INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 4)
Available online at: www.ijariit.com

A review paper on: Heart disease data set analysis using data mining classification techniques

Shreya Kalta
kaltashreya11@gmail.com
AP Goyal Shimla University, Shimla, Himachal Pradesh

Keshav Kishore
mails4keshav@gmail.com
AP Goyal Shimla University, Shimla, Himachal Pradesh

Aman Kumar
aman11304832@gmail.com
AP Goyal Shimla University, Shimla, Himachal Pradesh

ABSTRACT

The health care industry is one of the fastest growing industries in the 21st century. This is the era of increasing health problems and chronic diseases. The major chronic diseases faced the world over are cardio vascular diseases such as stroke and heart attacks. Heart disease is one of the common causes of death worldwide. According to WHO as many as, 17.9 Million people die of Cardio Vascular Diseases each year, 31% of all the deaths worldwide. Diagnosis of the disease is one of the most important tasks of medical science. Medical professionals need a decision support system for early prediction of heart diseases with good accuracy rate which can be achieved with the help of data mining techniques. The healthcare industry produces a large amount of data each day. Data mining helps in extracting hidden information and patterns from a large and complex database which is helpful in making decisions. The main objective of this research is to develop a heart disease prediction system by using data mining techniques with a good accuracy rate. Here we have a preprocessed data set consisting of 303 records and 14 predictors such as Gender, blood pressure, chest pain type, etc. as input for BPN and Decision Tree. In this research, we will compare two data mining algorithms: Decision tree and Back propagation network Algorithm and predict the presence or absence of heart disease in a patient. The algorithm with the highest accuracy rate will be considered good for heart disease prediction in hospitals.

Keywords— *Heart disease, Data mining, Decision tree, Back Propagation Network*

1. INTRODUCTION

There is the explosive growth of data from terabytes to petabytes, which indicates that "we are drowning in data, but starving for knowledge" a quote by John Naisbitt, Megatrends. Data mining is a technique used to extract knowledge and patterns from large databases consisting of past records used for future analysis and prediction. Data mining is used as an analysing tool in the process of Knowledge Discovery. The KDD process helps in exploring the facts and data in a pre-existing database which is helpful in generating new information for future purpose. The exposed knowledge can be used by the healthcare practitioners to get better quality of services in hospitals and affordability such as reduce the cost of medical tests. This can be achieved by installing a suitable decision support system [1]. Healthcare industry produces data in bulk and consists of treatment records of millions of patients [1]. It is complex data which consists of resource organization, disease diagnosis, patient records, etc. and is difficult to handle. Various data mining techniques have been used to handle the undiscovered patterns in the medical database. Doctor's knowledge and experience are helpful in designing data mining algorithms and are useful in carrying patient's data [2]. Accuracy of the model, time efficiency and prediction is the foremost concern of designing a model for heart disease prediction using predictive methodologies.

2. HEART DISEASE

The heart is one of the major organs of the human body. Our lives are completely dependent on the efficient working of the heart if the heart stops functioning it can lead to problems such as death. Heart disease is one such disease that affects the normal functionality of the heart. The term heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being [3]. Every age group even the young generation are getting affected by heart diseases [3]. Heart Disease can be discovered by some symptoms and signs indicated by the patient body. Symptoms of heart disease can be:

- Chest pain, Nausea, Indigestion, Vomiting
- Heartburn (choking feeling) and Stomach Pain
- Feeling Dizzy and Light Headed,
- Fatigue, Sweating, Pain in the Arms or below the breastbone
- Weakness or shortness of breath, irregular heartbeats

Major risk factors of heart disease

- Family history of heart disease
- Smoking, consumption of alcohol
- Use of tobacco
- High blood Cholesterol
- Poor diet
- High blood pressure
- Obesity
- Physical inactivity
- Hyper tension

The heart is one broad term. There are various components of a heart and the types of heart diseases affecting these components are also of different types. Such as stated below:

Table 1: Different types of heart disease [5]

