Heart Disease Prediction Using Machine Learning

Mr. S. S. Mhaske[1], Sakshi Ghatol[2], Vaishnavi Sontakke[3],
ENTC Department, MGICOET, Shegaon, Maharashtra, India

***Abstract*—This paper focuses on Heart diseases that are a leading cause of death worldwide, and early detection can significantly improve survival rates. Machine Learning (ML), a form of artificial intelligence, has become a valuable tool in healthcare for analyzing health data and identifying potential signs of disease. In this study, we developed an ML model to predict heart disease using the Cleveland heart disease dataset, applying a feature selection method to reduce the number of features while retaining the most important ones to enhance model performance.**

We trained multiple machine learning algorithms and compared their results. The Random Forest classifier outperformed the others, achieving 99.99% Sensitivity, 98.37% Specificity, 98.47% Accuracy, and an Area Under the Curve (AUC) of 94.48%. These results demonstrate that combining effective feature selection with Random Forest can generate highly reliable predictions for heart disease. This approach could assist healthcare professionals in detecting heart disease early and making better clinical decisions

***Index Terms*—Heart Disease, Machine Learning, Feature Selection, Cleveland Dataset, Health Prediction**

I. INTRODUCTION

According to the World Health Organization, heart disease is still the leading cause of death worldwide, making up almost one-third of all deaths each year. Even though there have been improvements in diagnosing and treating heart disease, the number of people affected continues to rise. Heart disease refers to a group of cardiac conditions including coronary artery disease (CAD), Arrhythmia, Heart Valve Disease, and Heart Failure that impair the heart's ability to pump blood effectively the most common type of heart disease is coronary artery disease, which happens when the arteries that supply blood to the heart become narrow or blocked due to plaque buildup.

This reduces blood flow to the heart and can lead to serious problems like heart attacks, irregular heartbeats (arrhythmias), or heart failure. Arrhythmias, which are caused by abnormal electrical signals in the heart, are the second most common heart condition. They can be mild or very dangerous. Certain types, like Atrial Fibrillation and Ventricular Fibrillation, increase the risk of stroke and sudden death.

Lifestyle factors such as poor diet, lack of exercise, and increased consumption of processed foods are major contributors to the rising incidence of heart disease. Early detection is crucial to prevent progression, but traditional clinical diagnostic methods are often invasive, time-consuming, and costly. Moreover, limited patient participation in clinical trials hampers early diagnosis efforts. As a result, there is a growing shift toward data-driven approaches that analyze patient health data to identify patterns and predict disease outcomes

Machine learning (ML) is a helpful tool for diagnosing heart disease. It can quickly and cost-effectively analyze data like age, cholesterol, blood pressure, and ECG to detect heart problems accurately. Although it doesn't replace doctors, ML improves diagnosis and helps doctors make better decisions. As more medical data becomes available and technology improves, ML will play a key role in spotting and preventing heart disease early.

II. PROPOSED MODEL

The proposed work focuses on predicting heart disease using four classification algorithms and evaluating their performance. Health professionals provide patient data, which the model uses to estimate the likelihood of heart disease. The study compares the effectiveness of each algorithm in making accurate predictions. Figure 1 shows the overall process flow. The goal is to assist healthcare professionals in early diagnosis and decision-making.

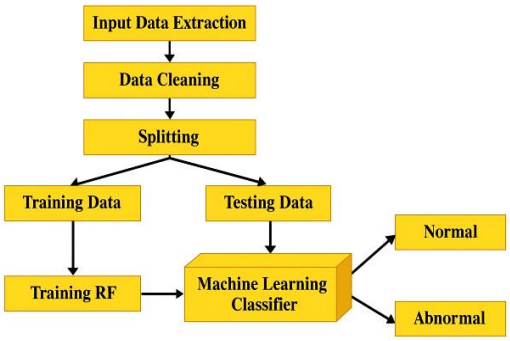


Fig. 1: Generic Model Predicting Heart Disease

A. Data Collection and Prepossessing

The online data set from the GitHub website is used, which has 1026 data samples, and these data samples are used to train the random forest classifier. There is total 14 attributes.

Attribute	Description	Range
Age	Age of the patient	29 to 77
Sex	0 = Female, 1 = Male	0, 1
CP	0 = typical angina 1 = atypical 2 = non-anginal 3 = asymptomatic	0, 1, 2, 3
RestBP	Resting blood pressure (in mm Hg)	94 to 200
Chol	Serum cholesterol in mg/dl	126 to 564
FBS	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)	0, 1
Heart Rate (Thalach)	Maximum heart rate achieved	71 to 202
Exang	1 = yes, 0 = no	0, 1
Oldpeak	ST depression induced by exercise relative to rest	0.0 to 6.2
Slope	0 = up-sloping 1 = flat 2 = down-sloping	0, 1, 2
CA	Number of major vessels colored by fluoroscopy	0 to 3
Thal	1 = normal, 2 = fixed defect 3 = reversible defect	1, 2, 3

Table 1: Attributes

B. Classification

Input features are analyzed using ML algorithms like Random Forest etc., The dataset is split 80/20 for training and testing to predict heart disease likelihood.

