ECG Signal Analysis for Cardiac Anomaly Detection

A

Project Report

Submitted for Partial Fulfilment of Requirements

for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

(Self Finance (2021-2025))

by

Devansh Kushwaha (2200520109102)
Rajat Kumar (2200520109105)

Under the supervision of

Dr. Aditi Sharma (Assistant Professor, CSE Dept.)

Mrs. Sonam Srivastava (Assistant Professor, CSE Dept.)

Department of Computer Science and Engineering

Institute of Engineering & Technology, Lucknow

DR. APJ ABDUL KALAM TECHNICAL UNIVERSITY, UTTAR PRADESH

June, 2025

Declaration

We hereby declare that this submission of project is our work and that to the best of our knowledge and belief it contains no material previously published or written by another person or material which to a substantial extent has been accepted for award of any other degree of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

We have not submitted this project report to any other institute for the requirements of any other degree.

Name: Devansh Kushwaha

Roll no. 2200520109102

Branch: CSE-SF

Signature:

Name: Rajat Kumar

Roll no. 2200520109105

Branch: CSE-SF

Signature:

Certificate

This is to certify that the project report entitled: **ECG Signal Analysis for Cardiac Anomaly Detection** submitted by **Devansh Kushwaha, Rajat Kumar** in the partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science & Engineering is a record of the bonafide work carried out by them under our supervision and guidance at the Department of Computer Science & Engineering, Institute of Engineering & Technology Lucknow.

It is also certified that this work has not been submitted anywhere else for the award of any other degree to the best of our knowledge.

(Dr. Aditi Sharma)
Assistant Professor, Department of Computer Science and Engineering,
Institute of Engineering & Technology, Lucknow

(Mrs. Sonam Srivastava)
Assistant Professor, Department of Computer Science and
Engineering, Institute of Engineering & Technology, Lucknow

Acknowledgement

We want to express our sincere gratitude to the following individuals and organizations for their invaluable contributions to the completion of this project: First and foremost, we extend heartfelt thanks to our supervisor, **Dr. Aditi Sharma**, for his unwavering support, guidance, and expertise throughout the project. His valuable insights and constructive feedback have greatly shaped the direction and outcomes of this work.

In addition, we are extremely grateful to our co-supervisor, **Mrs. Sonam Srivastava**, for providing her valuable guidance, which has contributed to the successful implementation of our project.

We are grateful to the **Department of Computer Science and Engineering, IET Lucknow,** for providing the necessary resources and facilities to conduct this research. Access to Project Lab, high-speed data connections, and research papers significantly contributed to the project's success.

We would also like to extend our appreciation to our colleagues who provided assistance and encouragement during various stages of this project. Their collaboration and brainstorming sessions helped us overcome challenges and provided valuable insights.

We want to thank the **Project Coordinator, Dr. Promila Bahadur**, who volunteered his time and willingly shared his experiences and perspectives, and without whom this research would not have been possible.

We would like to express our gratitude to the Project Evaluation Committee Members for their continuous support and encouragement throughout this project, their timely feedback and suggestions helped us improve our project quality, and to the authors of the resources and research papers we consulted, whose work provided the foundation for our research.

Lastly, we are deeply thankful to our friends and family for their love and support during this process. They supported us a lot in the completion of this journey.

Devansh Kushwaha

Rajat Kumar

Abstract

According to the World Health Organization, Cardiovascular diseases are a leading cause of mortality globally, often requiring early detection for effective intervention. Traditional ECG analysis for diagnosing cardiac anomalies relies on manual interpretation by specialists, which can be time-consuming and prone to errors in resource-limited settings.

The present work, developed by Devansh Kushwaha and Rajat Kumar, automates this process by analyzing 12-lead ECG images to classify conditions into four categories: Normal, Myocardial Infarction, History of Myocardial Infarction, and Abnormal Heartbeat.

Our proposed system, deployed on Render (https://anahata-ai.onrender.com), achieves 92.5% accuracy using an ensemble of Random Forest, SVM, and Neural Network models, with features extracted after PCA dimensionality reduction.

The use of individual models was also explored, yielding lower accuracy, while our threshold-based classification enhances reliability for early cardiac screening in non-critical applications.

Contents

Declaration

Certificate

Acknowledgements

Abstract

List of Figures

1 Introduction **1**
1.1 Introduction: 1
1.2 Diagnosis of Heart Diseases: 4
1.3 Current work in the field: 5
1.4 Problem Definition 6
1.5 Scope of the Project 6
1.6 Objectives 7
1.7 Relevance and Motivation of Project 7

2 Literature Review **8**
2.1 Related Work 8

3 Methodology **14**
3.1 Models and Algorithms 14
3.2 Detection and Classification: 19
3.3 Proposed System 26
3.4 System Flow 27
3.5 Integrated Data Flow Diagram for ECG Analysis 28

4 Experimental Results **29**
4.1 Experimental Results 29
4.2 Hyperparameter Tuning Results 32

5 Conclusion
5.1 Conclusion 33
5.2 Future Work 34

References **36**

Annexure **38**
A.1 Feature Extraction Code 38
A.2 Model Code, Hyperparameter Tuning & Model Evaluation Plot 40

Plagiarism Report **42**

Letter of Recommendation **43**

List of Figures

1.1 Normal vs Abnormal ECG 2
1.2 Labeled ECG Segments 4
1.3 Standard 13 Lead ECG Sample 5

3.1 ECG Signal Extraction Pipeline 18
3.2 SVM Multi-classifier Hyperplane 19
3.3 SVM Density Function 20
3.4 RBF Kernel Function 20
3.5 Sigmoid Curve for ECG Classification 20
3.6 K-NN classification with k=3 21
3.7 Gradient Boosting Model Structure 22
3.8 Soft Voting with Ensemble 23
3.9 5-Fold Cross Validation 24
3.10 Grid Search CV for Best Parameter 25
3.11 System Flow 27
3.12 Integrated Data Flow Diagram for ECG Analysis & Future Extension 28

4.1 Confusion Matrix 31
4.2 ROC Curve 32

A.1 Plagiarism Report 42

A.2 Letter of Recommendation 43

Chapter 1

Introduction

1.1 Introduction

The prevalence of cardiovascular conditions is growing worldwide, driven by shifts in contemporary lifestyles. This increase is notably evident in both industrialized and emerging nations, with a broadening impact across various age demographics. Younger and older people alike are facing heightened risks of heart-related irregularities.

Cardiovascular diseases, often termed heart diseases, encompass conditions that impair the heart's ability to pump blood effectively, leading to severe health complications [1]. These conditions disrupt blood circulation, which is critical for sustaining life, as the heart fails to deliver oxygen-rich blood to vital organs.

The World Health Organization reports that cardiovascular diseases cause more than 17 million deaths each year, positioning them as the top global killer [2]. A considerable share of these cases, including preventable events like myocardial infarction, could be mitigated with timely detection. Heart diseases are responsible for nearly 31 percent of global deaths.

In the age category of 30 years and older, heart diseases contribute to 45% of sudden cardiac arrests, 60% of acute myocardial infarctions, and 50% of heart failure cases, as per the Global Burden of Disease Study 2019. Even young adults aged 20-30 are increasingly affected, often due to lifestyle factors that exacerbate cardiac risks over time.

These trends highlight the urgent need for early detection to mitigate cardiac risks and prevent severe outcomes like heart attacks or heart failure. Early diagnosis through ECG analysis can enable timely interventions, reducing mortality rates. However, manual ECG interpretation is complex and error-prone, necessitating automated solutions for improved accuracy and accessibility.

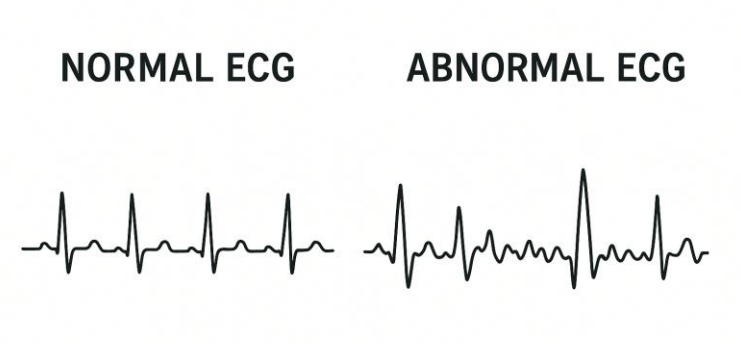

Fig 1.1 Normal ECG vs Abnormal ECG

These trends will, if unaddressed, lead to a significant rise in cardiovascular disease prevalence, increased mortality rates, and a greater need for advanced cardiac interventions. Treatments like bypass surgery or stenting carry risks, particularly for patients with comorbidities such as diabetes or hypertension. In some cases, despite interventions, cardiac function cannot be fully restored due to underlying damage from chronic conditions like heart failure. For a timely resolution of this issue, early detection of cardiac anomalies is essential. Since heart diseases are increasingly common, many are now aware of their most prevalent symptoms. Make it 2 lines more, and make sure below this, the signs and symptoms of failure.

For a timely resolution of this issue, early detection of cardiac anomalies is essential. Since heart diseases are increasingly common, many are now aware of their most prevalent symptoms.

Signs and Symptoms

1. Indications include chest discomfort (angina).
2. Difficulty breathing or shortness of breath.
3. Irregular or rapid heartbeats.
4. Persistent tiredness.
5. Light-headedness or vertigo.
6. Excessive perspiration.
7. Feeling of nausea.

Causes and Risk Factors:

1. High blood pressure.
2. High cholesterol.
3. Smoking.
4. Obesity.
5. Lack of sleep.

1.2 Diagnosis of Heart Diseases:

Diagnosing heart disease is a critical yet complex task in the medical field, often requiring specialized expertise. Traditionally, cardiologists analyze ECG records through manual check-ups at regular intervals, considering factors such as lifestyle, genetic history, and symptoms. Common diagnostic methods include:

1. **ECG Interpretation:** Doctors Manual ECG interpretation by doctors involves analyzing the duration, amplitude, and morphology of P, QRS, and T waves to detect abnormalities, such as ST elevation in myocardial infarction or irregular rhythms in arrhythmias like atrial fibrillation. This process is time-intensive and prone to human error, especially in high-volume clinical settings where rapid diagnosis is critical.