Arrhythmia	The heart beat is improper whether it may irregular, too slow or too fast.
Cardiac arrest	An unexpected loss of heart function, consciousness and breathing occur suddenly.
Congestive heart failure	The heart does not pump blood as well as it should, it is the condition of chronic.
Congenital heart disease	The heart’s abnormality develops before birth.
Coronary artery disease	The heart’s major blood vessels can damage or any disease occurs in the blood vessels.
High Blood Pressure	It has a condition that the force of the blood against the artery walls is too high.
Peripheral artery disease	The narrowed blood vessels which reduce the flow of blood in the limbs is the circulatory condition.
Stroke	Interruption of blood supply occurs damage to the brain.

Considering the death rate due to heart disease worldwide doctors need a decision support system for early diagnosis and accurate prediction of heart disease which has become a major task. The limitation with Medical professionals is that they can predict the probability of heart attack with an accuracy of up to 67% only. The below figure 1, illustrates the difficulties that will be arriving during diagnosis which leads to negative presumptions and unpredictable effects.

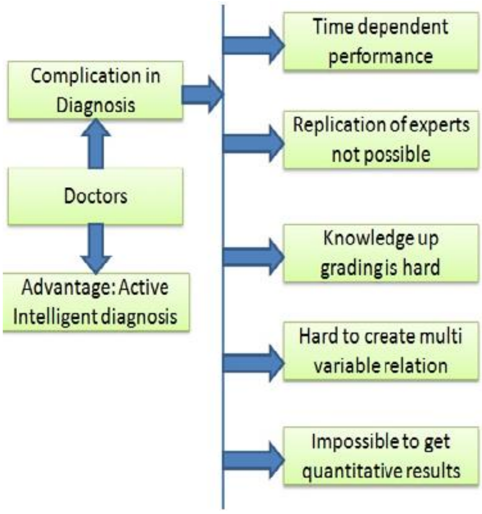
Complication in Diagnosis
Doctors
Advantage: Active Intelligent diagnosis
Time dependent performance
Replication of experts not possible
Knowledge up grading is hard
Hard to create multi variable relation
Impossible to get quantitative results

Fig. 1: Complexity in Diagnoses with Doctor

Data mining techniques are used to analyze heart related issues. Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease [6]. Different data mining techniques are used to analyze heart related issues such as decision tree, neural networks, random forest, KNN, etc. Using some data mining techniques heart disease prediction can be made simple by using various characteristic to find out whether the person suffers from a heart attack or not, and it also takes less time for the prediction and improve the medical diagnosis of diseases With good accuracy and minimizes the occurrence of heart attack [6].

3. RELATED WORK

Numerous researchers have been studying and focusing on methods to develop such systems that can predict and diagnose heart disease at an early stage using different data mining techniques for heart disease prediction with the help of different parameters and have achieved results with good accuracy. We have cited and reviewed many previously published research articles and drew a performance matrix based on their research work and approaches.

Amandeep Kaur and Jyoti Arora (April 2018) [13] of Desh Bhagat University, Punjab conducted a survey and analyzed different classification algorithms for heart disease prediction. They showed the relevance of Data Mining in Medical field and constructed steps towards applicable strategies in Disease Prediction. They studied various research works done by different people with some captivating procedures.

Poornima Singh, Sanjay Singh and Gayatri S Pandi Jain (Nov 2018) [14] Designed an effective heart disease prediction system (EHDPS) using data mining techniques. Two algorithms from the field of Artificial Neural Network, MLPNN (multilayer perception neural network) along with backpropogation algorithm to develop the EHDPS system and they took 15 medical parameters that can help in early prediction of the risk level of heart disease. The MLPNN model shows better result and assists the medical field experts in early diagnoses of the patient with heart disease. The experiment proves that by using neural networks the designed system can predict heart disease with 100% accuracy.

