Heart Disease Prediction Model Using Hybrid Classification

Harjinder Kaur
Research Scholar
Swami Vivekanand Institute of Engineering & Technology

Dr. Parteek garg
Head of Department
Swami Vivekanand Institute of Engineering & Technology

ABSTRACT - The data mining is the technique which can mine useful information from the rough data. The prediction analysis is the technique which can predict future possibilities based on the current information. This research work is based on the heart disease prediction. The heart disease prediction has various steps like pre-processing, feature extraction and classification. The hybrid technique is designed in this research work based on random forest and logistic regression. The random forest classifier is applied for the feature extraction and logistic regression is applied for the classification. The performance of proposed model is analyzed in terms of accuracy, precision and recall. It is analyzed that proposed model achieved accuracy upto 95 percent or the heart disease prediction

Keywords - Heart Disease prediction, MLP, Decision Tree, Naïve Bayes, Random forest, Logistic Regression

I. INTRODUCTION

The data is analyzed using the technology known as data mining in which the patterns are identified from data set using different data mining tools and techniques. By ensuring the least user input and efforts, the patterns are identified automatically through data mining. To handle decision making and forecast the future market trends, data mining is considered to be as a powerful tool. In different applications the data mining tools and techniques have been applied successfully. For dealing with the competitive environment in order to perform data analysis, several organizations have started using data mining commonly [1]. Various trends and patterns of market can easily be evaluated and quick and effective market trend analysis can be produced with the help of applying mining tools and techniques in various business applications. Following are the various techniques commonly used in data mining:

a. Association: One of the best known data mining technique in which the relationship among particular items of similar transaction is used to discover a certain pattern is known as association. For instance, the relationship of various attributes used for analysis in heart disease prediction is known through association technique. All the risk factors needed for disease prediction are used to sort out the patients suffering from such heart disease [2].

b. Classification: On the basis of machine learning, another classic data mining technique designed is classification. Every object in the data set is categorized into one of the predefined set of classes through classification. The various mathematical techniques are used in this method.

c. Clustering: The objects with similar property are clustered together to generate a meaningful cluster using an automatic approach known as clustering. The classes are defined by the clustering technique as well and the objects are placed in them. Further, in the predefined classes, the classification objects are assigned. For instance, it is possible to cluster the list of patients with similar risk factors using clustering when predicting heart disease. Therefore, the patients with high blood sugar and relevant risk factors can be separated [3].

d. Prediction: The relation among independent variables as well as dependent and independent variables are discovered by another data mining technique called prediction. For example, if the sale is considered as independent variable and profit is considered as a dependent variable in sales, the profit for future can be predicted by prediction analysis technique. Further, a fitted regression curve can be drawn for profit prediction on the basis of provided historical sale and profit data.

1.1. Prediction Analysis in Data Mining

For analyzing the historical data and information in efficient manner, various statistical trends and techniques ranging from machine learning and predictive modeling to data mining are used in prediction analysis technique. The predictions related

to any unknown future events are generated through this method [4]. For identifying any kinds of risks and opportunities, the prediction analysis helps in exploiting the patterns of historical business data depending upon the business aspect. For providing risk assessment or identifying any kind of potential threat, the relationships among various factors are captured. The important decision making steps are used to help in guiding the business. Referring to the prediction modeling and forecasting, the prediction analysis is defined sometimes. To perform forecasting, there are various prediction analysis models designed over the years. The three broader categorizations of these models are explained below:

a. Predictive Models: The relationship among various features present in the collected data is identified using the predictive models. The similarities among a group of units are assessed by this model. The presence of similar attributes that are being exhibited by a group of similar units is assured here [5].

b. Descriptive Models: The relationships among different attributes of unit are identified and quantified using these models. Then, the models are used for classifying these attributes into certain groups. With the ability of comparing and predicting the data as per the relationship existing among multiple behaviors of units, this model is different from other models.

c. Decision Models: The relationship among all the various data elements existing in the known dataset is identified and described through the decision models [6]. The known dataset on which the model is to be defined, the decision structure defined for categorization of known and predicted result and its classification are performed in this model. Depending upon the multiple attributes or features of dataset, the results of decisions are identified and predicted here [7].

II. LITERATURE REVIEW

Anjan Nikhil Repaka, et.al (2019) studied that there are several sources that are responsible for any kind of heart disease and data is collected from such sources [8]. This results in constructing the structure of database. To resolve the heart disease prediction related issues, the NB (Naive Bayesian) classification and AES (Advanced Encryption Standard) algorithm are considered which design the SHDP (Smart Heart Disease Prediction). An accuracy of around 89% is shown by the proposed approach in comparison to the existing Naïve Bayes approach which shown its level of improvement. As compared to other approaches, AES provided high security performance evaluation as well.

