Early Stage Cardiovascular Disease Prediction Using Machine Learning Techniques

Iliyas Ansari
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai
iliyas.ansari2022@vitstudent.ac.in

Dr. D. Kavitha
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai
kavitha.d@vit.ac.in

Abstract—**Machine Learning is employed across several spheres round the world. Machine Leaning plays a vital role in predicting presence/absence of movement disorders like heart diseases and a lot of. During this era, Individual area unit terribly busy and dealing difficulty so as to satisfy their materialistic wants and unable to pay time for themselves that results in physical stress and mental disturbance. Thus, Cardiovascular disease is incredibly common today. Significantly in urban areas owing to excess mental stress. As a result, Cardiovascular disease has become one among the foremost vital factors for death of men and girls. Within the medical field, predicting the heart disease has become difficult task. So, during this modern life, there is immediate need of a system which can analyze accurately the chance of obtaining heart disease. Predicting a cardiovascular disease in early stage can save several people's life. The most objective of this paper is to style a robust system that works expeditiously and can ready to predict the chance of getting heart disease accurately. Machine Learning (ML) has been showing a good help in creating selections and predictions from the massive amount of knowledge created by the aid industries and hospitals. The predictions model is projected with combos of various options and a numbers of classification techniques. We want to employ methods of machine learning to produce heart disease forecasts. K-Nearest Neighbors Classifier, Support Vector Machine, and Random Forest are a few of the techniques that we want to use. We will study approaches of cardiovascular disease prognosis that include more data at once. Clinical criteria such as sex, age, chest discomfort, cholesterol level, etc., are used to evaluate a person's chance of developing coronary heart disease.**

Keywords- Machine Learning, Heart Disease, Dataset, Decision Tree, Random Forest.

I. INTRODUCTION

Worldwide, heart disease is a frequent cause of death. According to studies undertaken by the World Health Organization, one in three fatalities may be attributable to cardiac issues [11]. Indeed, the heart, a critical organ that circulates blood throughout the body, is a sight to see. When the heart stops beating regularly, it may shut down the brain and other tissues, and a person can die in such a short period of time. For this reason, it is essential to be able to anticipate the development of such illnesses and take preventive actions as they progress.

When cutting-edge diagnostic instruments and highly trained medical professionals are not readily available, it is incredibly difficult to detect and treat cardiovascular issues. Heart disease is characterized by hypertension, angina, and elevated cholesterol levels, intensity, pain level, respiratory distress level, etc. Two separate sets of factors may induce cardiovascular problems. In one group, we have unchangeable traits such as age, gender, and genetic variables. The second group consists of modifiable cardiovascular disease causes. It is within one's capacity to manage risk factors [6]. In addition to genetic tendency, hypertension, and even certain medicines and drinks, [15] other causes of heart disease include hereditary predisposition and hypertension. Data processing, machine learning, deep learning, etc., are only a few of the latest automated methods used to identify health concerns, such as coronary heart disease. As a subfield of AI, Machine Learning can extract insights from enormous databases and generate predictions about fresh data using the same learning or instructional methods [8]. Both the human brain and the computer are able to learn from their mistakes in a similar manner [12]. Consequently, we have such a collection of datasets gathered through Kaggle. By analyzing this data, machines are trained to develop predictions, and the resultant model may be given user inputs to get correct results.

Individuals' computerized medical records include a plethora of information. Additionally, data scientists and researchers now have access to a multitude of tools that assist their involvement in the evolution of medical care due to technological breakthroughs. The data analysis enables us to trace the disease's roots and the healthcare professionals' involvement in increasing awareness and preventing it [3]. The angiography is by far the most common procedure for identifying heart problems. Nevertheless, angiography has advantages. This treatment is seldom utilized [6] because to the intricacy of the diagnostic process and the high expense of the surgery. These limitations push scientists to develop a more targeted method for predicting heart disease. Therefore, there is an urgent need for the development of an automated approach for identifying heart disease utilizing many human health markers.

II. LITERATURE SURVEY

Using Machine Learning and the Internet of Medical Things, Xiaoming Yuan et al. [1] are modelling cardiovascular disease. A Fuzzy-GBDT (gradient boosting tree) approach was created in order to simplify data and increase the generalizability of binary classification. After then, bagging and Fuzzy-GBDT were integrated to avoid overfitting. Following testing, their forecasts for both binary and multiple categories were very accurate and dependable. In the publication [2], we present MaLCaDD (Machine Learning-based Cardiovascular Illness Detection), a machine learning architecture that combines data balancing, feature extraction, and classification to produce more accurate and early heart disease predictions. By merging the findings of many Logistic Regression and KNN classifiers, they were able to increase the predictive ability of the ensemble.

