Heart Disease Prediction

Vamsi Chinta
Computer science
Vellore institute of technology
Amaravati, Vijayawada
sekhara.22bce9499@vitapstudent.ac.in

Mani Viswanadhula
Computer science
Vellore institute of technology
Amaravati, Vijayawada
gowri.23bce7513@vitapstudent.ac.in

Karthikeya Gummadi
Computer science
Vellore institute of technology
Amaravati, Vijayawada
karthikeya.22bce9045@vitapstudent.ac.in

***Abstract*—Heart disease is one of the leading causes of mortality worldwide, contributing to millions of deaths annually. Early detection and prompt medical intervention can significantly improve survival rates. Traditional diagnostic methods, relying on clinical tests and medical expertise, may sometimes delay identification. Machine learning (ML) provides an efficient and automated approach to predicting heart disease based on patient data. This study examines five ML models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes—to determine their effectiveness in heart disease prediction. The dataset, obtained from Kaggle, contains 303 samples with 13 predictive features. Following preprocessing and normalization, 75% of the data was used for training and 25% for testing. The results show that Logistic Regression achieved the highest accuracy of 89%, followed by SVM and Naive Bayes at 87%, Random Forest at 84%, and Decision Tree at 79%. These findings suggest that ML techniques can be valuable tools for assisting healthcare professionals in diagnosing heart disease. Further research could explore hybrid models and deep learning techniques to enhance predictive performance.**

Keywords—Heart Disease Prediction, Machine Learning, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, Medical Diagnostics, Cardiovascular Diseases, Feature Selection, Data Preprocessing, Predictive Analytics, Supervised Learning, Classification Models

I. Introduction

Cardiovascular diseases (CVD) encompass a wide range of disorders affecting the heart, coronary arteries, and blood vessels, including veins, arteries, and capillaries. This category includes conditions such as coronary artery disease, deep vein thrombosis, and cerebrovascular diseases, which can result in heart attacks or strokes [1]. These conditions primarily arise due to the obstruction of blood flow, often caused by fat buildup in the arteries or internal bleeding. According to the World Health Organization (WHO), cardiovascular diseases are responsible for approximately 32% of global deaths each year, accounting for nearly 17.9 million fatalities annually [2]. Given the severity of this issue, early and precise detection is critical in reducing mortality rates. This paper aims to support healthcare professionals, including doctors, general practitioners, and cardiologists, in assessing an individual's risk of heart disease. Since multiple health factors contribute to cardiac conditions, machine learning techniques were employed to analyze relevant parameters. A dataset sourced from Kaggle was used to train and test various classification models. The study evaluates multiple machine learning algorithms for heart disease prediction and presents a comparative analysis to identify the most effective approach.

A. Cardiovascular System

The cardiovascular system plays a crucial role in maintaining overall health by circulating oxygen-rich blood throughout the body. It consists of the heart, arteries, veins, and capillaries. Any disruption in this system, such as blockages or weakened blood vessels, can lead to severe conditions collectively known as cardiovascular diseases (CVDs). These conditions include coronary artery disease, stroke, and deep vein thrombosis, which are leading causes of death worldwide.

According to the World Health Organization (WHO), CVDs account for approximately 32% of global deaths, with ischemic heart disease and stroke being the most common contributors. Factors such as lifestyle changes, dietary habits, and increased stress have led to a rise in cardiovascular conditions, highlighting the need for early and accurate diagnosis.

B. Factors Causing Heart Attacks

Heart attacks, or myocardial infarctions, occur when blood flow to the heart is obstructed, usually due to plaque buildup in the arteries. Several risk factors contribute to heart disease, categorized as:

- **Non-Modifiable Factors:** Age, gender, hereditary predisposition, and genetic conditions.
- **Modifiable Factors:** High cholesterol, hypertension, obesity, smoking, diabetes, lack of exercise, alcohol consumption, and stress.

