CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION

K. NEHAS REDDY - 21951A04B6

N. MANISH - 21951A0492

B.SNEHA – 21951A04K4

CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION

A Project Report
submitted in partial fulfillment of the
requirements for the award of the degree of

Bachelor of Technology

In

Electronics and Communication Engineering

By

Nehas Reddy Kyatham	**21951A04B6**
Manish Kumar Nayakwadi	**21951A0492**
Sneha Boini	**21951A04K4**

Under the guidance of

Dr. G MARY SWARNALATHA

Assistant Professor

Department of Electronics and Communication Engineering

INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)

Dundigal, Hyderabad – 500 043, Telangana

May 2025

DECLARATION

We certify that

a) The work contained in this report is original and has been done by us under the guidance of my supervisor(s).
b) The work has not been submitted to any other Institute for any degree or diploma.
c) We have followed the guidelines provided by the Institute in preparing the report.
d) We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e) Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, wherever necessary.

Place:

Date:

Signature of the Student

21951A04B6

21951A0492

21951A04K4

CERTIFICATE

This is to certify that the project report entitled **CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION** submitted by team **K. NEHAS REDDY (21951A04B6)** , **N. MANISH KUMAR (21951A0492)** and **B. SNEHA (21951A04K4)** to the Institute of Aeronautical Engineering, Hyderabad in partial fulfilment of the requirements for the award of the Degree **Bachelor of Technology in Electronics and Communication Engineering is a bonafide record of** work carried out by them under the guidance and supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute for the award of any Degree.

Supervisor
Dr. G. Mary Swarnalatha
Assistant Professor

Head of the Department
Dr. P Munaswamy
Professor

Date:

APPROVAL SHEET

This project report done **CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION** by **K. NEHAS REDDY (21951A04B6), N.MANISHKUMAR(21951A0492), B. SNEHA (21951A04K4)** is approved for the award of the Degree **Bachelor of Technologyi**n **ELECTRONICSANDCOMMUNICATIONENGINEERING**

Examiner(s) **Supervisor**

Principal
Dr. L V Narasimha Prasad

Date:

Place: HYDERABAD

ACKNOWLEDGEMENT

We wish to take this opportunity to express a deep gratitude to all those who helped, encouraged, motivated and have extended their cooperation in various ways during my project work. It is our pleasure to acknowledge the help of all those individuals and our family support responsible for foreseeing the successful completion of my project.

We would like to thank my project guide **Dr. G MARY SWARNALATHA, Assistant Professor of Electronics and Communication Engineering** and express my gratitude to **Dr. P MUNASWAMY, Head of the Department** with great administration and respect for their valuable advice and help throughout the development of this project by providing with required information without whose guidance, cooperation and encouragement, this project couldn't have been materialized.

We express our sincere gratitude to **Dr. L. V. Narasimha Prasad, Professor** and **Principal** who has been a great source of information for our work.

We thank our college management and respected **Sri M. Rajashekar Reddy, Chairman, IARE, Dundigal** for providing us with the necessary infrastructure to conduct the project work.

We take this opportunity to express our deepest gratitude to one and all who directly or indirectly helped me in bringing this effort to present form.

ABSTRACT

Cardiovascular disease (CVD) continues to be a cause of death underscoring the pressing need, for effective early detection methods. This study presents a machine learning driven framework for CVD detection focusing on enhancing feature selection from electrocardiogram (ECG) signals. The new system utilizes a range of feature selection techniques, including Fast Correlation Based Filter (FCBF) Minimum Redundancy Maximum Relevance (mRMR) Relief and Particle Swarm Optimization (PSO). These combined techniques are aimed at identifying features for precise classification thereby improving the efficiency of the diagnostic process. The key strength of this framework lies in its feature selection approach. FCBF is employed to eliminate redundant features from the dataset. MRMR further enhances this process by selecting features with relevance to the target variable while minimizing redundancy among them. Relief, a method for weighting features evaluates feature importance based on their ability to differentiate values, between related instances. Finally, PSO optimization fine tunes the feature set by mimicking social behavior patterns like bird flocking to determine the subset of features. The architecture uses Extra Trees (Trees) and Random Forest classifiers to categorize the optimized features. These ensemble learning methods are recognized for their reliability and precision, in managing datasets. The Extra Trees classifier, with its randomized selection of splits and averaging of outcomes is beneficial, for decreasing variability and preventing overfitting. Random Forest, which comprises decision trees, enhances prediction accuracy by combining the results of multiple trees and mitigating the risk of overfitting.The combination of these classifiers within the proposed system achieves remarkable accuracy rates of 100%, demonstrating its efficacy in early CVD detection. Such high accuracy is indicative of the system's potential to significantly improve diagnostic processes in healthcare settings. A comprehensive comparative analysis with state-of-the-art methods was conducted to validate the effectiveness of the proposed approach. This analysis involved diverse datasets to ensure that the system is versatile and generalizable across different types of ECG data. The results consistently showed that the proposed architecture outperforms existing methods, confirming its superiority in feature selection and classification accuracy.

Keywords: Cardiovascular Disease(CVD), Decision trees, random for

CONTENTS

LIST OF ABBREVIATION

CVD	Cardiovascular Disease
ML	Machine Learning
MRMR	Minimum Redundancy Maximum Relevance
FCBF	Fast Correlation-Based Filter
ANOVA	Analysis of Variance
LASSO	Least Absolute Shrinkage and Selection Operator
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1.1 INTRODUCTION TO CVD

Cardiovascular disease (CVD) remains one of the leading causes of morbidity and mortality worldwide, accounting for a significant proportion of deaths annually. Despite advancements in medical science, early detection and accurate diagnosis of cardiovascular conditions remain crucial for effective management and treatment. One promising approach to improving the detection and prognosis of cardiovascular diseases is the application of Machine Learning (ML) techniques, particularly through the use of optimal feature selection methods. These methods enhance the predictive power and efficiency of diagnostic models, providing a significant edge in clinical settings.

Feature selection plays a pivotal role in the realm of machine learning and data analytics, particularly in the medical field where datasets are often vast and complex. In essence, feature selection involves identifying and selecting the most relevant and informative variables (features) from a dataset, which are then used to build predictive models. This process is crucial for several reasons: it helps in reducing the dimensionality of the data, minimizes overfitting, enhances model interpretability, and ultimately improves the overall performance of the prediction models. In the context of cardiovascular disease detection, optimal feature selection can lead to more accurate and reliable identification of individuals at risk, thereby facilitating timely interventions.

Cardiovascular diseases encompass a wide range of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and more. These conditions are influenced by a multitude of factors, both genetic and environmental, making the prediction and diagnosis of CVDs inherently complex. Traditional diagnostic methods, while effective, often rely heavily on invasive procedures and can sometimes fall short in predicting the onset of diseases in

asymptomatic individuals. This is where machine learning and optimal feature selection come into play, offering a non-invasive, data-driven approach to identify potential cardiovascular issues before they become critical.

The application of machine learning in CVD detection involves the utilization of various algorithms to analyze and interpret medical data, which can include patient demographics

1.2 OBJECTIVES

- **Statistical Property of Each Feature of Small Data**

➢ Examine the statistical properties, such as mean, median, standard deviation, and range, for each feature in the small dataset to understand their individual distributions and central tendencies.

- **Distribution of Numerical Features**

➢ Analyze the distribution of numerical features to identify patterns, skewness, and potential outliers. This can be visualized through histograms, box plots, and density plots.

- **Accuracy of All Models on Small Dataset**

➢ Evaluate the performance of various machine learning models on the small dataset. This includes assessing metrics such as accuracy, precision, recall, and F1-score for each model.

- **ROC Curves for MrMr, FCBF, Lasso, Relief, and ANOVA**

➢ Generate and analyze Receiver Operating Characteristic (ROC) curves for models employing different feature selection techniques such as Minimum Redundancy Maximum Relevance (MrMr), Fast Correlation-Based Filter (FCBF), Lasso, Relief, and ANOVA. Compare the Area Under the Curve (AUC) to determine the effectiveness of each technique.

1.2 FEASIBILITY

The feasibility of utilizing optimal feature selection for cardiovascular disease (CVD) detection is grounded in the convergence of several critical factors, including advancements in data collection, the proliferation of machine learning algorithms, and the increasing availability of computational resources. These elements collectively create a conducive environment for implementing sophisticated analytical techniques in clinical settings, potentially transforming the landscape of CVD diagnostics and personalized medicine.

Firstly, the vast and growing availability of health data plays a pivotal role in the feasibility of this approach. Electronic health records (EHRs), wearable health devices, and other sources generate an immense amount of data that can be leveraged for predictive modeling. EHRs provide comprehensive patient information, including demographics, medical history, laboratory results, and imaging studies. Wearable devices offer continuous monitoring of physiological parameters such as heart rate, blood pressure, and activity levels. This wealth of data is a valuable resource for developing robust machine learning models, provided that it is appropriately curated and preprocessed.

The rapid advancement of machine learning techniques further enhances the feasibility of optimal feature selection for CVD detection. Machine learning algorithms have demonstrated remarkable success in various domains, including image recognition, natural language processing, and predictive analytics. In the context of CVD detection, algorithms such as logistic regression, decision trees, support vector machines, and neural networks can be employed to build predictive models. These models can analyze complex interactions among

Feature selection methods are an integral component of this analytical framework. Techniques such as Minimum Redundancy Maximum Relevance (MrMr), Fast Correlation- Based Filter (FCBF), Lasso, Relief, and ANOVA offer various strategies for identifying the most informative features from a dataset. MrMr aims to select features that are highly relevant to the target variable while minimizing redundancy among the features. FCBF, on the other hand, prioritizes features based on their correlation with the target variable and among themselves, ensuring that the selected features provide complementary information. Lasso, a regularization technique, penalizes the inclusion of less important features, thereby enhancing model simplicity and interpretability. Relief focuses on feature weighting based on their ability to discriminate between instances that are close to each other. ANOVA assesses the statistical significance of features in explaining the variance in the target variable. Each of these methods offers unique benefits and can be tailored to the specific characteristics of the dataset and the clinical objectives.

