https://dx.doi.org/10.22161/ijaers.73.7 ISSN: 2349-6495(P) | 2456-1908(O)

Hybrid Machine Learning Techniques for Heart Disease Prediction

S. Sharanyaa[1], S. Lavanya[2], M.R. Chandhini[3], R. Bharathi[4], K. Madhulekha[5]

[1]Assitant Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India
[2,3,4,5]Department of Information Technology, Panimalar Engineering College, Chennai, India

Abstract*— *Diseases can affect people both physically and mentally, as contracting and living with a disease can alter the affected person's perspective on life. A disease that affects the parts of an organism, which isn't because of any immediate external injury. Diseases are often known to be medical conditions that are related to specific symptoms and signs. The deadliest diseases in humans are arteria coronaria disease (blood flow obstruction), followed by cerebrovascular disease and lower respiratory infections. Heart disease are most unpredictable and unexpectable. We can able to predict the heart disease using machine learning technique. The datasets are taken from UCI repository which is a public dataset. These trained dataset are used for the prediction. Techniques like Decision tree, Support Vector Machine, K Nearest Neighbor and Random Forest algorithms are used in the prediction of heart disease and hybrid of these algorithms provides 94 % accuracy.

Keywords*— *Cardiovascular disease (CVD), Decision tree, Support Vector Machine, K Nearest Neighbor and Random Forest

I. INTRODUCTION

Harmful deviation from the normal structural or functional state of an organism in the human body is named as disease. Generally we can predict the disease based on the symptoms. Healthcare industry faces the major issues like prediction the diabetes disease among several others. People with high blood glucose and cholesterol with damage of blood vessels will tend to develop a heart disease and other nerve diseases. The European Society of Cardiology (ESC) survey that 26 million adults worldwide were affected by heart disease and 3.6 million were diagnosed every year[7]. Machine learning techniques provides us the ability for automatic learning and experience without being explicitly programmed. Machine learning provides an objective opinion to improve efficiency, reliability and accuracy. Machine learning methods used for decision support achieves high accuracy of decisions and they recommend a deep understanding of decisions and the decision makers will trust machine learning methods. Methods for learning implicit, non-symbolic knowledge will provide better predictive accuracy. Methods for learning explicit, symbolic knowledge produce lots of comprehensible models. Hybrid machine learning models collaborate strengthens knowledge representation model types. In healthcare industry and medical platform, Collecting and analyzing the data is considered to be important. Through the machine learning concept, we can able to analyze and predict the data using several algorithms and techniques. Supervised machine learning algorithms have been the most leading method in the data mining field. This study aims to identify the key trends among several types of supervised machine learning algorithms and their performance, usage for disease risk prediction[6]. Managing diabetes involves a lots of issues and commitments like routine checking of blood pressure, blood sugar level and other health status. In this paper machine learning will predict the heart disease and non functional state of the heart using the necessary clinical data value. In classification method, the total dataset is divided into 70% of data for training and 30%of data for testing. The prediction of heart disease is based on machine learning algorithms like k-nearest neighbor algorithm (KNN), Decision tress algorithm, support vector algorithm(SVM), Random forest (RF) algorithm. [1].

II. LIRATURE SURVEY

Mohan et al. [1] Heart disease is one among the foremost significant causes of mortality within the world today. Prediction of disorder may be a critical challenge with the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting, making decisions and predictions from the massive quantity of knowledge produced by the healthcare industry. Various studies gives

https://dx.doi.org/10.22161/ijaers.73.7 ISSN: 2349-6495(P) | 2456-1908(O)

only a glimpse into predicting heart condition with ML techniques. In this paper, we apply machine learning technique to predict the heart disease with more accuracy. The prediction model is introduced with different combinations of features and a number of other known classification techniques. We produce an enhanced performance level with an accuracy level of 92% through the prediction model for heart disease with combination of hybrid random forest and linear model.

In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and feminine category and this ratio may vary consistent with the region also this ratio is taken into account for the people aged group 25-69. This doesn't indicate that the people with other age category won't be suffering from heart diseases. This problem may start in early ages also. Here in this paper, we have discussed various algorithms and tools used for prediction of heart diseases [2]. Data mining may be a technique performed on large databases for extracting hidden patterns by using hybrid methods from statistical analysis, machine learning and database technology. Further, the medical data processing is particularly a important research field to its performance within the development of varied applications in flourishing healthcare domain. In this work, three data processing classification algorithms like Random Forest, Decision Tree and Naïve Bayes are addressed and wont to develop a prediction system so as to analyze and predict the possibility of heart disease.

David, H et.al [3] , The main objective of this research is to use simplest classification algorithm to provide maximum accuracy when classification of normal and abnormal person is found. Thus prevention of the loss of lives at an earlier stage is feasible. The experiment is made up of the evaluation of the performance of algorithms with the assistance of heart condition benchmark datasets retrieved from UCI machine learning repository. It is found that Random Forest algorithm performs best with 81% precision in comparison to other algorithms for heart condition prediction.