Deepak Kumar Chohan and Dinesh C Dobhal [11] have used the Logistic Regression, Decision Tree, Support Vector Machine (SVM), Nave Bayes, Random Forest, and KNN algorithms to predict heart disease, with the Random Forest algorithm providing the most accurate data (98.53%) out of all the techniques employed. The authors of [4] used the XGBoost method to construct and evaluate models. The researcher has developed a novel method for accurately diagnosing heart abnormalities in real time by analysing short single-lead ECGs (9-45 seconds) (9-61 seconds). To detect and diagnose heart disease, the author of [3] used a combination of models including Logistic Regression, Support Vector Machine, k-Nearest Neighbors (KNN), Decision Tree, and Random Forest. Using the voting ensemble approach, researchers achieved 98.18 percent accuracy using Random Forest. Chunyan Guo et al. [5] use the Recursion Enhanced Random Forest with a more accurate linear model to detect heart defect. In addition, creating an Artificial Neural Network with a features selection and backpropagation learning technique for sickness classification. In [12], Akanksha Kumari and Ashok Kumar Mehta tried to predict heart disease using seven machine learning methods and aimed to enhance the performance of poorly performing models by using ensemble approaches such as AdaBoost and Voting Ensemble methodology. In [7], Mohammed Nowshad Ruhani et al. have trained their design utilising classification techniques such as Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Nave Bayes, Support Vector Machine, etc. Despite the fact that precision for numerous optimization techniques varies for a specific number of cases in the data source, SVM achieves the most accurate results with an accuracy of 91%. Instead of gathering datasets from internet databases like as Kaggle, UCI, etc., they individually obtained datasets from a number of Medical Organizations. Mihir J. Gaikwad et al. developed a model to predict cardiac disease using five ML approaches (Support Vector Machine, Random Forest, Gradient Boosting, Supply Regression, and Decision Tree Classifier) in [13]. The prognosis of each comparison to determine which is most compatible with the prognosis. D. P. Yadav et al. [14] have created a machine learning and optimization system to aid physicians. In [15], Likitha KN et al. study a variety of machine learning algorithms for estimating the resting area unit supply of internal organs. They used Machine Learning techniques and examined all of the characteristics, including age, chest pain, pulse rate (BP), gender, steroid, alcohol, and pulse.

Using Python and the Pandas package, Narendra Mohan et al. [6] attempted to predict cardiovascular disease. Initially, datasets will be divided into preparation and sample datasets for the length of this activity. Using a few important medical factors and four machine learning models (KNN, NB, LR, and RF), the researchers predicted in-person sickness using KNN, NB, LR, and RF. Ahmed Al Ahdal et al. [8] identified heart problems using six machine learning techniques. Using the robust machine learning algorithmic software random forest, M. Snehith Raja et al. [9] developed a cardiopathy prediction system that produces precise findings rapidly. Yu Lin investigated a Cleveland-based cardiovascular disease dataset in [10]. As part of the system training procedure, Logistic Regression, K-nearest Neighbors, Adaboost, CART, Random Forest, and XGBoost were used. Random Forest was by a significant measure the most accurate of these available models, with a prediction accuracy of 84.40%.

To help with center failure detection, Ashir Javeed et al. [16] developed a unique training strategy to combat overfitting, a problem plaguing the newly envisioned systems for forecasting heart illness. The projected outcome must overfit the available data in order to pass the tests. In order to construct a smart system with respectable accuracy on both testing and training data, the author came up with a new diagnostic method. Both the random search algorithm and the random forest algorithm are used in the proposed method for predicting coronary heart disease. Senthilkumar Mohan et al. [17] devised a strategy for enhancing the precision of cardiovascular disease predictions using machine learning techniques. It was found that the predictive approach had an accuracy of 88.7 percent. The proposed HRFLM method is a hybrid of the Linear Method and the Random Forest. Synthesized Minority Under Method Nearest Neighbor (SMOTE-ENN) is proposed by Norma Latif Fitriyani et al. in [18] to normalize an uneven training dataset, while Density - Based Spatial Clustering of Applications with Noise (DBSCAN) is proposed to find and exclude outliers. They used a technique called Extreme Gradient Boosting (XGBoost) to create their prognoses for heart failure.

