HEART DISEASE PREDICTION SYSTEM USING ANOVA, PCA AND SVM CLASSIFICATION

Kiranjeet Kaur

Student, CSE Department

Sri Guru Granth Sahib World University,

Fatehgarh Sahib, Punjab, India.

kkiranjeet99@yahoo.in

Lalit Mann Singh

Assistant Professor, CSE Department

Sri Guru Granth Sahib World University,

Fatehgarh Sahib, Punjab, India.

lalitmann19@gmail.com

Abstract— Heart disease is a term that assigns to a large number of healthcare conditions related to heart. These medical conditions describe the unexpected health conditions that directly control the heart and all its parts. The main objective of this research is to develop an efficient heart disease prediction system using feature extraction and SVM classifier that can be used to predict the occurrence of heart disease. The heart disease prediction system helps the physician and healthcare professionals as a tool for heart disease diagnosis. To protect the life of a patient from heart diseases there have to be quick and efficient prediction technique such as PCA with SVM classification technique is to be followed. This technique is widely used to validate the accuracy of medical data. By providing the effective treatments, it also helps to reduce the treatment costs.

Keywords— KDD, Heart Disease Prediction, Data Mining, Classifiers, PCA, Support Vector Machine.

I. INTRODUCTION

Data mining is a process of discovering/extracting the meaningful information from huge amount of data [4]. The data mining techniques are very beneficial to predicting the various diseases in the healthcare industry. Disease prediction plays most important role in the data mining [5].

Data Mining is a process of discovering interesting patterns and knowledge from huge amount of data. It refers to extracting or mining knowledge from large amount of data. As we are extracting knowledge it is also called knowledge mining from data or knowledge extraction or Knowledge Discovery from Data (KDD). As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge presentation. In this paper, we focuses on fundamental data mining concepts and techniques for discovering interesting patterns from data in healthcare industry.

Our paper's main objective is to learn the various data mining techniques which are used to predict the heart diseases. The purpose of the predictions in data mining is to help for discovering trends in patient data in order to improve their health. Our life is dependent on the proper working of heart because heart is the most important part of human body [4].

If heart's operation is not properly worked, then it will also affect the other parts of the human body [4]. In our daily life, heart diseases are the main cause of deaths.

There are a number of factors that causes the possibility of heart disease can be increased:

- Smoking,

- High cholesterol,

- High blood pressure,

- Use of alcohol,

- lack of physical activities,

- Obesity,

- Hyper tension etc.

Every year due to the heart diseases, World Health Organization (WHO) has estimated 12 million people died world-wide [6]. In 2012, WHO has estimated 17.5 million people died from Cardio Vascular Diseases (CVD).World Health Organization (WHO) estimated by 2030, almost 23.6 million will die due to the cause of heart diseases as written in [4]. So we make a heart disease prediction system using various data mining tools and techniques which were helpful in the healthcare industry and also used for practitioners for accurate heart disease diagnosis. Thus there is a beneficial way to predict the heart disease in the healthcare industry is an effective and efficient Heart Disease Prediction System (HDPS). In the healthcare industry, data mining plays an important role in order to predict the heart diseases using different data mining tools.

Data Mining and its techniques
Data Mining
Data mining is a process of extracting the meaningful information from huge amount of data [5].As we are extracting knowledge it is also called knowledge mining from data or knowledge extraction or Knowledge Discovery from Data (KDD). As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge presentation. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data [6]. The discovered knowledge is used by the healthcare professionals to improve the quality of service and it can also be used by the medical practitioners [9].

Data Mining Techniques
Classification
Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. It is a classic data mining technique which is based on the machine learning. Classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict the class labels for the given data) [9]. In the first step, we build a classification model based on the previous data. In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify the new data [9]. The data classification process:

(1)**Learning Step**: Training data are analysed by a classification algorithm.

(2)**Classification Step**: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Clustering
Cluster analysis is a process of partitioning a set of data objects (observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. Clustering is a task of grouping a set of data objects in such a way that objects in the same group (called clusters) are more similar (in some sense or another) to each other than to those in other groups (clusters) [9]. Clustering is sometimes called automatic classification. Here a critical difference between clustering and classification is that clustering can automatically find the groupings. In machine learning, Classification is known as supervised learning because the class label information is given. Clustering is known as unsupervised learning because the class label information is not present.

Decision Tree
A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The top-most node in a tree is the root node. It is the mostly used data mining technique and its model is quickly and easily understandable.

Prediction

It is a data mining technique which discovers relationship between dependent and independent variables [9].

