IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 10, May 2025

Impact Factor: 7.67

Cardiovascular Disease Prediction Using Machine Learning Models

Akshay Dukale, Omkar Daphale, Abhishek Khade
Department of Information Technology Engineering
NBN Sinhgad Technical Institutes Campus, Pune, India

Abstract: *Cardiovascular diseases are among the most essential reasons for death. Prediction of cardiovascular disease is a vital problem in the field of clinical data analysis. Machine learning and Artificial Intelligence are more hopeful in helping decide and predict from the huge data generated by healthcare. We have observed that various features have been utilized in recent advancements of the machine learning model. Here, we suggested machine learning methods for predicting cardiovascular disease based on features. The Cardiovascular system also includes a network of blood vessels, i.e., veins, arteries, and capillaries. These vessels supply blood throughout the body. Malfunctions in normal blood flow from the heart induce various forms of heart diseases which are generally referred to as cardiovascular diseases (CVD). Heart disease is the principal cause of death globally. 17.5 million total deaths worldwide due to heart attacks and strokes, as per a World Health Organization survey. Over 75% of cardiovascular disease deaths take place predominantly in low- and middle-income nations. Also, 80% of the deaths caused due to CVDs are due to stroke and heart attack.*

Keywords: Machine Learning (ML), Cardiovascular Diseases Prediction (CVD), Artificial Intelligence(AI), Decision tree classifier.

I. INTRODUCTION

The heart is a type of muscular organ that circulates blood through the body and is the core component of the body's cardiovascular system that also has the lungs. The cardiovascular system also includes a set of blood vessels, for instance, veins, arteries, and capillaries. The blood vessels transport blood throughout the body. Irregularities in normal blood circulation from the heart result in various kinds of heart diseases which are popularly referred to as cardiovascular diseases (CVD). Heart disease is the primary cause of death globally. The World Health Organization (WHO) survey states that 17.5 million total global deaths result from heart attacks and strokes. Over 75% of cardiovascular disease-related deaths take place primarily in middle-income and low-income nations. Also, 80% of the fatalities that happen because of CVDs are due to stroke and heart attack. Hence, the prediction of cardiac abnormalities at an early stage and devices for the prediction of heart diseases can save numerous lives and assist doctors in planning an effective treatment strategy that ultimately decreases the mortality rate due to cardiovascular diseases.

As a result of the emergence of advanced healthcare systems, a large amount of patient information is now available (i.e. ,Big Data in Electronic Health Record System) that can be utilized to design predictive models for Cardiovascular disease. Data mining or machine learning is a discovery technique for processing big data from a diversified viewpoint and condensing it into meaningful information. "Data Mining is a non-trivial extraction of implicit, previously unknown, and potentially useful information about data". Today, a huge volume of data regarding disease diagnosis, patients, etc., is created the healthcare industry. Data mining offers some techniques that uncover concealed patterns or similarities in data.

II. LITERATURE REVIEW

Health care awareness and technology developments have led to huge number of hospitals and health care centers. But still, the quality of health care service at an affordable cost is a challenging issue in developing countries has been explored by Anbarasi et al.[2], this fast-moving world, people want to live a very luxurious life, so they work like a machine to earn a lot of money and live a comfortable life. Therefore,in this race, they forget to take care of themselves has been explored by Asam Parveen et al.[5], KNN is one of the simplest and straightforward lazy learning data mining techniques. It is also called memory-based classification as the training samples need to be in memory at run time, and has been explored by M Akhil Jabbar et al.[4], Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes has been explored by C. Kalaiselvi et al.[6], The relationship between cardiovascular risk factors and cardiovascular disease is not linear, necessitating other studies to test new artificial intelligence methods, to assess more patients, and more cardiovascular risk factors have been explored by A. Sitar-Taut et al.[8]. . The primary goal is to compare the various classification methods like Decision Tree, KNN, and Naive Bayes. Subsequently, some performance measures of accuracy, precision, sensitivity, and specificity are tested.

III. PROPOSED METHOD

According to Figure 1, the system is implemented using five major steps in predicting cardiovascular disease.

