Heart Diseases Prediction

[1] Yamini Chouhan, [2] Burra Navya, [3] Amadagani Sandeep, [4] Kavati Ramesh
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
yaminichouhan_cse@siddhartha.co.in, 23tq1a5658@siddhartha.co.in, 23tq1a5608@siddhartha.co.in, 23tq1a5605@siddhartha.co.in,

Abstract

Heart disease is one of the leading causes of mortality worldwide, making early detection and diagnosis crucial for improving patient outcomes and reducing death rates. Traditional diagnostic methods are often expensive, invasive, and time-consuming, which can delay timely treatment. This project presents a machine learning-based approach for predicting heart disease using clinical and demographic patient data.

The dataset used in this study consists of 1,888 patient records with 14 significant medical features, including age, sex, chest pain type, cholesterol level, blood pressure, and heart rate. The methodology involves exploratory data analysis (EDA), data preprocessing, feature scaling, and the implementation of the Random Forest classification algorithm. The dataset is divided into training and testing sets to evaluate the model's performance and accuracy.

The Random Forest model achieved an accuracy of 95.5%, demonstrating strong predictive capability. Feature importance analysis indicates that factors such as chest pain type, number of major vessels, and thalassemia play a significant role in determining heart disease risk.

Overall, the system provides a reliable and efficient decision-support tool that can assist healthcare professionals in the early detection of heart disease, ultimately improving patient care and reducing mortality rates.

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for an estimated 17.9 million deaths each year, as reported by the World Health Organization (WHO). Early diagnosis and timely medical intervention are essential for improving patient survival rates and reducing the overall burden of heart-related illnesses. However, traditional diagnostic techniques such as angiograms, electrocardiograms (ECG), and stress tests are often expensive, invasive, and not easily accessible to all patients, particularly in rural or under-resourced areas.

Moreover, diagnosing heart disease is a complex process due to the involvement of multiple risk factors, including age, cholesterol levels, blood pressure, chest pain type, and lifestyle habits. These factors often exhibit non-linear relationships, making it difficult for clinicians to accurately assess risk using conventional methods alone. As a result, there is a growing need for intelligent, efficient, and non-invasive systems that can support medical professionals in identifying high-risk patients at an early stage.

With advancements in technology, machine learning has emerged as a powerful tool in the healthcare domain. By analyzing historical patient data, machine learning algorithms can identify hidden patterns and correlations that may not be easily detectable through traditional analysis. These models can provide accurate predictions and assist in decision-making, thereby improving diagnostic efficiency and patient care.

II. Literature Survey

Recent research in the application of machine learning for heart disease prediction has demonstrated significant advancements and promising results. Various studies have explored different algorithms and techniques to improve prediction accuracy and support clinical decision-making.

Banerjee, T., and Paçal, İ. (2025) conducted a comprehensive review highlighting the transition from traditional single classifiers to more advanced, optimized, and interpretable hybrid frameworks. Their study emphasizes the importance of model transparency and interpretability for effective integration into clinical practice.

Several research works have shown that ensemble learning methods such as Random Forest and XGBoost achieve high prediction accuracy, often exceeding 90%. These models are particularly effective in handling complex and non-linear relationships within medical data. Additionally, the integration of interpretability techniques such as SHAP (SHapley Additive Explanations) helps in understanding feature contributions, thereby increasing trust among healthcare professionals.

Hybrid deep learning models, including combinations of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have also been widely studied. These models are capable of capturing both spatial and temporal patterns in data, leading to improved performance compared to traditional machine learning algorithms.

III. System Analysis

System analysis focuses on understanding the requirements and workflow of the heart disease prediction system. The system takes clinical and demographic inputs such as age, sex, chest pain type, cholesterol level, blood pressure, and heart rate. These inputs are processed through data preprocessing steps like handling missing values, normalization, and feature scaling. Exploratory Data Analysis (EDA) is performed to identify patterns and relationships within the dataset. The processed data is then used to train a machine learning model. The system evaluates different algorithms and selects the most accurate one, such as Random Forest. The trained model predicts the likelihood of heart disease based on user input. The system ensures high accuracy and reliability in predictions. It is designed to provide quick and non-invasive diagnosis support.