II.   LITERATURE REVIEW

Abhishek Taneja et al. [1] in the year 2013 used data mining tool WEKA 3.6.4 in heart disease prediction system using J48 technique achieved 95.56% accuracy and using Naïve Bayes technique achieved 92.42% and using Neural Network achieved 94.85%. In this paper, various data mining techniques have been studied for the diagnosis of heart disease. The data for the study collected from PGI, Chandigarh which has a total of 15 attributes. A total of 7339 instances were trained. Of the 15, only 8 attributes were selected.

Mrs. G. Subbalakshmi et al. [2] in the year 2011 has developed a Decision Support in Heart Disease Prediction System (DSHDPS) using Naïve Bayes data mining modelling technique to discover the relationship between variables in database in the healthcare industry. This model could answer the complex queries. It is resulted out as the most effective model in the prediction of heart disease.

Mr Akhil Jabbar et al. [3] in the year 2012 proposed evolutionary algorithm for heart disease prediction. They used the genetic algorithm to predict the heart diseases in Andhra Pradesh population. They used the Association Rule Mining based on the sequence number and clustering for heart attack prediction.

Chaitrali S. Dangare et al. [4] in the year 2012 proposed Improved Study of Heart Disease Prediction System (HDPS) using data mining classification techniques. In this paper, data mining classification is based on a supervised machine learning algorithm. Tangara tool is used to classify the data and evaluated using 10 fold cross validation.

M.A.Nishara Banu and B.Gomathy et al. [5] in the year 2014 used C4.5 algorithm, MAFIA and K-means clustering in the year 2014 using 13 attributes in the data set achieving the 89percent accuracy.

Jyoti Soni et al. [6] in the year 2011 proposed a survey on current data mining techniques like Decision Tree (DT), k-NN, Neural Network, Naïve Bayes and classification via clustering (CVC) is done. Decision tree outperforms all the techniques. Genetic algorithm has been applied to pre-process the data set. After the pre-processing data set, accuracy of the decision tree and Naïve Bayes has been improved. Tangara tool has been used to classify the data and it is evaluated by using 10 fold cross validation.

Shadab Adam Pattekari et al. [7] in the year 2012 developed a prototype Heart Disease Prediction System (HDPS) using data mining modelling techniques, namely, Decision Trees, Naïve Bayes and Neural Networks. It is implemented in the web application. In this system, 15 attributes are used in the dataset for the heart diagnosis in Heart Disease Prediction System (HDPS) and in this user answers the predefined questions. Then it retrieves the hidden data from stored database and it compares the user's values with the trained data set. By providing the effective treatments, it also helps to reduce the treatment costs.

Sellappan Palaniappan et al. [8] in the year 2008 proposed a model of Intelligent Heart Disease Prediction System (IHDPS) that was built with the aid of Neural Network and data mining techniques namely, Decision Trees, Naïve Bayes was proposed. Intelligent Heart Disease Prediction System (IHDPS) is web-based, reliable, scalable, user-friendly and expandable.

M. Anbarasi et al. [9] in the year 2010 proposed Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. Originally 13 attributes involved in prediction of heart disease, proposed enhanced prediction of heart disease with feature subset selection using genetic algorithm using 10 attributes for predicting and data mining techniques after incorporating feature subset selection with high model construction time. Classification techniques are Naïve Bayes, Decision Tree and Classification by clustering. The genetic search starts with zero attributes, and an initial population with randomly generated rules. The dataset contains records with 13 attributes in each record.

Aqueel Ahmed et al. [10] in the year 2012 show the classification techniques in data mining and show the performance of classification among them. In this classification accuracy among these data mining tools has discussed. Aqueel Ahmed et al. used 11 attributes namely, age, sex, blood pressure, Cholesterol, Fasting blood sugar and chest pain type. In this paper data mining techniques namely, decision tree and SVM performs classification more accurately than the other methods and able to achieve the 91% accuracy.
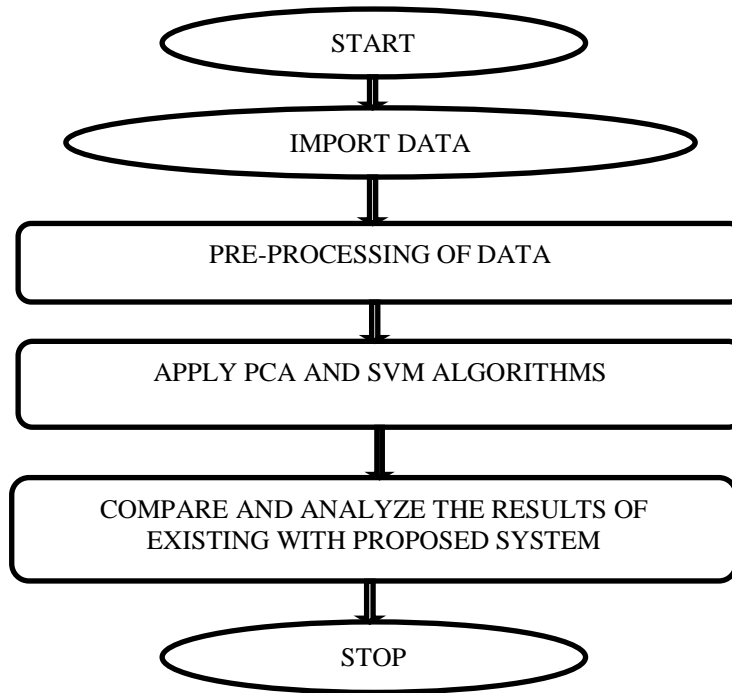
III.   PROPOSED WORK