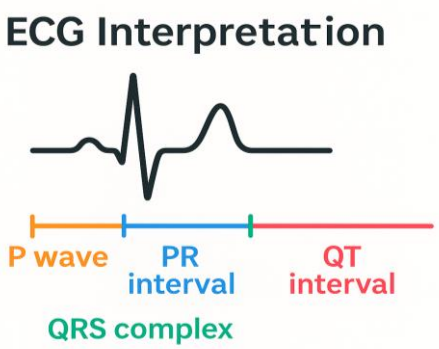

Fig 1.2 Labeled ECG Segments

2. **Echocardiogram:** Echocardiograms use ultrasound to visualize heart structures like chambers and valves, diagnosing conditions such as heart failure or valve regurgitation by evaluating wall motion and ejection fraction. Their effectiveness, however, depends on operator skill and equipment quality, requiring specialized training. In clinical practice, cardiologists and sonographers collaborate to ensure accuracy, especially in complex cases with subtle abnormalities. Advanced technologies like 3D imaging improve precision but highlight the need for continuous professional training.

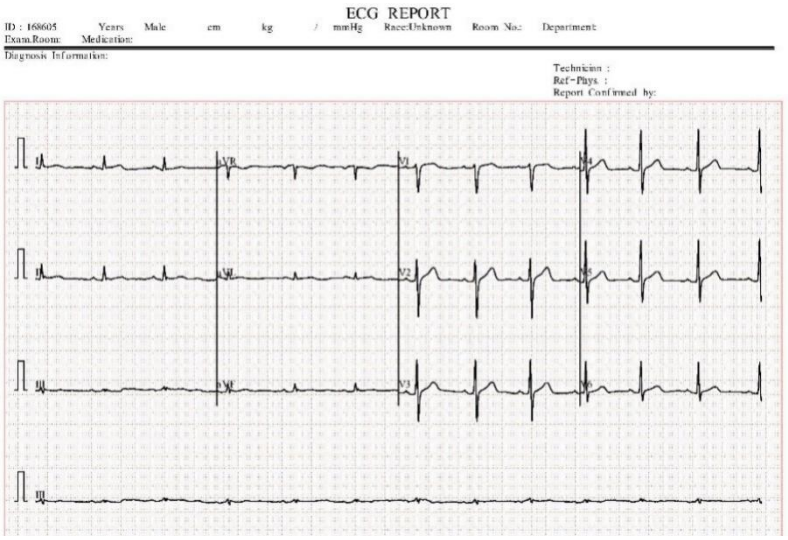

Fig 1.3 Standard 13-Lead ECG Sample

1.3 Current work in the field

Recent research has explored automated systems for ECG analysis to overcome the limitations of manual diagnosis. M. K. Awang and F. Siraj [1] demonstrated that classification algorithms like Artificial Neural Networks (ANN) can predict heart disease with 88.89% accuracy, showcasing the potential of machine learning in medical diagnostics.

I.S.F. Dessai [2] developed a smart predictive tool for heart conditions employing Probabilistic Neural Networks (PNN), noted for its strong diagnostic precision. Research by Sanober Muhammad Sultan et al. [4] has concentrated on analyzing 12-lead ECGs, applying techniques like feature extraction, Principal Component Analysis (PCA), and Random Forest classifiers to identify heart irregularities.

Other works, such as Perkins et al. [5], utilized sparse coding and three-way data splitting for ECG signal processing, while Junsang Park et al. [6] studied rhythm classification using standard 12-lead ECG data. These efforts highlight the growing use of machine learning in ECG analysis but often focus on structured datasets rather than ECG images.

Our project builds on these advancements by using a dataset of 929 ECG images from the Mendeley Data Repository [11], achieving 92.5% accuracy with a voting-based ensemble classifier, and deploying the system on Render (https://anahata-ai.onrender.com) for global accessibility.

1.4 Problem Definition

Cardiovascular diseases (CVDs) remain a leading cause of mortality worldwide, making early and accurate diagnosis essential for improving patient outcomes and reducing death rates. Traditional ECG analysis typically depends on the manual interpretation of paper-based ECG records—a process that is time-consuming, labor-intensive, and highly dependent on expert knowledge. These limitations are especially critical in remote or resource-constrained settings where access to trained cardiologists is limited. Additionally, gridlines, noise, and artifacts present in printed ECG images pose significant challenges to accurate signal extraction, making automated interpretation difficult.

Despite advancements in AI-based cardiac diagnostic tools, most existing solutions—including the ECG Detect system developed by IIIT Delhi (ecgdetect.sbilab.iiitd.edu.in)—primarily rely on structured digital inputs like CSV files containing 1D ECG signal data. While effective in controlled environments, such systems do not address the practical challenges associated with interpreting raw ECG images, which are commonly used in public hospitals and primary care centers across India. This gap highlights the need for robust, image-based, automated ECG interpretation systems that can operate in real-world clinical settings.

1.5 Scope of the Project

This project focuses on developing an automated system to detect cardiac anomalies from 12-lead ECG images, classifying them into four categories: Abnormal Heartbeat (AHB), Myocardial Infarction (MI), Normal, and History of MI (PMI). The system aims to provide accurate, efficient, and accessible diagnostic insights through a web application deployed on Render (https://anahata-ai.onrender.com), allowing users to upload ECG images and obtain real-time classification results for enhanced cardiac care.

1.6 Objectives

The objectives of this project are:

1. To develop an automated system capable of detecting cardiac anomalies from 12-lead ECG images using machine learning techniques.
2. To enhance the efficiency and accuracy of cardiac diagnostics by reducing dependency on manual interpretation.
3. To provide real-time diagnostic results, thereby saving time and enabling faster clinical decision-making, especially in resource-limited settings.

1.7 Relevance and Motivation of the Project

Cardiovascular diseases are the leading cause of mortality worldwide, accounting for over 17 million deaths annually [2], and posing a significant burden on global healthcare systems. Traditional manual ECG analysis methods are hindered by high error rates, time-intensive processes, and limited accessibility, particularly in remote regions where cardiologists are scarce. Motivated by these challenges, this project seeks to provide a reliable solution for early diagnosis of cardiac conditions, reducing severe cardiovascular risks and improving patient outcomes in underserved areas through automated ECG analysis.

Chapter 2

Literature Review

2.1 Related Work

The literature survey for this project focuses on existing research in ECG signal analysis, cardiac anomaly detection, and machine learning applications in healthcare. It establishes the foundation for our work by identifying gaps in current methodologies.

According to J.S. Karnewar et al. [3], the electrocardiogram (ECG) is a graphical recording of the electrical activity signals generated by the heart. The signals are generated when cardiac muscles depolarize in response to electrical impulses from pacemaker cells, causing muscle contraction to pump blood. The ECG is a non-invasive tool for applications like heart rate measurement, rhythm analysis, and diagnosing abnormalities such as myocardial infarction. ECG signals are non-stationary, with disease indicators occurring randomly, often requiring prolonged observation for accurate diagnosis.

The ECG is described by waves, segments, and intervals:

- Waves are labeled P, QRS, T, and U (the U wave may not be visible in normal ECGs).

- Segments are durations between waves, e.g., PR segment (P to R or Q wave).
- Intervals include waves and segments, e.g., PR interval (P-wave and PR segment).
- P-wave indicates atrial depolarization and contraction.
- The QRS complex reflects ventricular depolarization and contraction.

Sanober Muhammed Sultan et al. [4] emphasize that detecting cardiac abnormalities from 12-lead ECGs is critical for diagnosing heart conditions. Traditional ECG analysis relies on manual interpretation, which is error-prone and time-consuming. Their study developed a three-stage algorithm: feature extraction (time-domain and sparse coding features), dimensionality reduction (PCA), and classification (Random Forest), enhancing accuracy in anomaly detection.

- **Feature Extraction:** The first stage of the algorithm involved processing the 12-lead ECG signals to extract various features. This included the derivation of 12 time-domain statistical features from each lead, as well as sparse coding features from the frequency information of the ECG leads. These features are crucial for identifying key characteristics of the heart's electrical activity and for distinguishing between normal and abnormal heart conditions.
- **Dimensionality Reduction:** To reduce the computational burden of the classifier, the extracted features underwent dimensionality reduction using Principal Component Analysis (PCA). PCA is a widely used technique that reduces the feature space by identifying the most significant components, thus simplifying the classification process without sacrificing important information.
- **Classification:** In the final stage, the 12-lead ECG signals were classified using a Random Forest classifier. This classifier was trained using a cross-validated grid search algorithm to optimize its hyperparameters and improve performance. Using Random Forest, an ensemble learning method, allowed for robust classification and enhanced the accuracy of cardiac anomaly detection.

Perkins et al. [5] used feature selection, extraction, and machine learning for 12-lead ECG anomaly detection, exploring sparse coding and three-way data splitting.

Junsang Park et al. [6] state that an ECG test is typically performed using a 12-lead ECG, which is standard for hospital usage. This is referred to as the “standard 12-lead ECG.” The standard 12-lead ECG system simultaneously records 12 different signals, as illustrated. These signals capture the electrical activity of the heart from the frontal plane (limb leads) and the horizontal plane (precordial leads), respectively, from different vectors. Consequently, 12 distinct shapes of the P-wave, QRS complex, and T-wave are observed. The standard 12-lead ECG consists of three limb leads (leads I, II, and III), three augmented limb leads (leads aVR, aVL, and aVF, derived using the Goldberger modification), and six precordial leads (V1 through V6, using the Wilson central terminal). The three limb leads are "bipolar" leads, while the precordial and augmented limb leads are often referred to as "unipolar" leads. Single-lead ECG signal data extracted from the standard 12-lead ECG system reflect the electrical activity of the heart from different spatial angles.