Aditi Gavhane, et.al (2018) proposed an application through which the basic symptoms could be used to predict the vulnerability of a heart disease [9]. The accuracy and reliability of the machine learning algorithm known as neural networks was seen to the highest. Therefore, the proposed approach used this mechanism. The users were provided with a prediction result that had the state of a user the leading to CAD through MLP which was another machine learning algorithm. There has been a huge evolvement in the machine learning algorithms due to their recent advancements. Therefore, with its higher efficiency and accuracy, the proposed system used Multi Layered Perceptron (MLP). Depending upon the input provided by the users, nearly reliable output was given by the proposed algorithm. The awareness related to current heart status is increased with the increment in number of people using such systems. Thus, there will be reduction in number of people suffering from heart diseases.

Aakash Chauhan, et.al (2018) studied that in India, the number of people suffering from cardiovascular diseases is increasing [10]. The major cause of death in India in the upcoming years is predicted to be the coronary disease. So, reducing its impact is very important. Therefore, to identify the risk of heart disease in highly accurate manner, a heart disease prediction system was proposed in this research. A new system for heart disease prediction was designed using the data mining techniques. The frequent pattern growth association mining was applied on the dataset of patients to provide strong association rules. The data could be explored and the heart disease could be predicted accurately by the doctors using this proposed method.

C. Sowmiya, et.al (2017) proposed the evaluation of heart disease prediction by analyzing the potential of nine different classification techniques. There classification algorithms were used most commonly in various research studies [11]. To predict heart disease, this research focused on adapting the SVM and apriori algorithms. The medical profiles based on various factors were collected and used here. The patients that were more likely to get heart disease were predicted here. To detect and prevent heart disease, the medical society took

partial interest in this research. The research concluded that huge effectiveness and accuracy was achieved in comparison to the previous techniques, when applying proposed method.

Rashmi G Saboji, (2017) proposed a new framework in which the heart disease was predicted depending upon certain attributes by using the healthcare data [12]. Predicting the diagnosis of heart disease using small number of attributes was the major contribution of this research. The random forest was used on Apache Spark to provide a prediction solution. For deploying this solution on highly scalable landscape for insightful decision making, the proposed approach provided huge opportunity to the health care analysts. Around 98% of accuracy was achieved by applying this approach. In comparison to Naïve-Bayes classifier, the proposed approach using random forest classifier provided better outcomes.

III. RESEARCH METHODOLOGY

In human beings, the most important muscular organ known is heart. The blood is pumped through the blood vessels of circulatory system by the heart. So, every individual person's life completely depends on the heart. All the other organs of human body are also affected if the heart is affected by any kinds of disease. The computer based information is extracted from large sized databases using data mining. Several organizations have been using the data mining tools and techniques. In the healthcare field, the data mining tools are used for predicting the diseases. There are around 12 million people suffering from heart diseases are per the WHO reports. The maual records of heart patients are recorded in details in the medical organizations. Only electronic records are needed by the medical practitioner. Converting the data mining techniques into manual records is very easy for the data mining techniques. There are several risk factors based on which the heart diseases can occur in patients.

Following are the various phases of heart disease prediction:-

A. Data Acquisition: The data is collected from various clinical organizations to perform experiments.

B. Data preprocessing: For applying machine learning techniques such that completeness can be introduced and a meaningful analysis can be achieved on the data, the data preprocessing is performed. Initially, a numerical cleaner filter is used to mark the missing values in the data. By setting them to a defined default value, the numeric data that is too big or small is cleaned. A filter is then used to mark and detect the missing values and then replace them with the data distribution's mean value [6]. The performance of training model is improved by providing a clean and noise free data for the feature selection process and removing the irrelevant features from the dataset.

C. Feature selection: A subset of highly distinguished features is picked by feature selection to diagnose the disease. The discriminating features that belong to the available classes are selected by feature selection process. There are two phases of feature selection process. The attribute evaluator technique through which the features of dataset are evaluated based on the output class is the initial phase. The search method in which various combinations of features help in selecting an optimal set for classification problem is the second phase. In the proposed method, the random forest model is applied for the feature selection. The random forest model takes 100 as the estimator value and generates tree structure of the most relevant features. The random forest model selects the features which are most relevant or important for the heart disease prediction.

