INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
(Volume2, Issue4)
Available online at: www.Ijariit.com

Novel Approach for Heart Disease using Data Mining Techniques

Era Singh Kajal
Research Scholar
CBS Group of Institution, Jhajjar, Haryana

Ms. Nishika
Assistant Professor (CSE Department)
CBS Group of Institution, Jhajjar, Haryana

Abstract— Data mining is the process of analysing large sets of data and then extracting the meaning of the data. It helps in predicting future trends and patterns, allowing business in decision making. Presently various algorithms are available for clustering the proposed data, in the existing work they used K mean clustering, C4.5 algorithm and MAFIA i.e. Maximal Frequent Item set algorithm for Heart disease prediction system and achieved the accuracy of 89%. As we can see that there is vast scope of improvement in our proposed system, in this paper we will implement various other algorithms for clustering and classifying data and will achieved the accuracy more than the present algorithm. Several Parameters has been proposed for heart disease prediction system but there have been always a need for better parameters or algorithms to improve the performance of heart disease prediction system.

Keywords— Data Mining, Heart diseases, Classification, -Nearest Neighbour, LDA, SVM

I. INTRODUCTION

Data mining is the process of analyzing large sets of data and then extracting the meaning of the data. It helps in predicting future trends and patterns, allowing business in decision making. Data mining applications can answer business questions that take much time to resolve traditionally. Large amount of data which is generated for the prediction of heart disease is analysed traditionally and is too complicated and voluminous to be processed. Data mining provides the techniques and methods for the transformation of data into useful information for decision making. These techniques make the process fast and it takes less time for the prediction system to predict the heart disease with more accuracy. In the proposed work we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. .By Applying data mining techniques to heart disease data which needs to be processed, we can get effective results and achieve reliable performance which will help in decision making in healthcare industry. It will help the medical practitioners to diagnose the disease in less time and predict the probable complications well in advance. Identifying the major risk factors of Heart Disease categorizing the risk factors in an order which causes damages to the heart such as high blood cholesterol, diabetes, smoking, poor diet, obesity, hyper tension, stress, etc. Data mining functions and techniques are used to identify the level of risk factors to help the patients in taking precautions in advance to save their life. Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown patterns and trends in databases. This information is further used to build predictive models.

1.1 Causes of Heart Diseases

a. **High blood pressure:** When the heart pumps blood, the force of the blood pushes against the walls of the arteries causing pressure. If the pressure rises and stays high over the time it is called high blood pressure or hypertension which can harm the body in many ways i.e. increasing the risk of heart stroke or developing heart failure, kidney failure etc.
b. **High cholesterol:** Cholesterol is a waxy substance found in the fatty deposits in the blood vessels. Increase in the fatty deposits (high cholesterol) does not allow sufficient blood to flow in through the arteries causing heart attacks.
c. **Unhealthy diet:** Eating too much fast food increases blood pressure and cholesterol level causing the risk of heart attacks.
d. **Smoking:** It damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks.
e. **Lack of physical activity:** Lack of exercise increases cholesterol level in blood vessels which further increases the risk of heart attacks.
f. **Obesity:** Obese people are more likely to have high blood pressure, high cholesterol level and diabetes (increase in blood sugar level) which increases the risk of heart strokes in human body. Nowadays, data mining is gaining popularity in health care industry as this industry generates large amount of complex data about hospital resources, medicines, medical devices, patients, disease diagnosis etc. This complex data needs to be processed and analyzed for knowledge extraction which will further help in decision making and is also cost effective.

1.2 Prevention of Heart Diseases

Heart diseases can be prevented by living a healthy life style and by keeping a check on medical conditions.

1.2.1Healthy life style includes:

a. **Healthy diet:** A healthy meal can help preventing heart diseases and its complications. Eating plenty of fruits, salads, juice and fewer processed foods can help a lot in preventing heart diseases. Eating foods which are low in saturated fat, cholesterol, and Tran's fat and high in fiber can help prevent high cholesterol. Limiting salt in food helps to prevent high blood pressure and limiting sugar in food helps to control diabetes.
b. **Healthy weight:** Obesity or being overweight increases the risk of heart attack in a human body. One should maintain a healthy weight by avoiding fatty foods and also by consulting with dieticians and doctors who can help an obese person to maintain a healthy weight.
c. **Physical activity:** Physical activity helps to maintain a healthy weight and keeps the level of blood pressure and cholesterol low. For adults, doctors recommend 2hrs and 30 minutes of exercise and brisk walking or bicycling every week. For children and adolescents doctors recommend 1hr of daily physical exercise.
d. **No smoking:** Smoking of cigarettes increases the risk of heart attack. it damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks.
e. **Limited alcohol:** Intake of too much alcohol can raise the blood pressure level of human body which increases the risk of heart attacks. Men should not have more than 2 drinks per day and for women only.