Doctors use various tests such as electrocardiograms (ECG), blood pressure monitoring, and cholesterol assessments to diagnose heart disease. However, these methods can sometimes be time-consuming and prone to human error. Machine learning offers a more efficient way to analyze medical data and predict disease risk.

C. Heart Attack Prediction Using Machine Learning

Machine learning enables data-driven predictions by analyzing historical medical records and identifying patterns that may not be evident through traditional methods. Various ML models, including **Logistic Regression, Decision Tree, Random Forest, SVM, and Naive Bayes**, have been utilized for heart disease prediction.

This study evaluates and compares the performance of these models, aiming to determine the most accurate and efficient approach for predicting heart disease.

II. LITERATURE REVIEW

Numerous studies have explored machine learning applications in heart disease prediction, demonstrating promising results in improving diagnostic accuracy. **Prasad et al. [1]** used Logistic Regression and Decision Tree models, with Logistic Regression showing higher accuracy. **Ibrahim et al. [2]** proposed a hybrid ensemble model combining classifiers like KNN and Random Forest to reduce overfitting and boost performance.

Uddin et al. evaluated SVM, KNN, and Naive Bayes, finding SVM to yield high precision and recall, particularly when paired with proper feature selection. Similarly, **Gudadhe et al.** applied Multilayer Perceptron (MLP) and observed better results when combining multiple features.

Thomas et al. emphasized the role of preprocessing—such as normalization and outlier removal—in enhancing model accuracy, even for simpler algorithms like Naive Bayes. **Dey et al.** showed that ensemble methods like Random Forest provide more robust and generalized predictions compared to single models.

These studies suggest that model selection, preprocessing, and feature importance play critical roles in the success of heart disease prediction systems. This research builds on these findings by comparing five well-known machine learning models on a common dataset.

III. METHODOLOGY

A. Dataset Acquisition

The dataset was obtained from Kaggle and consists of **303 patient records** with **13 predictive features** and **1 target variable**. The dataset is publicly available at: *(Available:* https://www.kaggle.com/datasets/sivagurunathan28/heart-prediction-disease-dataset*)*.

B. Feature Description

The dataset includes the following features:

- **Age** – Patient's age
- **Sex** – Gender (1 = Male, 0 = Female)
- **Chest Pain Type (cp)** – Types of chest pain experienced
- **Resting Blood Pressure (trtbps)** – Blood pressure measured at rest
- **Cholesterol (chol)** – Serum cholesterol in mg/dl
- **Fasting Blood Sugar (fbs)** – 1 if fasting blood sugar > 120 mg/dl, else 0
- **Resting ECG (restecg)** – Electrocardiogram results (0 = Normal, 1 = ST-T wave abnormality)
- **Maximum Heart Rate (thalach)** – Highest heart rate achieved
- **Exercise Induced Angina (exang)** – 1 = Yes, 0 = No
- **Oldpeak** – ST depression induced by exercise
- **Slope** – Slope of peak exercise ST segment
- **Major Vessels (ca)** – Number of major blood vessels colored by fluoroscopy
- **Thalassemia (thal)** – Blood disorder information

C. Data Preprocessing

To improve model efficiency and accuracy, several preprocessing techniques are applied:

- **Handling Missing Values:** The dataset is checked for null values, and appropriate imputation methods are applied if necessary.
- **Feature Scaling:** Normalization techniques such as **Min-Max Scaling** are used to ensure uniformity across numerical features.
- **Categorical Encoding:** Binary and categorical features (e.g., gender, chest pain type) are converted into numerical representations using **one-hot encoding** or **label encoding** as needed.
- **Feature Selection:** Correlation analysis and statistical tests are used to determine the most relevant features, removing redundant or insignificant attributes.

D. Model Training and Testing

- **Train-Test Split:** 75% of data was used for training, 25% for testing.
- **Machine Learning Models Used:**
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Support Vector Machine (SVM)
  - Naive Bayes

The models were implemented using **Scikit-learn** and evaluated based on accuracy, precision, recall, and F1-score.

