Heart Disease Prediction Using Machine Learning

Ms. S.Leela Bharathi[1], Dr.S.Shahar Banu[2]
[1]Dept of Computer Applications
[2]Associate Professor, Dept of Computer Appliations
[1, 2] B.S.Abdur Rahaman Crescent Institute Of Science And Technology, India

Abstract- *In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease.The main objective of the project is to develop a system for heart disease prediction using machine learning based on several features, including age, sex, chest pain type, resting blood pressure, serum cholestoral, fasting blood sugar, and more.The heart disease dataset from the UCI machine learning repository is used in this study. The proposed work uses a variety of data mining approaches, including Logistic Regression, Random Forest and SVM to forecast the likelihood of Heart Disease. As a result, this research conducts a comparative analysis by evaluating the effectiveness of several machine learning methods. The trial results show that, when compared to other machine learning algorithms used, the Random Forest approach has the highest accuracy.*

Keywords- Logistic Regression, Random Forest, Support Vector Machine, Heart Disease Prediction, Machine Learning.

I. INTRODUCTION

The work that is proposed in this paper focuses mostly on different data mining techniques used to forecast cardiac disease. The primary organ of the human body is the heart. In essence, it controls the blood flow throughout our body. Any heart abnormality can be distressing to other body parts. Heart disease can be defined as any type of interruption to the heart's normal function. One of the main causes of death in the modern world is heart disease. Heart disease can be brought on by living a sedentary lifestyle, using tobacco products or alcohol, or eating a lot of fat, which can increase blood pressure. According to the World Health Organisation, more than 10 million people worldwide pass away each year as a result of heart disease. Only a healthy lifestyle and early identification can stop heart-related disorders.

Today's healthcare system faces major challenges in providing the highest quality services and precise, reliable diagnoses. Even though cardiac disorders have been identified as the leading cause of death worldwide in recent years, they are also the ones that can be effectively managed and controlled. The right moment of disease discovery determines how accurately a disease will be managed overall. The proposed strategy aims to identify these heart conditions early in order to prevent negative outcomes.

Records of a sizable collection of medical data compiled by medical professionals are available for analysis and knowledge extraction. The use of data mining techniques allows for the extraction of important and hidden information from the vast amount of available data. The medical database primarily contains discrete data. As a result, making decisions with discrete data is a challenging and complex undertaking. Data mining's subset of machine learning (ML) effectively manages massive, well-organized datasets. Machine learning can be used in the medical industry to diagnose, detect, and forecast a variety of diseases.The major objective of this study is to give medical professionals a tool to identify heart disease at an early stage. In turn, this will support giving patients effective care and averting negative outcomes. To uncover hidden discrete patterns and assess the provided data, machine learning (ML) plays a critical role. ML approaches aid in the early detection and prediction of cardiac disease after data processing. This study examines the effectiveness of different machine learning (ML) methods for early heart disease prediction, including Logistic Regression, Random Forest, SVM.

II. RELATED WORK

Using the UCI Machine Learning dataset, extensive research has been done to predict heart disease. Varied data mining approaches have been used to achieve varied accuracy levels, which are detailed below.

Avinash Golande and et. al.; studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

T.Nagamani, et al. have proposed a system which deployed data mining techniques along with the MapReduce

algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due to use of dynamic schema and linear scaling.

Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.

Anjan Nikhil Repaka, eatl., proposed a system in that uses NB (Naïve Bayesian) techniques for classification of dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease.

Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K-Nearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analysed for different number of attributes.

Nagaraj M Lutimath, et al., has performed the heart disease prediction using Naive bayes classification and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes.

The primary goal of the suggested system, which was developed after reading the aforementioned articles, was to develop a heart disease prediction system using the inputs listed in Table 1. In order to determine the best classification method that can be applied to the prediction of heart disease, we compared the accuracy of the three classification algorithms, namely Logistic Regression, Random Forest, and SVM.