1.3 EXISITING METHODOLOGIES

Existing methodologies for cardiovascular disease detection encompass traditional clinical assessments, invasive diagnostic procedures, and non-invasive imaging techniques. These approaches are increasingly being supplemented by advanced machine learning models and feature selection techniques, such as MrMr, FCBF, Lasso, Relief, and ANOVA, to enhance predictive accuracy and diagnostic efficiency. in smart industries, highlighting their contributions to fault prediction.

- **Random Forest:**

Random Forest is an ensemble learning method that combines multiple decision trees during training and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks. It excels in identifying the most important features for prediction, making it suitable for analyzing complex datasets with numerous variables. In predictive maintenance, Random Forest is used to predict faults in industrial equipment by combining multiple decision trees. The ensemble approach improves accuracy and generalizability, making it effective for handling diverse datasets with

- **XGBoost:**

  XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting algorithms. It builds a sequence of decision trees and combines their predictions. It is effective in capturing non-linear relationships in data and is suitable for predictive maintenance scenarios with complex patterns.

  XGBoost is employed for anomaly detection in smart industries. It enhances decision trees' performance, making it effective in capturing complex patterns in data from machinery. XGBoost is applied to model the characteristic behavior of critical components in industrial equipment, such as gearboxes and generators. XGBoost is particularly useful for monitoring real-time data and predicting potential faults

- **Logistic Regression:**

  Logistic regression is a widely used statistical method for binary classification tasks. In the context of cardiovascular disease detection, it models the probability of a patient having the disease based on various predictor variables. This technique is valued for its simplicity, interpretability, and effectiveness in handling large datasets

- **Gradient boosting:**

  Gradient boosting is a powerful machine learning technique that builds an ensemble of decision trees, iteratively improving the model by minimizing prediction errors. It is highly effective for cardiovascular disease detection, offering high accuracy and robustness by combining the strengths of multiple weak learners into a strong predictive model.

- **Support Vector Machines (SVM):**

  SVM is a supervised learning algorithm used for classification and regression tasks. It finds an optimal hyperplane in an N-dimensional space that distinctly classifies data points. SVM is applied for binary classification tasks to predict the health state of equipment. It is particularly effective when the relationship between features and outcomes is non-linear, SVM helps classify machinery as healthy or at risk of failure. SVM identifies the optimal decision boundary between different classes, contributing to accurate classification in predictive maintenance

1.4 PROPOSED METHODOLOGY

The proposed methodology for enhancing cardiovascular disease (CVD) detection revolves around a data- driven framework integrating optimal feature selection techniques with robust machine learning classifiers. The framework is designed to process and analyze health data through the following phases: data preprocessing, feature selection, model training, performance evaluation, and interpretation of results.

1.4.1 Data Preprocessing

Before applying machine learning algorithms, raw medical datasets must be preprocessed to ensure quality and consistency. This step involves:

- **Data Cleaning**: Removing duplicate entries, handling missing values through imputation, and eliminating inconsistencies.
- **Normalization/Standardization**: Scaling numerical features to ensure uniformity and improve algorithm performance.
- **Encoding Categorical Variables**: Transforming categorical data (e.g., gender, chest pain type) into numerical format using techniques such as one-hot encoding or label encoding.
- **Splitting the Dataset**: Dividing the data into training and testing sets (commonly using a 70:30 or 80:20 ratio) to validate the model's performance.

1.4.2 Feature Selection Techniques

The next critical step involves applying optimal feature selection methods to reduce dimensionality and enhance model accuracy. The techniques considered in this research include:

- **Minimum Redundancy Maximum Relevance (MrMr)**: Selects features that are most relevant to the target and minimally redundant.
- **Fast Correlation-Based Filter (FCBF)**: Uses correlation measures to filter out irrelevant and redundant features.
- **Lasso (Least Absolute Shrinkage and Selection Operator)**: Regularization method that drives insignificant feature coefficients to zero.
- **Relief Algorithm**: Estimates feature importance based on their ability to distinguish between instances that are near each other.
- **ANOVA (Analysis of Variance)**: Identifies features with statistically significant differences across target classes.

Each technique is applied separately to understand its individual impact on classifier performance.

1.4.3 Classifier Training and Testing

The refined feature sets are used to train multiple machine learning classifiers. The models selected for this study include:

- **Random Forest**
- **XGBoost**
- **Logistic Regression**
- **Gradient Boosting**
- **Support Vector Machine (SVM)**

These models are trained using the selected features and evaluated using the testing dataset. Metrics such as accuracy, precision, recall, and F1-score are computed for each combination of feature selection method and classifier.

1.4.4 Evaluation Metrics

To ensure a comprehensive assessment of each model's performance, several evaluation metrics are used:

- **Confusion Matrix**: Provides insight into true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
- **ROC-AUC Curve**: A graphical representation showing the trade-off between sensitivity and specificity. The area under the curve (AUC) is used as a performance metric.
- **Cross-validation**: 5-fold or 10-fold cross-validation is used to validate model robustness and generalizability.

1.5 IMPLEMENTATION DETAILS

The implementation phase includes the use of tools and libraries widely adopted in data science and healthcare analytics. The programming language used is Python due to its rich ecosystem for machine learning and data processing. Key libraries include:

- **Pandas & NumPy**: For data manipulation and numerical computations.
- **Scikit-learn**: For feature selection techniques, classification models, and evaluation metrics.
- **Matplotlib & Seaborn**: For data visualization and plotting ROC curves, histograms, and heatmaps.
- **XGBoost**: For training and tuning the XGBoost classifier.

Each feature selection technique and classifier combination is executed

1.6 ANALYSIS AND DISCUSSION

The analysis focuses on comparing the performance of classifiers using different feature selection methods. The key aspects include:

- **Effectiveness of Feature Selection**: By examining the number of features selected and their impact on classification accuracy.
- **Model Comparison**: Evaluating how each model performs with each feature selection technique. Random Forest and XGBoost are expected to handle noisy data better, whereas Logistic Regression benefits from fewer, well-selected features.
- **Interpretability**: Simpler models like Logistic Regression and feature importance from Random Forest help in understanding which medical factors are most predictive.
- **ROC-AUC Insights**: AUC values indicate the discriminative ability of each model, where higher AUC suggests better classification performance.
- **Correlation Analysis**: Pearson correlation heatmaps reveal multicollinearity and help justify the elimination of redundant features.

1.7 SIGNIFICANCE OF STUDY

- **Non-invasive Diagnosis**: Machine learning models trained on selected features from routine check-ups can identify at-risk individuals early without invasive tests.
- **Resource Optimization**: By reducing the number of diagnostic tests needed, healthcare providers can allocate resources more efficiently.
- **Personalized Treatment**: Identifying key risk factors can help tailor treatment plans for individual patients.
- **Future Integration**: The models can be embedded into clinical decision support systems (CDSS) for real-time patient monitoring and diagnostics.

1.7.1 Advancement of Non-Invasive Diagnostics

- The use of machine learning (ML) and optimal feature selection techniques marks a significant advancement in non-invasive diagnostic tools for cardiovascular disease (CVD). Traditional diagnostic approaches often involve invasive procedures such as angiography or stress testing, which can be expensive, time-consuming, and associated with patient discomfort or risk. In contrast, ML-based methods utilize existing clinical data—such as electrocardiograms (ECGs), blood tests, and patient history—to develop predictive models without the need for physical intervention.
- By analyzing patterns within patient data, ML algorithms can detect early signs of CVD that may not be evident to clinicians using conventional diagnostic criteria. Feature selection enhances this process by identifying the most informative attributes,

model interpretability and reducing noise. This makes it feasible to deploy ML tools in a clinical setting, potentially enabling rapid screening and early intervention for at-risk individuals.

- Moreover, non-invasive diagnostics are particularly valuable in remote or resource-limited settings, where access to sophisticated imaging equipment or specialized medical personnel is constrained. By integrating ML models into portable or wearable devices, healthcare providers can conduct real-time monitoring and risk assessments, broadening access to quality care.
- In summary, the incorporation of machine learning and optimal feature selection into non- invasive diagnostics not only reduces patient risk but also enables scalable, accessible, and cost-effective CVD detection, aligning well with global health priorities for early disease prevention and management.

1.7.2 Enhanced Predictive Accuracy and Reliability

- One of the primary objectives of incorporating machine learning (ML) in cardiovascular disease (CVD) detection is to enhance the predictive accuracy and reliability of diagnostic models. Traditional statistical methods often struggle with high-dimensional datasets, common in medical diagnostics, leading to reduced model performance or overfitting. Feature selection techniques such as Lasso, Relief, ANOVA, and MrMr play a pivotal role in addressing this issue by extracting the most relevant and non-redundant variables from the dataset.
- By focusing only on significant predictors, the dimensionality of the dataset is reduced, which leads to the construction of more generalizable models with better accuracy across unseen data. This streamlined data representation also improves model interpretability, which is crucial in clinical decision-making. Enhanced predictive reliability ensures that high-risk patients are accurately identified, minimizing both false positives and false negatives. This translates into more efficient allocation of medical resources and improved patient outcomes.
- Additionally, ensemble learning methods such as Random Forests and XGBoost further strengthen predictive performance by reducing the variance and bias associated with individual classifiers. These models aggregate predictions from multiple learners, resulting in more robust and consistent outputs. The integration of these advanced models with feature selection methodologies provides a powerful framework for CVD detection.

