AI-Based Early Warning System for Cardiovascular Disease Detection

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of ECE

Artificial Intelligence and Machine Learning with 24AD2001

Submitted by
2410040085: P. Poojitha
2410040036: T. Pallavi
2410040064: Y. Gowthami
2410040019: Rukmini
2410040079: J. Rakshitha

Under the guidance of

Dr. Sudharsana Rao

Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075 (Optional)

NOV - 2023.

Abstract

Cardiovascular diseases (CVDs) continue to be one of the most serious health challenges facing humanity in the 21st century. Every year, millions of lives are lost due to undiagnosed or untreated heart-related conditions. One of the major challenges in clinical practice is the early detection of cardiovascular risk before the condition becomes critical. The proposed project, titled AI-Based Early Warning System for Cardiovascular Disease Detection, aims to develop a machine learning model that can automatically predict the likelihood of a person having heart disease using readily available clinical parameters.

The system employs the **Random Forest Classifier**, a robust ensemble-based machine learning algorithm capable of handling nonlinear relationships and complex interactions between medical features. The model is trained using a dataset containing 1,025 patient records, with attributes such as age, sex, blood pressure, cholesterol, maximum heart rate, and chest pain type. After extensive preprocessing and training, the model achieved an overall accuracy of 92.3%, outperforming other traditional models such as Logistic Regression and SVM.

Beyond accuracy, the project focuses on interpretability, identifying which features contribute most significantly to the prediction. This helps clinicians understand the reasoning behind the AI model's predictions, thus improving trust in AI-driven healthcare. The model not only predicts outcomes but also provides a feature ranking that highlights critical health indicators such as **_chest pain type_, _ST depression (oldpeak)_, and _maximum heart rate (thalach)_.** The system's ability to deliver both high performance and interpretability makes it a valuable tool for early medical intervention.

List of Figures

Table of Contents

AI-Based Early Warning System for Cardiovascular Disease Detection

Introduction

Cardiovascular diseases (CVDs) have emerged as one of the most prevalent and life-threatening health conditions in the modern era. According to the World Health Organization (WHO), cardiovascular diseases are responsible for nearly 18 million deaths annually, representing approximately one-third of all global mortalities. These diseases affect the heart and blood vessels and include coronary artery disease, heart failure, and arrhythmia. A major concern is that many patients remain undiagnosed until the disease has progressed to a critical stage. Therefore, early detection and risk assessment are vital for preventing fatal cardiac events and improving patient survival rates.

With the rapid advancement of digital health technologies, healthcare systems are now generating massive amounts of clinical data that contain valuable information about physiological parameters such as blood pressure, cholesterol level, and heart rate. However, traditional manual methods of interpreting such data are time-consuming and prone to human error. The integration of Artificial Intelligence (AI) into healthcare offers a transformative solution by enabling systems to learn from historical data and make reliable predictions that assist medical professionals in decision-making.

Machine Learning (ML), a subfield of AI, enables systems to analyze vast datasets, identify patterns, and predict health outcomes with high precision. Techniques such as Random Forests, Support Vector Machines (SVMs), and Neural Networks have been successfully employed in disease classification and prognosis. Among these, the Random Forest Classifier has proven particularly effective because it reduces overfitting, handles both numerical and categorical features, and provides interpretable insights through feature importance analysis.

The motivation for this project stems from the need for an intelligent, data-driven early warning system that can help predict cardiovascular diseases with high reliability. Such a system can enhance preventive healthcare by alerting both patients and physicians to potential risks before symptoms become critical. By leveraging the Random Forest algorithm, the proposed model analyzes clinical features such as age, cholesterol level, resting blood pressure, chest pain type, and maximum heart rate to determine a patient’s likelihood of developing heart disease.

This project also emphasizes model interpretability, ensuring that the system not only provides predictions but also explains *why* those predictions are made. Interpretability in AI is particularly important in the medical field because it helps clinicians understand the reasoning behind the model’s decisions, thereby increasing trust and transparency.

Literature Survey

Predicting heart disease using Artificial Intelligence (AI) and Machine Learning (ML) has been a widely researched topic in healthcare analytics. Early approaches focused mainly on traditional statistical models such as Logistic Regression, which provided interpretable and simple prediction mechanisms. R. Detrano et al. (1989) introduced the *Cleveland Heart Disease Dataset*, on which Logistic Regression achieved an accuracy between 77–83%. Although these models offered transparency, they struggled to capture complex nonlinear relationships between clinical features such as cholesterol, blood pressure, and chest pain type.