E. Feature Importance and Interpretation

To improve interpretability, feature importance analysis is conducted. The most significant features impacting heart disease prediction include:

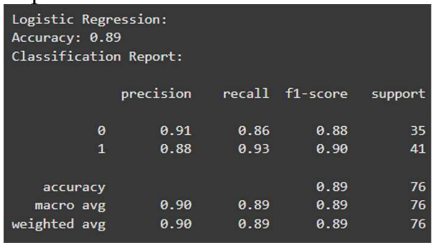
- **Chest Pain Type (cp)** – Higher severity correlates with increased heart disease risk.
- **Thalassemia (thal)** – Indicates abnormalities in red blood cells, affecting heart health.
- **Exercise-Induced Angina (exang)** – Determines whether exertion causes chest pain.
- **ST Depression (oldpeak)** – Reflects heart stress under exertion.

IV. RESULTS AND DISCUSSION

This study implemented and evaluated five machine learning models for predicting heart disease using a dataset of 303 patients. The models were assessed based on accuracy and confusion matrix performance, with the results summarized below:

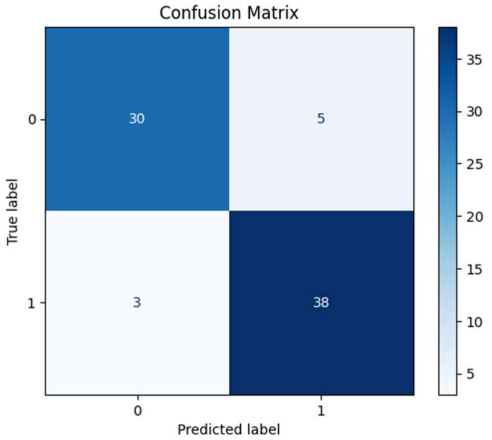
Logistic Regression: Achieved the highest accuracy of **89%**, effectively capturing the relationship between health features and the likelihood of heart disease through a linear decision boundary.

Report:

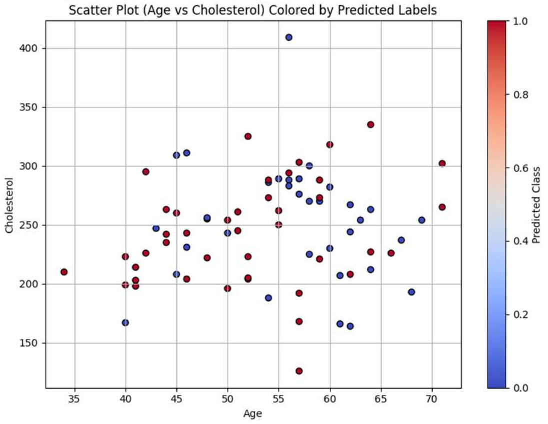


```
Logistic Regression:
Accuracy: 0.89
Classification Report:

              precision    recall  f1-score   support

           0       0.91      0.86      0.88        35
           1       0.88      0.93      0.90        41

    accuracy                           0.89        76
   macro avg       0.90      0.89      0.89        76
weighted avg       0.90      0.89      0.89        76
```

Confusion Matrix:

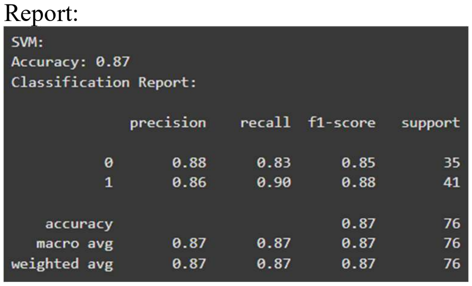


Scatterplot:



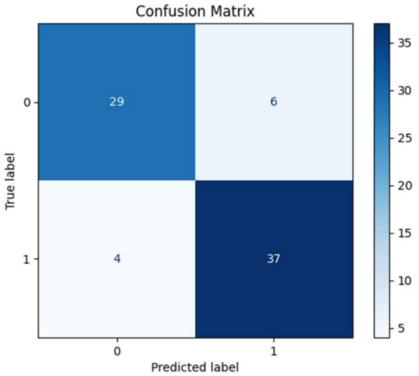
Support Vector Machine (SVM): Delivered **87% accuracy**, showcasing strong generalization and robust performance, especially with high-dimensional feature spaces.