Yekkala et.al [4]. Data is generated by the medical industry. Often this data is of very complex nature electronic digital records, handwritten scripts, etc. Since it is generated from multiple resources. Due to the

Complexity and large volume of this data necessitates techniques which will extract insight from this data in a quick and efficient way. These insights not only diagnose the diseases but also predict and may prevent disease. Heart disease or coronary artery disease (CAD) is one among the main causes of death everywhere in the planet. Comprehensive research using single data processing techniques haven't resulted in a suitable accuracy. Further research is being carried out on the effectiveness of hybridizing quite one technique for increasing accuracy in the diagnosis of heart disease. In this journal, the authors worked on datasets collected from the UCI repository, and used the Random Forest algorithm and Selection is done by using rough sets to accurately predict the occurrence of heart disease.

According to Reddy Prasad et.al [5], We are in a period of "Information Age" where the normal industry can pressure the rapid shift to the industrial revolution for industrialization, based on economy of information technology .Terabytes of data are produced and stored in a day-to day life due to rapid growth in Information Technology. The data which is collected is converted by data analysis by using various combinations of algorithms. The large amount of the information regarding the patients is generated by the hospitals like x-ray results , lungs results ,chest paining results, personal health records(PHRs), etc,. Some certain tools are used to extract the knowledge from the database for the detection of heart diseases . The main theme of this paper is that the prediction of heart diseases using machine learning techniques by summarizing the few current researches. During this paper the logistic regression algorithms is employed, so that the health care data which classifies the patients whether they are having heart diseases or not according to the information in the record. Also it will be able to attempt to use this data model which predicts the patient whether they are having heart condition or not.

III. METHODOLOGY

1. DATA PREPROCESSING

Heart disease data is pre-processed by removing noise and missing values. The datasets contains a total of 310 patient records where 7 records are with some missing values those 7 records have been removed from the dataset and remaining 303 patient records are used in preprocessing.

2. FEATURE SELECTION

From the total 13 attributes of the dataset, two attributes pertaining to age and gender are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records. Clinical data are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several (ML) techniques are used namely SVM, KNN, DT, RF Algorithm .The

https://dx.doi.org/10.22161/ijaers.73.7 *ISSN: 2349-6495(P) | 2456-1908(O)*

experiment was repeated with all the ML techniques using all 13 attributes.

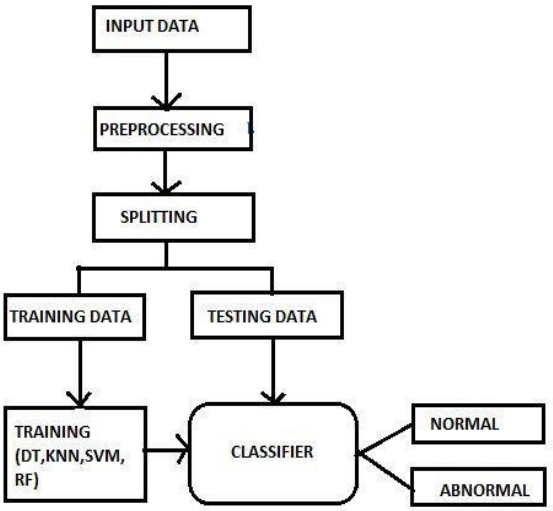

Fig 1. Experimental workflow with UCI dataset

3. CLASSIFICATION

The clustering of the dataset is done on the basis of the variables and criteria. Then the various classifiers are applied to each dataset in order to estimate its performance. The techniques are, Decision tree, Support Vector Machine, K Nearest Neighbor, Random forest algorithm.

3.1 DECISION TREE

For training samples of data the trees are constructed based on inputs. Regression and classification problems are solved using decision tree. This technique is performed on the basis of Top down divide and conquer approach. Tree pruning helps to remove irrelevant samples of data.

3.2 SUPPORT VECTOR MACHINE

SVM is said to be a supervised machine learning algorithm which may be used for classification or regression problems. It uses a way called the kernel trick to rework your data, then supported these transformations it finds an optimal boundary between the possible outputs.

3.3 RANDOM FOREST

A random forest algorithm is one among the foremost effective ensemble classification approach. This algorithm has been used in prediction and probability. The RF method consists of multiple decision trees. Each decision tree gives an information that indicates the decision about the class of the object. RF method blend bagging and random selection of

features. There are three different parameters in random forest are No. of the trees (n tree), Minimum node size and No. of features employed in splitting each node [9].

3.4 K NEAREST NEIGHBOR

K-Nearest neighbor (KNN) may be a simple, lazy and nonparametric classifier. KNN is preferred when all the features are linear. It is under supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. KNN is additionally called as case-based reasoning and has been utilized in many applications like pattern recognition, statistical estimation. Classification is obtained by identifying the closest neighbor to work out the category of an unknown sample. KNN is preferred over other classification algorithms because of its high convergence speed and ease [10].