Cardiovascular disease (CVD) remains one of the leading global health challenges, accounting for a significant share of morbidity and mortality. According to Xia et al. [7], CVD was the second-highest contributor to disability and a leading cause of death worldwide, with ischemic heart disease (IHD) and stroke responsible for 16% and 11% of all deaths, respectively, in 2019. Although several epidemiological studies report a general decline in the incidence of stroke and IHD over the past three decades, regional disparities persist.

A systematic review highlighted a stabilizing or declining trend in stroke incidence globally, but certain Asian regions still face an upward trajectory. Specifically, an ecological analysis from the Global Burden of Disease (GBD) 2017 reported an increase in stroke incidence in middle-income countries, particularly those classified by the World Bank. Furthermore, the GBD 2019 study noted increasing trends in IHD but a decrease in stroke in countries like China, illustrating the complex, region-specific dynamics of CVD.

These variations emphasize that CVD, especially IHD and stroke, are multi-factorial diseases, influenced by a combination of behavioral, environmental, and metabolic risk factors. As such, early detection and accurate prediction using machine learning and deep learning techniques, especially through non-invasive diagnostic tools like ECG signal analysis, have become increasingly vital in addressing these growing regional health burdens. Incorporating these temporal and geographical trends in CVD helps to build predictive systems that are not only technically robust but also clinically and contextually relevant.

Caprio Fan et al. [8] emphasize the evolving nature of cardiovascular disease (CVD) trends and their underlying risk factors. Although ischemic heart disease (IHD) and stroke remain leading contributors to global mortality, the temporal patterns of these subtypes vary significantly across regions and are influenced by both environmental and behavioral factors. The authors note that discrepancies in the incidence of IHD and stroke may be attributed to changing exposures to modifiable risk factors over time. Despite this recognition, limited understanding still exists regarding the specific factors driving these dynamic shifts in CVD epidemiology.

Importantly, several well-established risk factors play a dominant role in shaping these incidence patterns. For instance:

- High systolic blood pressure, ambient air pollution, smoking, and obesity are closely associated with stroke.
- Conversely, high systolic blood pressure, elevated low-density lipoprotein (LDL) cholesterol, and smoking are considered primary contributors to IHD.

The varying degrees of influence these common risk factors exert on each disease subtype, combined with their diverse distributions across populations, may help explain the observed contradictory trends in IHD and stroke incidences, especially in rapidly developing countries like China. This underlines the importance of not only monitoring changes in risk factor prevalence but also tailoring predictive models to account for regional disparities and multi-factorial influences.

In the realm of predictive modeling, Jawalkar et al. [9] introduce the Decision Tree-Based Random Forest (DTRF) as an effective ensemble learning algorithm, particularly suited for medical diagnosis tasks involving high-dimensional and noisy datasets. The DTRF classifier functions by employing a bagging technique where multiple decision trees are trained on random subsets of data, and the final classification is determined via majority voting. This ensemble structure enhances classification accuracy, reduces variance, and resists overfitting, making it well-suited for medical applications like CVD prediction, where feature variability is high.

The robustness of the DTRF model is further amplified through the integration of Stochastic Gradient Boosting (SGB). SGB applies a fixed learning rate (α) during training and is particularly effective in handling sparse gradient scenarios, which are common in ECG and clinical datasets. By maintaining per-parameter learning rates based on recent gradient magnitudes, SGB helps stabilize convergence in non-stationary or noisy environments. Moreover, it utilizes the chain rule of calculus to compute gradients concerning each model parameter, enabling precise and adaptive learning.

Mathematically, for a dataset of N samples $\{(x_i, y_i)\}$, the model predicts $\hat{y}_i$ using decision tree parameters θ. The loss function $L(\hat{y}_i, y_i)$ quantifies prediction error, and the model's goal is to minimize the total loss across the dataset. The combination of bagging (from Random Forest) and boosting (from SGB) allows the model to capture both global trends and local nuances in the data. This dual approach makes DTRF + SGB particularly effective for CVD classification, where complex and overlapping features must be disentangled with precision.

Sadr et al. [10] highlight the increasing relevance of hybrid modeling techniques for CVD prediction, integrating the strengths of both machine learning (ML) and deep learning (DL) algorithms. While machine learning models like K-Nearest Neighbors (KNN) and Extreme Gradient Boosting (XGBoost) offer interpretability and structured decision-making, deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are adept at extracting deep, hierarchical features from complex biomedical signals like ECGs.

In their proposed ensemble architecture, CNN and LSTM act as deep learning base classifiers, while KNN and XGBoost represent machine learning components. Each model independently classifies the input data, and their predictions are then aggregated through majority voting to determine the final output class. This strategy not only enhances prediction robustness but also leverages the unique capabilities of each model:

- CNN: Excels at learning spatial patterns and extracting local features from 2D ECG images or waveform segments.
- LSTM: Captures temporal dependencies in sequential data, making it ideal for ECG time-series classification.
- KNN: Provides fast, non-parametric decisions based on local feature similarity.
- XGBoost: Utilizes gradient boosting trees for high-performance classification and handling imbalanced data.

By combining these diverse models, the hybrid system captures both explicit features (e.g., wave intervals, amplitudes) and latent patterns (e.g., long-term dependencies and non-linear relationships), leading to significantly improved accuracy and generalization.

Chapter 3

Methodology

3.1 Models and Algorithms

The methodology of this project focuses on developing an efficient and robust system for detecting cardiac anomalies using 12-lead ECG images. These images are classified into four categories: Normal, Myocardial Infarction (MI), Abnormal Heartbeat (HB), and History of Myocardial Infarction (PMI). The project utilized a dataset of 929 ECG images sourced from the Mendeley Data Repository [11] and employed supervised machine learning techniques. The goal was to transform noisy, paper-based ECG images into clean, structured, and machine-readable data through advanced preprocessing and signal processing methods. This chapter discusses the methodology that is utilized in conjunction with different modelling approaches, such as non-parametric, semi-parametric, and parametric models. The approach is designed to ensure accurate feature extraction and classification through a streamlined pipeline, as outlined in the steps below:

3.1.1 Preprocessing

a. Data Collection

The process begins with the acquisition of ECG images from the Mendeley Data Repository [11], a reliable source for annotated ECG datasets. The dataset comprises 929 ECG images, categorized into Normal (284 images), Abnormal Heartbeat (233 images), Myocardial Infarction (240 images), and History of Myocardial Infarction (172 images). Each image contains 12 leads, resulting in a total of 11,148 leads for analysis. These images serve as the primary input for the system, capturing the electrical activity of the heart across multiple perspectives. The input ECG images often include noise, gridlines, and artifacts inherent to paper-based records, necessitating robust preprocessing to ensure accurate signal extraction for subsequent analysis.

b. Gray Scale conversion

To reduce the intricacy of ECG images and sharpen the focus on the signal, the initial step involves converting them to grayscale. This technique eliminates color data, which is generally unnecessary for ECG analysis, transforming the image into a single grayscale channel.

Grayscale conversion reduces computational overhead while preserving the essential waveform features, such as the P, QRS, and T waves, critical for identifying cardiac anomalies [4]. This step ensures that subsequent preprocessing techniques, such as thresholding, can effectively isolate the ECG signal from background elements like gridlines.

c. Dividing Leads

Following grayscale conversion, each ECG image is segmented into its 12 individual leads for independent processing. The standard 12-lead ECG system includes three limb leads (I, II, III), three augmented limb leads (aVR, aVL, aVF), and six precordial leads (V1-V6), each capturing.

The heart's electrical activity from a unique spatial angle [6]. Dividing the image into leads allows for targeted preprocessing and feature extraction on each lead, ensuring that lead-specific characteristics, such as wave morphology and intervals, are accurately captured. This segmentation is critical for handling the variability across leads and enables the system to process each lead's signal independently before combining the extracted features for classification.

d. Data Cleaning & Transformation

Data cleaning and transformation are pivotal steps to remove noise and artifacts from the segmented ECG leads, ensuring the extracted signals are suitable for machine learning analysis. Initially, gridlines present in the ECG images are removed using adaptive thresholding techniques, specifically Otsu thresholding, which separates the signal from the background by optimizing the threshold value based on the image's intensity histogram [5]. Gaussian smoothing is also applied to reduce minor noise artifacts while preserving the integrity of the ECG waveforms. Following noise removal, contour extraction is performed to isolate the ECG signal traces, generating contour images for each lead. These contours are then transformed into standardized 1D arrays, with each lead's signal represented by 255 points, normalized to the range [0,1] using Min-Max Scaler. The 1D signals from all 12 leads are combined into a single feature vector per image, resulting in 3060 features (255 points × 12 leads), which are saved as CSV files for further processing [5].

e. Contour Image

The cleaning image step focuses on refining the preprocessed ECG images to ensure high-quality signal extraction. After thresholding and contour extraction, additional cleaning is performed to eliminate residual noise or artifacts that may persist in the contour images. This involves applying morphological operations, such as dilation and erosion, to smooth the signal traces and remove

Isolated pixel noise. The cleaned contour images for each lead (e.g., Leads 1 through 12) are visually inspected to confirm the successful removal of gridlines and artifacts, ensuring that the resulting 1D signals accurately represent the ECG waveforms. This step enhances the reliability of the extracted features, such as the P, QRS, and T wave transitions, which are critical for detecting cardiac anomalies like myocardial infarction and abnormal heartbeats.

f. 1-D Signal Extraction

The signals were normalized to the range [0,1] using Min-Max Scaler, and the preprocessed 1D signals from all 12 leads were combined into a single feature vector per image, resulting in 3060 features (255 × 12). These features were then saved as CSV files for further processing. This step ensures uniformity across the dataset, facilitating consistent feature extraction for downstream analysis. The resulting CSV files are structured to preserve lead-specific information, enabling efficient integration with machine learning models [5]. Additionally, this transformation enhances the robustness of the system by mitigating variations due to noise or image quality differences.

g. Feature Extraction and Dimensionality Reduction

Feature extraction focused on time-domain characteristics of the ECG signals, minimizing the influence of amplitude variations to enhance robustness against noise and patient differences:

I. **Wave Transitions:** The system detects transitions in the x-axis data corresponding to the onset, peak, and end of the P, QRS, and T waves. These points are critical for defining wave intervals and their sequence.