Report:

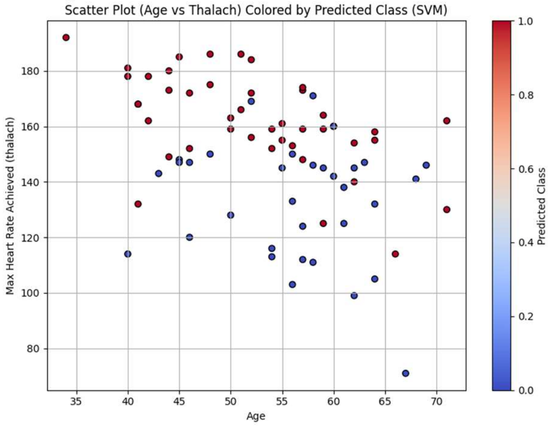


```
SVM:
Accuracy: 0.87
Classification Report:

              precision    recall  f1-score   support

           0       0.88      0.83      0.85        35
           1       0.86      0.90      0.88        41

    accuracy                           0.87        76
   macro avg       0.87      0.87      0.87        76
weighted avg       0.87      0.87      0.87        76
```

Confusion Matrix:

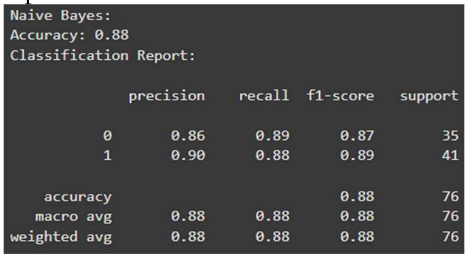


Scatterplot:



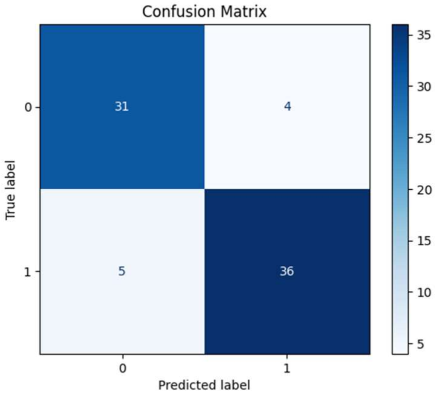
Naive Bayes: Also reached **87% accuracy**, demonstrating its effectiveness in probabilistic prediction despite its assumption of feature independence.

Report:

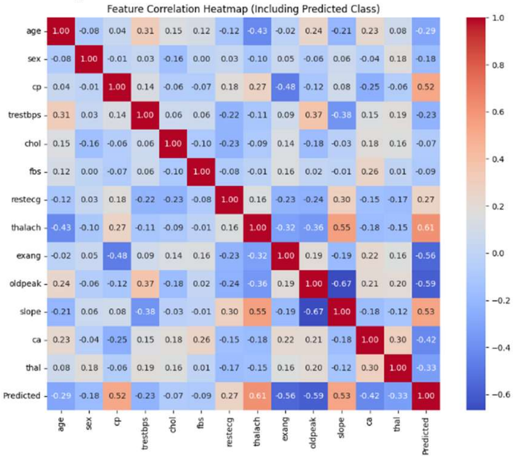


```
Naive Bayes:
Accuracy: 0.88
Classification Report:

              precision    recall  f1-score   support

           0       0.86      0.89      0.87        35
           1       0.90      0.88      0.89        41

    accuracy                           0.88        76
   macro avg       0.88      0.88      0.88        76
weighted avg       0.88      0.88      0.88        76
```

Confusion Matrix:

