Estimation of Prediction for Heart Failure Chances Using Various Machine Learning Algorithms

Ashwin Chavan[1], Bhairavi Chitnavis[2], Poorva Wadkar[3], Md. Ubaid Khan[4]

BE Students, Department of Computer Engineering, Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India

---

Abstract- The human heart is certainly the most important organ in our body. Our body cannot function normally if our heart cannot circulate enough blood to all our internal organs. Abnormalities in pumping blood causes heart failure. Though the term was coined in the 17th century it remains a global pandemic affecting over 26 million people worldwide. Hence predicting this deadly disease beforehand can do wonders for an individual as they can look after their health and fitness. Medical professionals find it difficult to come up with a scalable solution to predict the chances of heart failure. This is where advanced technologies like Machine learning can be used. With the help of Machine learning models, we can estimate if a person has chances of heart failure in the coming 10 years. In this study, we use a variety of machine learning methods to accurately predict heart failure. Here, we examined a dataset on heart failure that included significant pertinent medical data on 4238 patients. We have included the most crucial factors which play an important role in predicting if a person has chance of suffering from heart failure. We have implemented prediction models using various machine learning classification Algorithms. According to the findings of our study, in contrast with different machine learning algorithms, Random Forest had the greatest Accuracy = 96% as well as AUC = 99% when estimating the likelihood that patients would experience heart failure.

Keywords: Random Forest, machine learning model, heart failure prediction, Disease Prediction, Accuracy

I. INTRODUCTION-

Since a few years ago, the prevalence of cardiovascular diseases has been rising quickly throughout the world. Even though these diseases have been identified as the leading cause of death, they have also been identified as the most controllable and preventable diseases. Heart stroke is mostly brought on by artery obstruction. It happens when the heart cannot effectively pump blood throughout the body.

One of the major contributing factors to developing heart disease is high blood pressure. According to a report, 35% of people worldwide had hypertension between 2011 and 2014, which is a risk factor for heart disease. Like this, other variables like obesity, poor nutrition, high cholesterol, and insufficient exercise can result in heart disease. As a result, prevention is essential. It's essential to comprehend heart diseases to prevent them. The fact that almost 47% of fatalities take place outside of a hospital shows how frequently warning signs are disregarded.

The diagnosis of heart conditions is a big barrier. Finding out if someone has a heart condition or not might be difficult.

Even though there are devices that can forecast heart disease, they are either prohibitively expensive or inefficient at predicting the likelihood of heart disease in people. There has been extensive research done in this field because, according to a World Health Organization (WHO) report, only 67% of cardiac diseases can be predicted by medical professionals. In rural areas of India, there is a serious lack of access to hospitals and high-quality medical care. Only 58% of doctors in urban areas and 19% in rural areas have medical degrees, according to a 2016 WHO report.

To anticipate any heart disease in people, machine learning may be a promising option. Heart disorders are a serious challenge for medical science. Neural networks, decision trees, KNNs, and other methods can be used to forecast heart disorders. We will learn how to utilize Random Forest to find the accuracy for heart disease later in this study. Additionally, it demonstrates how ML will help us fight heart disease in the future.

II. RELATED WORK -

The literature is full of research studies on utilizing machine learning to diagnose cardiac problems. Here is a basic overview of that.

According to a study by Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, and Nidhi Lal that was released in January 2020 and titled "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," they achieved 87% accuracy using the logistic regression technique. According to their research, men tend

to be more susceptible to cardiovascular disease than women. Heart disease is also influenced by aging, daily cigarette consumption, and systolic blood pressure. The likelihood of CHD is not significantly altered by total cholesterol. The amount of HDL in the total cholesterol value may be the cause of this. The effects of glucose are equally insignificant. More data and the use of additional machine learning models, they said, might enhance the model [1].

A further study entitled "Heart failure survival prediction using machine learning algorithm: Am I Safe from Heart Failure?" is also available. Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, and Md Minhazur Rahman published this in 2022. They said that LightGBM, which has an excellent accuracy (85%) with AUC (93%), was the machine learning model that performed most successfully in their situation. The least reliable model, Decision Tree, scored 73.21% [2].

Aleyna Nur Karaoglu, Hasan Caglar, Ali Degirmenci, and Omer Karal published a study in 2021 titled "Performance Improvement with Decision Tree in Predicting Heart Failure". Three different ML methods are being considered in this study to forecast heart failure survival among patients. Here, the classification algorithms Decision Tree, KNN, and Logistic Regression are used. According to the testing findings, the Decision Tree algorithm scores optimum in terms of accuracy at 84.48%, and precision at 79% with an F1-score of 82%[3].