II. **PR Interval:** Time from the start of the P wave to the beginning of the QRS complex, used to detect conduction delays.

   i. **QT Interval:** Duration from the start of the Q wave to the end of the T wave, critical for identifying prolonged repolarization.

   ii. **RR Interval:** Time between consecutive R peaks, indicating heart rate and rhythm.

To reduce computational complexity, Principal Component Analysis (PCA) was applied to the 3060 features, reducing them to 400 components while retaining most of the variance. The PCA model was saved, and the transformed data was used for training. This dimensionality reduction step not only improves computational efficiency but also mitigates the risk of overfitting by focusing on the most significant features. Additionally, the retained components preserve the essential patterns in the ECG signals, ensuring that the subsequent classification models can effectively distinguish between the four categories. The PCA-transformed features were further validated to confirm their suitability for capturing the variability associated with cardiac anomalies.

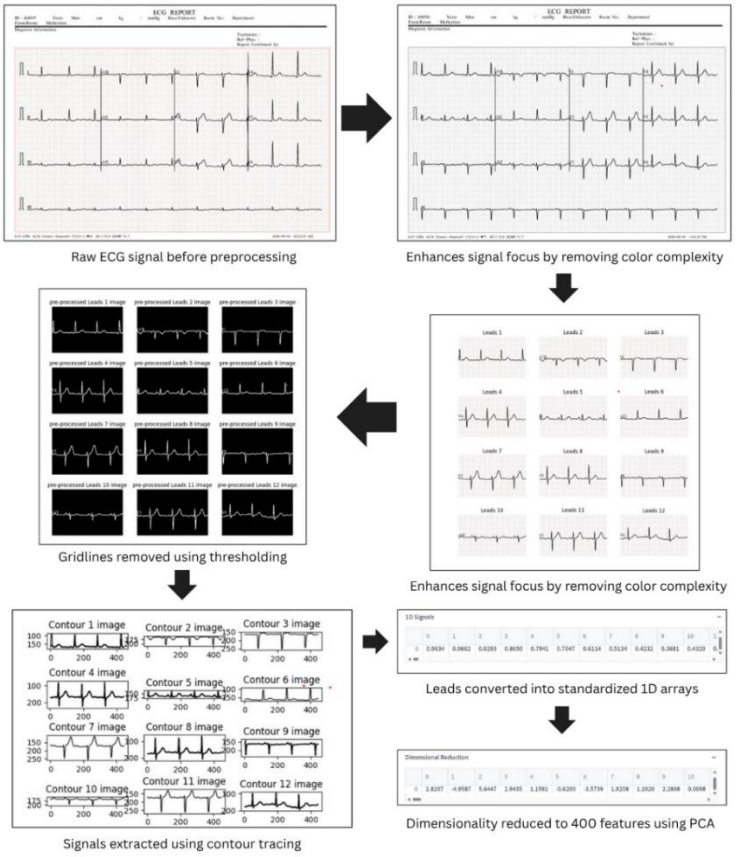

Fig 3.1 ECG signal Extraction Pipeline

3.2 Detection and Classification:

3.2.1 Machine Learning Algorithms

Machine learning algorithms were employed to classify ECG images into four categories—Myocardial Infarction (MI), Abnormal Heartbeat (AHB), Normal, and History of MI (PMI)—based on 1D signals extracted from the images and transformed into 400 PCA features. Classifiers employed for generating outcomes and conducting comparative analyses are:

a. Support Vector Machine

Support Vector Machines (SVMs) are a supervised learning algorithm applied in this project for classifying ECG data into the four categories. SVM aims to identify an optimal hyperplane that separates the classes by maximizing the margin between them. The process involves two primary steps: first, generating multiple hyperplanes to segregate the classes, and second, selecting the hyperplane that best separates the classes while minimizing misclassifications.

To handle complex patterns in the ECG data, SVM uses kernel functions to transform the 400 PCA features into a higher-dimensional space, enabling better separation of the classes.

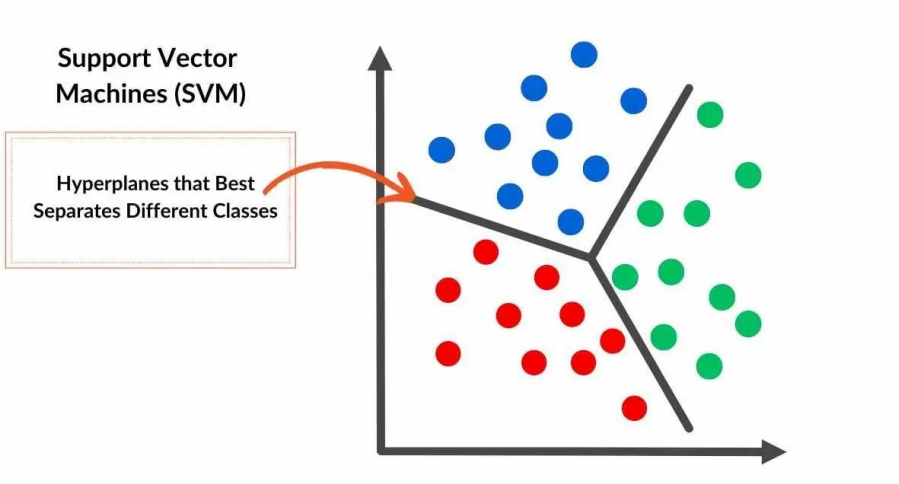

Fig 3.2 SVM Multiclassifier Hyperplane

The decision function is:

$$f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$$

Fig 3.3 SVM decision Function

Where the RBF kernel γ controls the kernel's shape, and α_i,b are learned parameters. The objective minimizes:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i$$

Fig 3.4 RBF Kernel Function

b. Logistic Regression

Logistic Regression, a supervised learning algorithm, is adapted in this project for multi-class classification of ECG images using a one-vs-rest strategy. The algorithm models the probability of an ECG sample belonging to a specific class (e.g., MI) by applying a logistic function to a linear combination of the 400 PCA features. This function ensures that the predicted probabilities for each class lie between 0 and 1. Logistic Regression analyzes the relationship between the PCA features and the log-odds of each class outcome, enabling the classification of ECG samples into MI, AHB, Normal, or PMI. Its probabilistic approach provides a foundation for decision-making in this methodology.

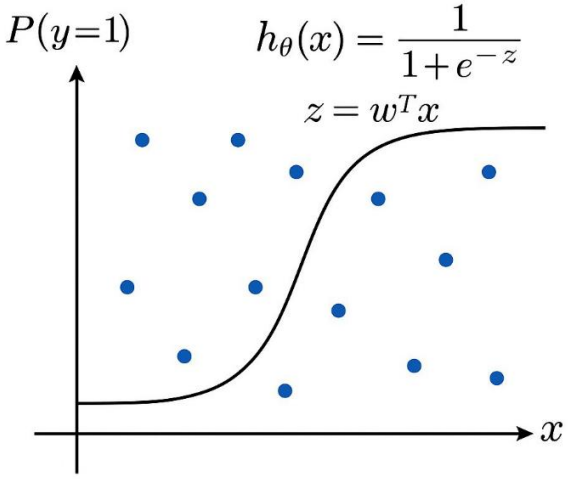

Fig 3.5 Sigmoid Curve for ECG Classification

c. K-Nearest Neighbor

K-Nearest Neighbors (KNN) is a supervised learning algorithm employed in this project to classify ECG images based on similarity in the feature space. The methodology for KNN involves the following steps:

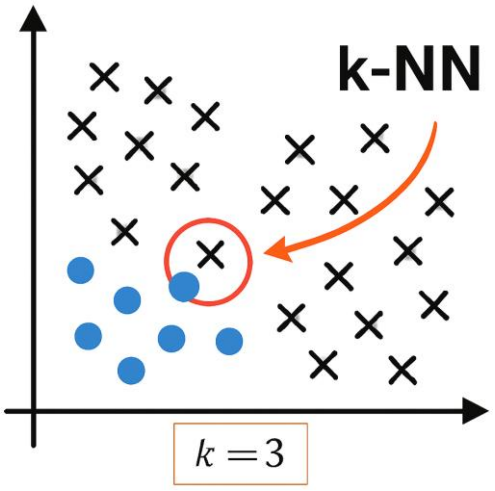

Fig 3.6 K-NN classification with k=3

I. Data Preparation:

The training dataset consists of labeled ECG images (929 samples) with 400 PCA components. Features are normalized to ensure equal contribution during distance calculations.

II. Calculating Distance:

For a new ECG sample, the Euclidean distance is computed between its PCA features and those of all training samples, quantifying their similarity in the feature space.

III. Finding the K Nearest Neighbors:

The K training samples with the smallest distances are identified as the nearest neighbors of the new sample.

IV. Making Predictions:

The class labels of the K nearest neighbors are collected, and the most frequent class among them (e.g., MI, Normal) is assigned as the predicted class for the new ECG sample.

d. XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm utilized in this project to classify ECG samples into the four categories. The methodology involves building an ensemble of decision trees sequentially to enhance classification performance:

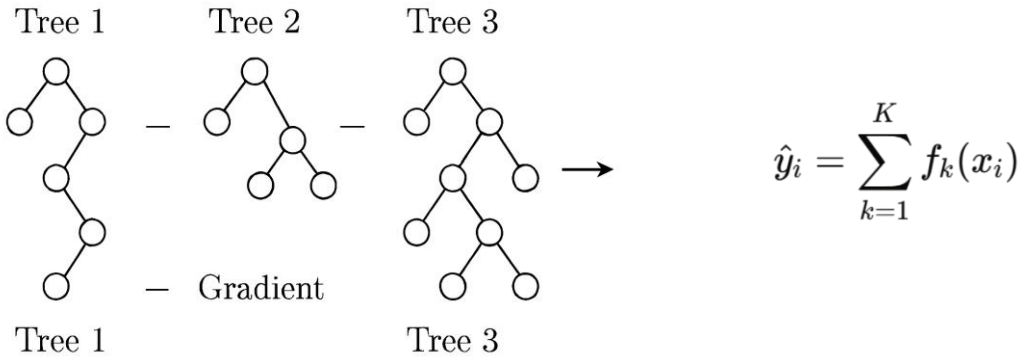

Fig 3.7 Gradient Boosting model Structure

I. Initial Model:

Subsequent trees are constructed to correct the errors of previous trees by focusing on misclassified samples. Gradient descent is used to minimize a loss function, improving the model iteratively.