III. DATASET

To train our machine learning (ML) model, we will acquire a dataset from Kaggle. This data collection consists of 1025 health records with 14 distinct attributes, such as age, gender, etc. In Table I, the characteristics and their descriptions are listed.

Attributes	Description	Type
age	Age of patients	Numeric
sex	Gender of patient: • 0 = female • 1 = male	Categoric
cp	Type of chest pain: • 0 = typical angina • 1 = atypical angina • 2 = non-anginal pain	Categoric

	• 3 = asymptomatic	
trestbps	The patient's resting blood pressure was measured in millimetres of mercury before to hospitalization (mm Hg).	Numeric
chol	Cholesterol level in mg/dl	Numeric
fbs	blood sugar > 120 mg/dl: • 0 = no • 1 = yes	Categoric
restecg	Conclusions from a resting electrocardiogram: • 0 = normal • 1 = ST elevation or depression of > 0.05mV and/or T wave inversions. • 2 = According to Estes' criteria, this score indicates that left ventricular hypertrophy is likely or present.	Categoric
thalach	heart rate	Numeric
exang	Symptoms of angina brought on by exercise: • 0 = no • 1 = yes	Categoric
oldpeak	Exertion-induced ST depression compared to resting ST elevation	Numeric
slope	The exercise-induced decline in the ST region: • 0 = upsloping • 1 = flat • 2 = downsloping	Categoric
Ca	Fluoroscopic vessel coloring (from 0-3)	Categoric
thal	here the outcomes of the thallium stress test: • 1 = normal • 2 = fixed defect • 3 = reversible defect	Categoric
target	Have heart disease: • 0 = no • 1 = yes	Categoric

Table 1: DESCRIPTION OF DATASET ATTRIBUTES

IV. PROPOSED METHODOLOGY

Our suggested method for the diagnosis of heart disease is not only basic in its operation, but also in its implementation and use. Figure 1 illustrates that the first stage in constructing our database is to collect previously acquired patient information. The characteristics of our dataset are shown in Table 1 below. This dataset was gathered using the Kaggle platform. After accumulating a dataset, we are verifying that no essential information is missing. If a value is missing from a record in our dataset, just the remaining records are used for preprocessing. The multiclass variable determines whether or not cardiovascular disease is present. Setting the value to 1 indicates heart illness, whereas setting the value to 0 indicates the absence of heart disease. In the data preprocessing step, medical records are converted into a diagnostic score. After cleaning and organizing the data, we will split it into two groups: 70% of the data will be used to train our model, while 30% will be used to test it.

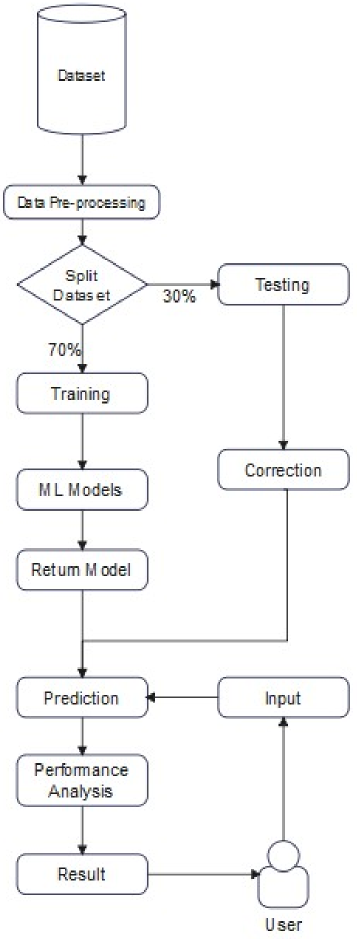

Fig 1: Process Chart

After training and testing we choose our model on basis of the model which gives the correct result with highest accuracy. Then after completion of model user just need to put input of each attributes, then our model will be able to give result with a best accuracy if user have heart disease or not.

V. MACHINE LEARNING ALGORITHM FOR DETECTING HEART DISEASE

The Kaggle dataset is preprocessed in order to detect any missing or redundant data. Before being given to the classifiers for processing and calculating the predicted accuracy, these values will be deleted and replaced with accurate ones, if feasible. Accepting and evaluating with the test data the results of the manufactured classifiers, the one with the highest accuracy, is permitted.

A. Heart Disease Prediction

Our primary purpose is to create a model that yields a high-precision result, which we will achieve through a variety of strategies. The medical agency's vast data production may be very beneficial for early illness identification.