Existing System

The existing system for heart disease diagnosis mainly relies on traditional medical methods and clinical expertise. Doctors use diagnostic tests such as ECG, stress tests,

and angiography to detect heart conditions. These methods are often time-consuming and may require expensive equipment. In many cases, diagnosis depends heavily on the experience and judgment of healthcare professionals. There is limited use of automated systems for early prediction. Traditional systems do not always consider multiple factors simultaneously in a data-driven manner. Additionally, access to advanced diagnostic tools may not be available in rural or remote areas. Some computerized systems exist but lack accuracy or real-time analysis capabilities. These systems are not always user-friendly or easily accessible. As a result, early detection of heart disease can be challenging.

Disadvantages of Existing System

- Expensive and invasive diagnostic procedures
- Time-consuming diagnosis process
- Dependence on doctor's experience and subjective judgment
- Limited accessibility in rural and remote areas
- Lack of early prediction systems
- Inability to efficiently analyze large amounts of data

Proposed System

The proposed system is a machine learning-based heart disease prediction system designed to provide early and accurate diagnosis. It uses clinical and demographic data such as age, sex, chest pain type, cholesterol level, blood pressure, and heart rate. The system preprocesses the data using techniques like normalization and feature scaling. A Random Forest algorithm is used to train the model due to its high accuracy and ability to handle complex datasets. The dataset is divided into training and testing sets to evaluate performance. The trained model predicts the likelihood of heart disease based on user input. The system is integrated into a user-friendly interface for easy interaction. It provides quick and non-invasive predictions. The model also identifies important features influencing the prediction.

Advantages of Proposed System

- Provides early detection of heart disease
- Non-invasive and cost-effective diagnosis
- High accuracy using machine learning algorithms
- Reduces dependency on manual diagnosis
- Quick and real-time predictions
- Can analyze multiple risk factors simultaneously

IV. Methodology

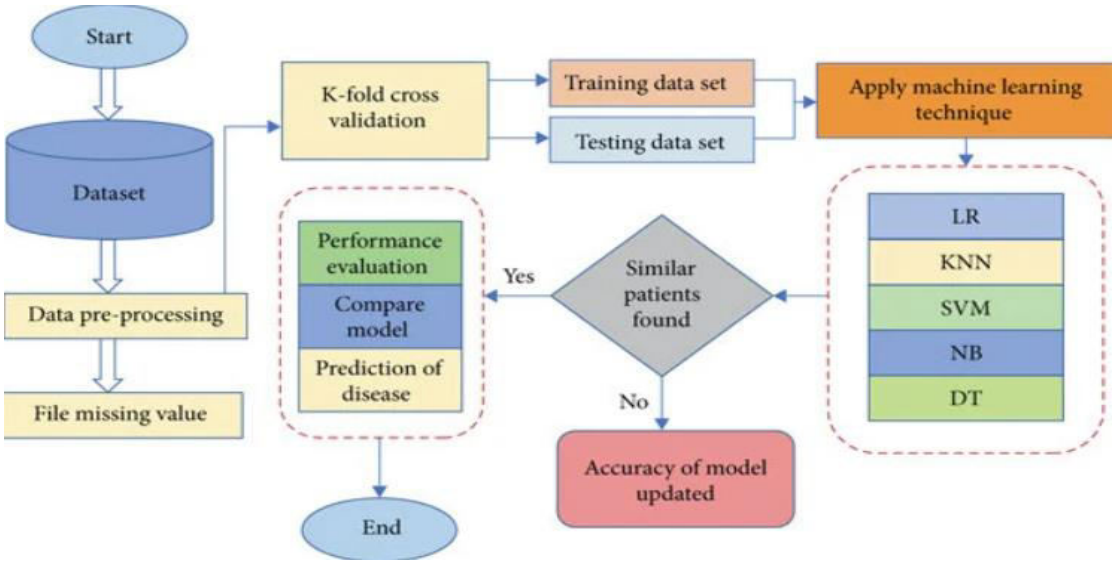
The methodology of the Heart Disease Prediction System involves several systematic steps to build an accurate and reliable model. Initially, the dataset containing patient medical records with features such as age, sex, chest pain type, cholesterol level, blood pressure, heart rate, and other clinical parameters is collected from a reliable source.