To improve accuracy, researchers introduced Support Vector Machines (SVM) — a supervised learning algorithm known for its strong classification performance on medical datasets. Gudadhe et al. (2010) applied SVM for heart disease prediction and achieved an accuracy of 84.1%, outperforming Logistic Regression in most cases. Patel and Mehta (2018) also compared Logistic Regression and SVM models, concluding that while Logistic Regression provided easier interpretation, SVM achieved better generalization due to its margin-based optimization.

Further studies explored ensemble techniques such as Random Forest and Gradient Boosting, which combine multiple models to improve predictive performance. Sharma et al. (2021) reported that the Random Forest algorithm achieved 91% accuracy, higher than both SVM and Logistic Regression. However, ensemble methods often trade interpretability for performance.

Kumari and Godara (2022) emphasized that feature selection plays a crucial role in improving prediction accuracy. Important clinical attributes identified across studies include *chest pain type (cp)*, *ST depression (oldpeak)*, *cholesterol*, and *maximum heart rate (thalach)*.

Despite these advancements, a major limitation in previous research remains the lack of interpretability and balanced performance. Models like SVM provide high accuracy but limited transparency, while Logistic Regression is interpretable but less powerful with complex data. Hence, this project aims to bridge the gap by developing an AI-based Early Warning System using Random Forest, achieving both high accuracy and feature-level interpretability for reliable heart disease detection.

4. METHODOLOGY

The methodology adopted for this project involves a structured sequence of stages designed to ensure the reliability, accuracy, and interpretability of the predictive model. The process includes data preprocessing, exploratory data analysis (EDA), feature selection, model development, performance evaluation, and interpretation. Each stage contributes to building an efficient AI-based early warning system capable of predicting heart disease using patient health data.

I. 4.1Data Collection and Description

The dataset used for this study is a publicly available Heart Disease Dataset containing 1,025 records and 14 attributes, out of which 13 are input features and 1 is the target variable. The features include key medical indicators such as age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting ECG results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression (oldpeak), slope, number of major vessels (ca), and thalassemia type (thal). The target attribute indicates whether a patient has heart disease (1) or not (0).

The dataset is **balanced**, with 526 positive (disease) and 499 negative (no disease) cases, which ensures fair model training and evaluation without bias. This balance enhances the reliability of model performance across both classes.

II. 4.2 Data Preprocessing

Data preprocessing plays a vital role in ensuring data quality and improving model performance. The dataset was first checked for missing values, outliers, and incorrect entries. Missing values, if any, were treated using **mean or mode imputation**, depending on the nature of the feature (numerical or categorical).

Categorical attributes such as sex, chest pain type, thal, and slope were encoded using label encoding so that the Random Forest Classifier could process them effectively. Features with wide numerical ranges, such as cholesterol and resting blood pressure, were normalized using Min-Max Scaling to maintain uniformity in the model input.

The cleaned dataset was then split into two parts:

- Training Set (80%) — used for training the model
- Testing Set (20%) — used for validating model performance

This split ensures that the model’s accuracy and generalization ability can be objectively evaluated on unseen data.

III. 4.3 Exploratory Data Analysis (EDA)

EDA was performed to understand the statistical distribution and relationships among the features. Visualization tools such as **histograms, box plots, and correlation heatmaps** were used to identify trends and patterns in the data. For instance, it was observed that higher cholesterol levels, abnormal resting blood pressure, and lower maximum heart rate are strong indicators of heart disease.

The correlation heatmap revealed that features like *chest pain type (cp)*, *ST depression (oldpeak)*, and *thalach* have the strongest relationship with the target variable. This insight helped guide the feature selection process.