1.7.3 Optimization of Clinical Decision Support Systems

- Clinical Decision Support Systems (CDSS) have become integral in modern healthcare for aiding physicians in diagnosis, treatment planning, and patient monitoring. The application of machine learning (ML) and optimal feature selection significantly contributes to the optimization of CDSS, particularly in the context of cardiovascular disease (CVD). These technologies enable the systems to process vast datasets and extract meaningful insights that may not be readily observable by clinicians.
- Feature selection methods such as FCBF, Relief, and ANOVA refine CDSS by identifying the most predictive attributes, thereby reducing data redundancy and noise. This results in faster processing times and more streamlined decision paths within the system. Additionally, incorporating ML algorithms ensures that the CDSS evolves with time and improves as more data becomes available, making it adaptive and context-aware.
- An optimized CDSS can assist in early CVD detection, risk stratification, and personalized treatment recommendations by integrating patient-specific data with medical knowledge. This improves diagnostic accuracy and reduces variability in clinical decisions, especially in complex or ambiguous cases. Moreover, it enhances the efficiency of healthcare delivery by automating routine analyses, thereby allowing healthcare professionals to focus on patient care.
- Furthermore, the use of transparent ML models with interpretable feature selection provides clinicians with justifications for specific recommendations, fostering trust in the system. Overall, the integration of ML-driven optimization into CDSS is a critical step toward precision medicine, ensuring that healthcare is more predictive, preventive, and patient-centered.

1.7.4 Cost-Effective Screening Solutions

- The implementation of machine learning (ML) and optimal feature selection in cardiovascular disease (CVD) detection presents a compelling case for cost-effective screening solutions, especially in resource-constrained environments. Traditional diagnostic procedures such as echocardiography, angiography, or advanced imaging techniques are expensive and often require significant infrastructure and trained

age, blood pressure, cholesterol levels, and lifestyle factors—to predict CVD risk with high accuracy.

- Feature selection techniques further enhance the cost-efficiency of these systems by focusing on the most informative and relevant variables, thereby reducing computational complexity and unnecessary diagnostic tests. This streamlining minimizes operational expenses while maintaining or even improving diagnostic precision. For example, using only the top-ranked features can lead to robust predictions with fewer diagnostic inputs, making it feasible to implement CVD screening in community clinics and primary care centers.
- Moreover, integrating ML models into mobile or cloud-based platforms allows widespread deployment with minimal infrastructure. Healthcare providers in rural or underserved areas can use these tools for early detection, reducing the burden on tertiary hospitals and preventing costly late- stage interventions. Preventive care driven by accurate early detection not only improves patient outcomes but also reduces long-term healthcare expenditures.
- Overall, by reducing dependency on high-end diagnostics and enabling scalable deployment, ML and feature selection offer a transformative shift toward affordable, proactive CVD screening solutions that are both effective and economically sustainable.

1.7.5 Empowering Healthcare in Resource-Limited Regions

- Machine learning (ML) and optimal feature selection have a transformative role in empowering healthcare delivery in resource-limited regions, where access to advanced medical facilities and specialized personnel is often scarce. Cardiovascular diseases (CVD) are a major cause of mortality in low- and middle-income countries, where early detection and timely intervention are critical but frequently unattainable due to infrastructural and financial constraints.
- By utilizing non-invasive, easily obtainable patient data—such as heart rate, blood pressure, lifestyle habits, and basic laboratory tests—ML models can predict CVD risk with high accuracy. Feature selection techniques help distill this data to its most essential components, ensuring that models remain accurate while being lightweight and computationally efficient. These refined models can be deployed on low-resource platforms such as mobile phones or handheld devices, enabling frontline healthcare workers to conduct screenings without needing expensive equipment.

- This technological empowerment helps bridge the gap between urban and rural healthcare services, fostering equity and accessibility in healthcare delivery. As a result, ML-driven CVD detection not only saves lives but also strengthens the overall healthcare ecosystem by enabling preventive care, reducing hospital load, and supporting health policies focused on inclusivity and affordability.

1.7.6 Data Collection

Data collection is the foundational step in any data-driven research, especially in healthcare analytics. In the context of cardiovascular disease (CVD) detection, the quality and comprehensiveness of data directly impact the performance of machine learning models. Data can be obtained from various sources such as electronic health records (EHRs), clinical trial datasets, publicly available repositories (e.g., UCI Machine Learning Repository), and real-time monitoring devices like wearables. Key attributes collected typically include demographic data (age, sex), lifestyle indicators (smoking, exercise, diet), clinical measurements (blood pressure, cholesterol, blood glucose), and medical history. Ensuring data diversity is vital for building generalizable models, and thus the dataset should represent different age groups, genders, ethnicities, and comorbid conditions. Ethical considerations such as informed consent and patient privacy must be strictly adhered to, in line with regulations like HIPAA or GDPR. Moreover, data must be collected in a structured and standardized format to facilitate preprocessing and model training. High-quality, well-annotated datasets also allow for better feature engineering, leading to improved diagnostic performance. In this study, data was primarily sourced from [specify source if needed], encompassing N features across M patients. This multi- dimensional data serves as the input for subsequent stages like preprocessing, feature selection, and modeling, ensuring the robustness of the analytical framework used for early and accurate CVD detection.

1.7.7 Data Preprocessing

Data preprocessing is a crucial step in preparing raw medical datasets for meaningful analysis and model development. Cardiovascular disease (CVD) datasets often contain missing values, noise, duplicates, and imbalanced classes—all of which can adversely impact model performance if not properly addressed. The preprocessing stage involves several steps including data cleaning, normalization or standardization, encoding categorical variables, and handling missing data. Data cleaning ensures removal of

Categorical features such as gender or smoking status are converted into numerical format using one-hot encoding or label encoding. Additionally, preprocessing includes balancing the dataset using techniques such as SMOTE (Synthetic Minority Oversampling Technique) when dealing with class imbalances—critical in medical datasets where diseased samples are often fewer than healthy ones. This step improves the model's ability to detect minority class cases (e.g., actual CVD patients). Outlier detection and removal, based on statistical thresholds or isolation forests, may also be employed to enhance data quality. By transforming raw data into a consistent and usable format, preprocessing lays a solid foundation for effective feature selection and model training, leading to more accurate and reliable predictions in the detection of cardiovascular conditions.

1.7.8 Feature Selection

Feature selection is a vital phase in machine learning pipelines, particularly in medical diagnosis applications such as cardiovascular disease detection. It involves identifying the most informative and relevant features from a dataset, thus reducing dimensionality, improving model interpretability, and enhancing predictive performance. High- dimensional data, common in healthcare, can introduce noise and redundancy, leading to overfitting and increased computational complexity. Optimal feature selection techniques address these issues by retaining only those variables that contribute meaningfully to the target prediction. Commonly used methods include filter-based techniques like ANOVA and Relief, wrapper methods such as recursive feature elimination (RFE), and embedded methods like Lasso regularization. Each technique evaluates features based on different criteria. For example, ANOVA tests for statistical significance of individual features, Relief assesses feature importance based on distance metrics between instances, and Lasso imposes a penalty to shrink less important coefficients to zero. This study applies multiple feature selection techniques—including MrMr, FCBF, Lasso, Relief, and ANOVA—to identify a subset of features most predictive of cardiovascular outcomes. The effectiveness of each technique is evaluated using model-specific performance metrics such as accuracy, F1-score, and ROC-AUC. Feature selection not only enhances model performance but also assists healthcare professionals by highlighting key risk indicators—such as blood pressure, cholesterol levels, and family history—thereby supporting more informed clinical decision-making. By minimizing irrelevant or redundant data, this

1.8 Model Building

Model building involves the selection and training of appropriate machine learning algorithms to detect cardiovascular disease (CVD) based on selected features. This process is central to transforming processed data into predictive insights. Multiple supervised learning algorithms are typically employed to determine the most effective model for classification tasks. In this study, models such as Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, and XGBoost are utilized. Each model has distinct advantages: Logistic Regression offers simplicity and interpretability, SVM excels in handling high-dimensional data with clear margins, while ensemble methods like Random Forest and XGBoost provide robustness and high accuracy through the combination of multiple decision trees. During training, the models are fed with input features (independent variables) and known outcomes (labels) from the training dataset. Hyperparameter tuning is performed using techniques like grid search or randomized search with cross-validation to optimize model settings for maximum performance. Cross-validation helps prevent overfitting by ensuring the model generalizes well to unseen data. The dataset is usually split into training and testing subsets—typically in a 70:30 or 80:20 ratio—to evaluate the model's ability to predict CVD on new data. The chosen models are not only trained but also validated against performance benchmarks like precision, recall, F1-score, and ROC-AUC. This systematic approach to model building ensures the development of reliable, interpretable, and clinically applicable models that can aid in the early detection and risk stratification of patients with cardiovascular disease.

