A MINI-PROJECT REPORT

ON

"ENSEMBLE LEARNING FOR HEART DISEASE PREDICTION"

BY

CHATURDHAN CHAUBEY (803)

MAHESH GAIKWAD (809)

VISHAL GAWALI (832)

Under the guidance of

Dr. Nilesh Bhelkar

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

Department of Artificial Intelligence & Data Science

University of Mumbai

March - 2025

CHAPTER 1

Introduction

1.1 Description

Heart-related diseases, also known as Cardiovascular Diseases (CVDs), have emerged as one of the leading causes of death globally, posing a significant threat to public health. In both India and the rest of the world, the growing rate of CVDs has become a serious concern. The heart, being a vital organ that pumps blood and sustains the body's functionality, demands timely diagnosis when any abnormality arises. Predicting heart disease at an early stage can significantly improve treatment effectiveness and patient survival rates.

To improve diagnosis, the healthcare industry is turning to **Machine Learning (ML)** and **Data Analytics**. These technologies help analyze large volumes of patient data to uncover patterns that assist in predicting diseases. Algorithms such as **Random Forest**, **Support Vector Machine (SVM)**, and **Artificial Neural Networks (ANN)** are increasingly used to support medical professionals in detecting heart disease risks.

This project applies an **Ensemble Learning** approach, combining **Logistic Regression**, **Random Forest**, and **SVM** models through a **Voting Classifier**. By leveraging multiple models, the system achieves better accuracy and reliability in heart disease prediction. This AI-based solution aims to support early diagnosis and contribute to smarter, data-driven healthcare systems.

1.2 Objectives

The primary objectives is:

1. To develop a reliable and accurate machine learning-based model that can predict the likelihood of heart disease using patient health parameters.
2. To apply and compare different ML algorithms such as Logistic Regression, Random Forest, and Support Vector Machine to evaluate their individual performance.
3. To implement an ensemble learning approach (Voting Classifier) that combines multiple models for improved prediction accuracy and robustness.
4. To visualize and analyze performance metrics such as confusion matrices and accuracy scores, enabling better understanding of model effectiveness.

CHAPTER 2

Technologies Used

The following technologies and libraries were used for the implementation of this project:

2.1. Python

Python is the core programming language used in this project. It provides extensive support for API integration, image processing, and automation of tasks required for generating AI-powered images. Python's simplicity and vast ecosystem of libraries make it ideal for AI and machine learning projects.

2.2. Jupiter Notebook

Jupiter Notebook serves as the interactive coding environment where users can run the program, enter prompts, and view generated images in real time. It provides a flexible interface for development, debugging, and visualization of results.

2.3. Pandas

Pandas is a powerful Python library used for data manipulation and analysis. It provides easy-to-use data structures like Series and DataFrames. Pandas is widely used for cleaning, transforming, and analyzing structured data.

2.4. NumPy

NumPy (Numerical Python) is the foundational package for numerical computing in Python. It provides high-performance arrays and tools to perform mathematical operations on them. NumPy supports vectorization, broadcasting, and advanced indexing. It's highly optimized for performance and is used as the base for many scientific libraries..

2.5 Scikit-learn

Scikit-learn is a popular Python library for machine learning and data mining. It provides simple and efficient tools for data classification, regression, clustering, and Built on top of NumPy, SciPy, and matplotlib, it integrates well with other scientific libraries. It includes many pre-built models like Logistic Regression, SVM, Random Forest, andmore.