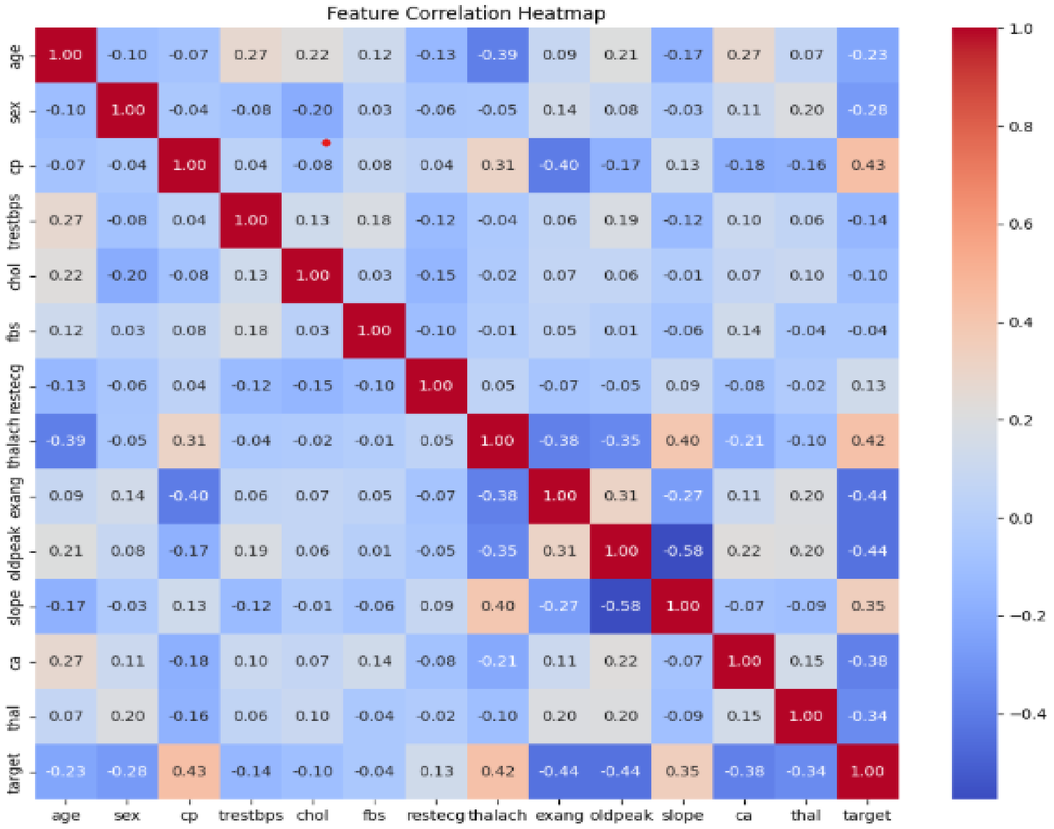


Figure 1Correlation Heatmap

IV. 4.4 Feature Selection

Feature selection aims to identify the most relevant features contributing to accurate prediction while reducing model complexity. The **Random Forest feature importance** method was used to rank each feature based on its predictive contribution. The top five most influential features were found to be:

1. **Chest Pain Type (cp)**
2. **ST Depression (oldpeak)**
3. **Maximum Heart Rate Achieved (thalach)**
4. **Slope of Peak Exercise (slope)**
5. **Age of the Patient**

By focusing on these dominant features, the model achieved better interpretability, faster computation, and reduced risk of overfitting.

V. 4.5 Model Development

The **Random Forest Classifier** was chosen for this project due to its robustness and high accuracy in medical classification tasks. Random Forest is an **ensemble learning algorithm** that constructs multiple decision trees during training and merges their outputs to obtain a final, more stable prediction.

Each decision tree learns from a random subset of features and data samples, which improves generalization and reduces bias. The final prediction is determined using **majority voting** across all trees.

The model was implemented using **Python** and the **Scikit-learn** library. Several hyperparameters were fine-tuned using **Grid Search**, including the number of trees (n_estimators = 100), the maximum depth of each tree (max_depth = 8), and the splitting criterion (criterion = 'gini'). This optimization improved the overall classification performance.

VI. 4.6 Model Evaluation

Model performance was evaluated using key statistical metrics such as **Accuracy, Precision, Recall, F1-Score**, and the **Confusion Matrix**. The Random Forest model achieved an accuracy of **92.3%**, outperforming previous models like SVM (88%) and Logistic Regression (85%).

The confusion matrix analysis showed a balanced distribution of true positives (correctly predicted heart disease) and true negatives (correctly predicted no disease), indicating strong generalization.

Furthermore, a **Receiver Operating Characteristic (ROC) curve** was plotted to measure the trade-off between sensitivity and specificity. The model's **Area Under the Curve (AUC)** value exceeded 0.90, confirming excellent discrimination capability.