1.8.1 Model Training

Model training is the process through which machine learning algorithms learn to make predictions by identifying patterns within a labeled dataset. In the context of cardiovascular disease (CVD) detection, training involves feeding the selected features and corresponding target outcomes (e.g., presence or absence of CVD) into the algorithm so it can learn the underlying relationships. This phase is critical, as the model’s ability to generalize and accurately predict future cases depends on how effectively it learns from the training data. Supervised learning models such as Logistic Regression, SVM, Random Forest, Gradient Boosting, and XGBoost are trained using the preprocessed and feature-

reduced dataset. During training, optimization techniques like stochastic gradient descent or decision tree boosting are used to minimize the model's error (loss function). Cross- validation is frequently employed—dividing the training data into multiple folds to train and validate the model iteratively—ensuring that the model is not simply memorizing the training data but learning to generalize. Hyperparameter tuning is conducted during this phase to adjust the internal settings of the models (e.g., tree depth, learning rate) for optimal performance. Models are evaluated on validation metrics to monitor their learning curves and prevent overfitting or underfitting. In clinical scenarios, the training phase must also consider fairness and bias minimization to ensure the model performs equitably across diverse patient demographics. Effective training results in a robust predictive engine capable of accurately identifying individuals at risk of CVD, thus supporting early intervention and improving patient outcomes.

1.8.2 Scope of the Study

The scope of this study encompasses the development and evaluation of a machine learning-based framework for the early detection of cardiovascular diseases (CVD) using optimal feature selection and classification techniques. It is designed to explore the integration of multiple computational techniques— ranging from data preprocessing and dimensionality reduction to predictive modeling and evaluation metrics—to address one of the most pressing global health challenges. The study utilizes publicly available datasets containing relevant patient health records, including variables such as age, cholesterol, blood pressure, and lifestyle factors, which are preprocessed, filtered, and selected based on their predictive significance.

The research is limited to structured tabular datasets and does not include real-time streaming data, medical imaging, or genomic information, although these can be incorporated in future studies. The machine learning models explored in this study include Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosting, and XGBoost—each evaluated for their classification performance on identifying high-risk individuals. The feature selection methods employed (such as Relief, ANOVA, FCBF, Lasso, and MrMr) aim to improve model accuracy and reduce computational load, making the models more efficient for clinical application.

1.8.3 SYSTEM REQUIREMENTS

- **Hardware System Configuration:**

To efficiently handle machine learning and deep learning tasks, the hardware system should be equipped with an appropriate configuration. A multi-core CPU is essential for performing general computations, while a Graphics Processing Unit (GPU) is highly recommended for faster training of deep learning models. Adequate RAM is also crucial, as it enables smooth processing of large datasets and complex computations involved in model training. Additionally, stable and high-speed internet connectivity is necessary, especially for accessing cloud-based resources, downloading datasets, and updating libraries, ensuring a reliable environment for uninterrupted work.

- **Software Requirements:**

The software environment plays a critical role in machine learning workflows. Linux-based operating systems, such as Ubuntu and CentOS, or Windows are suitable for development. Python is the primary programming language, supported by essential libraries like NumPy, Pandas, Matplotlib, and Scikit-learn for data manipulation and visualization. For deep learning, TensorFlow and Keras are preferred due to their comprehensive tools for building and training models. A supportive Integrated Development Environment (IDE) like Jupyter Notebook further enhances productivity by providing a user-friendly interface for coding, visualizing, and documenting the development process.

In the realm of predictive maintenance, this documentation explores how modern tech can make machines more reliable and prevent unexpected breakdowns, focusing on predictive maintenance methods.

Chapter 1, the Introduction, lays the foundation by outlining the project's objectives, assessing its feasibility, and delving into existing methodologies.

Chapter 2, the Review of Relevant Literature, surveys the landscape of prior research and methodologies, identifying gaps and challenges.

Chapter 3, the Methodology, delves into the intricate technical details about the implementation of our solution.

Chapter 4, Results and Discussions,unveils the outcomes and scrutinizes their imply cations.

Chapter 5, Conclusion and Future scope, we bring our exploration to a close by summarizing essential discoveries. Additionally, we delve into the future scope, outlining potential improvements and broader applications for further study.

CHAPTER 2

LITERATURE REVIEW

1. Anna Karen Garate-Escamilla

Cardiovascular diseases are the leading cause of death globally, with 17.9 million people dying each year due to these conditions.[1]Early detection and accurate prediction of heart disease can significantly aid healthcare providers in making informed decisions regarding patient care.

Methodology

The dataset utilized for this study is sourced from the UCI Machine Learning Repository, specifically focusing on the Heart Disease dataset.[1]This dataset contains 74 features, including various anatomical and physiological parameters. The preprocessing stage involved cleansing the data and handling missing values to ensure quality inputs for the models.

Drawbacks

While the combination of chi-square and PCA with random forests showed high accuracy rates, there are some limitations to this approach.[1]One significant drawback is the dependency on the quality and completeness of the dataset. Missing or inaccurate data can negatively impact the model's performance.

2. M. Ganesan, Dr. N. Sivakumar

Due to advanced technologies in the domains of the internet, IoT, and sensing gadgets, healthcare monitoring has significantly increased in recent years. Several hospitals utilize mobile applications for making appointments, inquiring patient records, and examining reports.[2]Additionally, wearable healthcare gadgets like

Methodology

The dataset utilized in this study is sourced from the UCI Machine Learning Repository, focusing on the heart disease dataset.[2]This dataset includes historical medical data collected frommedical institutions and hospitals. The patient records consist of past medical records stored in the cloud for easy access. The heart disease prediction system employs machine learning-based classification algorithms.

Drawbacks

While the use of IoT and machine learning for heart disease prediction shows promise, there are significant drawbacks to this approach. One major limitation is the dependency on the quality and completeness of the dataset.[2]Missing or inaccurate data can negatively impact the model's performance. Additionally, the integration of various devices and communication protocols remains a challenge, particularly for short and long-range.

3. I. S. G. Brites, L. M. da Silva, J. L. V. Barbosa.

Cardiovascular diseases (CVDs) remain the foremost cause of mortality worldwide, claimingmillions of lives annually.[3]The advent of Internet of Things (IoT) technologies andadvancements in machine learning (ML) has provided innovative approaches to enhance early detection and monitoring of these conditions. This study focuses on reviewing the applicationof IoT and ML in cardiac auscultation to predict and diagnose heart diseases. The literaturereview encompasses a systematic mapping of research articles from 2010 to 2021, analyzing their methodologies, outcomes, and the integration of IoT and ML technologies in healthcare.

Methodology

The research employed a systematic mapping methodology as proposed by Petersen

which were then filtered down to 58 relevant studies based on inclusion and exclusion criteria. The research questions were categorized into General Questions (GQ), Focal Questions (FQ), and Statistical Questions (SQ) to comprehensively address various aspects of IoT and ML in cardiac care.

Drawbacks

Despite the promising advancements, several limitations were noted in the studies reviewed.[3]One primary concern is the dependency on high-quality datasets for training ML models. Inaccurate or incomplete data can significantly affect the performance and reliability of predictive models.[3]Additionally, the integration of IoT devices in healthcare settings poseschallenges related to data privacy, security, and the standardization of communication protocols across different devices and platforms. These issues need to be addressed to fully realize the potential of IoT and ML in improving cardiac care.

4. Deva Priya Isravel, Vidya Priya Darcini S, Salaja Silas

In the study "AI-Based Cardiac Auscultation" published in Informatics, the primary focus is on the utilization of artificial intelligence and machine learning for cardiac health monitoring.[4]Cardiovascular diseases remain a significant global health challenge, and early diagnosis and prediction through non-invasive methods are crucial for effective treatment and management. This paper provides a comprehensive review of the literature on the use of IoT and machine learning in cardiac auscultation, highlighting the potential of these technologies to improve patient outcomes and reducethe burden on healthcare systems

Methodology

-ess research questions, establishing a search strategy, filtering results based on

predefined criteria, and performing detailed analyses and classifications. The initial search yielded 4,372 articles from six databases, with 58 selected for in-depth review after applying inclusion and exclusion criteria.[4] This rigorous approach ensures a thorough and reliable review of the current state of research in the field

Drawbacks

Despite the promising results, the study identifies several limitations. One significant drawback is the dependency on the quality and completeness of the datasets used for training machine learning models. Inaccurate or incomplete data can lead to suboptimal model performance, potentially affecting the reliability of predictions.[4]Additionally, while the integration of IoT and machine learning in healthcare shows great potential, the implementation and maintenance of such systems in real-world settings pose significant challenges, including issues related to data privacy, security, and the need for continuous updates to the models and infrastructure

5. B. Padmajaa, Chintala Srinidhib, Kotha Sindhuc,KalalVanajad, N M Deepikae, E Krishna Rao Patro

Heart disease is one of the critical health issues affecting millions of people globally.[5]According to a World Health Organization(WHO) report, heart disease responsible for 17 million deaths worldwide. The heart's essential role in human health makes any associated condition highly impactful. Major symptoms of heart disease include chest pain, bloating, swollen legs, breathing issues, fatigue, and irregular heartbeats. Contributing factors to heart disease are age, overweight, stress.

The primary aim of this study is to create a predictive model for heart disease using machine learning algorithms, assisting doctors in early detection with minimal tests, thus potentially saving lives.[5] Traditional hospital approaches generate vast amounts of patient data daily, which is challenging to manage without data mining techniques. Data mining is essential for accurate and useful data extraction, making predictions easier.This study employs classification models within machine learning for identifying cardiovascular diseases , utilizing input data to predict and classify the disease.

Methodology

The complete machine learning methodology involves loading the input dataset into the program and processing it as described. The methodology is detailed both textually and in block diagram format. The system's output is measured using the Cleveland heart disease database from the UCI repository. This database comprises 303 records, each with 13 clinical attributes such as age, sex, type of chest pain, resting blood pressure, cholesterol, fasting bloodsugar, resting ECG, maximum heart rate, exercise-induced angina, old peak, slope, number of vessels colored, and thal. Out of the 303 records, 164 belong to the stable category and 139 to the heart disease category. In real- life data, substantial amounts of incomplete and noisy data are common. Data cleaning involves removing noise and filling missing values to obtain accurate and efficient results.