B. Medical Strategies for Treating Heart Disease

KNN, SVM, DT, and RF are used to classify the properties of the dataset, which include age, sex, cholesterol level, etc. The input dataset was separated into a training dataset containing 70% of the data and a testing dataset containing 30% of the data. After training a model, it is evaluated on a separate dataset.

(i) Random Forest (RF):

Random Forest mixes the average results of several decision trees that each reflect a distinct subset of the data to improve the quality of the model. The algorithm's accuracy increases with a bigger dataset.

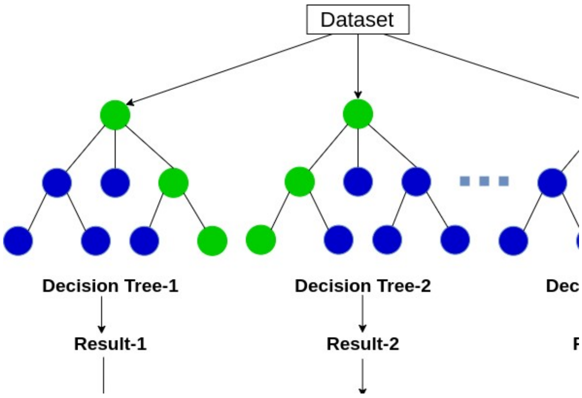

Fig 2: Random Forest

(ii) Support Vector Machine (SVM):

Similar to DT, SVM may be used to address classification and regression problems. SVM has numerous uses, however classification is its principal function. The SVM approach generates a hyperplane or line that categorizes the data. We may use a nonlinear dataset as we are constructing this model for health data. Support Vector Machine might thus be a potential option.

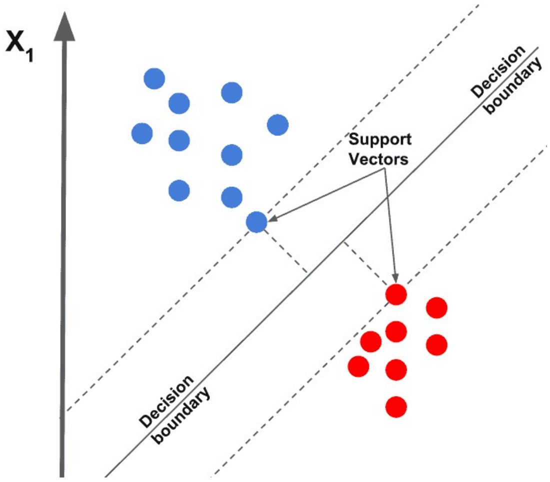

Fig 3: SVM

(iii) K- Nearest Neighbor (KNN):

In addition to its classification applications, KNN is also useful for regression analysis. For the KNN method to function, each data point must be considered a neighbour. This strategy includes locating the k nearest data points in the training set to the data point where a target value is missing, and then assigning the mean of the values discovered to the missing values.

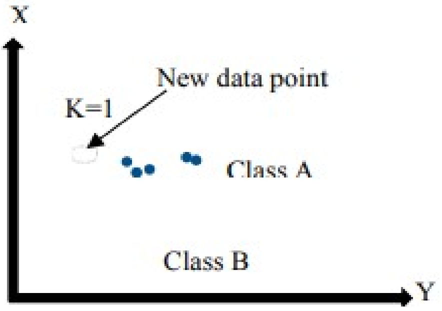

Fig 4: KNN

VI. RESULT ANALYSIS

The Confusion Matrix may be used to assess the efficiency of classification methods. It depicts the predicted and actual outcomes graphically. The (n x n) matrix displays the frequency of accurate and incorrect predictions.

	Actual: NO	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

Fig 5: Confusion Matrix

- True Negative (TN): Model has predicted the disease No, and in real the person is not suffering from heart disease.
- True Positive (TP): Model has predicted the disease Yes, and in real the person is suffering from heart disease.
- False Negative (FN): The model has predicted the disease No, but in real the person is suffering from heart disease.
- False Positive (FP): The model has predicted the disease Yes, but in real the person is not suffering from heart disease.

After completion of confusion matrix, the confusion matrix will help us to find the accuracy level, error rate, etc. of our results.