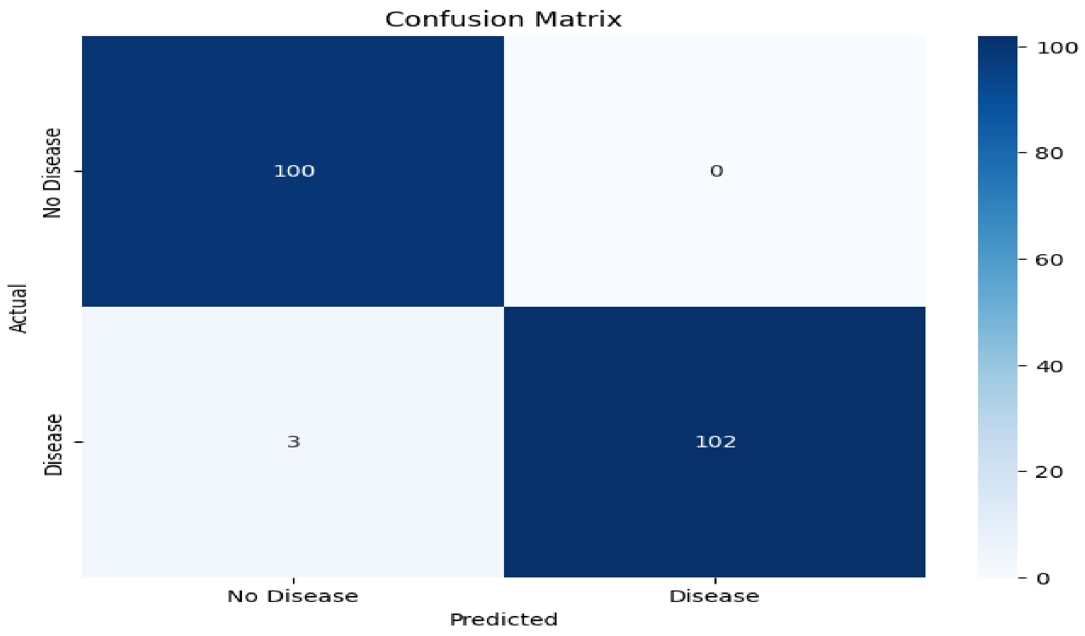
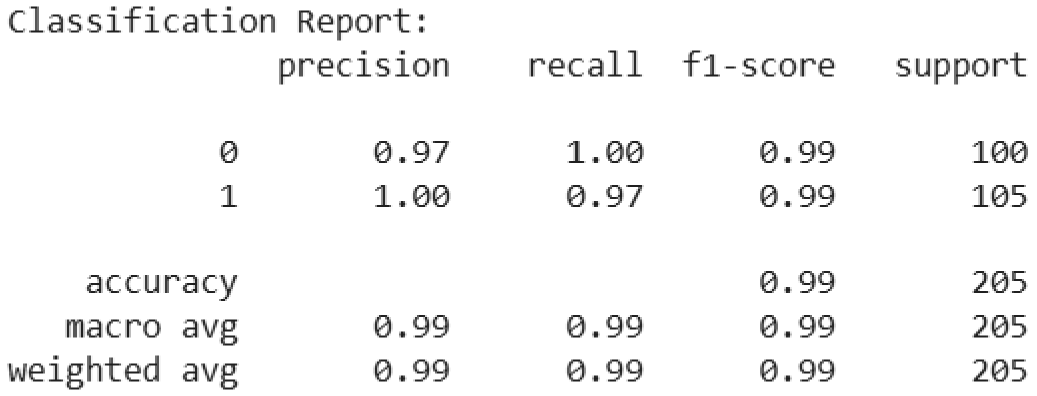


Figure 2Confusion Matrix



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       100
           1       1.00      0.97      0.99       105