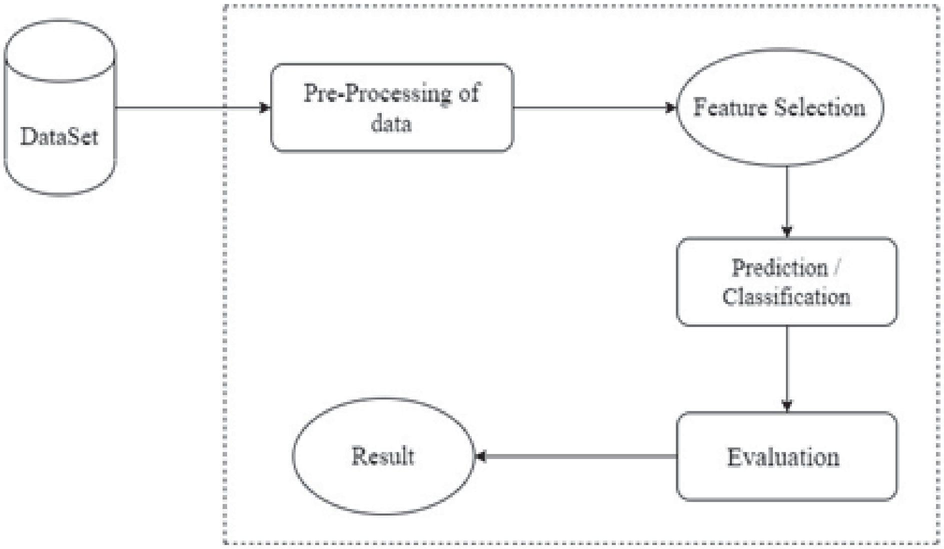

Fig. 1. Experiment Workflow with the dataset

IV. METHOD EXPLANATION

A. Data Description

There are important features of input classified as Attribute Name.

1. Attribute features for factual information from patients.

TABLE I: FEATURE DESCRIPTION

Sr no	Attribute Name	Description
1	Age	Age in years
2	Sex	(1 = male; 0 = female)
3	Cp	Chest Pain
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	Resting electrocardiographic results

8	Thalach	Maximum heart rate achieved
9	Exang	Exercise-induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Target	1 or 0

Features used in the implementation are:
Age, Sex, Chest pain, Cardiovascular disease (Presence), Trestbps, Restecg, FBS, Thalach, Exang, Oldpeak, Slope, Ca, Diastolic blood pressure, and Cholesterol are the features of the dataset used in the model for a description refer (table no.1).

B. Data Pre-processing

The dataset is consists of 14 rows and 300 columns (patient records). After removing similar records, the remaining patient records were used. For the features of the provided dataset, the binary classification and the multiclass parameter are proposed. To examine the presence or absence of heart disease, the multiclass parameter is used to The value 1 indicates that the patient has heart disease. State 0 specifies that the heart disease is absent in the patient. The medical records are transformed into detection values during the pre-processing. Feature Selection and Adding a New Feature

A new feature is that different types of algorithms are appended to the dataset. Logistic regression, KNN , random forest, and Decision tree algorithms are used pre-process the data. This file contains all the pre-processing functions needed to process all input documents and texts. First,we read the train, test,and validation data files,then performed some preprocessing like tokenizing, stemming,etc. There are some exploratory data analyses performed like response variable distribution and data quality checks like null or missing values etc.

C. Experimental setup for evaluation

We have split the dataset into an 80:20 ratio. That is, the training set size is 80%, and the testing set size is 20% of the entire dataset. The training set is used to develop a model, while the testing set is utilized to assess the predictive model's performance. Different models are evaluated using various metrics. Common evaluation metrics for heart disease prediction include accuracy, precision, recall, and F1-score. There are different types of testing:unit testing, integration testing, functional testing, System testing, and White box testing.

D. Classification

Here we have built all the classifiers for heart disease detection. The extracted features are fed into different classifiers. We have used Naive-Bayes, Logistic Regression, Linear SVM, Stochastic gradient descent, and Yolo Classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the F1 score and checked the confusion matrix.

After fitting all the classifiers, the 2 best-performing models were selected as candidate models for heart disease classification. We have performed parameter tuning by implementing GridSearchCV methods on these candidate models and chosen the best-performing parameters for this classifier.

Finally selected model was used for heart disease detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidf Vectorizer to see what words are most important in each of the classes.

We have also used Precision-Recall and learning curves to see how training and test sets perform when we increase the amount of data in our classifiers.

The following algorithms are used :-

(1) Logistic Regression: It is a predictive analysis technique that is used when the target variable is dichotomous (binary). The logistic Regression model explains the relationship between one dependent binary variable and one or more independent variables. It predicts the probability of the target value.