I. Random Forest

Random Forest is an ensemble method that combines the predictions of multiple decision trees.

Let {T1(x), T2(x),… Tn(x)} be the individual decision trees in the forest, where Ti(x) is the prediction of the i-th tree for input x.

Majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

For regression:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} T_i(x)$$

II. Decision Tree

A Decision Tree represents decisions and their possible consequences in a tree-like format. It works by dividing data based on feature values and making predictions by following a path from the root to a leaf. It is simple, interpret-able, and requires minimal data preprocessing.

III. Logistic Regression

Logistic Regression is widely used for solving binary classification problems. It applies the logistic (sigmoid) function to map predicted values between 0 and 1. When working with multiple independent variables, it remains a strong performer in classification scenarios.

IV. K-Nearest Neighbors (KNN)

KNN is an intuitive, instance-based algorithm that classifies new inputs by examining the 'K' closest data points in the feature space. The majority label among these neighbors determines the class. It is best suited for small datasets and problems with fewer features.

V. Support Vector Machine (SVM)

SVM identifies the best decision boundary (hyperplane) that separates different classes. It excels in high-dimensional data and scenarios with a clear margin of separation. With the help of kernel functions, it can also solve non-linear classification problems effectively.

VI. Gradient Boosting

Gradient Boosting is a powerful ensemble method that sequentially builds models, where each model aims to fix the errors of the previous one. Often using decision trees as base learners, it achieves high accuracy and is frequently applied in practical machine learning solutions.

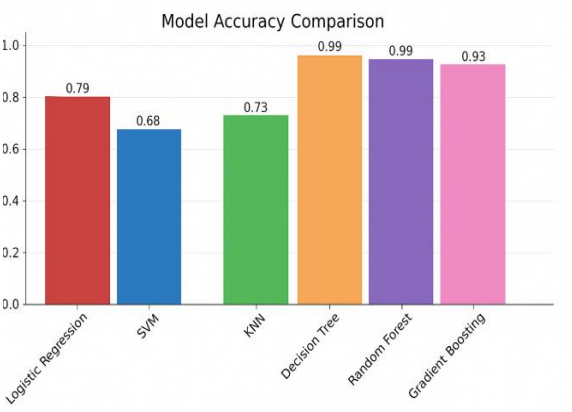


Fig .2: Model accuracy

Fig.3 : Output Image

According to datasets that we have load above accuracy of various model we have got, The bar chart shows the accuracy of various models used for heart disease prediction. Random Forest and Decion Tree achieve the highest accuracy of 99%, followed by Gradient Booster 93%. So that we have consider Random Forest model for Further analysis.

III. TOOLS AND TECHNOLOGY USED

i. Python

Python is a high-level, general-purpose programming language known for its readability and flexibility. It supports multiple paradigms, such as procedural, object-oriented, and functional programming, making it ideal for machine learning, data science, and automation.

ii. Kaggle

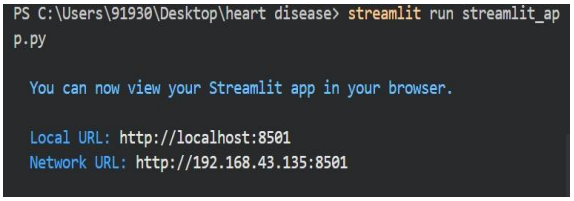
Kaggle is a community-driven platform for data science and machine learning practitioners. It offers a wide range of public datasets, competitions, cloud-based notebooks, and tools for sharing code and collaborating with others.

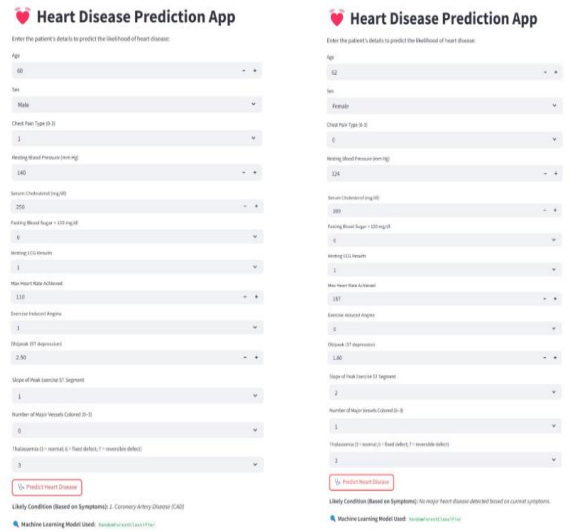
iii. Visual Studio Code (VS Code)

VS Code is a lightweight, open-source code editor developed by Microsoft. It supports a wide range of programming languages and is highly extensible through plugins, making it a powerful tool for coding, debugging, and building machine learning models locally.

iv. Streamlit

Streamlit is an open-source Python framework used to create interactive web applications for data science and machine learning. It allows developers to turn data scripts into shareable web apps quickly, with minimal effort and no front-end experience required.