A study titled "Heart Disease Prediction Using Random Forest Algorithm" was released in March 2022 by Kompella Sri Charan and Kolluru S S N S Mahendranath. By evaluating the accuracy scores of Decision Tree, Random Forest, SVM, Ada boost, and Gradient Boosting algorithms, the study determined the most effective machine learning method for the identification of heart illnesses. The outcome shows that the Random Forest algorithm, which has an accuracy rating of 92.16% in the forecasting of heart disease, is the most effective. The literature demonstrates that there is, above all, room for improvement and more potential for the model[9].

In the paper "Heart Disease Diagnosis Using Data Mining Technique", by Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, and Hena M, multiple attributes were incorporated in K-means algorithms, MAFIA algorithm, as well as Decision tree classification, applying a data mining technique for the treatment of heart disease. They concluded that Decision Tree possesses greater effectiveness by employing 14 attributes, after applying generalized linear models[8].

In another study with the title "Predictions in Heart Disease Using Techniques of Data Mining "Published by Monika Gandhi and Dr. Shailendra Narayan Singh, their methods use concealed patterns to find the best course of action for organizations providing healthcare[10].

III. DATASET INFORMATION-

We have used the dataset from previous cardiovascular research of people living in the Massachusetts town of Framingham which is available on the Kaggle website. The objective of the project is to predict whether the patient has a 10-year risk of developing future coronary heart disease (CHD)[11]. This dataset contains a total of 4238 medical records of heart patients who were monitored throughout their therapy. The profile of each patient contains a total of 16 medical features. All the features are explained in TABLE I.

TABLE I Dataset attribute names and information

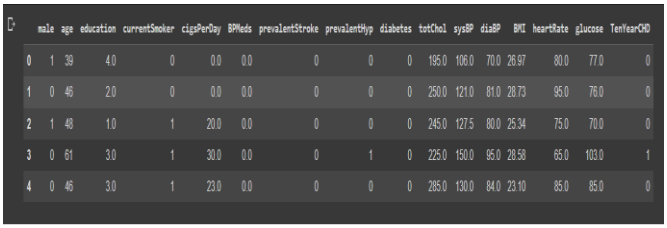
Attribute names (Feature variables)	**Attribute information**
Male	Male or Female
Age	Patient's Age
Education	0: Less than High School and High School degrees, 1: College Degree and Higher
currentSmoker	Whether a patient is a current smoker or not
cigsPerDay	Per day Cigarette intake
BPMeds	Blood Pressure medications
prevalentStroke	Current Strokes
prevalentHyp	Current Hypertension
Diabetes	Whether the patient has diabetes or not
totChol	Patient's Total Cholesterol Level
sysBP	Systolic blood pressure
diaBP	Diastolic blood pressure
BMI	Body Mass Index
heartRate	Heart rate of the patient
glucose	Patient's Glucose Level
TenYearCHD [Target]	If the patient has heart disease or not (binary)

In this dataset total of 15 features are independent and 1 target feature is dependent. age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, and glucose

are independent features. Whereas TenYearCHD is a target variable.

IV. METHODOLOGY-

The first step in model building is data collection or data extraction followed by dataset pre-processing. Fig. 1 shows the first 5 records of the dataset.



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Fig 1: Dataset

Once the dataset is cleaned and transformed into the proper format, classifiers like Logistic Regression, Random Forest, Support Vector Machine, Kernal-Support Vector Machine, and LightGBM are employed for training and testing on the dataset. k-fold cross-validation is used to compute the performance of the model. Finally, all models were evaluated and compared to find the best model for heart failure prediction.

Dataset pre-processing

The dataset has been pre-processed by different Feature Engineering and Feature selection techniques. Inside **Feature Engineering** we have performed ***Exploratory Data Analysis***,

--Identifying Numerical and Categorical Features

--Finding Missing Values

--Detecting Outlier

--Data Cleaning

Also, the Imbalanced Dataset is Balanced and Outliers are removed.

