Heart Disease Prediction Using Machine Learning Techniques

Abhinav kumar Rai
Galgotias College
Greater Noida, India

Aayush Gupta
Galgotias College
Greater Noida, India

Aashi Gupta
Galgotias College
Greater Noida, India

Abstract - Heart disease, often known as cardiovascular disease, encompasses a broad range of heart-related disorders and has emerged as the leading cause of mortality during the last few decades everywhere in the globe. Numerous risk factors are linked to cardiovascular disease, and timely access to accurate, reliable, and commonsense methods for diagnosing the condition early is essential for effective disease management. Data mining is a common strategy for dealing with large datasets in the healthcare business. To assist healthcare practitioners in creating accurate heart illness predictions, researchers study huge, complex medical datasets using a range of data mining and machine learning methodologies. Naive Bayes, decision trees, K-nearest neighbour, and random forest are just a few of the supervised learning algorithms utilised in the construction of a model for the identification of cardiovascular disease risk factors. It makes use of the pre-existing Cleveland database of UCI's repository of heart disease patients. There are 303 Occurrences and 76 characteristics in the collection. Only 14 of these 76 attributes are actually tested, but they're crucial for proving the efficacy of various algorithms. This study's ultimate goal is to foresee the likelihood that people may acquire heart disease. According to the findings, K-nearest neighbour yields the best accuracy.

Keywords—Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Heart Disease Prediction

I. INTRODUCTION

Conditions affecting the heart are often referred to as "heart disease." According to the latest data from the World Health Organisation, cardiovascular illnesses are now the top cause of mortality globally, accounting for 17.9 million fatalities per year. High cholesterol, obesity elevated triglyceride levels, hypertension, and other similar conditions all contribute to an increased risk of heart disease. However, as time goes on, more and more research data and hospital patient records become accessible. Patients' medical histories may be accessed via several publicly available channels, and studies can be undertaken to determine which of several available computerized methods is most suited to making ar accurate diagnosis and preventing the progression of this deadly illness. Learning and artificial intelligence are Playing a huge role in the medical industry. Medical agencies all across the world gather data on a wide range of health issues. To extract value from data, a variety of machine learning algorithms may be applied. However, there is a lot of data being collected, and much of it is noisy. Machine learning technologies quickly explore datasets that are just too huge for human brains to manage. As a consequence, These algorithms have recently shown significant promise in predicting the existence or absence of heart-related disorders. Several machine learning and deep learning models could be used for disease detection to identify or predict outcomes. Machine learning techniques provide a comprehensive examination of genetic data. Knowledge of pandemic forecasts may be taught on models, and medical data can be changed and analyzed more extensively to improve predictions.

RELEATED WORK

Lot of work has been carried out to predict heart disease using UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques which are explained as follows.

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 [1]. 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 [2] which deployed data mining techniques along with the Map Reduce 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 [3]. Rapid Miner tool was used which resulted in higher accuracy compared to Mat lab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM

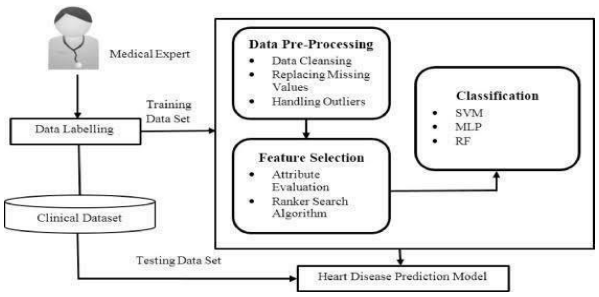
classification algorithms were compared. Decision tree algorithm had the highest accuracy.

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 analyzed for different number of attributes [5].

II. PROPOSED MODEL

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis the objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. 1 shows the entire process involved.

Fig:Architectur Diagram of the proposed model



A. Data Collection and Preprocessing

Despite the fact that the Heart Disease Dataset contains four databases, For this investigation, only the UCI Cleveland dataset was utilized. Only 14 of the 76 properties in this database have been reported to have been used in any experiments [9]. Consequently, we employed the pre-processed UCI Cleveland dataset, which is available on the Kaggle platform, for our inquiry. You can see a complete rundown of all 14 criteria that will be utilized in the next research in Table 1.

TABLE I. FEATURES SELECTED FROM DATASET

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	*Age*- represent the age of a person	Multiple values between 29 & 71
2.	*Sex*- describe the gender of person (0-Female, 1-Male)	0,1
3.	*CP*- represents the severity of chest Pain patient is suffering.	0,1,2,3
4.	*Rest BP*-It represents the patient's BP.	Multiple values between 94& 200
5.	*Chloe*-It shows the cholesterol level of the patient.	Multiple values between 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 heart beat of patient *Exang*- used to identify if there is an	Multiple values from 71 to 202
10	*Old Peak*-describe patient's depression level	Multiple value 0 to 6.2
11	*Slope*- describes patient condition during peak exercise. It is divided into three segments(Up sloping, Flat, Down sloping)	1,2,3
12	*CA*- Result of fluoroscopy.	0,1,2,3
13	*Thal*- test required for patient suffering from pain in chest or difficulty in breathing .There are 4 kind of value which represent Thallium test	0,1,2,3
14	*Target*-It is the final column of the Dataset. It is class or label Colum. It represents the number of classes in dataset .this database has binary classification's. Two classes(0,1).In class-0|represent there is less possibility of heart disease whereas represent high chances of heart disease whereas depends on other 13 attribute	0,1

B. Classification

A number of ML algorithms [12] take attributes from Table 1 as input, There are many classification techniques employed, including Random Forest, Decision Tree, Logistic Regression, and Naive Bayes. Eighty percent of the input data are in the training dataset, while the remaining wenty percent are in the test dataset. The model is trained using the training dataset. The efficacy of the trained model is evaluated by means of a testing dataset. Accuracy, precision, recall, and sensitivity are only a few of the metrics that are used to evaluate and study algorithm performance.scores on the F-test, are further upon below.in this study, we investigate the following sets of algorithms.