The collected data undergoes preprocessing, which includes handling missing values, removing duplicates, and encoding categorical variables. Feature scaling techniques such as StandardScaler are applied to normalize the data and improve model performance. After preprocessing, Exploratory Data Analysis (EDA) is conducted to understand data distribution, correlations, and key patterns.

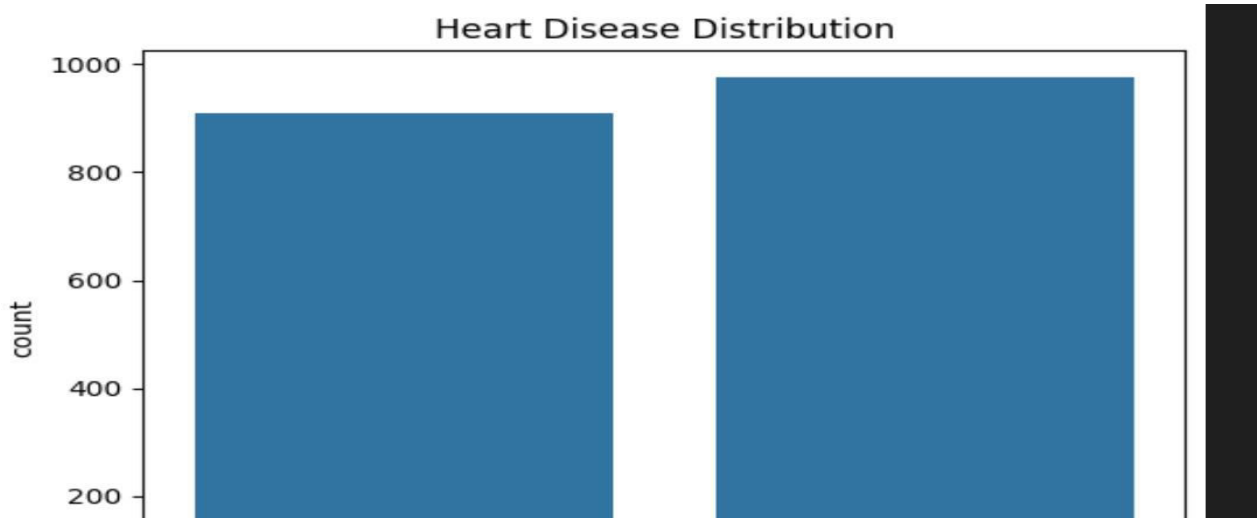
System Analysis

The system architecture defines the structure and workflow of the Heart Disease Prediction System. The process begins with the User Interface, where users (patients or healthcare professionals) enter clinical data such as age, sex, chest pain type, cholesterol, and blood pressure. The input data is then sent to the Data Preprocessing Module, which performs operations like cleaning, encoding, and feature scaling.