Inside **Feature selection**, we have analyzed the Correlation between all the features with the help Heatmap shown in Fig. 2

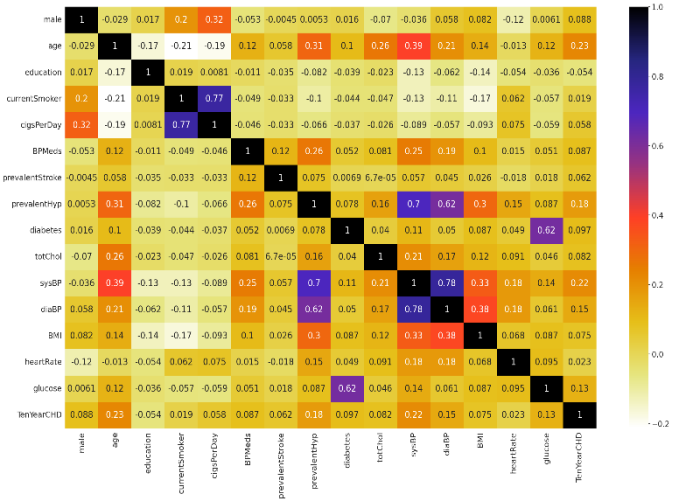


Fig 2: Correlation matrix of the heart failure dataset

Validation method

In Machine Learning Cross Validation is used to evaluate the performance of a trained model on unseen data. We have selected the K-fold cross-validation method for our dataset. Here data sample is divided into 10 folds, i.e., k=10.

Model Development

We have developed total of 5 machine-learning models with algorithms like Logistic Regression, Random Forest, Support Vector Machine, Kernal-Support Vector Machine, and LightGBM. *By training them with our pre-processed dataset, we compared their performance based on different Evaluation Metrics. Out of 5 models, we found that the model with the Random Forest algorithm is performing well in prediction.*

Model Evaluation

We have evaluated our all models based on the following metrics,

--Accuracy

-- Confusion Matrix

-- Precision

-- Recall

-- F1 Score

-- AUC-ROC

V. RESULT AND DISCUSSION-

Table 2 compares the performance results for the chosen machine learning algorithms, including LightGBM, Random Forest, Support Vector Machine, and Kernal Support Vector Machine. Following that, we examined the model's performance, as depicted in Figure 3.

We have listed the Train Accuracy, Test Accuracy, Precision, Recall, and F1-Score data in Table 2 for tracking model performance. After studying the dataset, it is seen that men have a high risk of heart disease than women. After inspecting the dataset more it is seen that the increase in factors like age, cigsPerDay, and heartrate shows high chances of an increase in heart disease.

Table II: - Result of various machine learning based on different measurement

ML Algorithms	SVM	LR	K-SVM	LightGBM	RF
Train Accuracy	0.67	0.67	0.67	**0.95**	**1.0**
Test Accuracy	0.66	0.67	0.67	**0.85**	**0.96**
Precision	0.68	0.69	0.72	0.84	0.93
Recall	0.71	0.69	0.58	0.92	0.99
F1-Score	0.7	0.69	0.64	0.88	0.96
AUC Score	0.732	0.713	0.735	0.945	0.99

According to the observation in Figure 3, Random Forest has the best accuracy (96%), whereas SVM has the lowest accuracy (66%). In addition, LightGBM has also performed well with an accuracy of 85%. And from Figure 4, it is seen that the ROC Curve of Random Forest is good and close to 1.0 which tells us that our model is good.

Finally, it can be said that the aim of the project is completed as it has beaten the existing system results

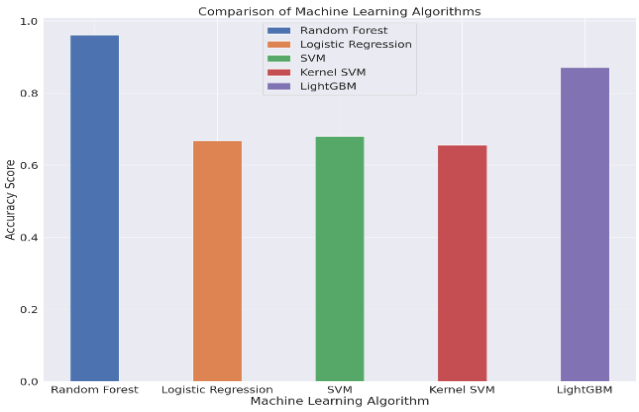


Fig. 3: Examination of accuracy for heart disease prediction using machine learning algorithms.