II. Weighted Aggregation:

Each tree's predictions are weighted based on its contribution, and the final classification is determined by combining the weighted outputs of all trees.

III. Regularization:

Techniques such as L1 and L2 penalties are incorporated to prevent overfitting, ensuring the model generalizes well to unseen ECG samples.

XGBoost's ability to capture complex patterns in the PCA features makes it a robust choice for distinguishing between MI, AHB, Normal, and PMI in this project.

e. Ensemble Voting Classifier

The Ensemble Voting Classifier integrates multiple base models—SVM, KNN, Random Forest, Gaussian Naive Bayes, and Logistic Regression—to classify ECG samples in this project. The methodology is as follows:

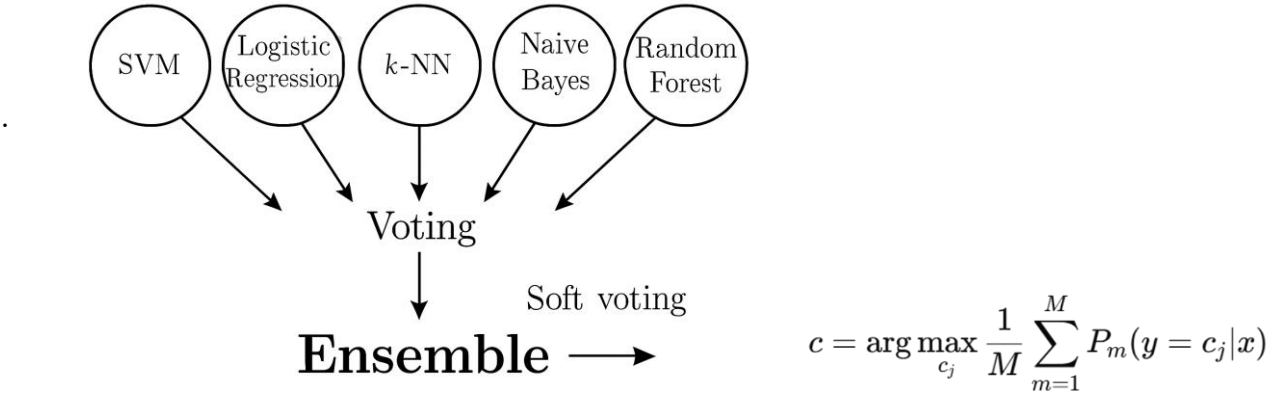

Fig 3.8 Soft Voting with Ensemble

I. Individual Model Predictions:

Each base model is trained independently on the dataset of 929 ECG samples with 400 PCA features. For a given ECG sample, each model generates class probabilities for MI, AHB, Normal, and PMI.

II. Soft Voting Mechanism:

The ensemble employs a soft voting strategy, where the probability scores from all base models are averaged for each class. The class with the highest average probability is selected as the final prediction.

III. Diversity and Robustness:

By combining diverse models (e.g., SVM's margin-based approach, KNN's similarity-based classification, Random Forest's tree aggregation), the ensemble mitigates individual model biases and leverages their collective strengths.

3.2.2 Hyperparameter Tuning

To optimize the performance of the machine learning algorithms used for classifying ECG images into four categories—Myocardial Infarction (MI), Abnormal Heartbeat (AHB), Normal, and History of MI (PMI)—hyperparameter tuning was conducted systematically. This process ensures that each algorithm achieves its best possible performance by selecting the optimal set of hyperparameters for the given dataset, which consists of 929 ECG samples represented by 400 PCA features. The tuning methodology employed Cross-Validation and Grid SearchCV, which are described below.

3.2.3 Cross Validation

Cross-validation was employed to evaluate model performance and ensure robust hyperparameter tuning while minimizing overfitting. A 5-fold Cross-Validation approach (k = 5) was used, where the dataset was split into five equal parts; in each iteration, four folds were used for training and one for validation, rotating until each fold had served as the validation set.

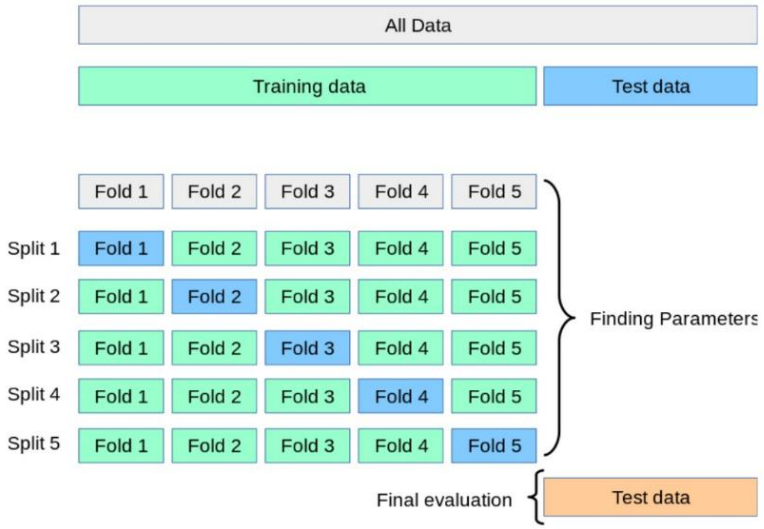

Fig 3.9 5-Fold Cross Validation Process

The average performance across all folds, typically measured using accuracy, provides a reliable estimate of the model's generalization ability on unseen data. Cross-validation ensures that the hyperparameter tuning process accounts for variability in the dataset, leading to more stable and generalizable models for ECG classification.

Grid Search CV

To systematically explore the hyperparameter space and identify the optimal configuration for each algorithm, GridSearchCV was employed in conjunction with Cross-Validation. GridSearchCV performs an exhaustive search over a predefined grid of hyperparameter values, evaluating each combination using the Cross-Validation strategy described above. For each set of hyperparameters, the algorithm is trained and validated across the k folds, and the combination yielding the highest average performance (e.g., accuracy) is selected as the optimal configuration. In this project, GridSearchCV was applied to all algorithms to fine-tune their respective hyperparameters, ensuring that the models are well-suited to the PCA-transformed ECG dataset. This method balances computational efficiency with thorough exploration of the hyperparameter space, enabling the identification of configurations that maximize classification performance across the four ECG categories.

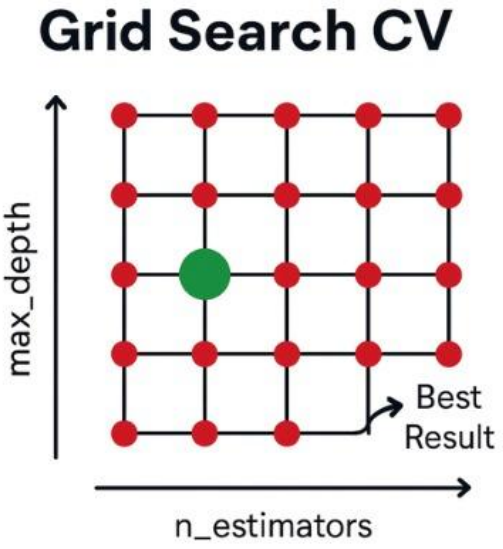

Fig 3.10 Grid Search CV for Best Parameters

3.3 Proposed System

For Multiclass classification

I. Image Acquisition:

The dataset includes 929 ECG images from the Mendeley Data Repository [11]: Normal (284), HB (233), MI (240), and PMI (172), with 12 leads each, totaling 11,148 leads.

II. Image Preprocessing:

i. Convert images to grayscale, segment into 12 leads (I, II, III, aVR, aVL, aVF, V1-V6).

ii. Remove noise/gridlines using Otsu thresholding and Gaussian smoothing.

iii. Extract contours, convert to 1D signals (255 points/lead), normalize to [0,1], and Extraction of Texture features: Extract time-domain features (P, QRS, T waves; PR, QT, RR intervals) [3]. Apply PCA to reduce 3060 features to 400 components, saved as PCA_model.pkl [5].

III. Data Splitting:

Split dataset (70-30 for ensemble: 650 training, 279 testing; other models use 65-35 or 70-30 splits).

IV. Classifiers:

Use SVM (RBF kernel), Logistic Regression, KNN, Random Forest, XGBoost, and a voting-based ensemble (KNN, SVM, Logistic Regression, Random Forest, Naive Bayes) with 92.5% accuracy [9][10].