Fig. 1- Flow diagram of proposed system

IV.   EXPERIMENTAL RESULTS

We have implemented our proposed algorithm in Matlab version 2014a, over the operating system windows 10, with a processor speed 2.5 GHz. In table 1 we show the attribute we have taken into consideration and results shown in table 2 from the results we can easily analyse that the proposed PCA and SVM classification algorithm outperformed the rest of the algorithm present in the literature with our proposed algorithm we have achieved an accuracy of 99.6%.

Table 1- Heart disease attributes

Id.	Attribute
1.	Age
2.	Blood cholesterol
3.	Blood pressure
4.	Hereditary
5.	Smoking
6.	Alcohol intake
7.	Physical activity
8.	Diabetes
9.	Diet

Id.	Attribute
10.	Obesity
11.	Stress
12.	Sex

Table 2- Prediction accuracy between data mining techniques

Data Mining	Precision	Recall	Accuracy
K-mean based MAFIA	78	64	74%
K-mean based MAFIA with ID3	80	84	84%
K-mean based MAFIA with ID3 and C4.5	82	89	89%
Proposed	97.3	98.54	99.6%

In the fig. 2, we show the comparison of accuracy in the existing system and proposed system. In this fig., the existing system achieves an accuracy of 89%, but in the proposed system we have achieved an accuracy of 99.6%.
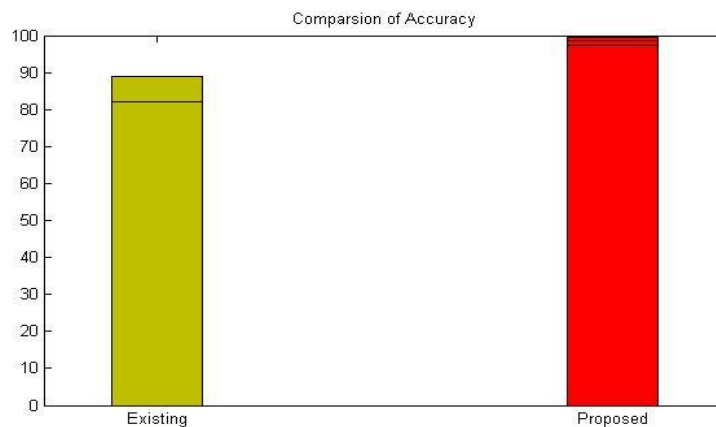

Fig. 2- Accuracy of existing system and proposed system

V.   CONCLUSION

We have analysed the heart disease prediction on the basis of data mining techniques. From the above study it has been observed that PCA AND SVM classification perform well for Heart disease prediction. Also the accuracy for PCA with REP classifier is good but in terms of accuracy PCA with SVM performs better than other classifier.

In future, we can use other classifier techniques with PCA to improve the accuracy and testing time of result. Or we can use combination of other classifier techniques with other dimension reduction techniques to improve the accuracy and testing time of result.

REFERENCES

1. Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques", Oriental Scientific Publishing Co., India, 2013.

2. G. Subblakshmi, "Decision Support in Heart Disease Prediction System Using Naïve Bayes", Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976 – 5166, Vol.2, No.2, April - May 2011.

3. M. Akhil jabbar, Dr. Priti Chandra, Dr. B.L Deekshatulu, "Heart Disease Prediction System Using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies", 2012.

4. Chaitrali S. Dangre and Dr. Mrs. Sulbha 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, June 2012.

5. M.A. Nishara Banu and B.Gomathy,"Disease Forecasting System Using Data Mining Methods", International Conference on Intelligent Computing Applications, 2014.

6. Jyoti Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, Vol.17, No.8, March 2011.

7. Shadab Adam Pattekari and Asma Parveen, "Prediction System for Heart Disease Prediction System Using Naïve Bayes", International Journal of Advanced Computer and Mathematical Sciences", ISSN: 2230 – 9624, Volume-3, Issue-3, pp.290– 294, 2012.

8. Sellappan Palanappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8, No.8, August 2008.

9. M. Anbarasi, E. Anupriya, N.ch.s.n.Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology,2010.

10. 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.

11. Cleveland dataset from http://archive.ics.uci.edu.