A. Accuracy**:**

It defines the how much our model predicts the result correctly. It is the ratio of total true prediction to total prediction.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

B. Error rate:

In other words, it quantifies the degree to which our model was off in its prediction. It represents the proportion of incorrect predictions relative to all predictions.

$$\text{Error Rate} = \frac{FP+FN}{TP+TN+FP+FN}$$

C. Precision:

It defines the accuracy of positive prediction. It is the ratio of actual true prediction to total positive prediction.

$$\text{Precision} = \frac{TP}{TP+FP}$$

D. Recall:

Recall shows the proportion of valid findings that were properly labelled. In other words, it measures how accurate a forecast was compared to the total number of good results.

$$\text{Recall} = \frac{TP}{TP+FN}$$

VII. CONCLUSION

This study focuses on the challenges surrounding the diagnosis of cardiovascular disease, such as the difficulty and complexity of establishing the diagnosis and the patient's high cost for the diagnostic test. This research presented an automated strategy for predicting heart illness based on machine learning methods. The dataset, which is accessible on Kaggle, has 1,025 patient records and 14 distinct attributes. To learn and classify based on these traits, ML techniques such as K-Nearest Neighbors Classifier, Support Vector Machine, and Random Forest have been used. In the future, we may construct a website that allows visitors to self-evaluate their risk of acquiring heart disease by contributing their own data to the model.

REFERENCES

[1] Xiaoming Yuan, Jiahui Chen, Kuan Zhang, Yuan Wu and Tingting Yang “A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT” IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 18, NO. 3, MARCH 2022

[2] Aqsa Rahim, Yawar Rasheed, Farooque Azam, Muhammad Waseem Anwar, Muhammad Abdul Rahim, And Abdul Wahab Muzaffar “An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases” IEEE Access VOLUME 9, 2021

[3] Sarria E. A. Ashri, M. M. El-gayar, And Eman M. EL-Daydamony “HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm” IEEE Access VOLUME 9, 2021

[4] Dimitris Bertsimas, Luca Mingardi, Bartolomeo Stellato “Machine Learning for Real-Time Heart Disease Prediction” IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 25, NO. 9, SEPTEMBER 2021

[5] Chunyan Guo, Jiabing Zhang, Yang Liu, Yaying Xie, Zhiqiang Han, Jianshe Yu “Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform” IEEE Access VOLUME 8, 2020

[6] Narendra Mohan, Vinod Jain, Gauranshi Agrawal, “Heart Disease Prediction Using Supervised Machine Learning Algorithms” Oct 22-23, 2021

[7] Mohammed Nowshad Ruhani Chowdhury, Ezaz Ahmed, Md. Abu Dayan Siddik, Akhlak Uz Zaman, “Heart Disease Prognosis Using Machine Learning Classification Techniques” Apr 02-04, 2021

[8] Ahmed Al Ahda, Dr. Deepak Prashar, Manik Rakhra, Ankita Wadhawan, “Machine Learning-Based Heart Patient Scanning, Visualization, and Monitoring” 2021 International Conference on Computing Sciences (ICCS)

[9] M.Snehith Raja, M.Anurag , Ch.Prachetan Reddy, NageswaraRao Sirisala, “Machine Learning Based Heart Disease Prediction System” Jan 27-29

[10] Yu Lin Khoury College of Computer Science, Northeastern University Boston, The United States “Prediction and Analysis of Heart Disease Using Machine Learning” 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence

[11] Deepak kumar chohan and Dinesh C Dobhal, “A Comparison Based Study of Supervised Machine Learning Algorithms for Prediction of Heart Disease” 1st International Conference on Computational Intelligence and Sustainable Engineering Solution (CISES-2022)

[12] [Akanksha Kumari, Ashok Kumar Mehta, “A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms” Aug 28-29, 2021

[13] Mihir J. Gaikwad, Prathmesh S. Asole, Prof. Leela S. Bitla, “Effective Study of Machine Learning Algorithms for Heart Disease Prediction” Jan 21-22, 2022

[14] D.P.Yadav, Prabhav Saini, Pragya Mittal, “Feature Optimization Based Heart Disease Prediction using Machine Learning” Oct 22-23, 2021

[15] Likitha KN, Nethravathi R, Nithyashree K, Ritika Kumari, Sridhar N, Venkateswaran K, “Heart Disease Detection using Machine Learning Technique” Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021) IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

[16] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour “An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection” IEEE Access VOLUME 7, 2019

[17] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava “Effective Heart Disease Prediction Using Hybrid

Machine Learning Techniques” IEEE Access VOLUME 7, 2019

[18] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, Jongtae Rhee “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System” IEEE Access VOLUME 8, 2020