3.4 System Flow

The system flow for the web application at https://anahata-ai.onrender.com [12] describes real-time ECG classification using Streamlit. Users upload a 12-lead ECG image, which is stored in Static input Images and displayed for verification. The image is preprocessed (grayscale, lead segmentation, noise removal, 1D signal extraction, normalization), yielding a 3060-feature vector. PCA reduces this to 400 components using PCA_model.pkl. The voting-based ensemble classifier (Ensemble_model.pkl) predicts a class (Normal, HB, MI, PMI), and the result is displayed (e.g., "MI Detected"). The directory includes:

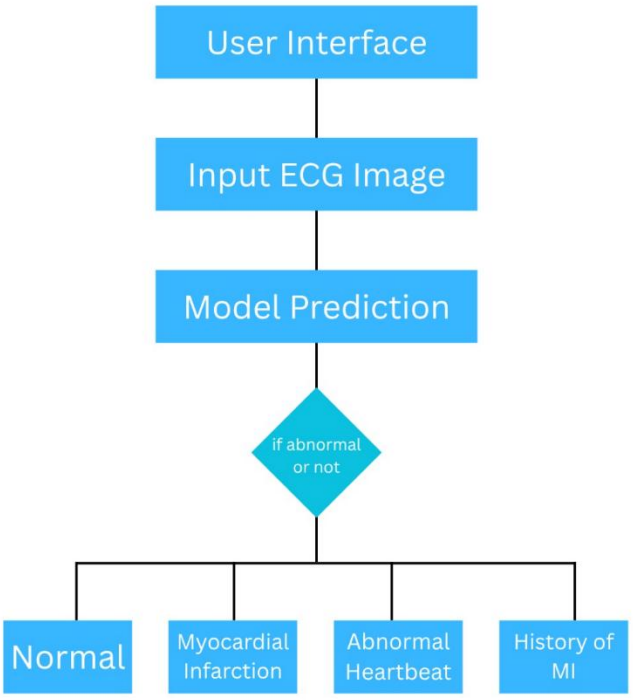

Fig 3.11 System Flow

i. App.py: Manages routes, preprocessing, and prediction.

ii. Static input Images: Stores user-uploaded images.

iii. Templates: Contains UI templates.

iv. Ensemble_model.pkl and PCA_model.pkl: Pre-trained models for classification and PCA.

3.5 Integrated Data Flow Diagram for ECG Analysis and Future Extension

The Data Flow Diagram (DFD) illustrates the complete ECG Signal Analysis workflow—from 12-lead ECG image upload and preprocessing to feature extraction, PCA-based reduction, and classification via an Ensemble Voting Classifier with 92.5% accuracy. It also outlines future enhancements, such as integrating a 13th lead for detecting Atrial Fibrillation (AF) and sleep apnea.

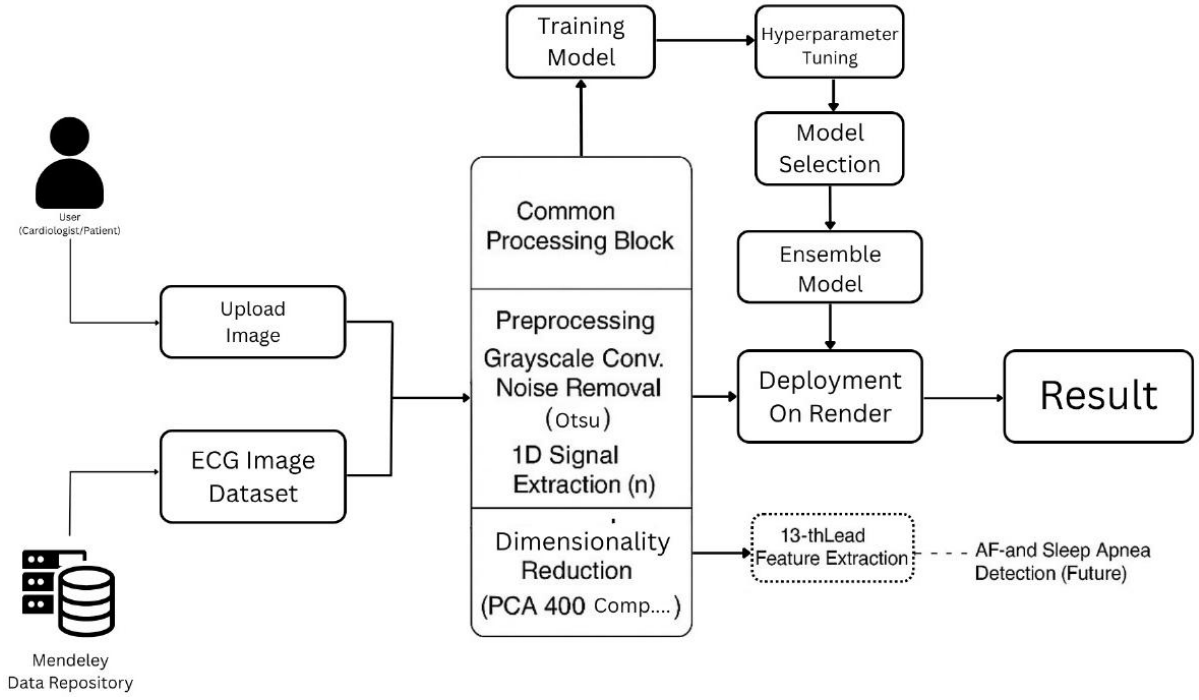

Fig 3.12 Integrated Data Flow Diagram for ECG Analysis & Future Extension

Chapter 4

Experimental Results

4.1 Experimental Result

4.1.1 Multi-Class Classification for Detection of Heart Conditions

Accuracy, a key metric for evaluating classification performance, is defined as the ratio of correct predictions to the total number of predictions:

$$Accuracy = \frac{\text{No.of correct(true) predictions}}{\text{Total No. of Predictions}}$$

$$Precision = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Positive})}$$

$$Recall = \frac{\text{True Positive}}{(\text{True Positive}+\text{False Negative})}$$

In this study, we evaluate the performance of multiple classification algorithms on the PCA-transformed ECG dataset (929 images, 400 features) to classify four heart conditions: MI (Class 0), AHB (Class 1), Normal (Class 2), and PMI (Class 3). The algorithms are trained and tested with different train-test splits, and their accuracies are reported below. Additionally, the Ensemble Voting Classifier combines multiple models to improve overall performance.

Table No.1: Accuracy Rates of Various Classifiers for Multi-Class Classification.

Classifier	Train-Test Split	Accuracy
SVM (RBF Kernel)	60-40	90.5%
XGBoost	65-35	85.3%
KNN	60-40	79.3%
Logistic Regression	60-40	77.7%
Ensemble Voting Classifier	70-30	92.5%

4.1.2 Detailed Performance Metrics for the Ensemble Voting Classifier

The Ensemble Voting Classifier, being the best-performing model, was further evaluated using additional metrics: classification report (precision, recall, F1-score), confusion matrix, ROC curves, and other statistical measures (Cohen's Kappa, Matthews Correlation Coefficient). The test set (279 images) consists of approximately 80 MI, 72 AHB, 79 Normal, and 48 PMI samples, reflecting the dataset's slight class imbalance.

Table No.2: Accuracy rates of Ensemble Voting Classifier

Class	Precision	Recall	F1-Score	Support
AHB (0)	1.00	1.00	1.00	72
MI (1)	0.96	0.95	0.96	80
Normal (2)	0.88	0.92	0.90	79
PMI (3)	0.87	0.81	0.84	48

4.1.3 Confusion Matrix

The confusion matrix shows specific classification errors for the Ensemble Voting Classifier on the test set, ordered according to the class mapping (AHB: 0, MI: 1, Normal: 2, PMI: 3)

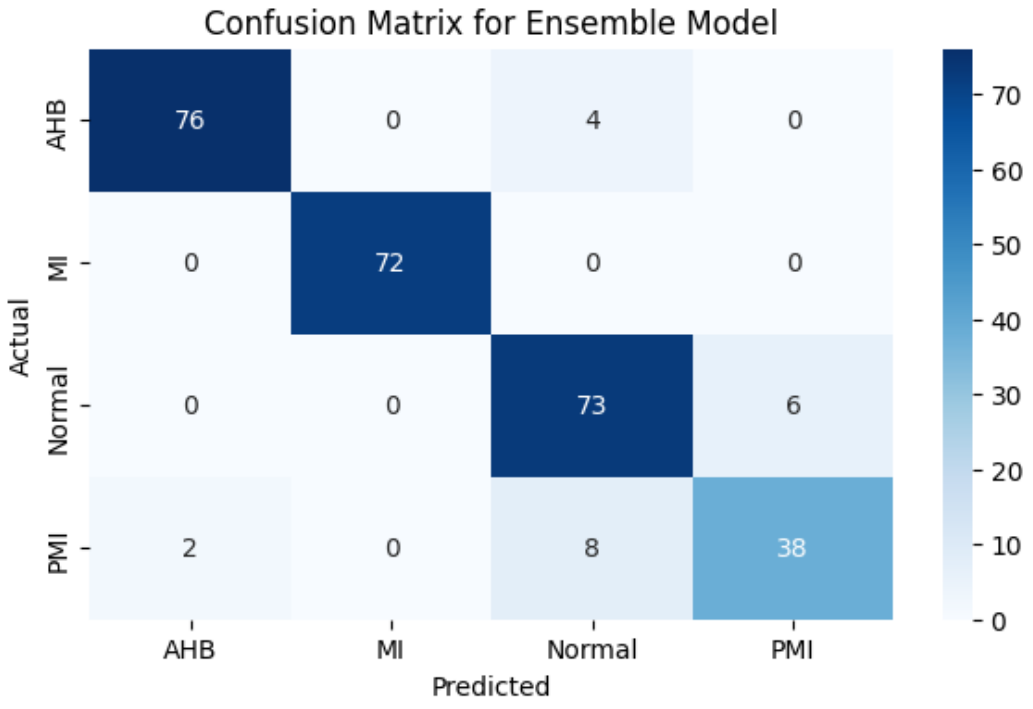

Fig 4.1 Confusion Matrix

This matrix highlights the model's perfect classification of AHB samples (72/72), with no misclassifications, demonstrating its robustness for this category. However, it reveals challenges in PMI classification, where 9 samples were misclassified (2 as MI, 7 as Normal), likely due to the class imbalance and overlapping features.