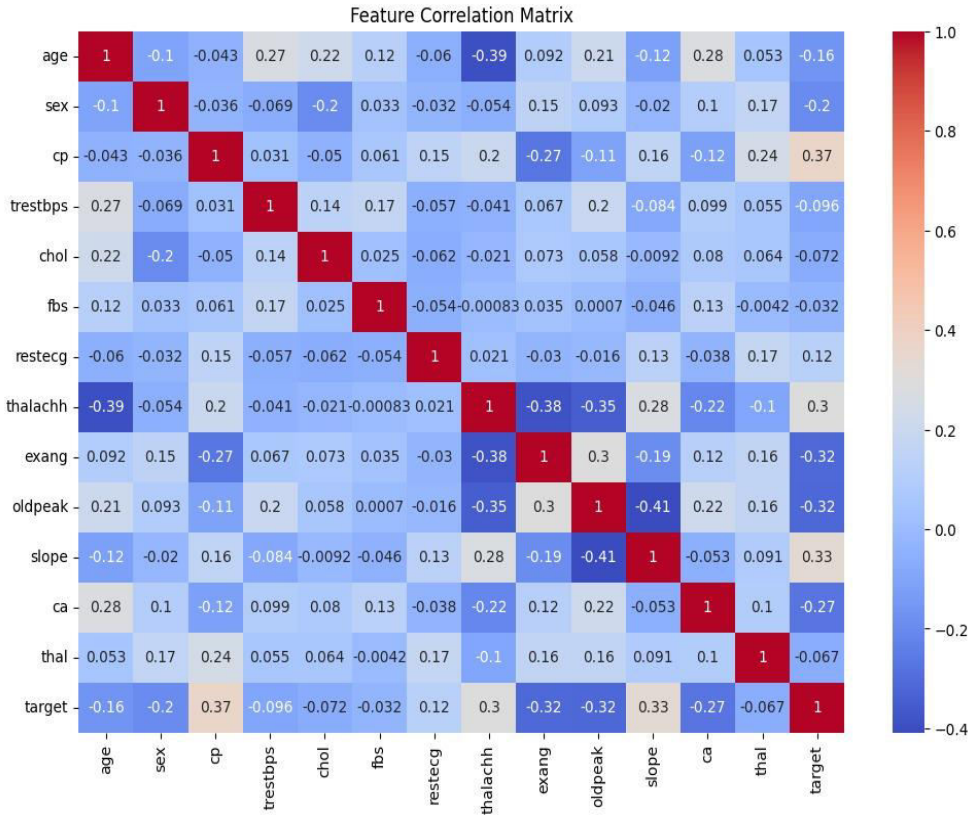
The processed data is forwarded to the Machine Learning Model (Random Forest), which has been trained on historical patient data.