III. PROPOSED MODEL

The proposed work examines the three classification methods stated above and does performance analysis to forecast heart disease. This study's goal is to accurately determine whether a patient has heart disease. The medical expert inputs the values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. The full procedure is depicted in Fig. 1.

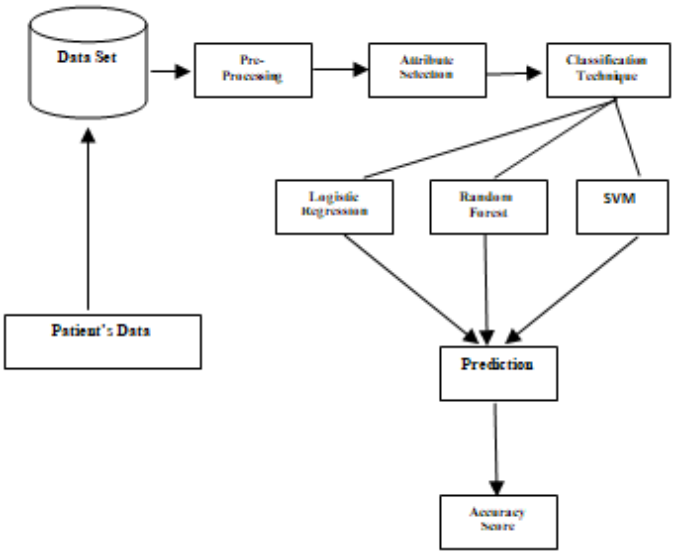

Fig 1:Model Predicting Heart Disease

A. Data Collection and Preprocessing:

The dataset used was the Heart disease Dataset which is a UCI Cleveland dataset. The dataset has 303 rows and 14 attributes. We have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below.

TABLE I: FEATURES IN THE DATASET

Sl.No	Attribute Description	Distinct values of Attribute
1.	Age- represent the age of a person	Multiple valuesbetween 29 & 71
2.	Sex- describe the gender of person (0-Feamle, 1-Male)	0,1
3.	CP- represents the severity of chestpain patient is suffering	0,1,2,3
4.	RestBP-It represents the patient's BP	Multiple value between 94& 200
5.	Chol-It shows the cholesterol level ofthe patient.	Multiple valuesbetween 126 & 564

6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max hear beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1or else no=0	0,1
10.	OldPeak- describes patient's depression level.	Multiple value between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3
12.	CA- Results of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent thallium test.	0,1,2,3
14.	Target- It is the final column of the dataset. It is class or label column. It represents the number of classes in dataset. This dataset has binary classification i.e, two classes (0,1). In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "" or "1" depends on another 13 attributes.	0,1

B. Classification:

The attributes listed in Table 1 are used as input by many ML algorithms, including SVM classification, Random Forest, and Logistic Regression. Eighty percent of the input dataset is used as training data, and the remaining twenty percent is used as test data. The dataset used to train a model is referred to as the training dataset. The performance of the trained model is evaluated using the testing dataset. The accuracy score is used to compute and analyse the performance for each algorithm, as will be explained further. The different algorithms explored in this paper are listed as below.

i. Logistic Regression:

Logistic Regression is a classification algorithm mostly used for binary classification problems. Instead of fitting a straight line or hyperplane, the logistic regression algorithm uses the logistic function to constrain the output of a linear equation to the range between 0 and 1. Because there are 13 independent variables, logistic regression is effective for categorization.

ii. Random Forest:

Random Forest algorithms are used for classification as well as regression. The data is organised into a tree, and predictions are based on that tree. Even with a substantial number of record values missing, the Random Forest algorithm can still produce the same results when applied to huge datasets. The decision tree's generated samples can be preserved and used to different sets of data. Create a random forest in the first step of random forest, and then use the classifier produced in the first stage to generate a prediction.

iii. Support Vector Machine:

SVM is a supervised learning technique that may be applied to both classification and regression applications. The main goal of SVM is to identify the hyperplane that can most effectively discriminate between the various classes in the dataset. With the use of various kernel functions, such as polynomials, radial basis functions (RBF), or sigmoid, SVM is capable of handling data that cannot be separated linearly. Training and prediction are the two phases of SVM. The SVM classifier uses these parameters to forecast the class labels for the fresh input data when the method is in the training stage, where it learns the best hyperplane parameters.