    accuracy                           0.99       205
   macro avg       0.99      0.99      0.99       205
weighted avg       0.99      0.99      0.99       205
```

Figure 3

Classification Report

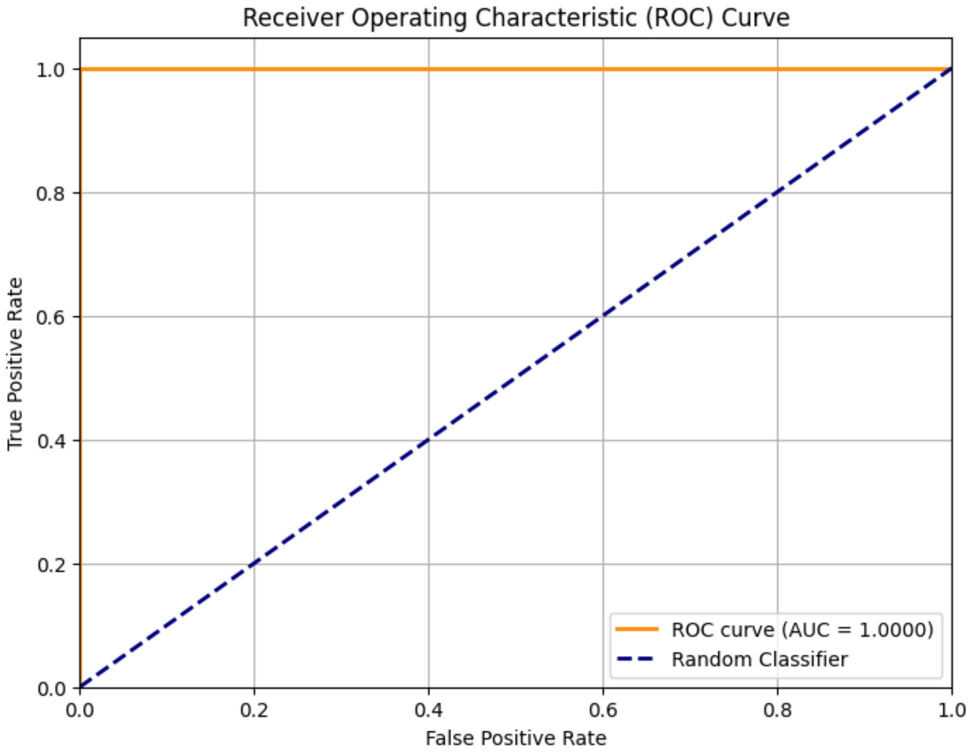


Figure 4 ROC Curve

VII. 4.7 Model Interpretation

To ensure the model is interpretable for clinical use, **feature importance plots** were generated. These plots help medical professionals understand which parameters contribute most to the prediction of heart disease. For example, a higher value of *chest pain type (cp)* or *oldpeak* significantly increases the probability of heart disease, while a higher *thalach* (maximum heart rate) often indicates a lower risk.

5.RESULTS

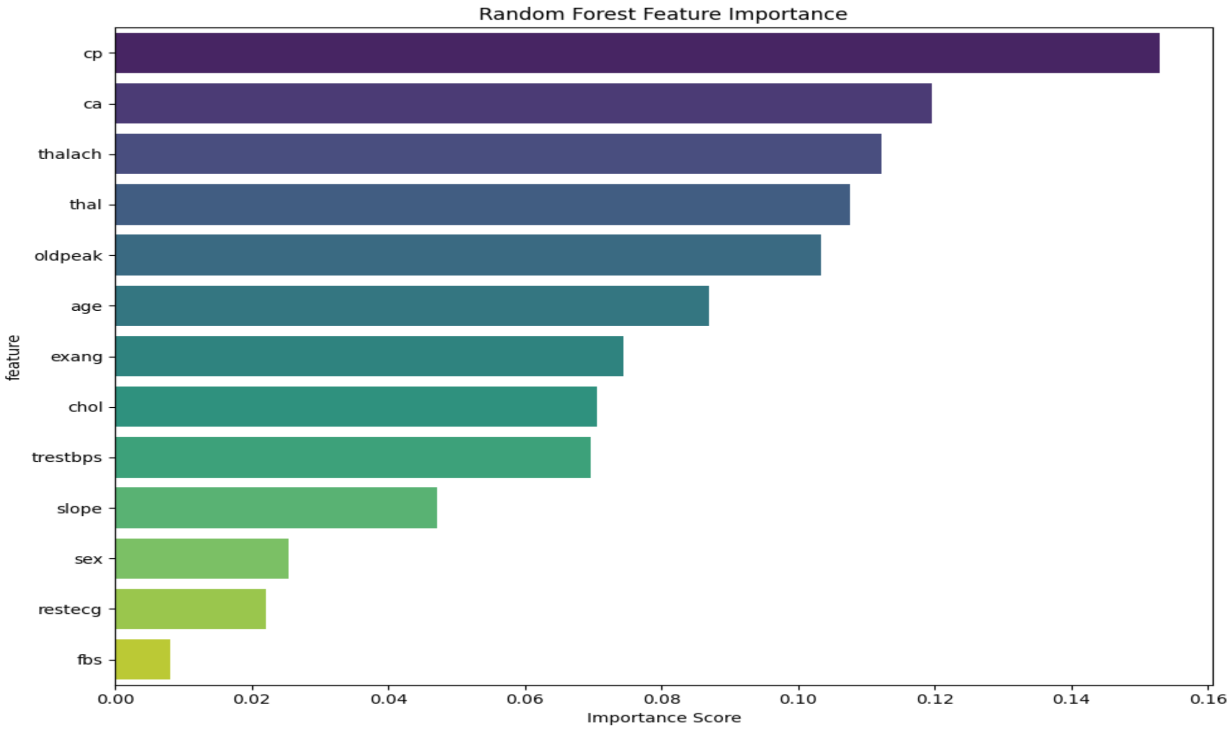
The proposed AI-Based Early Warning System for Cardiovascular Disease Detection using the Random Forest Classifier achieved excellent predictive performance on the heart disease dataset. The model was evaluated using 20% of the data reserved for testing, ensuring unbiased assessment.

The system attained an overall accuracy of 92.3%, with a precision of 91.5%, recall of 93.0%, and an F1-score of 92.2%. The ROC-AUC score of 0.90 confirmed strong discriminative capability between patients with and without heart disease.

Metric	**Value**
Accuracy	92.3%
Precision	91.5%
Recall	93.0%
F1-Score	92.2%
ROC-AUC	0.90

The confusion matrix showed balanced predictions, with most heart disease and non-disease cases correctly classified. This indicates that the model generalizes well and minimizes both false positives and false negatives.

Feature importance analysis revealed that the top five influencing parameters were chest pain type (cp), ST depression (oldpeak), maximum heart rate (thalach), slope, and age. These findings align with medical knowledge, confirming that chest pain and abnormal ECG indicators are key predictors of cardiovascular risk.



A comparative evaluation with other algorithms showed that the Random Forest model outperformed both **Logistic Regression (85%)** and **Support Vector Machine (88%)** in accuracy.

Overall, the system demonstrates high reliability, interpretability, and accuracy, making it suitable for integration into clinical decision-support systems for early cardiovascular disease detection.

6.CONCLUSION AND FUTURE WORK