1.2.2 Treating medical conditions:

a. **Check cholesterol level:** Health care provider should test cholesterol level at least once every 5 years. If one is having high cholesterol level, medications and a healthy lifestyle can help reduce the risk of heart disease.
b. **Control blood pressure:** Blood pressure should be checked at regular basis as blood pressure has no symptoms. If a person does not have blood pressure then the blood pressure of the person should be checked at least once in 2 years. If a person is suffering from blood pressure, the health care team should check the blood pressure more frequently to be safe. One should also take medication from the medical practitioner if necessary and the amount of sodium should also be lowered in diet.
c. **Manage diabetes:** If suffering from diabetes, one should monitor blood sugar levels carefully. One should consult from a medical practitioner for the medication and a dietitian for diet. One should avoid too of sugar in diet to avoid the risk of heart strokes.

d. **Take medicine:** Taking medicines for high cholesterol, diabetes, blood pressure on time and following doctor's instructions carefully can help avoiding the risk of heart diseases. Never stop taking the medicines without consulting the doctor.

e. **Consult with health care team:** Consulting regularly with health care team regarding the health problems with which one is suffering from and the life style changes that one needs to make helps to avoid the risk of heart disease.

1.3 Data Mining techniques

a. **Association:** It is the best known and well researched method for data mining. Association is also called relation technique because patterns which are discovered from the dataset are based on the relationship between the items. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent [1].

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer or client behaviour.

For example, when association technique is used in heart disease prediction system, it tells us the relationship between all the attributes and sort out all the patients with all the risk factors which are required for the heart disease predictions [2].

b. **Classification:** It is a data mining technique which is used to classify each item in a data set into one of predefined set of classes or groups. It is a classic data mining technique which is based on machine learning. As with the most data mining solutions, classification comes with a degree of certainty. It might be probability of the object belonging to the class or it might be some other measure of how closely the object resembles other examples from that class. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups. Goal of classification is to build a concise model that can be used to predict a class of records whose class label is not known [1]. Example of classification is illustrated below in the following figure 1.1

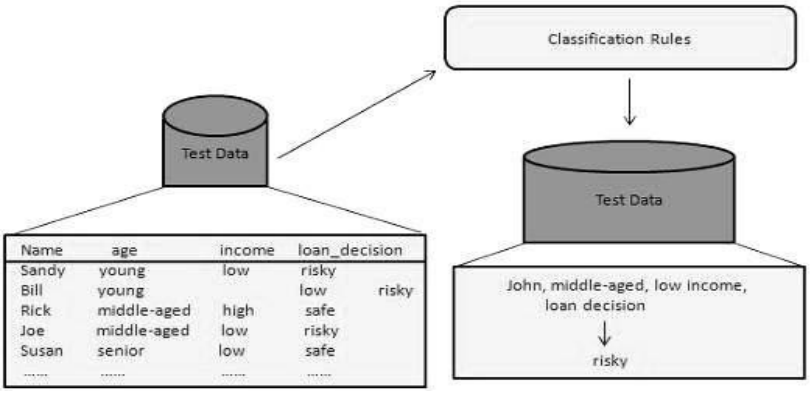

Fig 1.1 Example of Classification

c. **Clustering:** A data mining technique that creates cluster of objects having similar characteristics is known as clustering. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign labels to the group. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster

members, dense areas of the data space, intervals or particular statistical distributions. There is a slight difference between clustering and classification. Clustering defines classes and put objects in them while classification assigns objects into predefined classes. Clustering helps to make clusters or list of patients having same risk factor [1].