IV. RESULTANDANALYSIS

The results obtained by applying Random Forest, Logistic Regression and Support Vector Machine are shown in this section. The accuracy score is the metric used to evaluate the algorithm's performance. The pre-processed dataset is used in the experiment to carry out the experiments, and the algorithms mentioned earlier have been examined and used. The accuracy score is a frequently used metric in machine learning to assess how well a categorization model is performing. To obtain the accuracy score, the predicted class labels generated by the classification model are compared to

the true class labels for a set of test data. The accuracy score is then calculated as the ratio of the number of correct predictions to the total number of predictions made.

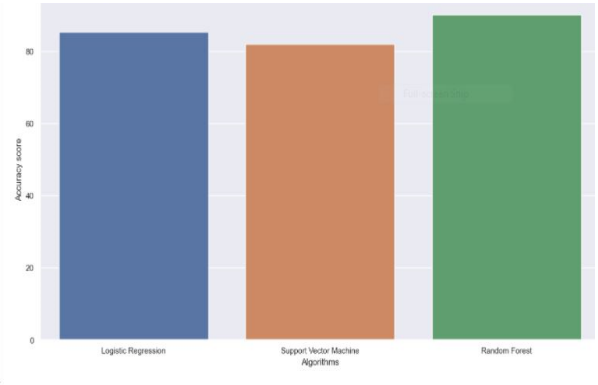

TABLE II: ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Accuracy Score
Logistic Regression	85.25%
Random Forest	90.16%
Support Vector Machine	81.97%

V. CONCLUSION

It is essential to create a system that can forecast heart diseases precisely and effectively given the rise in fatalities caused by heart diseases. The goal of the study was to identify the most effective ML algorithm for heart disease identification. Using data from the UCI machine learning repository, this study compares the accuracy scores of Logistic Regression, Random Forest, and Support Vector Machine for predicting heart disease. The outcome of this study shows that the Random Forest algorithm is the most effective algorithm for predicting heart disease, with an accuracy score of 90.16%.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 provide better results and help health professionals in predicting the heart disease effectively and efficiently.

VI. ACKNOWLEDGEMENT

First and Foremost, I am thankful to the B.S.Abdur Rahaman Crescent Institute Of Science And Technology, Department of Computer Applications and Dr.S.Shahar Banu, Associate Professor, Department of Computer Applications, B.S.Abdur Rahaman Crescent Institute Of Science And Technology. A special word of gratitude to Dr.S.Pakkir Mohideen, Head of Department, Department of Computer Applications, B.S.Abdur Rahaman Crescent Institute Of Science And Technology, for his continued guidance and support for my project work.

REFERENCES

[1] AvinashGolande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Frankllin, "Design And Implementation Heart Disease Prediction Using Navies Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Humanheart Disease Prediction System using Data Mining Techniques", International Conference on Circuit using Data Mining Techniques', International Conference on Circuit Power and ComputingTechnologies, Bangalore,2016.

[6] Nagaraj M Lutimath,ChethanC,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[8] SayaliAmbekar, Rashmi Phalnikar,"Disease Risk Prediction by Using Convolutional Netural Network", 2018 Fourth International Conference Convolutional Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field, in Machine Learning Paradigams,2019,pp.71-99.

[10] JafarAlzubi, AnandNayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018.

[11] FajrIbrahemAlarsan., and MamoonYounes'Analysis and classification of heart diseases using heartbeat

features and machine learning algorithm', Journal Of Big Data, 2019;6:81.

[12] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/a