D. Classification: To categorize the given features for performing disease prediction, the selected features are mapped to the training model. As a multi-class problem, the classification is formulated and then among the four various classes, the clinical data is categorized. Here, a kind of heart disease is represented by each separate class. The logistic regression model is applied for the classification. The logistic regression takes input of the extracted features. The logistic classifier is the probability based classification model which calculates the probability and based on the probability data can be classified into certain classes. In the research work, two classes are defined which are heart disease and no heart disease. It means that which persons have probability of heart disease and which don't have probability of heart disease. The logistic regression takes input of the extracted features. In the research work, two classes are defined which are heart disease and no heart disease. It is a kind of regression which is capable for forecasting the probability of happening of an event for which data is robust to a logistic function. Like the various types of regression analysis, various predictive variables which are either numerical or categorical utilize in LR. The hypothesis of LR is described as:

$$h_\theta(x) = g(\theta^T x)$$

In which the function g denotes the sigmoid function which is expressed as:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The sigmoid function contains special properties which provide the values in range [0,1]. The cost function for LR is defined as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)} \log\left(h_e\left(x^{(i)}\right)\right) - \left(1 - y^{(i)}\right) \log(1 - h_e\left(x^{(i)}\right))]$$

The minimum of this cost function is discovered in ML using a built-in function known as $fmin\_bfgs^2$, suing which the best parameters θ is found for the LR cost function provides a fixed dataset that has x and y as values. It means that which persons have probability of heart disease and which don't have probability of heart disease.

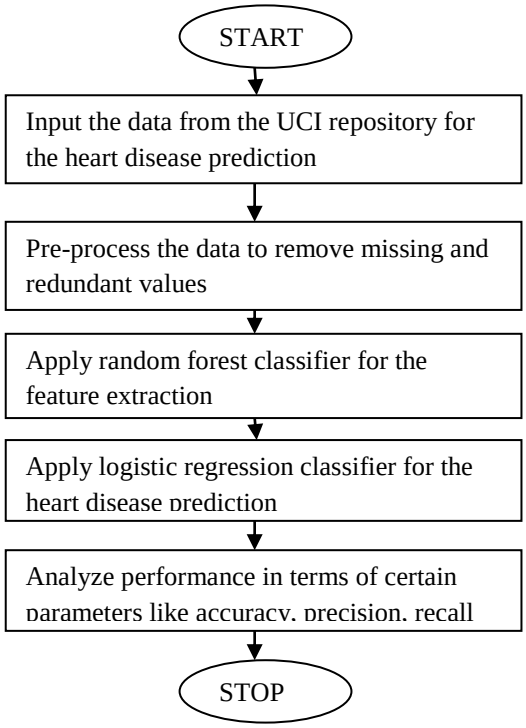

Figure 1: Proposed Methodology

IV. RESULT AND DISCUSION

The chevaland dataset has been widely used for the heart disease prediction. This dataset has the 76 attributed but for the experiment purpose only 14 attributes are used widely. The 14 attributes are age, sex, cp, trestbps, chol. Fbs, restecg, thalach, exang, oldpeak,slope,ca,thal,num and predicted attribute.

In this research work, various models are implemented and compared for the heart disease prediction. The decision tree, naïve bayes, Multilayer perception, Ensemble classification method which is combination of random forest, naïve bayes, baysian belief models, proposed models are compared in terms of accuracy, precision and recall

Table 1: Accuracy Analysis

Models	Accuracy
Decision Tree	75.41 percent
Naïve Bayes	83.61 percent
Multilayer perceptron	83.61 percent
Ensemble Method	85.25 percent
Proposed Method	95.08 percent

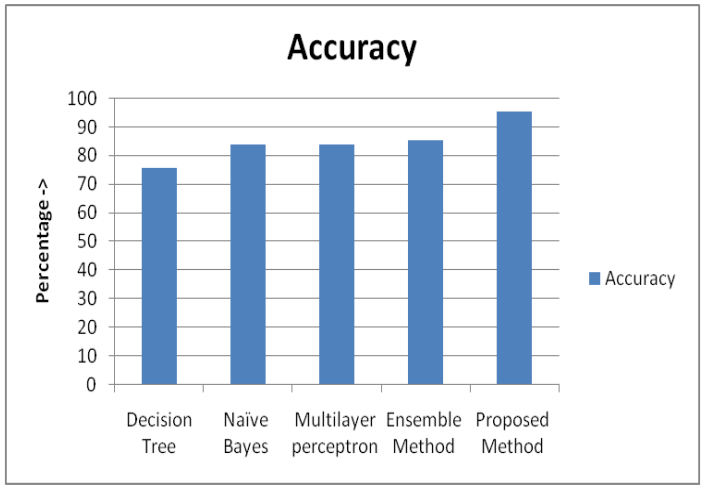

Figure 2: Accuracy Analysis

As shown in figure 2, the various models like decision tree, naïve bayes, multilayer perceptron, ensemble and proposed models are compared in terms of accuracy. It is analyzed that accuracy of proposed model is approx 95 percent which is

maximum as compared to other models for the heart disease prediction.