Fig 1.2 Example of Clustering

d. Neural network: It is a set of input/output units and each connection has a weight present on it. During the learning phase, network learns by adjusting the weights so as to be able to predict the correct class labels of the input tuples. Neural network have remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued input or outputs [1].

e. Decision tree: It is the most used data mining techniques and its model is easily understandable. The root of the decision tree is a simple question or condition that has multiple answers. Each answer leads to a set of questions or conditions which helps to determine the data so that we can take a final decision based on it [1].

f. Regression: Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation [1].

II. LITERATURE REVIEW

Sujata Joshi et al. (2015) [11] used the classification based data mining techniques are applied to healthcare data. This research focuses on the prediction of heart disease using three classification techniques namely Decision Trees, Naïve Bayes and K Nearest Neighbour. The results show that KNN has highest accuracy as expected since KNN remembers all the instances. But when used for prediction the Decision Tree performs well when compared to other two methods for the given heart dataset.

M.A. NisharaBanu et al. (2014) [12] proposed the frequent patterns could be classified using C4.5 algorithm as training algorithm using the concept of information entropy. The results showed that the designed prediction system is capable of predicting the heart attack successfully. In this work, heart disease prediction system was developed using clustering and classification algorithms to predict the effective risk level and accuracy of the patients.

Monali et al. (2014) [13] used C4.5 algorithm, Multilayer Preceptor and Naïve Bayes algorithm using WEKA data mining tool in the heart disease prediction system. Her objective was study and analysis of data mining algorithms for healthcare decision support system.
Abhishek et al (2013) [14] in the year 2013 used data mining tool Weka 3.6.4he in heart disease prediction system usingJ48 technique achieved 95.56% accuracy and using Naive Bayes achieved 92.42%.

Chitra R.et al.(2013) [15] developed the computer aided heart disease prediction system that helps the physician as a tool for heart disease diagnosis. From the analysis it is concluded that neural network with offline training is good for disease prediction in early stage an d good performance can be obtained by pre-processed and normalized dataset.

Indira S. FalDessai et al.(2013) [16] used decision tree and Naïve Bayes in the year 2013 and achieved 84 percent accuracy and used neural network and achieved 80 percent accuracy in heart prediction system.
Ms. Ishtake et al. (2013) [17] developed a prediction system for heart diagnosis using decision tree, Neural Network and Naive Bayes techniques using 15 attributes in the year 2013.

Jesmin et al. (2013) [18] in the year 2013 in heart prediction system used Naïve Bayes and achieved 92.08 percent of accuracy. Then again he used SMO, AdaBoostM1, J48 and PART achieving 96.04%.

Aqueel Ahmed et al. (2012) [19] show the classification techniques in data mining and show the performance of classification among them. In this classification accuracy among these data mining has discussed. In this decision tree and SVM perform classification more accurately than the other methods and was able to achieve 91% accuracy.

Chaitrali S. et al.(2012) [20] developed a heart disease prediction system in the year 2012 using data mining and artificial neural network technique. From the ANN, a multilayer perception neural network along with back propagation algorithm is used to develop the system.

Chang-Sik Son and Yoon-Nyun Kim, et al. (2012) [21] have proposed a decision making model which provides critical factors and knowledge associated with Congestive Heart Failure (CHF). The accurate diagnosis of heart disease characteristics was quite difficult in emergency room patient. The accurate diagnosis of heart disease made use of Rough Sets (RS) and decision trees. RS-based model and Logistic Regression (LR) were two subset necessary factors to differentiate CHF patients with risk factor were founded among 72 laboratories. 10-fold cross-validation was conducted by RS and LR-based decision model and showed the usefulness of proposed system. The result of RS-based model was consistently better than LR-based model after the comparison of accuracy, sensitivity, specificity, positive predictive value and negative predictive value in both models.

Debabrata Pal and K.M. Mandana, et al (2012) [22] have proposed a method to detect coronary artery disease at early stage by designing expert system. Because CAD affects millions of people and early detection of this disease is still a challenge for prevention. The knowledge acquisition and knowledge representation techniques were the two methods used to avoid uncertainty present in medical domain. It was prevented by creating rules from the doctors and fuzzy expert system. The implementation of the system was done using object oriented analysis and design. The rules provided by medical expert predicted the patient's risk status of CAD. Organization of rules were focused using the concept of modules, meta-rule base, rule address storage in tree representation and rule consistency checking for efficient search of large number of rules focused on the organization of created rules. In CAD risk computation it leads to 95.85% sensitivity and 83.33% specificity.

