different accuracy.

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Data Mining Techniques</th>
<th>Accuracy (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>ANFIS &amp; KNN</td>
<td>80</td>
</tr>
<tr>
<td>2.</td>
<td>K-means, MAFIA &amp; C4.5</td>
<td>89</td>
</tr>
<tr>
<td>3.</td>
<td>Neural Network</td>
<td>92.8</td>
</tr>
<tr>
<td>4.</td>
<td>Fuzzy classifier &amp; fuzzy inference system</td>
<td>98.88</td>
</tr>
<tr>
<td>5.</td>
<td>K-means, GA &amp; SVM</td>
<td>98.79</td>
</tr>
<tr>
<td>6.</td>
<td>Neural &amp; GA</td>
<td>89</td>
</tr>
<tr>
<td>7.</td>
<td>Decision tree</td>
<td>71.43</td>
</tr>
<tr>
<td>8.</td>
<td>ANFIS</td>
<td>98.55</td>
</tr>
</tbody>
</table>

Table II shows an accuracy of different data mining approaches to diagnose the diabetes and heart disease. And same analysis of accuracy is shown in figure 2.

VI. CONCLUSION AND FUTURE WORK

Data mining plays an important and decisive role in medical research. Data mining tools would be a valuable asset for diabetes and heart disease researchers because it can uncover and expose hidden knowledge from a huge amount of clinical related data which significantly help to improve the quality of health care for patients. The recognition and prediction can give a warning to patient at an early stage, where some medications and precautionary action can facilitate the patient to increase the period of patient’s healthy life. After completing review of the above papers, Fuzzy Classifier and Fuzzy Inference System give the better accuracy with compare to other approaches. In future, work can be
extended by using GA or SVM based classifier with ANFIS for better results.

References