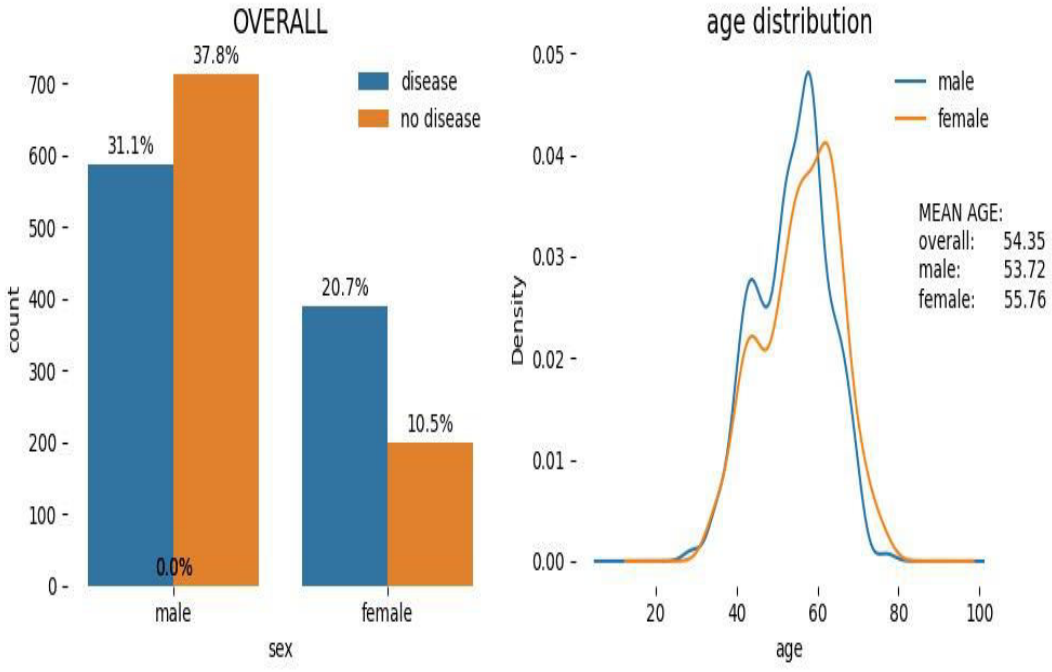
V. Result and Output



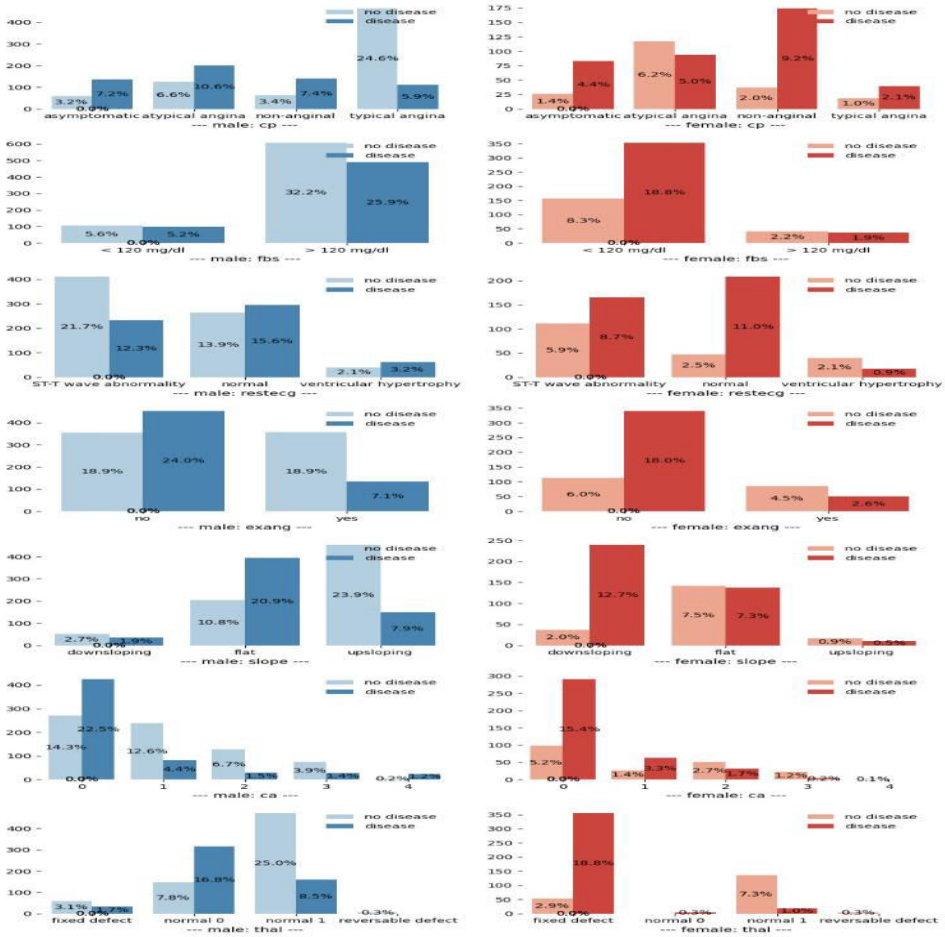
Feature Correlation Matrix(Heatmap)