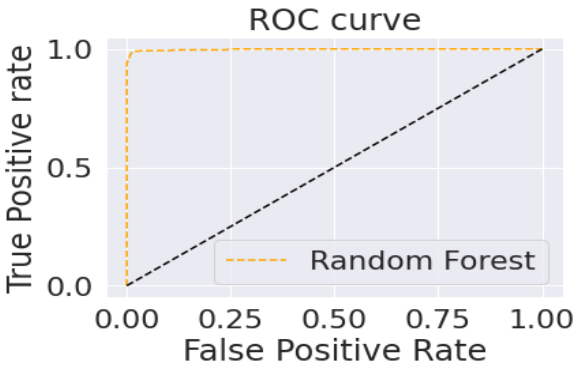


Fig 4. AUC-ROC Graph of Random Forest

Which had an accuracy of 87%.It can be said that the overall performance of Random Forest in terms of Accuracy and AUC Score is good compared to other algorithms.

VI. CONCLUSION-

Heart problems may become more prevalent than they are now to a maximal level. Heart conditions are challenging. Unfortunately, this sickness claims the lives of many people each year. Manually calculating the likelihood of developing heart disease based on the above-mentioned risk factors is challenging. This software can anticipate heart problems for any non-medical employee, which will save doctors time. The fact that this work only applies classification techniques and algorithms to the prediction of heart disease is one of its main shortcomings. To increase the precision of heart disease prediction, implementation is still a work in progress.

VII. FUTURE WORK

The future scope of heart disease prediction using machine learning is very promising, and there are several areas where this technology can make a significant impact.

Wearable devices and other sensors can collect real-time data on a person's heart rate, blood pressure, and other vital signs. Machine learning algorithms can be used to analyze this data and provide real-time alerts to doctors and patients if any concerning trends or abnormalities are detected. Nowadays most of the data is computerized and everything is on the cloud which can be accessed, analyzed, and use for model training. Again, we can use advanced techniques and algorithms which help to make predictions in less time complexity and with more accuracy.

VIII. REFERENCES-

[1] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104210.

[2] M. Mamun, A. Farjana, M. A. Mamun, M. S. Ahammed and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?" 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2022, pp. 194-200, doi: 10.1109/AIIoT54504.2022.9817303.

[3] A. N. Karaoglu, H. Caglar, A. Degirmenci, and O. Karal, "Performance Improvement with Decision Tree in Predicting Heart Failure," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 781-784, doi: 10.1109/UBMK52708.2021.9558939

[4] T. P. Pushpavathi, S. Kumari, and N. K. Kubra, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 915-923, doi: 10.1109/ICICT50816.2021.9358733.

[5] C. De Silva and P. Kumarawadu, "Performance Analysis of Machine Learning Classification Algorithms in the Case of Heart Failure Prediction," 2022 International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, 2022, pp. 1160-1165, doi: 10.1109/IWCMC55113.2022.9824214.

[6] D. Mehta, A. Naik, R. Kaul, P. Mehta, and P. J. Bide, "Death by heart failure prediction using ML algorithms," 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2021, pp,-1-5,doi: 10.1109/ICNTE51185,2021.9487652.

[7] X. Sang, Q. Z. Yao, L. Ma, H. W. Cai, and P. Luo, "Study on survival prediction of patients with heart failure based on support vector machine algorithm," 2020 International Conference on Robots & Intelligent System (ICRIS), Sanya, China, 2020, pp. 636-639, doi: 10.1109/ICRIS52159.2020.00160.

[8] S. Babu et al., "Heart disease diagnosis using data mining technique," 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017, pp. 750-753, doi: 10.1109/ICECA.2017.8203643.

[9] K. Sri Charan,K.S S N S Mahendranath, M.Thirunavukkarasu,' Heart Disease Prediction Using Random Forest Algorithm', International Research Journal of Engineering and Technology (IRJET),Volume: 09 Issue: 03 Mar 2022 ,e-ISSN: 2395-0056,p-ISSN: 2395-0072

[10] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015, pp. 520-525, doi: 10.1109/ABLAZE.2015.7154917.

[11]https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data

IX. Biographies–

Ashwin Chavan, Final Year BE Student, Department of Computer Engineering, Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India

Bhairavi Chitnavis, Final Year BE Student, Department of Computer Engineering, Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India

Poorva Wadkar, Final Year BE Student, Department of Computer Engineering, Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India

Md. Ubaid Khan, Final Year BE Student, Department of Computer Engineering, Dr. D. Y. Patil College of Engineering, Ambi, Pune, Maharashtra, India