Evanthia E. Tripoliti et al (2012) [23] have proposed a dynamic determination of the number of trees in random forests algorithm, a computerized diagnosis of diseases based on sorting. They have addressed the dynamic purpose of the optimum number of fundamental classifiers making up the random forests. Their proposed technique is different from most of the techniques presented in the literature. They dogged the number of classifiers during the growing procedure of the forest. Their proposed technique produces an ensemble not only accurate but also diverse ensures the two essential properties that distinguish an ensemble classifier. Their technique is derived from online fitting procedure and it is calculated using eight biomedical datasets and five versions of random forest algorithm.

III. EXPERIMENTAL RESULTS

The results for the proposed algorithm are given in this chapter.To analyse the performance of proposed technique following Qualitative parameters are used**:**

a. PRECISION: Precision is how many selected items are relevant. It is a ratio of true positive to the sum of true positive and false positive.

b. RECALL: Recall is how many relevant items are selected. It is a ratio of true positive to the sum of true positive and false negative.

c. ACCURACY: Accuracy predicts the class label correctly and tells how well the value of the predicted attribute for a new data can be guessed.

Fig 1.3 Start page

Fig 1.3 shows the panel where we start for the implementation of the algorithms.This contains two buttons.One is start which helps to start the application and other is quit which helps to quit from the application once we get the accuracy between the existing and the proposed system.

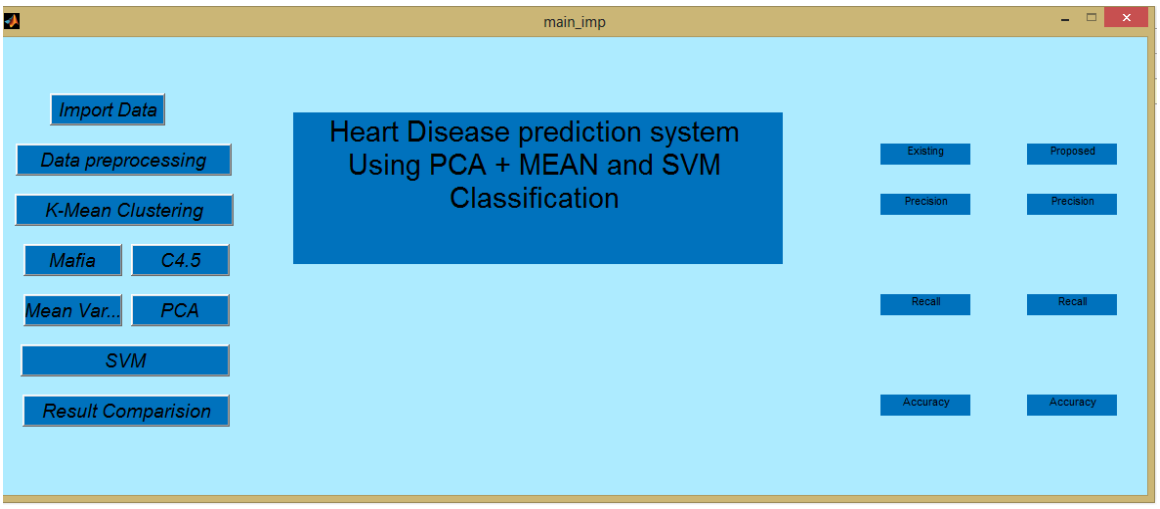

Fig 1.4 Main GUI for the proposed system

The above fig. 1.4 shows different buttons where we move in a given order.