IV. RESULT EVALUATION

The prediction of disease is developed using 13 features and 4 classifiers to improves the accuracy of the models. The highest accuracy is achieved by K-nearest neighbor classification method when compared with existing methods.

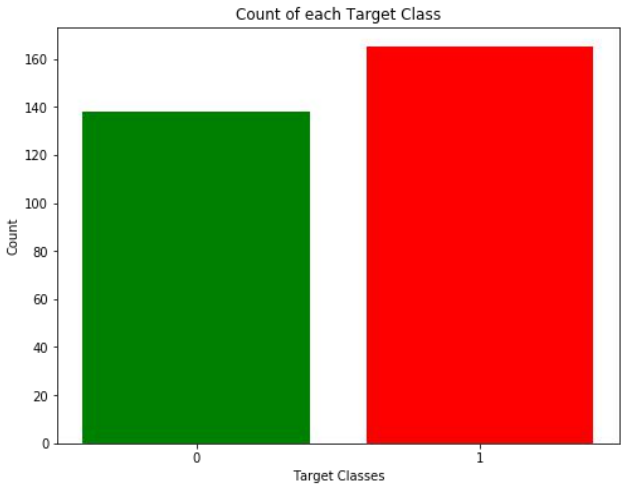

Fig 2. Target Class

```
# Evaluating using accuracy_score metric
from sklearn.metrics import accuracy_score
accuracy_logreg = accuracy_score(Y_test, Y_pred_logreg)
accuracy_knn = accuracy_score(Y_test, Y_pred_knn)
accuracy_svc = accuracy_score(Y_test, Y_pred_svc)
accuracy_nb = accuracy_score(Y_test, Y_pred_nb)
accuracy_dectree = accuracy_score(Y_test, Y_pred_dectree)
accuracy_ranfor = accuracy_score(Y_test, Y_pred_ranfor)
```

```
# Accuracy on test set
print("Logistic Regression: " + str(accuracy_logreg * 100))
print("K Nearest neighbors: " + str(accuracy_knn * 120))
print("Support Vector Classifier: " + str(accuracy_svc * 100))
print("Naive Bayes: " + str(accuracy_nb * 100))
print("Decision tree: " + str(accuracy_dectree * 100))
print("Random Forest: " + str(accuracy_ranfor * 100))
```

```
Logistic Regression: 77.04918032786885
K Nearest neighbors: 88.52459016393443
Support Vector Classifier: 75.40983606557377
Naive Bayes: 68.85245901639344
Decision tree: 68.85245901639344
Random Forest: 73.77049180327869
```

Fig 3. Performance of each classifier

https://dx.doi.org/10.22161/ijaers.73.7 ISSN: 2349-6495(P) | 2456-1908(O)

Table 1. Feature information of the dataset.[8]

Sno	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic]	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Target	Class Attribute	0 or 1

V. CONCLUSION AND FUTURE WORK

Identifying the processing of raw healthcare data heart information will help in the long term saving of human lives and early prediction of the abnormalities in heart conditions. Machine learning methods were used in the process of raw data and provide the prediction of the disease and health status of the patient. Heart disease prediction is one of the challenging process in the medical field. Using this project the mortality rate can be drastically controlled if the disease is detected. The proposed hybrid approach is used to combine the characteristics of fuzzy logic and k-nearest neighbor algorithm which provides 94% accuracy. This method proved the accuracy of highest prediction rate. The further course of this research can be performed with the mixture of deep learning techniques to achieve better prediction in accuracy.

REFERENCES

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554.

[2] Avinash Golande, Pavan Kumar T (June 2019) Heart disease prediction using effective machine learning techniques.

[3] David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques. *ICTACT Journal on Soft Computing*, *9*(1).

[4] Yekkala, I., & Dixit, S. (2018). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, *3*(1), 1-12.

[5] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa(Feb 2019) Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning

[6] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1-16.

[7] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.

[8] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

[9] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. *Journal of Network and Innovative Computing*, *4*(2016), 175-184.

[10] Jabbar, M. A. (2017). Prediction of heart disease using k-nearest neighbor and particle swarm optimization.

https://dx.doi.org/10.22161/ijaers.73.7 *ISSN: 2349-6495(P) | 2456-1908(O)*

[11] Sharanyaa, S., and K. Sangeetha. "Blocking Adult Account in OSN's Using Iterative Social Based Classifier Algorithm."

[12] Sharanyaa, S., Abitha, P., Karthikeyan, B., Buvaneswari, B., & Sumithra, M. Identification of dysphonia related to parkinson's disease using parametric and non parametric models.

[13] Buvaneswari, B., & Reddy, T. K. L. (2019). High performance hybrid cognitive framework for bio-facial signal fusion processing for the disease diagnosis. Measurement, 140, 89-99.