A Novel Approach to Predict Heart Disease Using Machine Learning Algorithm: A Survey

Amit Singh[1] Amrit Suman[2] Ankush Gupta[3] Dayashankar Yadav[4] Kumud Saxena[5]
[1,2,3,4]Student [5]Head of Department
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]NIET, Gr. Noida, U.P, India

Abstract— According to a recent survey by WHO Organization every year 17.5 million people die each year. It shall increase to 70 million by the year 2030. The medical professionals that have been working in the field of heart disease have their own limitations and they can also save the lives of many people by predicting the chances of heart rate disease up to 65% accuracy. Seeing the current epidemic scenario doctors need a support system for better and precise prediction of heart diseases. Machine Learning opens many possible doors for better prediction of heart related diseases. Paper provides a lot of information about state of art methods in Machine learning and deep learning. An analytical comparison has been provided to help the new researchers working in this field.

Keywords: Machine Learning, Heart Disease, Naïve Bayes, Decision Tree, Neural Network, SVM and Deep Learning

I. INTRODUCTION

Heart disease has created a lot of serious concerns among many researches, one of the major challenges they face in heart disease is correct detection and finding presence of it inside a human body. There are various medical instruments available in the market for predicting heart disease but there are two major problems in them which are as follows:

1) Firstly they are very much expensive.
2) Secondly they are not efficiently able to calculate the chance of heart disease in humans.

According to the latest survey conducted by WHO, the medical professionals are able to correctly predict only 65% of heart diseases so there is a vast scope of research in the area of predicting Heart related disease in human.

With advancement in computer science, it has brought vast opportunities in different areas, medical science is one of the fields where the instrument of computer science can be used extensively. Application areas of computer science vary from metrology to ocean engineering. Medical science also used some of the major available tools in computer science. In the last decade artificial intelligence has gained its moment because of advancement in computation power. Machine Learning is one such tool which is widely utilized in different domains because it doesn't require different algorithms for different dataset. Reprogrammable capacities of machine learning bring a lot of strength and open various doors of opportunities for areas like medical science. Machine learning could be a better choice for achieving high accuracy for predicting not only heart disease but also other diseases because this vary tool utilizes feature vector and its various data types under various condition for predicting the heart disease, algorithms such as Naive Bayes, Decision Tree, KNN, Neural Network, are used to predicate risk of heart diseases each algorithm has its specialty such as Naive Bayes used probability for predicting heart disease, whereas decision tree is used to provide classified report for the heart disease, whereas the Neural Network provides opportunities to minimize the error in prediction of heart disease. All these techniques are using old patient records for getting predictions about new patients. This prediction system for heart disease helps doctors to predict heart disease in the early stage of disease resulting in saving millions of lives.

This survey paper is dedicated for wide scope survey in the field of machine learning technique in heart disease. Later part of this survey paper will discuss various machine learning algorithms for heart disease and their relative comparison on the various parameter. It also shows future prospects of machine learning algorithms in heart disease. This paper also does a deep analysis on utilization of deep learning in the field of predicting heart disease.

II. LITERATURE REVIEW

Various Researchers have contributed for the development in this field. Prediction of heart related diseases based on machine learning algorithms is a very difficult case for researchers. Our goal in this project is to bring out kinds of work by different authors and researchers. Amit Singh, Dayashankar Yadav have illustrated about how the datasets available for heart disease are generally raw in nature which is highly redundant and inconsistent. There is a need for pre-processing of these data sets. In this phase high dimensional data set is reduced to low data set. They also show the extraction of crucial features from the data set because there are every kind of such features available. Selection of important features reduces work of training the algorithm and hence results in reduction of time complexity and space complexity. Performance is measured by running algorithms (Bayes Net and SMO) on data sets collected from WEKA software and then compared using predictive accuracy, ROC curve, ROC value. An optimization of feature has been done to achieve higher classification efficiency in Decision Tree. It is an approach for early detection of heart related diseases by utilizing a variety of features. This kind of approach can also be utilized for other spheres of research. Other than decision trees, random forest and Knn were adopted for achieving the goal of perfect detection of heart disease in human beings. After going through the majority of state of art technique we have pointed out certain loopholes existed in them. Some of them are discussed below:

1) There is a wide need for a more robust algorithm which can minimize the noise in the dataset because medical dataset may consist of various types of redundancy and noise in them.
2) Recently with advancement in deep learning there could be a chance to enhance efficiency and accuracy for detection of heart disease.

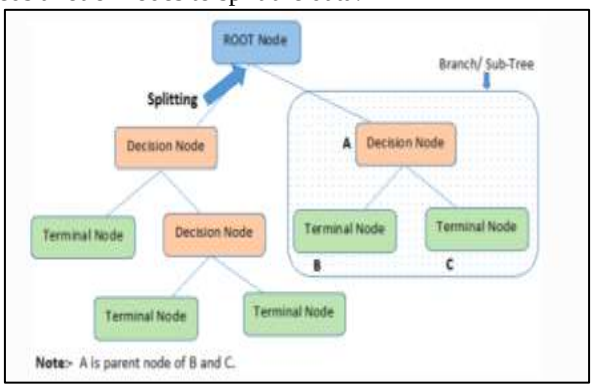
3) Dimensionality of the medical dataset is very high. These put ergs to find such algorithms which can compress and reduce higher dimensionality, resulting in gaining execution time.