CHAPTER 3

Implementation details

The figures below represent the system architecture for heart disease prediction

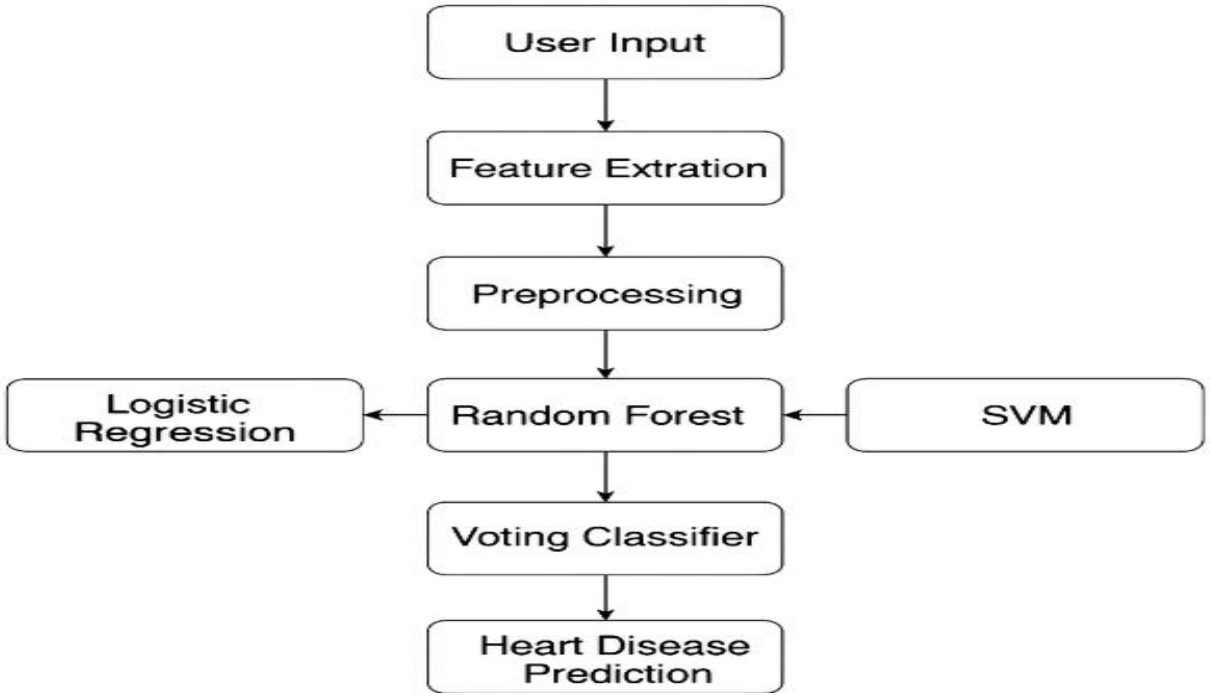

Fig 3.1 System Architecture

A. User Input

The prediction process begins with **collecting user data**, which consists of **various physiological and clinical parameters** relevant to heart health. These may include age, sex, blood pressure, cholesterol levels, ECG results, heart rate, and other significant markers such as Troponin levels. This stage is critical because the **quality and completeness of input data** directly influence the accuracy of the prediction. Input can be collected manually by a clinician or through electronic health records (EHRs).

B. Feature Extraction

Once data is gathered, the next step is to extract meaningful features from it. **Feature extraction** is the process of selecting the most relevant variables from raw input that will help the model make better predictions. This might involve:

1. Selecting attributes like age, blood pressure, etc.

2. Transforming or encoding categorical features (e.g., gender)
3. Creating new features using domain knowledge (e.g., BMI from weight and height)
4. Good feature extraction ensures that the model focuses on the most informative parts of the data and removes noise or irrelevant information.

C. Data Preprocessing

Before feeding the data into machine learning models,

1. Missing values are imputed or removed.
2. Categorical variables are encoded using techniques like one-hot or label encoding.
3. Feature scaling is applied using methods like MinMaxScaler or StandardScaler to normalize the data, especially important for models like SVM or Logistic Regression.
4. The dataset is split into training and testing sets, ensuring that the model is evaluated on unseen data.

D. Model Training

This system utilizes three different machine learning models, each trained separately on the same preprocessed dataset.

1. Logistic Regression:
   A statistical method used for binary classification problems. It models the probability that a patient belongs to a specific class (heart disease or not) based on a linear combination of input features. It is simple, interpretable, and often effective on linearly separable data.
2. Random Forest:
   An ensemble of decision trees that combines the results of multiple trees to improve accuracy and prevent overfitting. Each tree is trained on a random subset of the data and features, and their predictions are aggregated. It handles non-linear data and is robust to noisy or missing data.

3. Support Vector Machine (SVM):

   A powerful classifier that finds the optimal hyperplane to separate the data into two categories. It performs well in high-dimensional spaces and is effective in complex scenarios, especially when classes are not linearly separable.