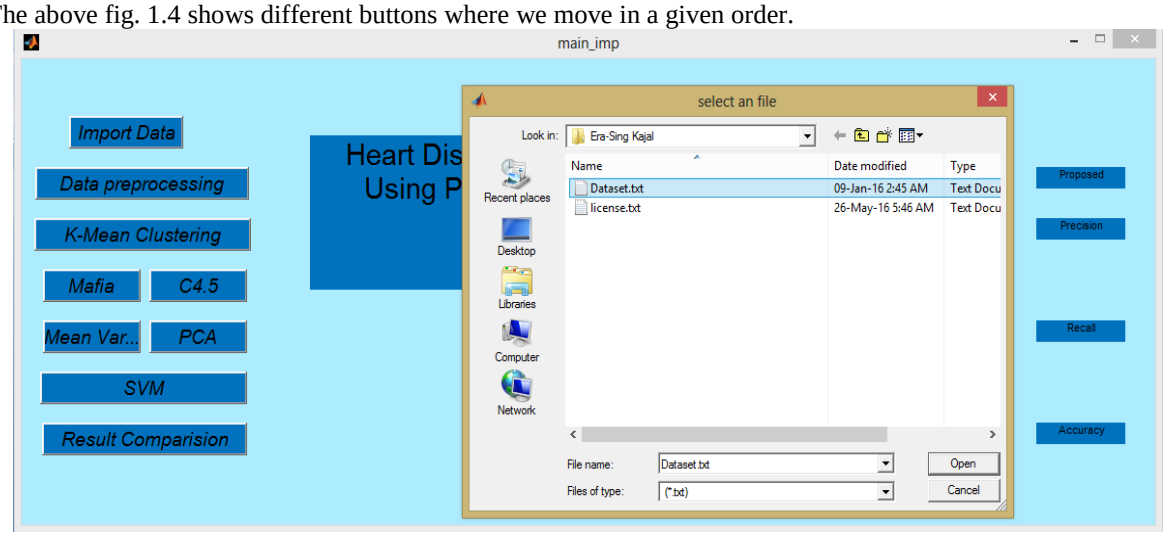

Fig 1.5 Importing data

This is where we import the data set as show in fig 1.5. As soon as the data is imported,thepreprocessing of the data is done.After that we click on various buttons which are meant for different algorithms.

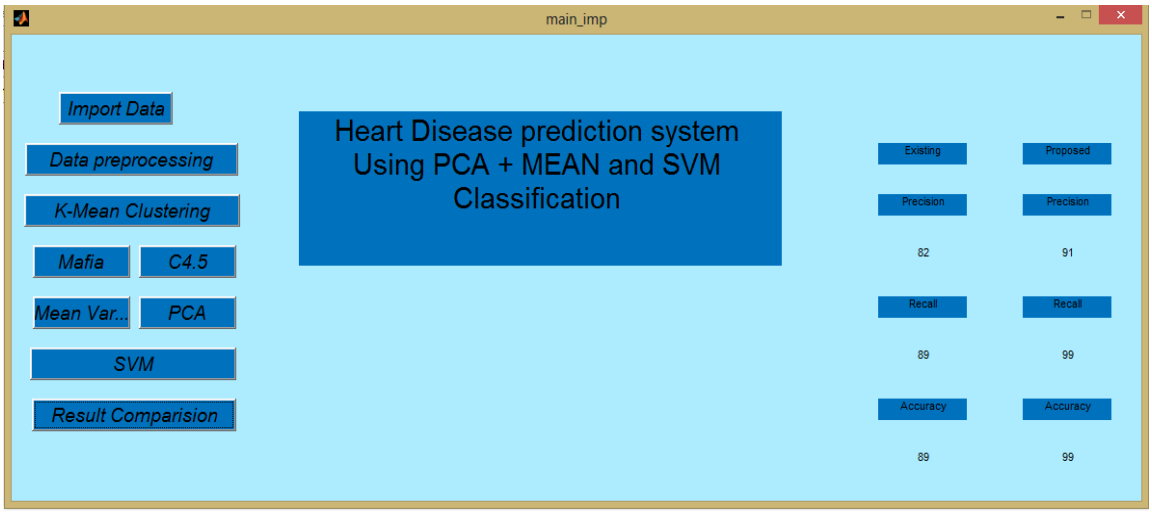

Fig 1.6 Result Comparison

In fig. 1.6 finally when we click on result comparison we get the accuracy between the existing and the proposed system.

DISCUSSIONS

The objective of this proposed work is to have greater accuracy, as high precision and recall metrics.For the implementation of our proposed algorithm I have used Matlab version 2015, with i7 processor with ram of 8gb having processor speed 2.7ghz, for fast optimization of our algorithm we initialized Matlab pool using the 'local' profile, nevertheless as we can see that the output performance of various classifier in figure the performance of improved SVM (support vector machine) Outperformed the rest of the classifier as it give the 100% correct classification rate for no, and 85.4% correct classification rate for yes, and over all classification rate is 94.8% which is the highest accuracy achieved present in the literature.