I. Random Forest

It has been suggested that Random Forest techniques might be useful in both classification and regression. The data is organized into a tree structure, and then inferences are drawn from it. Since the Random Forest method maintains consistency even when sizable chunks of record values are missing, it is useful for large datasets. The decision tree samples might be kept and utilized once again with new data. Random forest is a two-stage process that begins with the creation of a classifier and ends with a prediction based on that classifier..

II. Decision Tree

The decision tree approach resembles a flowchart, with a center node standing in for the characteristics of the dataset and branches radiating out to reflect the final outcome. Decision trees are preferred over other methods of data analysis due to their speed, accuracy, clarity, and little need for preprocessing. Class label predictions in Decision Trees are made at the tree's base. The value of the root property is contrasted with the value of the attribute on the record. The comparison result decides whether the Corresponding or Jump route is taken.

III. Logistic Regression:

A common use of the classification method Logistic Regression is in solving issues of binary classification. In the logistic regression method, the logistic function is used to fit data instead of a line or a hyperplane.

function that takes a linear equation's output and squeezes it into the range [0,1]. Logistic regression works well for classifying data with 13 independent factors.

Naive Bayes

The Naive Bayes algorithm is founded on the principles of the Bayes rule. A critical assumption for classification is the independence of dataset attributes. Under the condition of independence, it can be predicted quickly and with high accuracy. The computation of the posterior probability of event A involves the division of the ratio P(A/B)[10], which denotes the prior probability of the occurrence of event B.

$$P(A/B) = P(B/A)P(A)/P(B) \quad (1)$$

▫▫. RESULT AND ANALYSIS

Several examples of machine learning algorithms include Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. The algorithm's effectiveness is measured using accuracy, precision, recall, and the F-measure. The precision measure (given in equation (2)) quantifies the accuracy of a positive analysis. Recall [specified in equation (3)] is the percentage of correct responses. The F-measure is a measure of accuracy (equation (4)).

$$\text{Precision} = (TP)/(TP + FP) \quad (2)$$

$$\text{Recall} = (TP)/(TP + FN) \quad (3)$$

$$\text{fMeasure} = (2 * Presicion * Recall)/(Presicion + Recall) \quad (4)$$

- TP True positive: the patient has the disease and the test is positive.
- FP False positive: the patient does not have the disease but the test is positive.
- TN True negative: the patient does not have the disease and the test is negative.
- FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms. The accuracy score obtained for Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques[12] is shown below in

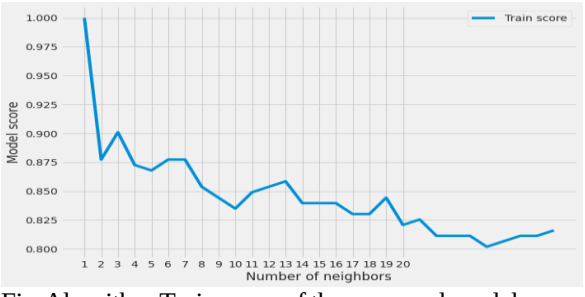


Fig:Algorithm Train score of the proposed model

TABLE II. ACCURACY OF THE ALGORITHMS WE USED

SI.NO	Model	Training Accuracy %	Testing Accuracy %
1	Logistic regression	79.5	76.8
2	K-nearest neighbor	80.4	80.48
3	Support vector Machine	83.40	87.91
4	Decision Tree Classifier	65.34	67.21
5	Random Forest Classifier	88.30	85.22

References

[1] Avinash Golande, 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 Franklin, Design And Implementation Heart Disease Prediction Using Naives Bayesian, *International Conference on Trends in Electronics and Information(ICOEI 2019).*

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', *International Conference on Circuit Power and Computing Technologies,Bangalore*,2016.

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

[7] A. K and A. S. Singh, "Detection of Paddy Crops Diseases and Early Diagnosis Using Faster Regional Convolutional Neural Networks," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 898- 902, doi: 10.1109/ICACITE51222.2021.9404759.

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

[9] Sayali Ambekar, Rashmi Phalnikar,—Disease Risk Prediction by Using Convolutional Neural Network‖,2018 Fourth International Conference on Computing Communication Control and Automation. [10] 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 Paradigms, 2019, pp. 71–99.

[11] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018

[12] Fajr Ibrahem Alarsan., and Mamoon Younes _Analysis and classification of heart diseases using heartbeat features and machine learning algorithms‘,Journal Of Big Data,2019;6:81.

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

[14] A. Chanchal, A. S. Singh and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1- 5, doi: 10.1109/ICRITO51393.2021.9596228.

[15] A. K, D. D, A. Lakhanpal, K. Manoj Sagar, K. Murugan and A. Shanker Singh, "Discover Pretend Disease News Misleading Data in

Social Media Networks Using Machine Learning Techniques," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 784-788, doi: 10.1109/ICACITE51222.2021.9404648.