Table 2: Precision Analysis

Models	Precision
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	85 percent
Ensemble Method	86 percent
Proposed Method	95 percent

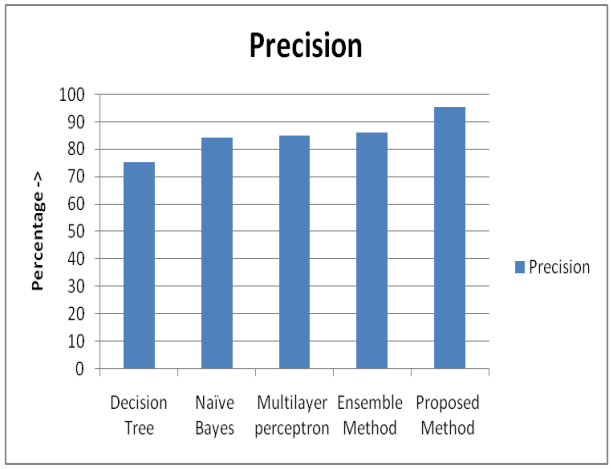

Figure 3: Precision Analysis

As shown in figure 3, the various models like decision tree, naïve bayes, multilayer perceptron, ensemble and proposed models are compared in terms of precision. It is analyzed that precision of proposed model is approx 95 percent which is maximum as compared to other models for the heart disease prediction.

Table 3: Recall Analysis

Models	Precision
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	84 percent
Ensemble Method	85 percent
Proposed Method	95 percent

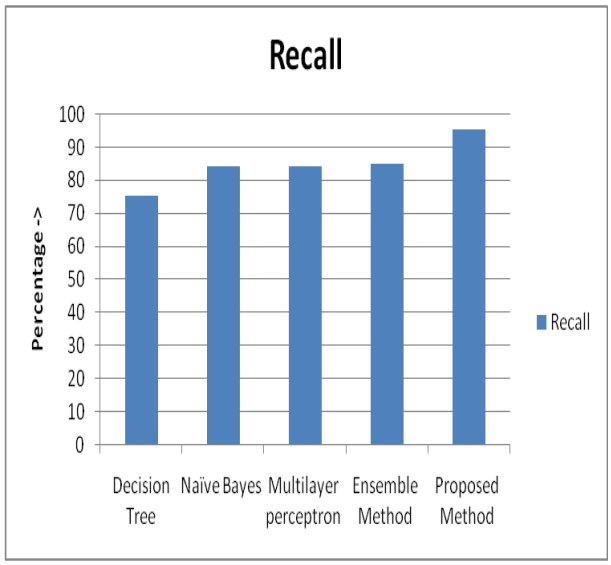

Figure 4: Recall Analysis

As shown in figure 4, the various models like decision tree, naïve bayes, multilayer perceptron, ensemble and proposed models are compared in terms of recall. It is analyzed that recall of proposed model is approx 95 percent which is maximum as compared to other models for the heart disease prediction.

V. CONCLUSION

In this work, it is concluded that heart disease prediction is the major challenge due large number of attributes. The various models are tested for the heart disease prediction like decision tree, naïve bayes, multilayer perceptron, ensemble classifier. The novel model is designed for the heart disease prediction which is the combination of random forest and logistic regression. The random forest is applied for the feature extraction and logistic regression is used for the classification. The precision, recall and accuracy of the proposed model is achieved upto 95 percent.

VI. REFERENCES

[1] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.

[2] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival Causes After Out of

Hospital Cardiac Arrest using Data Mining Method", Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

[3] W.J. Frawley and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[4] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[5] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[6] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[7] Niti Guru, Anil Dahiya and Navin Rajpal, "Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[8] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian", 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)

[9] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning", 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

[10] Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", 2018, 4th International Conference on Computational Intelligence & Communication Technology (CICT)

[11] C. Sowmiya, P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques", 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)

[12] Rashmi G Saboji, "A scalable solution for heart disease prediction using classification mining technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

[13] Ankita Dewan, Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning", 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

[15] M. A. Jabbar, Shirina Samreen, "Heart disease prediction system based on hidden naïve bayes classifier", 2016 International Conference on Circuits, Controls, Communications and Computing (I4C)

[16] Purushottam, Kanak Saxena, Richa Sharma, "Efficient heart disease prediction system using decision tree", 2015, International Conference on Computing, Communication & Automation

[17] Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", 2018, 4th International Conference on Computational Intelligence & Communication Technology (CICT)

[18] C. Sowmiya, P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques", 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)

[19] Rashmi G Saboji, "A scalable solution for heart disease prediction using classification mining technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)