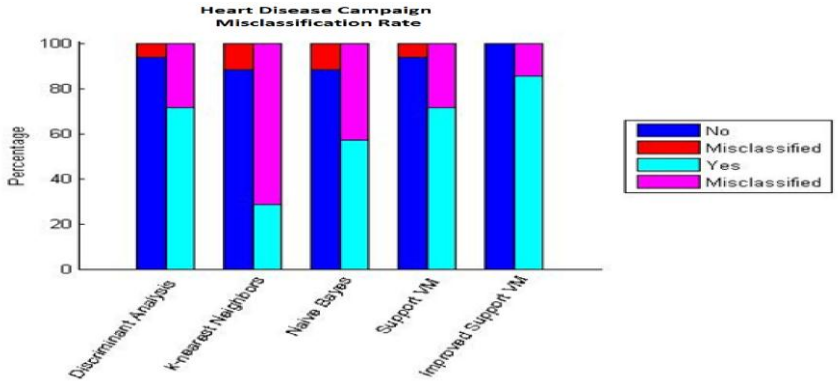

Fig 1.7 Heart Disease Campaign Classification Rate

Table 1.1 Prediction accuracy between various data mining techniques

Data mining technique	Precision	Recall	Accuracy

K-mean based MAFIA	0.78	0.64	0.74
K-mean based MAFIA with ID3	0.8	0.84	0.84
K-mean based MAFIA with ID3 and C4.5	0.82	0.89	0.89
Proposed	0.91	0.99	0.99

Fig 1.8 Prediction of Precision between various data mining techniques

Precision is the fraction of retrieved instances that are relevant to also called positive predictive value. In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query. Above fig 1.8 shows the comparison between the existing and proposed Precision. In the above fig 1.8 it is clear that proposed method is give near about 100% Precision which is very effective as compare with existing.

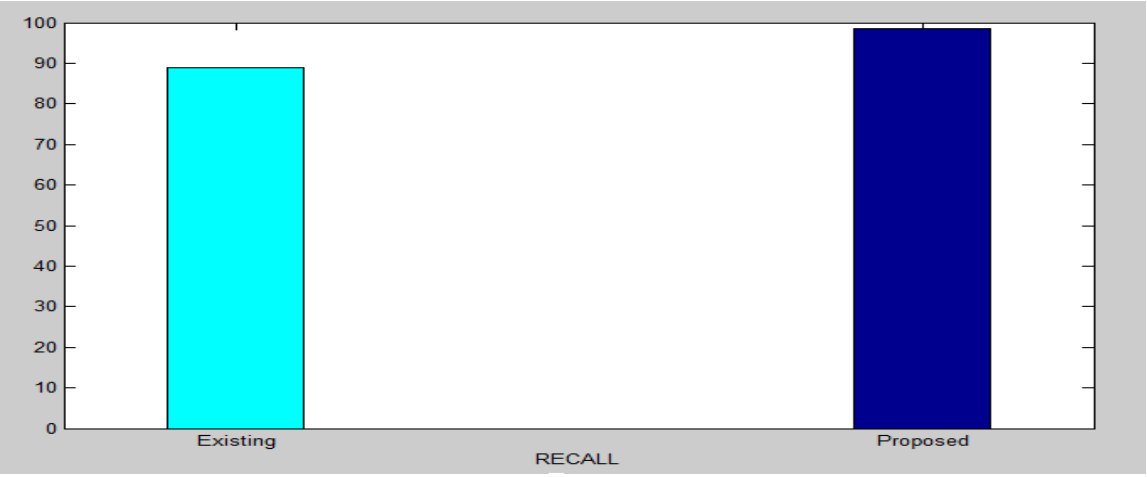

Fig 1.9 Prediction of Recall between various data mining techniques

Recall is the fraction of relevant instances that are retrieved or also called sensitivity. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. Above fig 1.9shows the comparison between the existing and proposed recall. In the fig 1.9it is clear that proposed method is give near about 100% recall which is very effective as compare with existing.

Fig 1.10 Prediction of Accuracy between various data mining techniques

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.Above fig 5.7 shows the comparison between the existing and proposed accuracy. In the fig 1.10 it is clear that proposed method is give 100% accuracy which is very effective as compare with existing.