E. Voting Classifier (Ensemble Learning)

After training individual models, their predictions are combined using an ensemble learning technique called a Voting Classifier:

1. In soft voting, each model predicts probabilities for each class.
2. These probabilities are averaged across all models.
3. The final class is chosen based on the highest average probability.

Ensemble methods often outperform individual models by reducing variance, bias, and improving generalization. This combination allows the system to leverage the strengths of each algorithm while compensating for their individual weaknesses.

F. Heart Disease Prediction (Positive or Negative Class)

1. Positive Class (1) – At Risk of Heart Disease

This means that the model has identified that the patient's health parameters indicate a significant likelihood of having or developing heart disease.

1. It is a positive detection for heart disease risk.
2. A value of 1 is returned by the model.

2. Negative Class (0) – "Not at Risk of Heart Disease"

This means that, based on the given input data, the patient is not currently showing signs of heart disease or is at low risk.

1. It is a negative detection for heart disease risk.
2. A value of 0 is returned by the model.
3. However, this does not guarantee permanent safety — regular checkups and healthy habits are still important.

CHAPTER 4

Results

The Heart Disease Prediction system using ensemble learning combines Logistic Regression, Random Forest, and SVM to improve prediction accuracy. While each model offers unique strengths, the Voting Classifier merges their outputs based on probability scores, resulting in better overall performance. This approach reduces false negatives and enhances reliability, which is vital for early disease detection. Confusion matrix heatmaps show that the ensemble model outperforms individual models in terms of sensitivity and accuracy. Overall, the system serves as a powerful, data-driven tool to support healthcare professionals in predicting heart disease risks effectively.the implementation result are shown in below fig 4.1. and 4.2

Positive:-The Person has Heart Disease

Nagative:-The Person does NOT have Heart Disease

```
]: # Collect user input for prediction
print("Please input the following features:")
input_features = {}

for col in X.columns:
    value = float(input(f"{col}: "))
    input_features[col] = value

input_df = pd.DataFrame([input_features])
input_scaled = scaler.transform(input_df)
prediction = voting_clf.predict(input_scaled)
print("Predicted Class:", le.inverse_transform(prediction)[0])
```

```
Please input the following features:
age: 54
gender: 1
impluse: 58
pressurehight: 117
pressurelow: 68
glucose: 443
kcm: 5.8
troponin: 0.359
Predicted Class: positive
```

Fig: -4.1

Positive:-The Person has Heart Disease

Nagative:-The Person does NOT have Heart Disease

```
# Collect user input for prediction
print("Please input the following features:")
input_features = {}

for col in X.columns:
    value = float(input(f"{col}: "))
    input_features[col] = value

input_df = pd.DataFrame([input_features])
input_scaled = scaler.transform(input_df)
prediction = voting_clf.predict(input_scaled)
print("Predicted Class:", le.inverse_transform(prediction)[0])
```

```
Please input the following features:
age: 64
gender: 1
impluse: 66
pressurehight: 160
pressurelow: 83
glucose: 160
kcm: 1.8
troponin: 0.012
Predicted Class: negative
```

Fig: -4.2

CHAPTER 5

Conclusion & Future Work

5.1 Conclusion

This project highlights the effectiveness of ensemble learning for heart disease prediction using a combination of Logistic Regression, Random Forest, and SVM within a Voting Classifier. The ensemble model demonstrated improved accuracy and reliability compared to individual models, making it a strong candidate for clinical decision support. By leveraging patient health data and advanced machine learning techniques, this system provides a practical and scalable approach for early detection of cardiovascular risk, supporting timely medical intervention and enhancing healthcare outcomes.

5.2 Future Work

It opens up more powerful and insightful directions, the following enhancements can be implemented:

1. **Integration with Real-time Health Monitoring Devices**: The model can be integrated with wearable devices and health monitoring systems to continuously assess heart health and provide instant alerts in case of risk.
2. **Web or Mobile Application Deployment**: A user-friendly interface can be developed for patients or healthcare professionals to input parameters and get instant predictions along with actionable suggestions.
3. **Explainable AI (XAI)**: Future models can incorporate explainability to help users understand why a certain prediction was made, increasing transparency and trust in AI-based diagnostics.
4. **Expansion to Multi-class Diagnosis**: The model can be enhanced to predict not just binary outcomes (Positive/Negative) but different stages or types of cardiovascular diseases for deeper diagnosis.