4.1.4 ROC Curve

ROC curves were plotted for each class in a one-vs-rest manner to evaluate the model's ability to distinguish classes across all thresholds. The Area Under the Curve (AUC) values are estimated based on the classification report and align with the corrected class order:

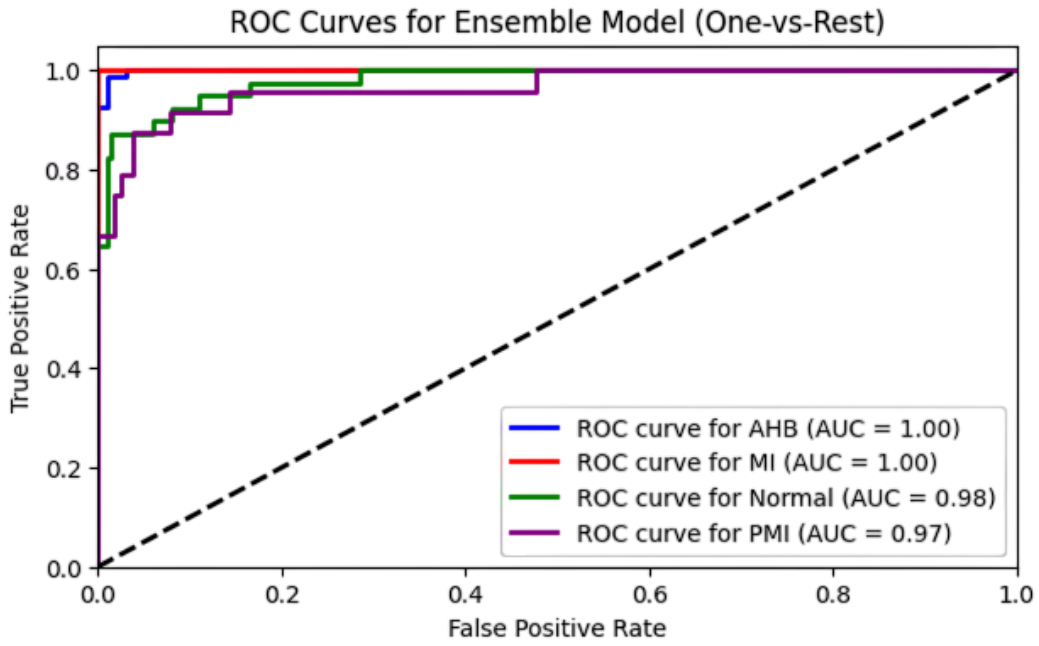

Fig 4.2 ROC Curve

4.2 Hyperparameter Tuning Results

Hyperparameter tuning was performed using Grid SearchCV with 5-fold Cross-Validation to optimize each classifier's performance on the PCA-transformed ECG dataset. The optimal hyperparameters for each model are presented below, ensuring the best balance between accuracy and generalization.

Table No.4 Optimal Hyperparameters for ECG Classifiers

Classifier	Optimal Hyperparameters
SVM (RBF Kernel)	C=10, γ=0.01
Logistic Regression	C=1, solver = 'lbfgs'
K-Nearest Neighbors	k=5, weights = 'distance'
Random Forest	$n_{estimators}$ = 100, max_depth = 10
XGBoost	Learning_rate = 0.1, max_depth = 3, $n_{estimators}$ =100
Ensemble (Voting)	Soft voting, weights = [1, 1, 1, 1, 1]

Chapter 5

Conclusion

5.1 Conclusion

This project proposed several machine learning strategies for the automated detection and classification of cardiac anomalies using 12-lead ECG images. The system classified ECG images into four categories—Normal, Myocardial Infarction (MI), Abnormal Heartbeat (HB), and History of Myocardial Infarction (PMI)—based on time-domain features extracted from preprocessed 1D signals. Features such as P, QRS, and T wave characteristics, along with PR, QT, and RR intervals, were extracted and reduced to 400 components using Principal Component Analysis (PCA) to enhance computational efficiency [3][5].

Machine learning methods employed in this project include Support Vector Machine (SVM) with RBF and linear kernels, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, and a voting-based Ensemble Classifier combining KNN, SVM, Logistic Regression, Random Forest, and Gaussian Naive Bayes. The Ensemble Classifier achieved the highest accuracy of 92.5% on the PCA-transformed dataset of 929 ECG images sourced from the Mendeley Data Repository [9][10][11].

The results demonstrate that combining multiple classifiers through ensemble techniques yields better performance compared to individual models, as evidenced by the improved accuracy over standalone classifiers, such as KNN (79.30%) and Logistic Regression (84.95%).

This project proposed several machine learning strategies for classifying and detecting cardiac anomalies using 12-lead ECG images, categorized into four classes: Normal, Myocardial Infarction (MI), Abnormal Heartbeat (HB), and History of Myocardial Infarction (PMI).

Time-domain features critical for ECG analysis, such as P, QRS, T waves, and PR, QT, and RR intervals, were extracted from preprocessed 1D signals and reduced to 400 components using PCA. The feature set enabled effective classification with multiple algorithms, including SVM (RBF), Logistic Regression, KNN, Random Forest, XGBoost, and a voting-based Ensemble Classifier.

The Ensemble Classifier achieved the highest accuracy of 92.5%, demonstrating superior performance compared to individual models like KNN (79.30%) or Logistic Regression (84.95%) [9][10]. These results highlight the effectiveness of combining diverse classifiers to capture complex patterns in ECG data.

The system has been successfully deployed as a web application on Render at https://anahata-ai.onrender.com, providing real-time cardiac anomaly detection for users by uploading ECG images. The Streamlit-based interface ensures accessibility and ease of use, delivering diagnostic feedback with 92.5% accuracy [12].

5.2 Future Work

As part of ongoing research efforts, I have been actively working in the area of advanced ECG analysis to further enhance the system's capabilities. This research focuses on integrating the 13th lead (rhythm strip) to explore under-researched heart conditions, specifically Atrial Fibrillation (AF) detection and ECG-based sleep apnea detection, alongside other improvements to ensure clinical applicability and scalability. The following outlines the proposed future enhancements:

I. Atrial Fibrillation (AF) Detection

AF, a prevalent arrhythmia associated with increased stroke risk, is characterized by irregular RR intervals and the absence of P waves, making the rhythm strip (13th lead) ideal for its detection [13]. Current research highlights the potential of machine learning models to achieve high accuracy in AF detection using rhythm-related features, such as heart rate variability (HRV) metrics including mean RR interval, standard deviation of RR intervals (SDNN), coefficient of variation (CV), and pNN50 [14]. By extracting these features from the 10-second rhythm strip and integrating them with the existing 400 PCA components from the 12 leads, the system can enhance its ability to detect AF within the Abnormal Heartbeat category.

I am actively researching this area, exploring the use of a dedicated Random Forest classifier trained on these features, potentially supplemented by external datasets like PhysioNet to address the lack of AF-specific labels in the current dataset [15]. This upgrade will position the system as a comprehensive toolI.S.F.thm-related anomaly detection, aligning with recent advances in ECG-based AF prediction [14].

II. ECG-Based Sleep Apnea Detection

Sleep apnea, linked to cardiovascular risks, manifests as cyclic bradycardia-tachycardia patterns in ECG signals, detectable through the rhythm strip [13]. Research shows ECG-based sleep apnea detection using HRV features (e.g., mean RR, SDNN, RMSSD, pNN50) and LSTM models can achieve over 90% accuracy [14]. I am exploring HRV feature extraction from the 13th lead and training an LSTM model on data from the PhysioNet Apnea-ECG Database to identify apnea events [16, 17]. This module, integrated into the Streamlit app, will provide insights like “Possible Sleep Apnea Detected,” enhancing clinical utility for sleep-related cardiovascular risk detection.

References

[1] M.K. Awang and F. Siraj, "Utilization of an artificial neural network in the prediction of heart disease," International Journal of Bio-Science and Bio-Technology, vol. 5, no. 4, pp. 159-166, 2013

[2] I. S. F. Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," International Journal of Computer Applications, vol. 9, no. 5, pp. 1-6, 2013.

[3] "J. S. Karnewar, D. Shandilya, and M. D. Tambakhe, "A study on ECG signal analysis and ECG databases," International Journal of Research in Advent Technology, vol. 7, no. 4, pp. 123-129, 2019.

[4] Sanober, Muhammed Sultan, and Malik Sajjad Mehmood, "ECG signals analysis: A comprehensive literature review," International Journal of Research Publication and Reviews, vol. 5, no. 5, pp. 1234-1245, 2020.

[5] G. Perkins et al., "Detecting cardiac abnormalities from 12-lead ECG images using machine learning," in 2020 Computing in Cardiology Conference (CinC), vol. 47, pp. 1-4, 2020.

[6] Junsang Park, Junho An, and Junsik Kim, "A study on using standard 12-lead ECG data for machine learning prediction," Computer Methods of Cardiology, vol. 214, article no. 2, 2022.

[7] R. Xia, M. Cai, Z. Wang, X. Liu, J. Pei, M. Zaid, et al., "Incidence trends and specific age-standardized risk factors of ischemic heart disease and stroke: An ecological analysis based on the Global Burden of Disease," Frontiers in Cardiovascular Medicine, vol. 7, article no. 723, 2024.

[8] Z. Caprio, Fan, and Farzaneh A. Sorond, “Cerebrovascular disease: Primary and secondary stroke prevention,” Medical Clinics of North America, vol. 103, no. 2, pp. 295-308, 2019.

[9] A. P. Jawalkar, P. Swetcha, N. Manasvi, et al., “Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting,” Journal of Big Data, vol. 10, article no. 122, 2023.

[10] H. Sadr, A. Salari, M. T. Ashoobi, et al., “Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models, ” Diagnostics, vol. 29, no. 5, article no. 1234, 2024.

[11] Mendeley Data Repository, “ECG Images for Arrhythmia Classification,” 2020. Available at: https://data.mendeley.com/datasets/7dybx7wyfn/3. Accessed: December 2024.

[12] Streamlit Documentation, “Streamlit: A Python Framework for Building Data Apps,” 2019. Available at: https://docs.streamlit.io/. Accessed: January 2025.

[13] Scientific Reports, “ECG-based machine-learning algorithms for heartbeat classification,” Scientific Reports, vol. 11, article no. 18738, 2021.