IV. CONCLUSIONS

Medical related information is highly voluminous in nature in the healthcare industry. It can be derived or retrieved from various sources which are not entirely applicable in this feature.In this work, heart disease prediction system was developed using classification algorithms through Matlab data mining tool to predict effective and better accurate results regarding whether the patient is suffering from heart disease or not. As the heart disease patients are increasing world-wide each year and huge amounts of data is available for research, where different data mining techniques are used in the diagnosis of heart disease. So, different techniques used have shown different accuracies depending upon the number of attributes taken and tool used for implementation.

ACKNOWLEDGMENT

Every success stands as a testimony not only to the hardship but also to hearts behind it. Likewise, the present work has been undertaken and completed with direct and indirect help from many people and I would like to acknowledge all of them for the same.

REFERENCES

[1] DivyaKundra, Er. NavpreetKaur, Review On Prediction System For Heart Diagnosis Using Data Mining Techniques, Ijlret, Volume 1 Issue 5, October 2015, Pp 09-14

[2] V. Ramya And M.Ramakrishnan, Mining Association Rules Using Modified Fp- Growth Algorithm, International Journal For Research In Emerging Science And Technology, 2016

[3] Teh Sin Yin, Data Mining For Robust Tests Of Spread, November 2008

[4] Gaurav Saini, A Survey On Web Research For Data Mining, International Journal Of Advance Research, 2014.

[5] Keyur J Patel, Review of Data Mining: Techniques, Applications and Issues, Paripex, Volume : 3 | Issue : 5 | June 2013

[6] Akin Ozcift and ArifGulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms", Journal of Computer Methods and Programs in Biomedicine, Vol.104, PP.443-451, 2011.

[7] Aqueel Ahmed, Shaikh Abdul Hannan,"Data Mining Techniques to Find Out Heart Diseases", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[8] Bahadur Patel, Ashish Kumar Sen D P, Shamsher Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level", International Journal Of Engineering And Computer Science ISSN: 2319-7242, Page No. 2663-2671, Volume 2 Issue 9 Sept, 2013.

[9] Carlos Ordonez, Edward Omiecinski, Mining Constrained Association Rules to Predict Heart Disease, IEEE. Published in International Conference on Data Mining (ICDM), p. 433- 440, 2001.

[10] Chih-Lin Chi and W. Nick Street, et al, "A decision support system for cost-effective diagnosis", Journal of Artificial Intelligence in Medicine, Vol.50, PP. 149-161, 2010.

[11] Sujata Joshi and Mydhili K. Nair, Prediction of Heart Disease Using Classification Based Data Mining Techniques, 2015.

[12] M.A.NisharaBanu, B.Gomathy, Disease Forecasting System Using Data Mining Methods, International Conference on Intelligent Computing Applications, 2014.

[13] MonaliDey, SiddharthSwarupRautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies(2014).

[14] Abhishek Taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[15] Chitra R and Seenivasagam V, "review of heart disease prediction system using data mining and hybrid intelligent techniques", issn: 2229-6956(online) ictact journal on soft computing, july 2013, volume: 03, issue: 04, 2013.

[16] Indira S. FalDessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, 2013.

[17] IshtakeS.H ,Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research,2013

[18] JesminNahar and Tasadduq Imam et al," Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013.

[19] Aqueel Ahmed, Shaikh Abdul Hannan,"Data Mining Techniques to Find Out Heart Diseases", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[20] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, Vol. 47, No. 10, pp. 0975 – 888, 2012.

[21] Chang-Sik Son and Yoon-Nyun Kim, et al, "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, Vol.45, PP. 999–1008,2012.

[22] Debabrata Pal and K.M. Mandana, et al, "Fuzzy expert system approach for coronary artery disease screening using clinical parameters", journal of knowledge based system, Vol.36, PP.162-174, 2012.

[23] Evanthia E. Tripoliti and Dimitrios I. Fotiadis et al, "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm", Journal of IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 4, July 2012.

[24] Jaya Rama Krishnaiah.V.V.,D.V.ChandraSekhar and K.Ramchand H Rao, "Predicting the Heart attack symptoms using Biomedical data mining techniques", The International Journal of Computer Science & Applications, Volume 1, No. 3, May 2012, pp. 10-18.