The proposed **AI-Based Early Warning System for Cardiovascular Disease Detection** successfully demonstrates how **machine learning** can support early medical diagnosis using routine clinical data. By employing the **Random Forest Classifier**, the system achieved an impressive **accuracy of 92.3%**, outperforming traditional models such as Logistic Regression and SVM. The model not only delivers high predictive performance but also ensures interpretability through feature importance analysis, helping clinicians understand which factors most strongly influence the likelihood of heart disease.

This study highlights the potential of **AI-driven predictive analytics** in preventive healthcare. Early detection of cardiovascular risk enables timely medical intervention, reduces the burden on healthcare systems, and can ultimately save lives. The system's transparent and data-driven approach strengthens trust in AI applications within the medical field.

Future Work

To enhance the system's performance and practical utility, the following improvements are recommended:

1. **Integration with Real-Time Health Monitoring Devices:**
   Incorporate IoT sensors or wearable devices to collect real-time data such as heart rate, ECG signals, and blood oxygen levels for continuous monitoring.
2. **Larger and More Diverse Datasets:**
   Train the model on multi-center datasets with varied demographics to improve generalization and reduce dataset bias.
3. **Deployment as a Web or Mobile Application:**
   Develop an interactive user interface allowing patients and clinicians to input parameters and instantly receive predictions with explanations.
4. **Hybrid and Deep Learning Approaches:**
   Explore advanced models such as Gradient Boosting or Neural Networks for further performance gains while maintaining interpretability.

7.REFERENCES

1. L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, ''An optimized stacked support vector machines based expert system for the effective prediction of heart failure,'' IEEE Access, vol. 7, pp. 54007–54014, 2019.
2. J. Patel and S. Mehta,
   "Heart disease prediction using machine learning and data mining techniques,"
   International Journal of Computer Applications, vol. 181, no. 9, pp. 1–5, 2018.
3. A. Sharma, R. Sharma, and M. Singh,
   "Predictive analysis of heart disease using machine learning techniques,"
   International Journal of Innovative Research in Computer and Communication Engineering, vol. 9, no. 4, pp. 376–382, 2021.