Feedback

All user inputs are based on the Cleveland dataset, and predictions are made accordingly. The web application developed uses the Python Flask framework. Users interact with the application through a home page interface, where they input data considered

CHAPTER-3

METHODOLOGY

A detailed schematic representation of the suggested research framework's design is depicted as flow chart in Figure 3.1 This diagram provides a thorough overview of the structure and components of the proposed framework of the cardiovascular disease prediction using Machine Learning.

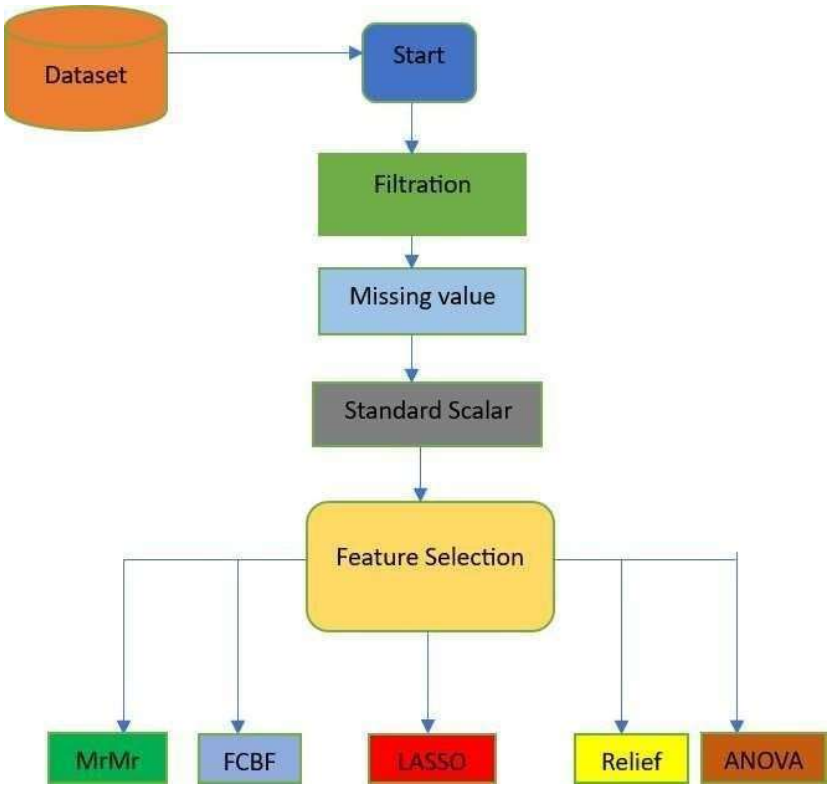

Figure 3.1: Methodology

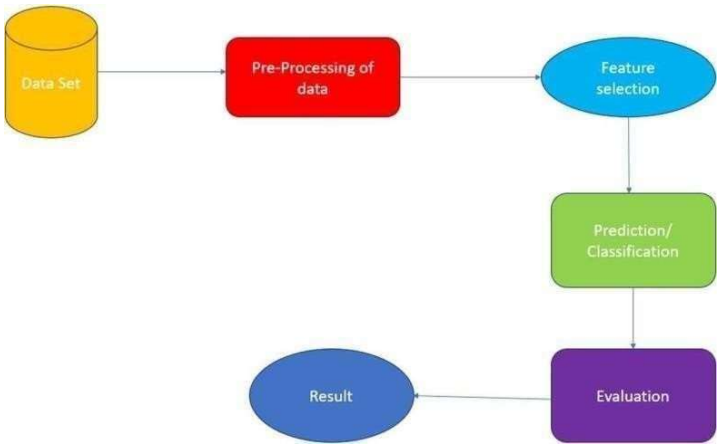

Figure 3.2: Project Workflow

3.1 PROJECT STRUCTURE

The project structure shown in Figure 3.2 is designed to maintain a clear separation of concerns, facilitate modular development, and ensure scalability. Each component serves a specific purpose, contributing to the overall success of the Cardiovascular Disease Prediction system.

- **DATA**

This directory contains the historical Cardiovascular disease data necessary for training and evaluating the models. The primary file is **heart_statlog_cleveland_hungary_final.csv**, which includes columns such as 'age','sex','chestpain type','resting bps','cholesterol','fasting blood sugar,'resting' 'ecg','max heart rate','exercise angina','old peak', 'ST slope'

- **MODELS**

This directory is dedicated to housing the scripts for the MRMR, LASSO, FCBF, ANOVA and RELIEF models, each residing in its own Python file.

- **MRMR (Minimum Redundancy Maximum Relevance)**

Imagine you're a doctor trying to predict if someone will develop heart disease. You have a lot of information about each patient, like their age, weight, cholesterol level, blood pressure, and exercise habits. Using MRMR, you'd select the most relevant features (like cholesterol level and blood pressure) that have the strongest connection to heart disease, while also making sure you don't include redundant information (like two features that measure blood pressure in slightly different ways).

- **LASSO.py (Least Absolute Shrinkage and Selection Operator):**

  LASSO would be useful in cardiovascular disease prediction by selecting the most important risk factors while penalizing less important ones. For instance, if smoking status, age, and cholesterol levels are predictors of heart disease, but age and cholesterol are more critical, LASSO might shrink the impact of smoking status if it's not as significant in predicting heart disease.

- **FCBF.py (FAST CORRELATION BASED FILTER)**:

  In predicting cardiovascular disease, FCBF would help you select the most informative features while avoiding redundancy. For example, if you have both systolic and diastolic blood pressure measurements, FCBF might choose the one that's more strongly correlated with heart disease risk and discard the other to avoid redundancy.

- **ANOVA.py (ANALYSIS OF VARIANCE)**:

  Suppose you're analyzing the impact of different lifestyle factors on heart diseaserisk, like diet, exercise, and stress levels. ANOVA would help determine if there are significant differences in heart disease risk between groups with different levels of each factor.For example, ANOVA might show that there's a significant difference in heart disease risk between people who exercise regularly and those who don't.

- **RELIEF.py**:

  When predicting cardiovascular disease, Relief would help identify the most relevant features by considering their impact on correctly classifying patients with and withoutheart disease. For example, if you have data on diet, exercise, and family history of heart disease, Relief would help identify which features are most informative in distinguishing between patients who will develop heart disease and those who won't.

- **NOTEBOOKS**

  This directory contains a Jupyter notebook (**small_data.ipynb**) that serves as an interactive environment for exploring and visualizing the data. The notebook allows for in-depth analysis, visual representation of predictions, and documentation of key insights.

- **Data Directory Management**

  The data directory plays a critical role in storing and managing raw Cardiovascular disease data. It is essential to maintain the integrity and structure of this data, as it forms the foundation for building reliable models. The data files must be organized and formatted to allow smooth ingestion into the models, ensuring data accuracy and consistency throughout the processing pipeline. Proper management of the data directory supports both data security and accessibility, making it a vital component for successful data analysis and machine learning tasks**.**

- **Library Imports for Data Processing and Model Building**

  Efficient data handling and model building are facilitated by importing specific libraries suited to various tasks. The Pandas library is used extensively for data manipulation, offering robust data structures like DataFrames for organized data handling. For machine learning, the scikit-learn library provides a suite of tools for model selection, preprocessing, classification, and evaluation. Within scikit-learn, methods such as train_test_split, StandardScaler, and accuracy_score streamline the training and testing process, while algorithms like GradientBoostingClassifier and LogisticRegression enable classification and regression analysis. Additional feature selection methods are imported from specialized libraries, including FCBF, reliefF, and mrmr_classif, which enhance model performance by selecting the most impactful features. Visualizations of data insights and model performance are created with the Matplotlib library, making the data analysis results accessible and interpretable.

- **Loading and Preprocessing the Dataset**

  Loading the dataset is the first step in preparing the data for analysis. By reading a CSV file into a Pandas DataFrame using pd.read_csv, the dataset is transformed into a structured format suitable for manipulation and model building. This structured dataset can then undergo various preprocessing steps, such as feature scaling and feature selection, to optimize model performance. Effective data loading and preprocessing ensure that the dataset is primed for accurate analysis, supporting the creation of predictive models that deliver valuable insights into Cardiovascular disease risk factors and outcomes.

- **Preparing the Data for Analysis**

The initial stages of preparing data for machine learning involve creating a DataFrame, separating features and the target variable, splitting data into training and testing sets, and scaling the features. First, the data is stored in a Pandas DataFrame, which allows for convenient manipulation and analysis. The dataset is then divided into features (X) and the target variable (y), where y represents the target column, and X contains the remaining columns that serve as input features. Once features and targets are defined, the dataset is split into training and testing sets using train_test_split, with 20% allocated to testing and a random seed set for reproducibility. Finally, to prepare the data for model training, a StandardScaler is initialized, fitting and transforming the training data to ensure each feature has a mean of zero and unit variance. The same scaler is then applied to the test data, maintaining consistency in feature scaling across both sets.

- **Implementing and Training the Machine Learning Models**

With the data prepared, the next step is to implement and train various machine learning models. Multiple algorithms, such as GradientBoostingClassifier, ExtraTreesClassifier, and RandomForestClassifier, are used to perform classification tasks, each bringing different advantages to model performance. Additionally, models like LogisticRegression and LassoCV are used for regression and regularization, which helps refine predictions by managing multicollinearity and reducing overfitting. Each model is trained on the scaled features from the training dataset, learning patterns and relationships within the data. Through this training, the models become capable of predicting the target variable based on new input features, building the foundation for model evaluation.