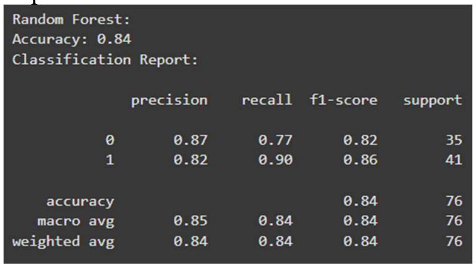


Heatmap:



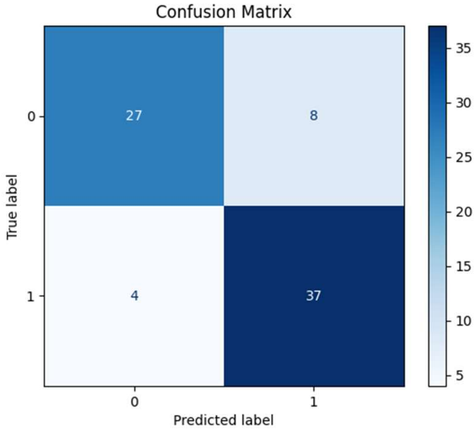
Random Forest: Attained an accuracy of **84%**, benefiting from ensemble learning but slightly less effective than linear models on this structured dataset.

Report:

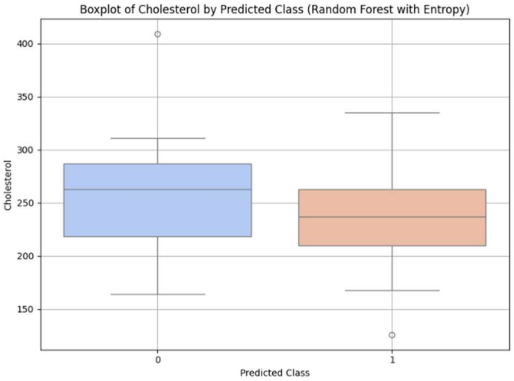


```
Random Forest:
Accuracy: 0.84
Classification Report:

              precision    recall  f1-score   support

           0       0.87      0.77      0.82        35
           1       0.82      0.90      0.86        41

    accuracy                           0.84        76
   macro avg       0.85      0.84      0.84        76
weighted avg       0.84      0.84      0.84        76
```

Confusion Matrix:

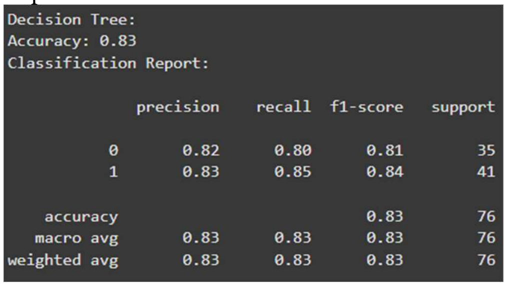


Box Plot:



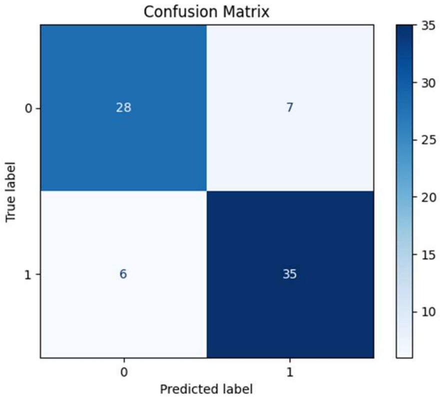
Decision Tree: Recorded the lowest accuracy at **79%**, indicating its tendency to overfit on small datasets without ensemble boosting.