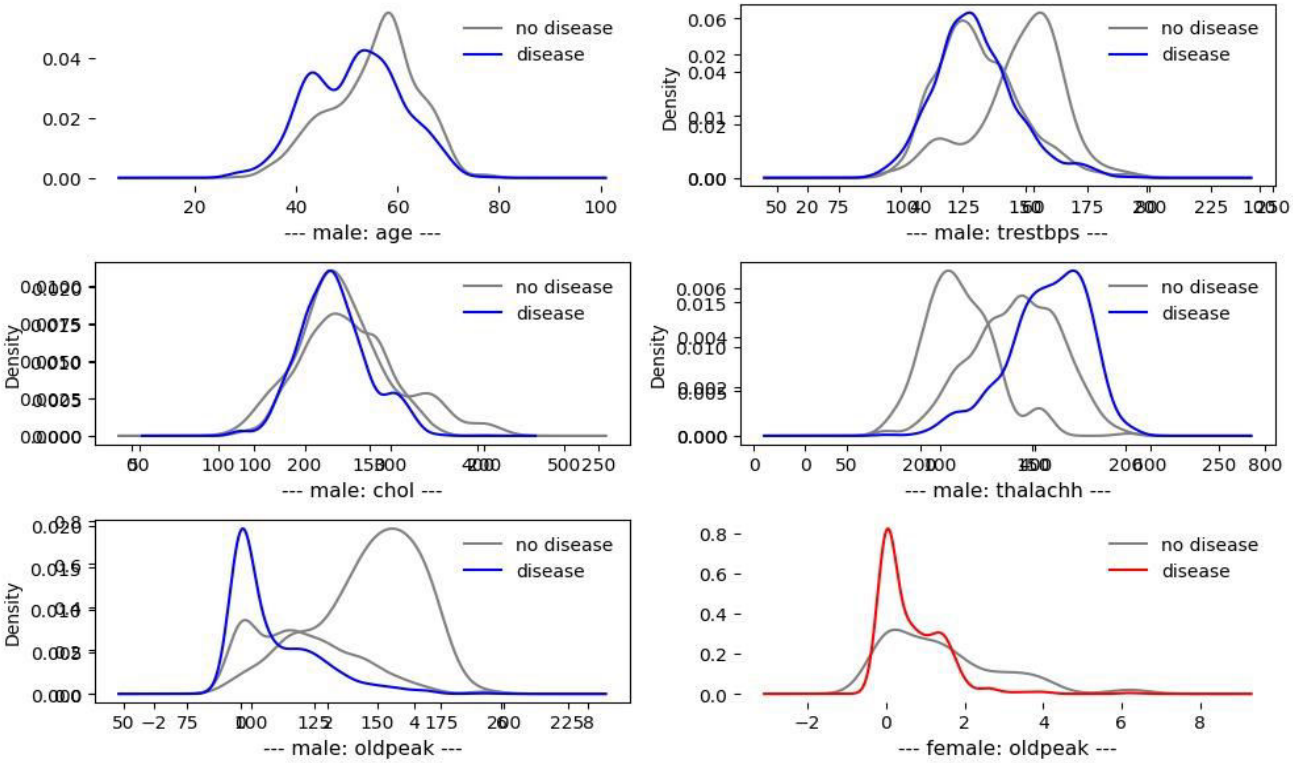
Overall and Gender-wise Age Distribution



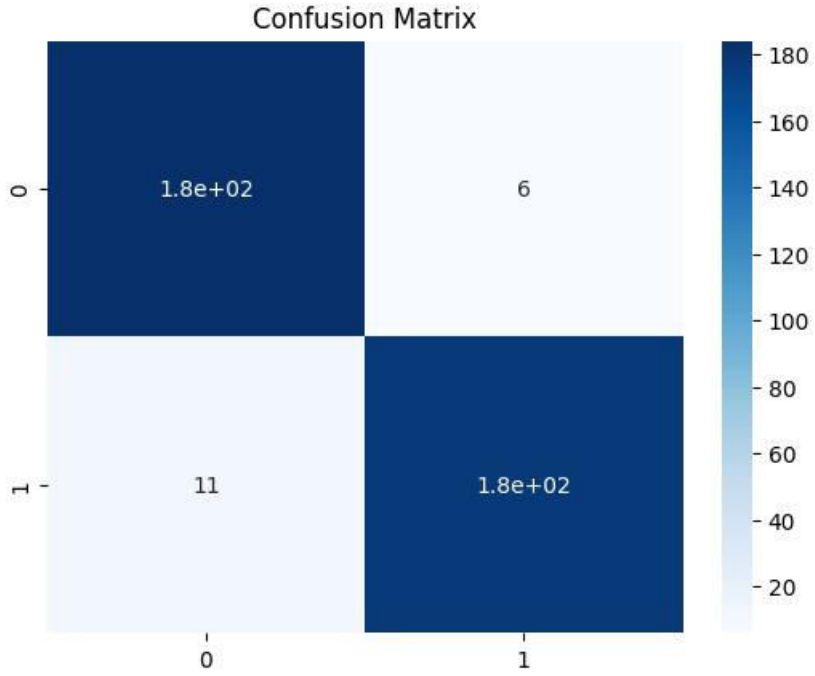
Categorical Feature Analysis (Male vs Female)