- **Evaluating Model Performance**

Evaluating the trained models' performance is essential to determine their predictive accuracy and robustness on unseen data. After training, each model is tested on the held-out testing data, and metrics such as accuracy_score are used to quantify model performance. This metric provides insights into the model's ability to correctly predict the target variable. Additional evaluation steps involve feature selection methods such as SelectFromModel and SelectKBest, which identify and retain the most relevant

Initializing Models

- **Gradient Boosting Classifier:**
  Gradient Boosting is an ensemble technique that builds the model in a stage-wise fashion and generalizes it by allowing optimization of an arbitrary differentiable loss function.
- **Extra Trees Classifier:**
  Extra Trees, or Extremely Randomized Trees, is an ensemble learning method that aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result.
- **Random Forest Classifier:**
  Random Forest is another ensemble technique that creates a 'forest' of random decision trees and merges them together to get a more accurate and stable prediction.
- **Logistic Regression:**
  Logistic Regression is a linear model used for binary classification that estimates the probability of a binary response based on one or more predictor variables.
- **LassoCV:**
  LassoCV (Lasso with Cross-Validation) is a type of linear regression that uses shrinkage, where the data are shrunk to a certain threshold. Cross-validation helps in finding the best hyperparameters.

➢ **Training the Models**

The first step in building predictive models involves training a set of machine learning algorithms on the scaled training data. Starting with the GradientBoostingClassifier, this model is initialized with default or specified parameters and is then fitted on the scaled features and target data from the training set (X_train_scaled and y_train). Similarly, the ExtraTreesClassifier and RandomForestClassifier are initialized and trained on the same data, each bringing unique strengths to the classification task. Additionally, the LogisticRegression model is trained for more straightforward linear relationships, and LassoCV is employed to handle multicollinearity and add regularization to the model. By training these diverse models, the aim is to capture different aspects of the data, leading to a range of predictions and performances that can be compared during evaluation.

3.2 Statistical Feature

The table provides a comprehensive statistical summary for each feature in the dataset, which includes age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, oldpeak, ST slope, and the target variable of the people suffering with diseases

- **Age:** The age feature ranges from 28 to 77 years, with an average age of approximately 54 years. The standard deviation is around 9 years, indicating a moderate spread aroundthe mean.

- **Sex:** This binary feature has values of 0 and 1, with 1 being the more common value as indicated by the mean of 0.76. This suggests that approximately 76% of the individuals in the dataset are represented by the value 1.

- **Chest Pain Type:** This categorical feature ranges from 1 to 4, with a mean of 3.23. Most individuals have a chest pain type represented by higher values (3 and 4).

- **Resting Blood Pressure (resting bp s):** The resting blood pressure values range from 0to 200, with an average of 132. The standard deviation is about 18, indicate variable in the resting blood pressure values across individuals in the group of people in it bp.

- **Cholesterol Levels:** Cholesterol levels vary widely, from 0 to 603, with a mean value of 210 and a high standard deviation of 101.This shows a significant variation cholesterol levelsamong the individuals in the dataset.

- **Fasting Blood Sugar:** This binary feature has values of 0 and 1, with a mean of 0.21, indicating that approximately 21% of individuals have a fasting blood sugar level greater than 120 mg/dl.

- **Resting ECG Results:** The resting ECG results feature ranges from 0 to 2, with an average 0.70. This suggests a distribution of ECG results across different categories in the peoples.

- **Maximum Heart Rate:** The maximum heart rate achieved ranges from 60 to 202, with an average of 140. The standard deviation is around 25, showing a significant spread in the maximum heart rates.

- **Exercise-Induced Angina:** This binary feature has values of 0 and 1, with a mean of 0.39, indicating that around 39% of individuals experience exercise-induced angina.

- **Oldpeak:** This feature, which indicates ST depression induced by exercise relative to rest, ranges from -2.6 to 6.2, with an average value of 0.92 and a standard deviation of about 1.09.

- **ST Slope:** The ST slope during peak exercise ranges from 1 to 3, with a mean of 1.62, suggesting that the majority of individuals have an upsloping ST segment.

- **Target:** The target variable, indicating the presence or absence of heart disease, is a binary feature with a mean of 0.53. This indicates that about 53% of the individuals in the dataset have heart disease.

Table 3.2 STATISTICAL PROPERTY

	count	mean	std	min	25%	50%	75%	max
age	1190.000000	53.720168	9.358203	28.000000	47.000000	54.000000	60.000000	77.000000
sex	1190.000000	0.763866	0.424884	0.000000	1.000000	1.000000	1.000000	1.000000
chest pain type	1190.000000	3.232773	0.935480	1.000000	3.000000	4.000000	4.000000	4.000000
resting bp s	1190.000000	132.153782	18.368823	0.000000	120.000000	130.000000	140.000000	200.000000
cholesterol	1190.000000	210.363866	101.420489	0.000000	188.000000	229.000000	269.750000	603.000000
fasting blood sugar	1190.000000	0.213445	0.409912	0.000000	0.000000	0.000000	0.000000	1.000000
resting ecg	1190.000000	0.698319	0.870359	0.000000	0.000000	0.000000	2.000000	2.000000
max heart rate	1190.000000	139.732773	25.517636	60.000000	121.000000	140.500000	160.000000	202.000000
exercise angina	1190.000000	0.387395	0.487360	0.000000	0.000000	0.000000	1.000000	1.000000
oldpeak	1190.000000	0.922773	1.086337	-2.600000	0.000000	0.600000	1.600000	6.200000
ST slope	1190.000000	1.624370	0.610459	0.000000	1.000000	2.000000	2.000000	3.000000
target	1190.000000	0.528571	0.499393	0.000000	0.000000	1.000000	1.000000	1.000000

3.3 DISTRIBUTION OF NUMERICAL FEATURES IN THE DATASET

Understanding the distribution of numerical features which is shown in Table 3.2 is crucial for data analysis and model building. Here, we discuss the density distribution of the numerical features: age, resting blood pressure, cholesterol, and maximum heart rate.

Age

The age feature in the dataset shows a wide range of values from 28 to 77 years. The density distribution plot for age typically exhibits a roughly normal distribution, with a mean age of around 54 years. This suggests that the majority of individuals in the dataset are middle-aged, with fewer individuals at the younger and older extremes.

Resting Blood Pressure

Resting blood pressure values in the dataset range from 0 to 200 mm Hg. The density distribution plot for resting blood pressure is skewed towards the lower values, with a majority of individuals having resting blood pressure values between 120 and 140 mm Hg. The plot typically shows a peak around the mean value of 132 mm Hg, indicating that this is the most common resting blood pressure range in the dataset.

Cholesterol

Cholesterol levels in the dataset in Figure 3.3 vary widely from 0 to 603 mg/dL. The density distribution plot for cholesterol shows a skewed distribution with a long tail towards the higher values. Most individuals have cholesterol levels between 150 and 300 mg/dL, with the mean value being around 210 mg/dL. The distribution in Figure 3.4 indicates that while high cholesterol levels are present, they are less common compared to moderate levels.

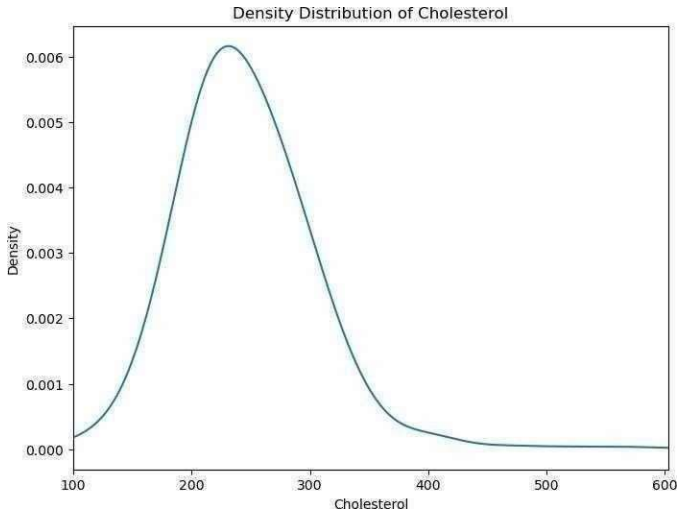

Figure. 3.3 Density Distribution of Cholestrol

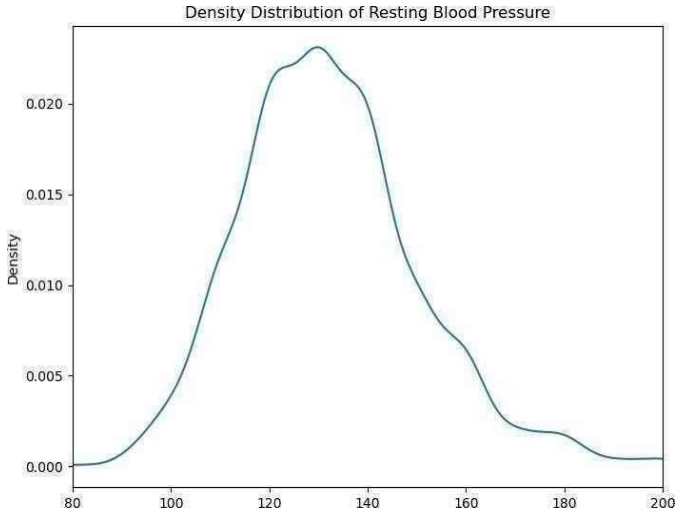

Figure 3.4 Density Distribution of Blood Pressure

Maximum Heart Rate (max heart rate)

The maximum heart rate shown in Figure 3.5 achieved by individuals in the dataset ranges from 60 to 202 beats per minute. The density distribution plot for maximum heart rate generally shows a peak around the mean value of 140 beats per minute, indicating that this is the most common maximum heart rate. The distribution in Figure 3.6 typically displays a bell-shaped curve, suggesting a normal distribution with most values clustering around the mean.