(2) k-Nearest Neighbors algorithm: k-Nearest Neighbors algorithm is a classification algorithm. The class of a particular data point is determined based on the class which is most common among its k nearest neighbors where k is a small positive integer.

(3) Naïve Bayes: It is a set of supervised learning algorithms based on applying Bayes' theorem. The naïve assumption being the conditional independence between every pair of features.

(4) Neural Networks: A neural network is a series of algorithms that recognizes underlying relations in a training data set through a process that vaguely mimics the way of working of the human brain. Basically, neural networks are system of neurons that are either artificial or organic in nature. Neural network adapts to the changing input which allows it to generate the best output.

(5) Decision Tree Classifier: This classifier model applies a Decision Tree as a predictive model. It organizes the characteristics (tree branches) to inferences about the target value (tree leaves). The classification trees are the tree models in which the target parameter can acquire a finite set of values. In these three frameworks, the class labels are signified by the leaves, and the branches describe the concurrences of features that guide those class labels. The regression trees are the decision trees in which the target parameter can take the continuous values (generally real numbers).

(6) Random forest: Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.It is used for both classification and regression tasks.

E. Evaluation:

The classification models were developed using 13 features. The train and test accuracies were calculated for each model. The evaluation of the model was based on seven different classifiers. After Classifying both models. We can use different types of machine learning models.

F. Result:

The decision tree classifier tested with the highest accuracy. To find out which technique predicts cardiovascular disease with more accuracy we used different algorithms such as Neural Networks, K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree Classifier, XGB classifier, and LGBM classifier and their result.

The accuracy score achieved using K-Nearest Neighbors is: 79.0 %

The accuracy score achieved using Decision Tree is: 77.0 %

The accuracy score achieved using Linear Regression is: 43.57096901526326 %

The accuracy score achieved using Random Forest is: 83.0 %

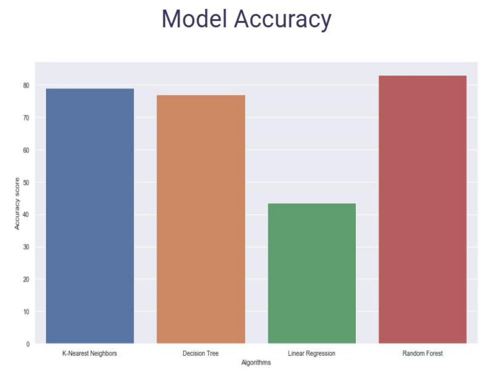

Fig.2 . Model Accuracy of Algorithms.

V. CONCLUSION

Heart disease prediction is essential as well as challenging work in the medical Field. Nevertheless, the mortality rate can be reduced if the disease is recognized at the initial stages, and precautions and proper treatment are possible. This paper illustrates various automated computerized Cardiovascular Disease Prediction methodologies, which can be performed by Supervised Learning plus Classification and Regression methods. The algorithms are tested using various features.

In this project, we introduce the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Yolo Classifications and Logistic Regression: we have analyzed that the Yolo Classifications has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Yolo Classifications by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

VI. ACKNOWLEDGMENT

In the proposed models, for some algorithms, the testing accuracy is slightly greater than the training accuracy. Generally, the testing accuracy should be less than that of the training accuracy. Test data is the data unseen by the model, and train data is the data that the model uses to train itself. Also,note that the difference is minimal.

REFERENCES

[1] P .K. Anooj, ―Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40. Computer Science & Information Technology (CS & IT) 59.

[2] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[3] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[4] M Akhil Jabbar, BL Deekshatulu, Priti Chandra," Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, 11, 2013

[5] Shadab Adam Pattekari and Asma Parveen," PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, Vol 3, Issue 3, 2012, pp 290-294.

[6] C. Kalaiselvi, PhD, "Diagnosis of Heart Disease Using K-Nearest Neighbor Algorithm of Data Mining", IEEE, 2016

[7] Keerthana T. K., " Heart Disease Prediction System using Data Mining Method", International Journal of Engineering Trends and Technology", May 2017

[8] A. Sitar-T˘ aut, D. Zdrenghea, D. Pop, and D. Sitar-T˘ aut, "Using machine learning algorithms in cardiovascular disease risk evaluation," Age, vol. 1, no. 4, p. 4, 2009.

[9] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classifica tion using machine learning algorithms Optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019.

[10] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology

[11] Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".