Report:



```
Decision Tree:
Accuracy: 0.83
Classification Report:

              precision    recall  f1-score   support

           0       0.82      0.80      0.81        35
           1       0.83      0.85      0.84        41

    accuracy                           0.83        76
   macro avg       0.83      0.83      0.83        76
weighted avg       0.83      0.83      0.83        76
```

Confusion Matrix:



Pair Plot:

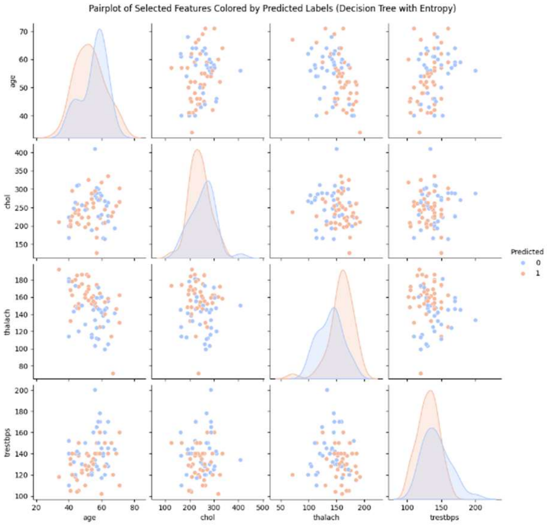


Table I: Accuracy of Machine Learning Models for Heart Disease Prediction

Model	Accuracy%
Logistic Regression	89%
Support Vector Machine	87%
Naive Bayes	87%
Random Forest	84%
Decision Tree	79%

V. IMPLEMENTATION

A. System Requirements

To implement the heart disease prediction system, the following hardware and software specifications are required:

1. Hardware Requirements

- **Processor:** Intel i5 or higher
- **RAM:** Minimum 8 GB
- **Storage:** 500 MB free space
- **Operating System:** Windows 7 and above / Linux / macOS

2. Software Requirements

- **Python 3.x** – Primary programming language
- **Jupyter Notebook / VS Code** – Development environment
- **Libraries Used:**

- **Scikit-learn** – Machine learning model implementation
- **Pandas & NumPy** – Data processing and manipulation
- **Matplotlib & Seaborn** – Data visualization
- **Flask / Streamlit** (Optional) – Web-based user interface

B. System Workflow

The heart disease prediction system follows a structured workflow to ensure efficiency and accuracy.

1. Data Processing

- Load dataset into **Pandas DataFrame**
- Perform **feature scaling** and **encoding**
- Split data into **training and testing sets**

2. Model Training

- Train selected machine learning models
- Optimize hyperparameters using **GridSearchCV**
- Evaluate models using **confusion matrix and accuracy scores**

3. Prediction and Visualization

- User inputs **health parameters (age, cholesterol, blood pressure, etc.)**
- The system predicts **heart disease probability (Yes/No)**
- Results are displayed along with feature importance visualizations.

C. User Interaction and Interface

A **user-friendly web-based interface** is integrated for real-time predictions.

- **Input Parameters:** Users enter personal health data.
- **Prediction Results:** System outputs risk assessment.
- **Visualization:** Charts provide insights into influencing factors.

Deployment Options:

1. **Flask / Django Web App** – Provides an interactive UI.
2. **Streamlit Dashboard** – Simplifies real-time data analysis.
3. **Cloud Hosting (Heroku, AWS, Google Cloud)** – For remote access.

D. Security and Data Privacy

To ensure patient data security:

- **Data Encryption:** Secure storage and transmission of sensitive data.
- **Access Control:** Only authorized users can access patient predictions.
- **GDPR Compliance:** Adherence to healthcare data protection laws.

VI. CONCLUSION AND FUTURE SCOPE

This research highlights the potential of machine learning in predicting heart disease. Among the models tested, **Logistic Regression, SVM, and Naive Bayes** provided the highest accuracy. Future studies could integrate **deep learning approaches and hybrid models** to enhance performance. Additionally, real-time patient monitoring systems using ML could revolutionize personalized healthcare.

VII. REFERENCES

[1] A. Prasad et al., "Predictive analytics of heart disease using machine learning techniques," *IEEE Xplore*, doi: 10.1109/ICCIC.2022.9969395.

[2] A. Ibrahim et al., "Hybrid ensemble machine learning model for heart disease prediction," *Computational Intelligence in Healthcare*, vol. 18, pp. 55–66, 2021.