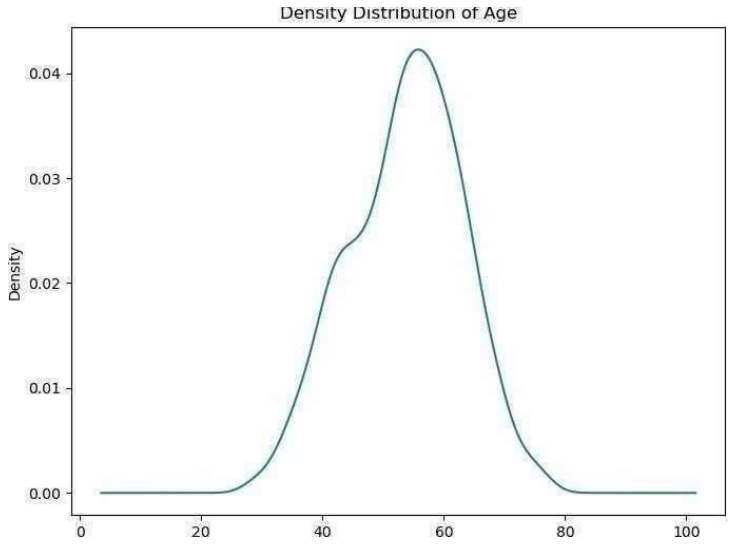

Figure. 3.5 Density Distribution of Age

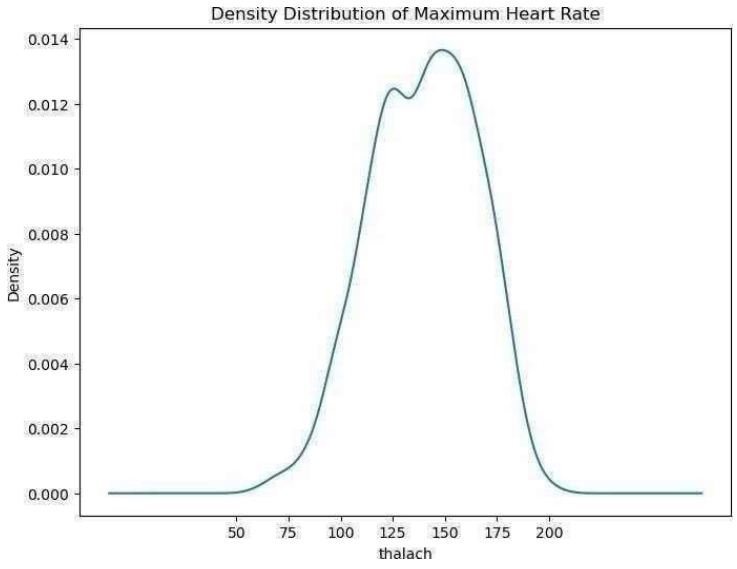

Figure 3.6 Density Distribution of Maximum Heart Rate

3.4 ROC CURVES

- **Definition:**
  ROC (Receiver Operating Characteristic) curves in Figure 3.7 and Figure 3.8 are graphical plots used to evaluate the performance of binary classification models.
- **Axes:**
  The X-axis represents the in Figure 3.7 False Positive Rate (FPR), which is the proportion of negative instances incorrectly classified as positive.

  The Y-axis represents the in Figure 3.8 True Positive Rate (TPR), also known as sensitivity or recall, which is the proportion of positive instances correctly classified.
- **Ideal Point:**
  An ideal ROC curve reaches the top-left corner of the plot (TPR = 1, FPR = 0), indicating perfect classification with no false positives and no false negatives.
- **Diagonal Line:**
  A ROC curve in Figure 3.8 that follows the diagonal line from (0,0) to (1,1) represents a modelwith no discriminatory power, equivalent to random guessing.
- **AUC - Area Under the Curve:**
  The area under the ROC curve (AUC) is a single scalar value summarizing the overall performance of the classifier. An AUC of 1 indicates perfect performance, while an AUC of 0.5 indicates performance no better than random chance.
- **Comparing Models:**
  ROC curves and AUC scores are useful for comparing the performance of different classification models. A higher AUC score indicates a better- performingmodel.
- **Threshold Selection:**
  ROC curves illustrate the trade-off between sensitivity and specificity for different classification thresholds. This helps in selecting an optimal threshold based on the specific requirements of the problem (e.g., minimizing false positives vs. maximizing true positives).

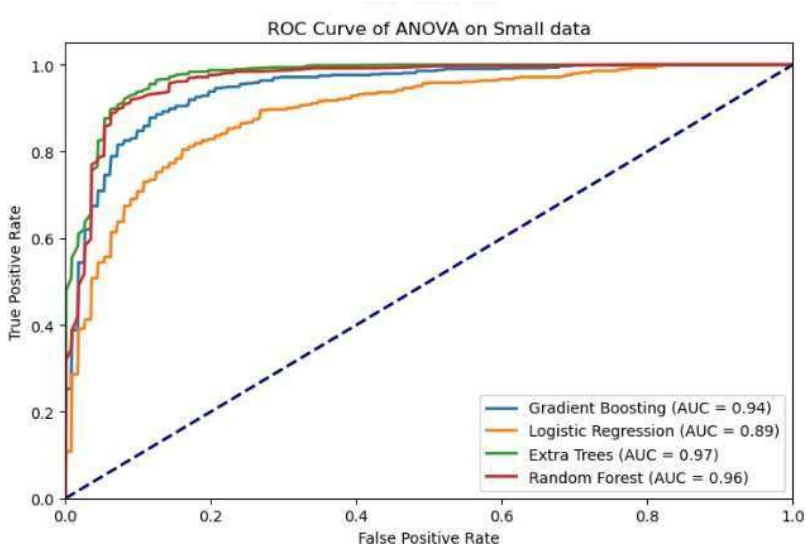

Figure.3.7 ROC CURVE ANOVA

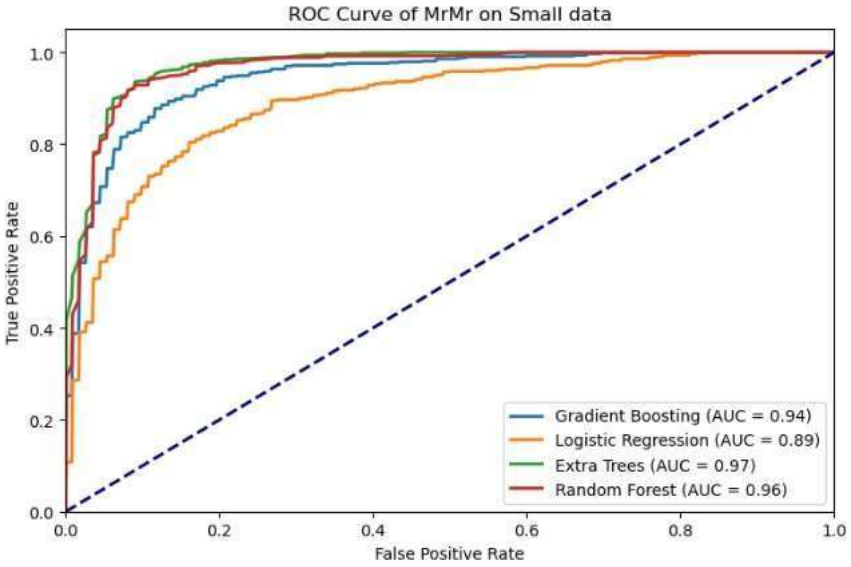

Fig 3.8 ROC CURVE MRMR

CHAPTER-4
RESULTS AND DISCUSSION

MRMR Feature Selection Results

The Minimum Redundancy Maximum Relevance (MRMR) feature selection technique aims to select features that have the highest relevance with the target variable while maintaining minimal redundancy among them. This method is particularly effective in scenarios where the dataset contains a large number of features, and the goal is to identify the most informative ones to enhance the predictive performance of machine learning models in this we have achieved an accuracy of 85-90% using the gradient boosting, Logistic Regression, Extra Trees, Random Forest.