Abhishek Rairikar, Vedant Kulkarni, Vikas Sabole and Anuradha Lamugunde (2017) [15] used three data mining data mining classification techniques namely Decision trees, Naive Bayes and KNN methods to predict heart disease with increased accuracy. They proposed a system that uses 13 attributes such as gender, blood pressure, cholesterol, etc. and also extracted and fragmented forms from heart attack data warehouse using a methodical approach to foresee the probability of a patient having a heart attack.

Tulay Karayilan, Ozkan (2017) [16] proposed a system in which they designed a Multilayer Perceptron Neural Network using 13 clinical data which were obtained from Cleveland dataset. They used the Back-Propagation Algorithm in this system to predict whether the patient has heart disease or not. The accuracy rate obtained from the study is 95%.

Sanjay Kumar Sen (June 2017) [17] used four classifiers to test their predictive performance on the same dataset from the UCI machine learning repository. He concluded that Naïve bayes classifier is the best out of Support Vector Machine, Decision Tree, and K-Nearest Neighbor if had minimum misclassification and highly correct classification rate. However, their performance was found to be closely competitive with a very slight difference.

C. Sowmiya and Dr. P. Sumitra (2017) [18] analyzed different classification techniques such as decision tree, Naïve Bayesian, Neural network, SVM, KNN usage in data mining to detect heart disease. Further, they proposed an algorithm Apriori algorithm and SVM for heart disease prediction. They analyzed all types of diagnosis for heart disease and proved that classification based techniques were more effective and accurate than previous ones.

Bhavini Bhatia and Vamika Razdan (2017) [19] proved that the decision tree method is more accurate than KNN and can be used in heart disease prediction analysis. They observed that the classification method of data mining techniques is highly efficient and accurate for knowledge discovery. Further for normalization of data they used Z-score, Min-Max, and Decimal Scaling and found that Min-Max approach is most accurate.

4. DATA MINING TECHNIQUES USED FOR PREDICTION

4.1 Back propagation networks

An Artificial Neural Network (ANN) is a computational model based on the structure and functions of biological neural networks such as the human brain. ANN's are considered nonlinear statistical data modeling tools where the complex relationships and patterns between inputs and outputs [9]. In ANN all the layers are interconnected and they consist of neurons. There are three layers i.e. Input, Hidden, Output layer which arrange the Neurons. The input layer consists of input values associated with weights and a function that sums up the weights; at the hidden layer, the functions are taking place. The hidden neurons are able to learn the pattern in data during the training phase and mapping the relationship between input and output pairs [6] and there are several unseen layers present in a neural network. Each neuron in the hidden layer uses a transfer function to process data it receives from input layer and then transfers the processed information to the output neurons for further processing using a transfer function in each neuron [6]. The final layer is the output layer which produces output, one node for which class [10]. There is a connection between each layer & weights are assigned to each connection. The primary function of neurons of the input layer is to divide input xi into neurons in the hidden layer. The neuron of a hidden layer adds input signal xi with weights wji of respective connections from the input layer. The output Yj is a function of $Yj = f (\Sigma wji xi)$ Where f is a simple threshold function such as sigmoid [6].

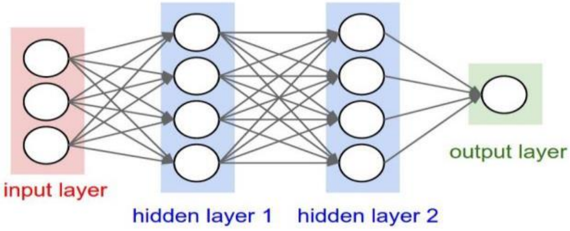

Fig. 2: Artificial Neural Network

Perceptron is the mathematical model of the neuron. It is a single neuron connected to inputs and it has only one output. It is used to classify linear problems. It has two parts sum function and transfer function. Figure 3 Perceptron Model showing x1, x2, x3 inputs with corresponding weights w1, w2, w3. There are two parts of the perceptron one is sum function and another one is step function as the activation function [11].