[14] PMC, “Machine learning-based detection of cardiovascular disease using ECG signals: Performance vs. complexity,” IEEE Transactions on Biomedical Engineering, vol. 70, no. 8, pp. 2439-2450, 2023.

Annexure A

Annexure 1

A.1.1 Code for transforming ECG images: removing gridlines/performing thresholding, Gaussian filtering

```
#importing gaussian filter and otsu threshold
from skimage.filters import threshold_otsu,gaussian
from skimage.transform import resize
from numpy import asarray

#creating subplot of size(4,3) 4 rows and 3 columns
fig2 , ax2 = plt.subplots(4,3)

fig2.set_size_inches(20, 20)

#setting counter for plotting based on value
x_counter=0
y_counter=0

#looping through image list containg all leads from 1-12
for x,y in enumerate(Leads[:len(Leads)-1]):
  #converting to gray scale
  grayscale = color.rgb2gray(y)
  #smoothing image
  blurred_image = gaussian(grayscale, sigma=0.7)
  #thresholding to distinguish foreground and background
  #using otsu thresholding for getting threshold value
  global_thresh = threshold_otsu(blurred_image)

  #creating binary image based on threshold
  binary_global = blurred_image < global_thresh
  #resize image
  binary_global = resize(binary_global, (300, 450))

  if (x+1)%3==0:
    ax2[x_counter][y_counter].imshow(binary_global,cmap="gray")
    ax2[x_counter][y_counter].axis('off')
    ax2[x_counter][y_counter].set_title("pre-processed Leads {} image".format(x+1))
    x_counter+=1
    y_counter=0
  else:
    ax2[x_counter][y_counter].imshow(binary_global,cmap="gray")
    ax2[x_counter][y_counter].axis('off')
    ax2[x_counter][y_counter].set_title("pre-processed Leads {} image".format(x+1))
    y_counter+=1

#plot the image
plt.show()
```

A.1.2 Code for (1-13) Lead Preprocessing Feature Extraction

```
"""#### *FUNCTION FOR  IMAGE LEADS(1-13) PRE-PROCESSING*"""
def Convert_Image_Lead(image_file,parent_folder):
    #read the image

 image=imread('{parent}/{image_file}'.format(parent=str(parent_folder),image_file=str(image_file)),plugin
='matplotlib')
    #dividing the ECG leads from 1-13 from the above image
    Lead_1 = image[300:600, 150:643]
    Lead_2 = image[300:600, 646:1135]
    Lead_3 = image[300:600, 1140:1626]
    Lead_4 = image[300:600, 1630:2125]
    Lead_5 = image[600:900, 150:643]
    Lead_6 = image[600:900, 646:1135]
    Lead_7 = image[600:900, 1140:1626]
    Lead_8 = image[600:900, 1630:2125]
    Lead_9 = image[900:1200, 150:643]
    Lead_10 = image[900:1200, 646:1135]
    Lead_11 = image[900:1200, 1140:1626]
    Lead_12 = image[900:1200, 1630:2125]
    Lead_13 = image[1250:1480, 150:2125]

    #list of leads
    Leads=
[Lead_1,Lead_2,Lead_3,Lead_4,Lead_5,Lead_6,Lead_7,Lead_8,Lead_9,Lead_10,Lead_11,Lead_12,Lead_13]

    #folder_name to store lead_images
    folder_name= re.sub('.jpg', '',image_file)

    #loop through leads and create seperate images
    for x,y in enumerate(Leads):
      fig , ax = plt.subplots()
      #fig.set_size_inches(20, 20)
      ax.imshow(y)
      ax.axis('off')
      ax.set_title("Leads {0}".format(x+1))
      if (os.path.exists(parent_folder+'/'+folder_name)):
        pass
      else:
        os.makedirs(parent_folder+'/'+folder_name)

      #save the image
      plt.close('all')
      plt.ioff()

 fig.savefig('{parent}/{folder_name}/Lead_{x}_Signal.png'.format(folder_name=folder_name,x=x+1,parent=par
ent_folder))

    extract_signal_leads(Leads,folder_name,parent_folder)
```

Annexure 2

A.2.1 Ensemble Voting Classifier Code

```
# Importing required modules
from sklearn import linear_model, tree, ensemble
from sklearn.naive_bayes import GaussianNB
import xgboost
from xgboost import XGBClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import cohen_kappa_score, matthews_corrcoef
from sklearn.metrics import roc_curve, auc, precision_recall_curve, average_precision_score
from sklearn.preprocessing import label_binarize
import pickle

# Input
X = final_result_df.iloc[:, 0:-1]
# Target
y = final_result_df.iloc[:, -1]

# Create train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)import
matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Stacking of ML Models
eclf = VotingClassifier(estimators=[
    ('SVM', SVC(probability=True)),
    ('knn', KNeighborsClassifier()),
    ('rf', ensemble.RandomForestClassifier()),
    ('bayes', GaussianNB()),
    ('logistic', LogisticRegression()),
], voting='soft')
```

A.2.2 Hyperparameter Tuning for Ensemble Code

```
# Hyperparameter Tuning using gridSearch
params = {
    'SVM__C': [1, 10, 100],
    'SVM__gamma': [0.1, 0.01],
    'knn__n_neighbors': [1, 3, 5],
    'rf__n_estimators': [300, 400],
}

grid = GridSearchCV(estimator=eclf, param_grid=params, cv=5)
voting_clf = grid.fit(X_train, y_train)

print(grid.best_params_)
y_pred = voting_clf.predict(X_test)


# Compute and print metrics
Voting_Accuracy = voting_clf.score(X_test, y_test)
print("Accuracy: {}".format(Voting_Accuracy))
print(classification_report(y_test, y_pred))
print(voting_clf.best_params_)
```

A.2.3 Model Evaluation Plot Code

```
# Compute and display the confusion matrix
# The confusion matrix is computed with the class order [0, 1, 2, 3] (AHB, MI, Normal, PMI)
cm = confusion_matrix(y_test, y_pred)
print("\nConfusion Matrix:")
print(cm)

# Visualize the confusion matrix as a heatmap
# Update labels to match the class order [0, 1, 2, 3] (AHB, MI, Normal, PMI)
plt.figure(figsize=(7, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['AHB', 'MI', 'Normal', 'PMI'],
yticklabels=['AHB', 'MI', 'Normal', 'PMI'])
plt.title('Confusion Matrix for Ensemble Model')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()

# Compute Cohen's Kappa
kappa = cohen_kappa_score(y_test, y_pred)
print("Cohen's Kappa: {:.4f}".format(kappa))

# Compute Matthews Correlation Coefficient
mcc = matthews_corrcoef(y_test, y_pred)
print("Matthews Correlation Coefficient: {:.4f}".format(mcc))

# Compute ROC Curve and AUC for each class (one-vs-rest)
# Binarize the test labels with the class order [0, 1, 2, 3] (AHB, MI, Normal, PMI)
y_test_bin = label_binarize(y_test, classes=[0, 1, 2, 3])
y_score = voting_clf.predict_proba(X_test)

fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(4):
    fpr[i], tpr[i], _ = roc_curve(y_test_bin[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])

# Plot ROC curves
# Update class_names to match the class order [0, 1, 2, 3] (AHB, MI, Normal, PMI)
plt.figure(figsize=(7, 4))
colors = ['blue', 'red', 'green', 'purple']
class_names = ['AHB', 'MI', 'Normal', 'PMI']
for i, color in enumerate(colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=2, label=f'ROC curve for {class_names[i]} (AUC =
{roc_auc[i]:.2f})')
plt.plot([0, 1], [0, 1], 'k--', lw=2)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curves for Ensemble Model (One-vs-Rest)')
plt.legend(loc="lower right")
plt.show()
```

Plagiarism Report

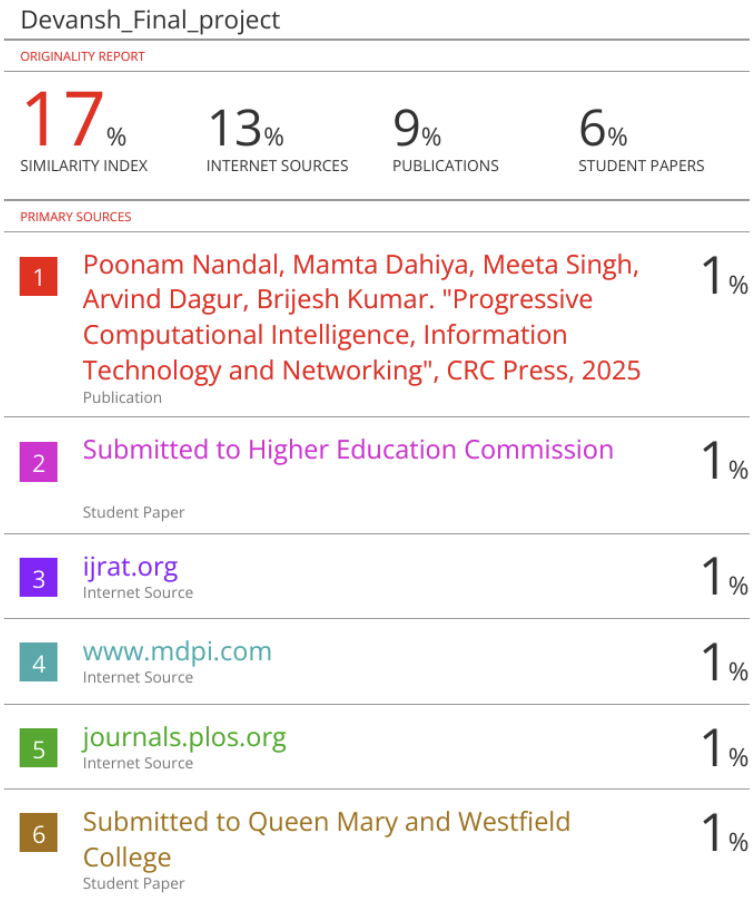

Fig Plagiarism Report