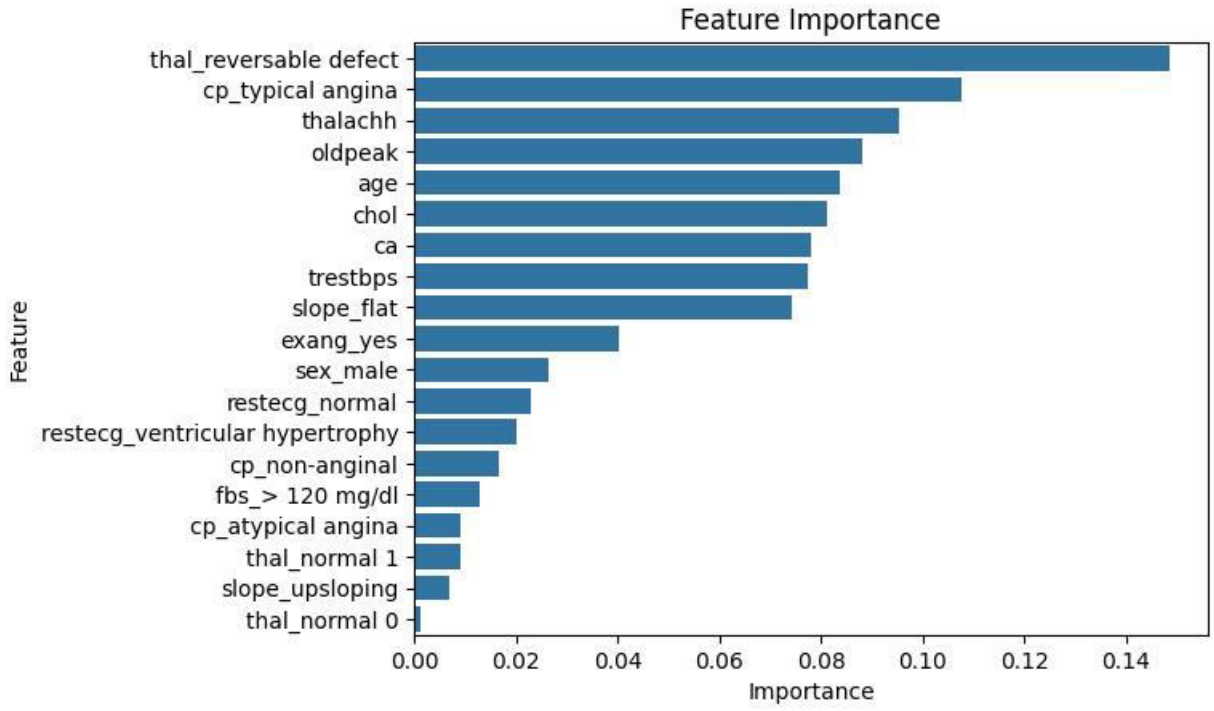
Numerical Feature Distribution by Gender and Target



Random Forest Model Accuracy Output



Feature Importance



VI. Conclusion

This project successfully demonstrated the effective application of machine learning techniques for the prediction of heart disease. Through comprehensive exploratory data analysis and preprocessing, a robust Random Forest classifier was developed using a dataset of 1,888 patient records with relevant clinical features.

The model achieved an impressive accuracy of 95.5%, along with high precision and recall, as validated through performance metrics such as the classification report and confusion matrix. Feature importance analysis further enhanced the interpretability of the model by identifying key predictors such as chest pain type, number of major vessels, and thalassemia.

The results clearly indicate that machine learning, particularly ensemble methods like Random Forest, can provide a reliable, non-invasive, and efficient approach for early detection of heart disease. This system can assist healthcare professionals in making informed decisions and improving patient outcomes.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time
Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis
of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial

Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie
recommender using collaborative filtering and sentiment analysis," *International Research
Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and
Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and
Communication Engineering (ICCSCE)*, Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.