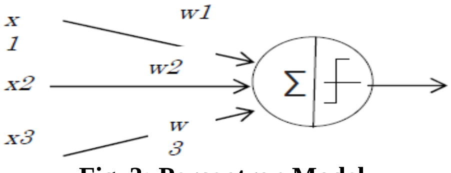

Fig. 3: Perceptron Model

Back propagation algorithm is a procedure for training multilayer Artificial Neural Network. Back propagation maps the input to the best output. This calculates the slope of a loss function of the optimization method. Optimization method means the choice of best elements from many essentials. To compute the loss gradient function the output of each input value should be known. Loss function maps an event or values of one or more variable to a real number [10]. Back propagation algorithm looks for the minimum of error function using the method of gradient descent. The combination of weights which minimizes the error function is considered to be a solution to the learning problem. The output value of the multilayered feed forward neural network algorithm is then calculated using back propagation neural network algorithm to minimize the error function [12]. While propagating backwards, the back propagation algorithm makes sure to change the weight values of the neurons according to the error gradient descent function. The weight values are recalculated until the input layer is reached [12].

4.2 Decision trees

The decision tree is a hierarchical tree like structure used for building classification models and decision analysis. The goal of the decision tree is to predict the value of a target variable by creating a model which uses several input variables. Such tree like structure makes it simple to debug and handle. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in trees [7]. The root nodes are the top most node, leaf nodes represent a class label and the outcome of the test is represented by branches of the tree. From source nodes to leaf nodes a path is constructed, that holds the class prediction for that tuple [10]. A Decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one and by following these rules, appropriate conclusions can be reached [9].

In the medical field, decision trees determine the sequence of attributes. First, it produces a set of solved cases. Then the whole set is divided into a training set and testing set, where a training set is used for the induction of a decision tree. While the testing set is used to find the accuracy of an obtained solution [8]. The decision tree has shown good accuracy in heart disease diagnosis.
Decision Tree is easy to recognize and it can handle both numeric and categorical dataset with good accuracy as it follows a greedy approach (top-down). It is a supervised learning method. It can represent complex decision boundaries and can handle multidimensional data. Decision tree uses different algorithms like ID3 (Iterative Dichotomizer), C4.5 (successor of ID3), CART (Classification and Regression Tree) [10]. For attribute selection, some measures are used such as Information Gain, Gain Ratio

The algorithm for the decision tree is given below [7]:

Step 1: Identify the information gain for the attributes in the dataset.
Step 2: Sort the information gain for the heart disease datasets in descending order.
Step 3: After the identification of the information, gain assign the best attribute of the dataset at the root of the tree.
Step 4: Then calculate the information gain using the same formula.
Step 5: Split the nodes based on the highest information gain value.
Step 6: Repeat the process until each attribute are set as leaf nodes in all the branches of the tree.

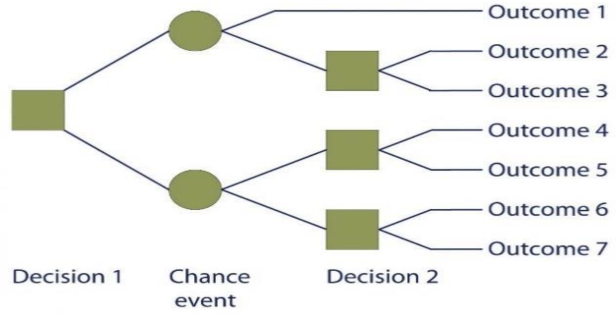

Fig. 4: Decision Trees

5. CONCLUSION

The objective of this study was to summarize the different data mining techniques available to predict heart disease. Compare the prediction performance of these algorithms and find the best method to apply to the proposed system to predict heart disease. We can use relevant feature selection methods for greater accuracy and high performance. Data mining is of great knowledge for the medical dataset and can come up with several treatment methods for heart disease patients. As we come to know from the literature survey that only a marginal success has been achieved till now in the making of a successful heart disease prediction model and there is a need for improvement which can be achieved by making a hybrid model which can predict the heart disease at an early stage with good accuracy.