The image shows a Heart Disease Prediction App interface where users input patient health data to predict the risk of heart disease. It displays two different patient scenarios.

Left side (Male, Age 60): The model predicts a high risk of coronary artery disease (CAD) based on symptoms and inputs.

Right side (Female, Age 62): The model finds no major signs of heart disease based on the current inputs.

Both predictions were made using the Random Forest Classifier, as indicated at the bottom of each panel. The app uses key health indicators like blood pressure, cholesterol, heart rate, and more to analyze and give results. Females generally have a lower risk of heart disease before menopause due to the protective effects of estrogen, which helps maintain healthy blood vessels and cholesterol levels. This natural hormone offers some cardiovascular protection compared to males of the same age.

IV. CONCLUSION

In summary, this project introduced a heart disease prediction method utilizing Logistic Regression, Random Forest, SVM, and K-Nearest Neighbors (KNN). Additionally, Gradient Boster was applied to predict heart disease and perform a comparative analysis to determine the most effective algorithm. The models were evaluated using the Heart Disease Prediction dataset from the Cleveland UCI repository, sourced from Kaggle. In comparison with traditional machine learning techniques for heart disease prediction, Random Forest & Decision Tree achieved the highest accuracy at 99%, followed by Gradient boosting at 93%, KNN at 73%, and SVM at 68%. it outperformed all other algorithms, achieving the highest accuracy of 99%. Therefore, while the proposed approaches can assist cardiologists in diagnosing heart disease, the Random Forest model stands out as the most effective for this task.

REFERENCE

[1] Ahmad, B., Chen, J., & Chen, H., *"Feature selection strategies for optimized heart disease diagnosis using ML and DL models"*, arXiv preprint arXiv:2503.16577, https://arxiv.org/abs/2503.16577

[2] Yi, J., Yu, P., Huang, T., & Xu, Z., *"Optimization of Transformer heart disease prediction model based on particle swarm optimization algorithm"*, arXiv preprint arXiv:2412.02801, 2024. [https://arxiv.org/abs/2412.02801

[3] Banday, M., Zafar, S., Agarwal, P., Alam, M. A., & Abubeker, K. M., (2024). *Early Detection of Coronary Heart Disease Using Hybrid Quantum Machine Learning Approach*, arXiv:2409.10932. https://arxiv.org/abs/2409.10932

[4] Dorraki, M., Liao, Z., Abbott, D., et al., *"Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data: A Prospective UK Biobank Study"*, JACC: Advances, vol. 3, no. 9_Part_2, 101180, 2024. [Online]. Available: https://www.jacc.org/doi/10.1016/j.jacadv.2024.101180

[5] Hu, Y., Chen, J., Hu, L., et al., *"Personalized Heart Disease Detection via ECG Digital Twin Generation"*, arXiv preprint arXiv:2404.11171, 2024. https://arxiv.org/abs/2404.11171

[6] Mulani, A. O., Sayyad Liyakat, K. K., Warade, N. S., et al., *"ML-powered Internet of Medical Things Structure for Heart Disease Prediction"*, Journal of Pharmacology and Pharmacotherapeutics, 2025. https://journals.sagepub.com/doi/10.1177/0976500X241306184

[7] Alshraideh, M., et al., *"Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital"*, Applied Computational Intelligence and Soft Computing, 2024, Article ID 5080332. : https://onlinelibrary.wiley.com/doi/full/10.1155/2024/5080332

[8] NHS England, *"NHS in England to trial AI tool to predict risk of fatal heart disease"*, The Guardian, 2024. https://www.theguardian.com/society/2024/oct/23/nhs-england-trial-ai-tool-aire-heart-disease

[9] The Guardian, *"Algorithm could help prevent thousands of strokes in UK each year"*, The Guardian, 2024. : https://www.theguardian.com/society/2024/dec/28/algorithm-could-help-prevent-thousands-of-strokes-in-uk-each-year

[10] Islam, M. S., et al., *"Proposed CardioTabNet, a hybrid transformer model for heart disease*

prediction using tabular data", arXiv preprint, 2025.

[11] Yi, J., et al., *"Optimization of Transformer Model using Particle Swarm Optimization for improved heart disease prediction"*, arXiv preprint, 2024.