III. MACHINE LEARNING ALGORITHM FOR HEART DISEASE PREDICTION

Machine learning is a widely used artificial intelligence tool in all major sectors of application programming with advancement in processing power machines for learning about new technologies and studies.

IV. DECISION TREE

Decision tree is a graphical representation of a specific decision situation that uses a tree like model of decisions and their possible consequences. It is a way to display an algorithm that only contains conditional control statements. It can be used as a predictive model, the main component of the decision tree involves root, nodes, and branching decisions. There are few approaches for building trees such as ID3, CART, CYT, C5.0 and J48 has used the approaches to classify the dataset using J48, similarly have compared the decision trees with classification output of various other algorithms. Decision trees are used in those areas of medical science where multiple parameters are involved in classification of data sets. In heart disease where number of parameter affect patient such as blood pressure, blood sugar, age, sex, genetic and other factor. By seeing the decision tree, the doctor can clearly identify the most effecting feature among all the parameters. They can also generate the most affecting features in the mass of the population. Decision trees are based on entropy and Information gain, clearly signifies the importance of a dataset. Drawback of decision trees is that it suffers from two major problems over fitting and it is based on greedy methods. Overfitting happened due to a decision tree split dataset aligned to the axis; it means it needs a lot of nodes to split the data.



V. K- NEAREST NEIGHBOR ALGORITHM (KNN):

KNN is a slow supervised learning algorithm that takes more time to get trained classification like algorithm. It is divided into two step training from data and testing it on new instances. The K Nearest Neighbor working principle is based on assignment of weight to the each data point which is called as neighbor. In K Nearest Neighbor Algorithm, distance is calculated for training dataset for each of the K Nearest data points. Then classification is done on the basis of majority of votes. There are three types of distances needed to be measured in KNN(Euclidean, Manhattan, Minkowski) distance in which Euclidean will be considered most one. The following formula is used to calculate their distance:

$Eucledian\ Distance = D(x, y)$

$= (xi - yi)2ki = 1$

Where,

k = number of cluster,

x , y = coordinate sample spaces

$Manhattan\ distance = (xi - yi)ni = 1$

Where,

x , y are coordinates

Minkowski $distance$:

$Min = (|xi - yi|\ p)\ 1/p$

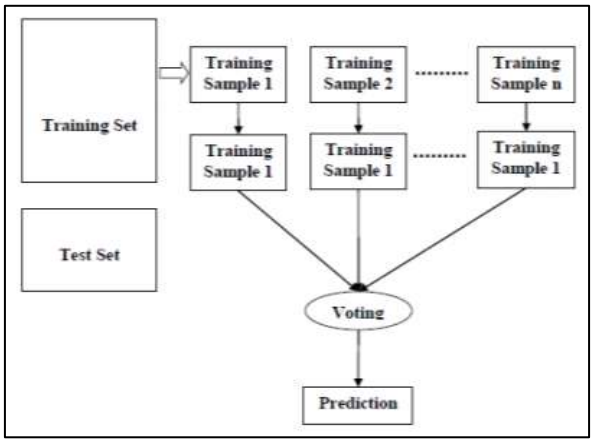
Selection of k values plays a very crucial role, if the k value is large then it is precise and less noisy. The algorithm for KNN is defined in the steps given below:

1. D represents the samples used in the training and k denotes the number of nearest neighbour.
2. Create super class for each sample class.
3. Compute Euclidian distance for every training sample
4. Based on majority of class in neighbour, classify the sample

VI. RANDOM FOREST:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. We can understand the working of the Random Forest algorithm with the help of the following steps −

1) Step 1 − First, start with the selection of random samples from a given dataset.
2) Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
3) Step 3 − In this step, voting will be performed for every predicted result.
4) Step 4 − At last, select the most voted prediction result as the final prediction result.



VII. CONCLUSION

The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to predict the result more efficiently.

REFERENCES

[1] Krishnaiah V., Narsimha G., Chandra N. Subhash (2016): Heart Disease Prediction System Using Data Mining Techniques and Intelligent Fuzzy Approach , CSE, Research Scholar, JNTUH Dept. of CSE, Hyderabad Research Scholar, JNTUH(Pg- Vol-136).
[2] Tanja Abhishek, Jain S.A.(2013): Heart Disease Prediction System Using Data Mining Techniques, CSE , Ambala City, India, An International Open Free Access, Peer Reviewed Research Journal .
[3] VidyaPeetham Amrita Vishwa, Kasavanahalli, Carmelaram P.O. (2015):An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease, CSE , (Pg 23-32).
[4] Mohammad A. M. Abushariah, Assal A. M. Alqudah, Omar Y. Adwan, Rana M. M. (2014): Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro- Fuzzy Inference Systems(ANFIS),(Pg-Vol.7).
[5] Pouriyeh Seyedamin, Vahid Sara, Sannino Giovanna, Arabnia Hamid, Gutierrez Juan, A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease, CSE, University of Georgia, Athens,
[6] Babič František, Olejár Jaroslav(2017). Predictive and Descriptive Analysis for Heart Disease Diagnosis , Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, (pp. 155–163 DOI: 10.15439/2017F219 ISSN 2300-5963 ACSIS).
[7] Methaila Aditya, Kansal Prince, Arya Himanshu(2014). EARLY HEART DISEASE PREDICTIONUSING DATA MINING TECHNIQUES, CSE,Pankaj Kumar Netaji Subhas Institute of Technology, India, CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014 DOI : 10.5121/csit.2014.4807.