6. REFERENCES

[1] K.Gomathi and D. S.. (February 2016) "Heart Disease Prediction Using Data Mining Classification." International Journal for Research in Applied Science & Engineering Technology. vol. 4(II),
[2] S. Mukherjee and A. S. . (JanUARY 2019) "Intelligent Heart Disease Prediction using Neural Network." International Journal of Recent Technology and Engineering, vol. 7(5),
[3] H. B. F. D. a. S. A. Belcy. (OCTOBER 2018) "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES." ICTACT JOURNAL ON SOFT COMPUTING, vol. 9(01),
[4] N. A. (10th October 2013) "Backpropogation neural network for prediction of heart disease." Journal of Theoretical and Applied Information Technology, vol. 56(1),

[5] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra. (September 2018) "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach." International Journal of Computer Applications. vol. 181(18),

[6] P. M. S., F. K., and V. Ravichandran. (June 2017) "Comparing Data Mining Techniques Used For Heart Disease Prediction." International Research Journal of Engineering and Technology (IRJET). Vol. 4(6), p.1161-1167.

[7] H. B. F. D. a. S. A. Belcy. (OCTOBER 2018) "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES." ICTACT JOURNAL ON SOFT COMPUTING, vol. 9(01),

[8] Vaghela, C., N. Bhatt, and D. Mistry, A Survey on Various Classification Techniques for Clinical Decision Support System. International Journal of Computer Applications, 2015. 116(23).

[9] A. Hazra, S. K. Mandal, A. Gupta, and A. M. a. A. Mukherjee. (2017) "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review." Advances in Computational Sciences and Technology ISSN 0973-6107 © Research India Publications. Vol. 10(7), p.2137-2159.

[10] V. R. Suma, S. Renjith and S. A. a. M. V. Judy. (March 2016) "Analytical Study of Selected Classification Algorithms for Clinical Dataset." Indian Journal of Science and Technology. vol. 9(11),

[11] S. M. AWAN, M. U. RIAZ, and A. G. KHAN. (SeptemberDecember 2018) "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK." VFAST Transactions on Software Engineering. Vol. 13(3), p.102-112.

[12] A. A. Shinde, R. M. Samant, A. S. Naik, S. N. Kale, and S. A. Ghorpade. (2017) "Heart Disease Prediction System using Multilayered Feed Forward Neural Network and Back Propagation Neural Network." International Journal of Computer Applications. vol. 166(7),

[13] A. Kaur and J. Arora. (MarchApril 2018) "Heart Disease Prediction Using Data Mining Techniques: A Survey." International Journal of Advanced Research in Computer Science. vol. 9(No. 2),

[14] P. Singh, S. Singh, and G. S. P. Jain. (2018) "Effective heart disease prediction system using data mining techniques." International Journal of Nanomedicine. P.121-124.

[15] Abhishek Rairikar, Vedant Kulkarni, Vikas Sabole and Anuradha Lamugunde (2017)l "Heart Disease Prediction Using Data Mining Techniques", International Conference on Intelligent Computing and Control

[16] Tulay Karayilan, Ozkan (2017) "Prediction of Heart Disease Using Neural Network", (UBMK'17) 2nd International Conference on Computer Science and Engineering

[17] S. K. Sen. (June 2017) "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms." International Journal of Engineering and Computer Science. Vol. 6(6), p.21623-21631.

[18] C. Sowmiya and Dr. P. Sumitra (2017) "Analytical Study of Heart Disease Diagnosis Using Classification Techniques", IEEE International Conference On Intelligent Techniques In Control, Optimization And Signal Processing 2017

[19] Bhavini Bhatia, Vamika Razdan (August 2017) "Comparative Analysis of Heart Disease Dataset using KNN and Decision Tree Classification", International Journal of Engineering Research in Computer Science and Engineering. vol. 4(8),