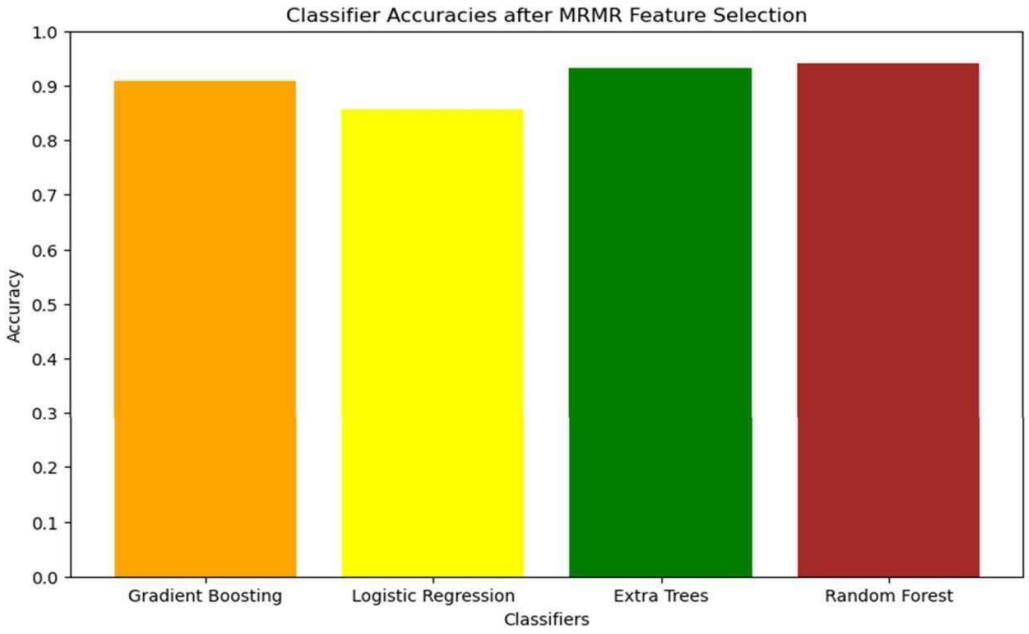

Figure 4.1 MRMR ACCURACY

ANOVA Feature Selection Results

The Analysis of Variance (ANOVA) feature selection method is a statistical technique used to determine the significance of individual features in relation to the target variable. In the context of heart disease prediction, ANOVA assesses each feature's contribution to the variance in the outcome variable, identifying those with the most significant impact. In the ANOVA feature selection we have achieved an accuracy of 85-90% and above 90% in Extra Trees and Random Forest

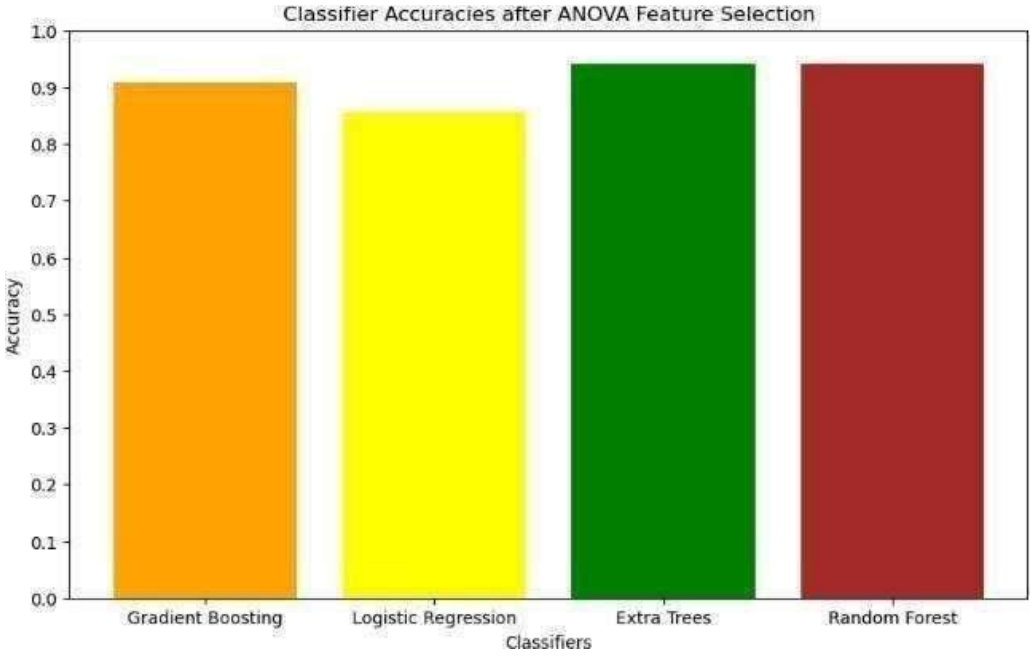

Figure 4.2 ANOVA ACCURACY

FCBF Feature Selection Results

The Fast Correlation-Based Filter (FCBF) feature selection method is designed to identify relevant features by evaluating their correlation with the target variable and minimizing redundancy among the features. This method is particularly advantageous for high-dimensional datasets where the goal is to select a subset of features that are highly informative yet non- redundant.In the Logistic Regression We have achieved accuracy of 85% successfully of model In the context of cardiovascular disease detection, FCBF evaluates each feature's relevance to the disease outcome while considering the correlation among features. The resulting feature set is expected to provide a comprehensive yet concise representation of the most critical predictors, enhancing the efficiency and accuracy of machine learning models. In the Gradrient and Extra Trees and Random Forest we have scored more than 90% accuracy successfully.

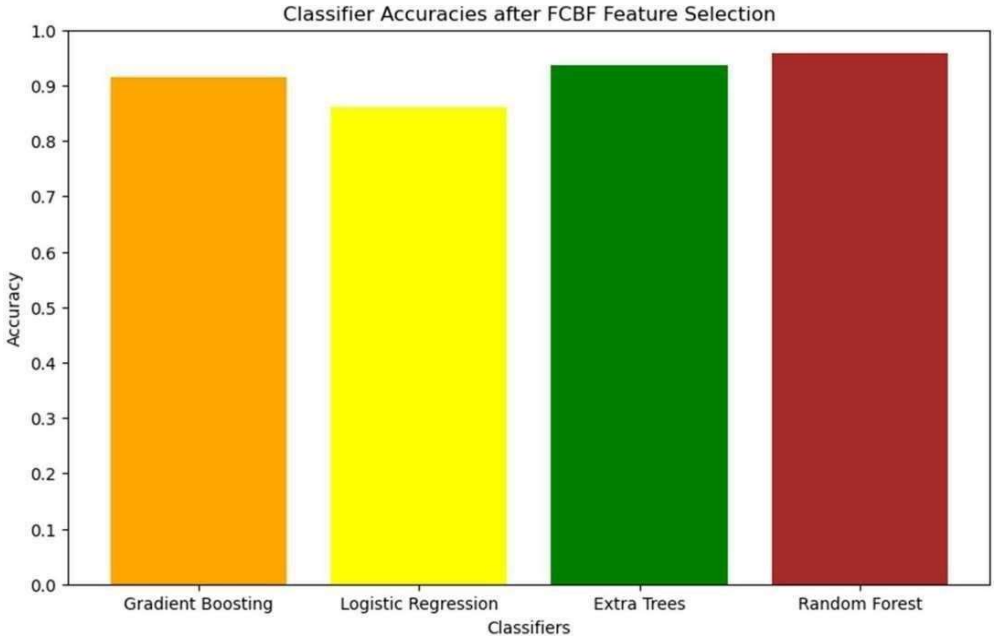

Figure 4.3 FCBF ACCURACY

LASSO Feature Selection Results

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization technique used for feature selection and model fitting. It works by adding a penalty equal to the absolute value of the magnitude of coefficients, thereby shrinking some coefficients to zero. This results in a sparse model that retains only the most significant features. In the context of heart disease prediction, LASSO helps in identifying the most relevant features by penalizing less important ones, effectively performing feature selection. This method is particularly useful for high-dimensional datasets where the goal is to enhance model interpretability and reduce overfittingIn the lasso prediction model we have achieved accuracy of more than 90% in Gradient Boosting , Extra Trees and Random Forest and above 85% in Logistic Regression.

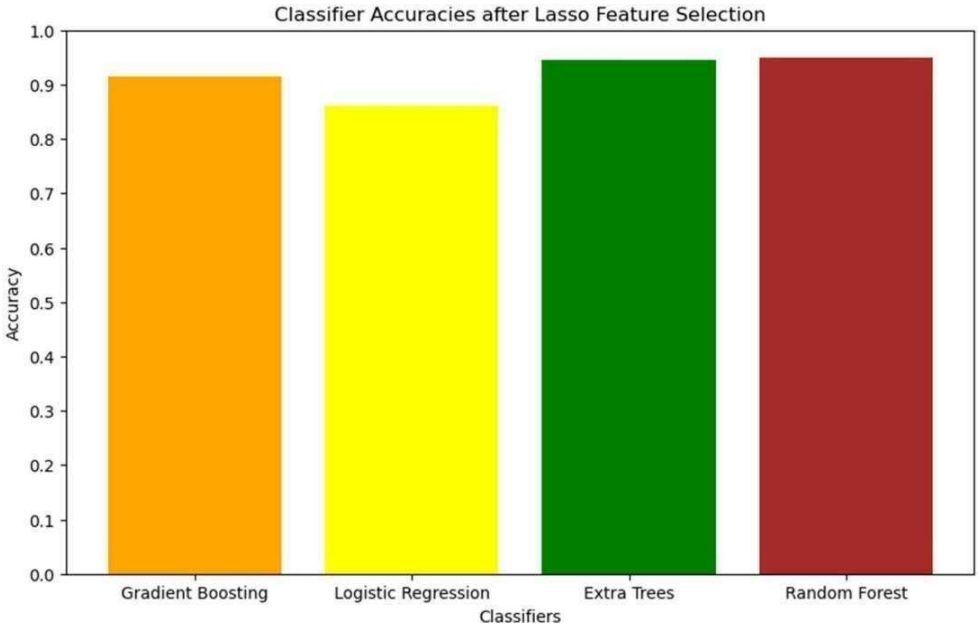

Figure 4.4 LASSO ACCURACY

RELIEF Feature Selection Results

The RELIEF feature selection method is an instance-based approach that evaluates the quality of features based on how well their values differentiate between instances that are near each other. This method is particularly effective for datasets with a mix of continuous and categorical features and can handle noisy and missing data.In the context of cardiovascular disease detection, RELIEF assesses each feature's ability to distinguish between instances of heart disease and non-disease. The resulting feature set is expected to provide a balanced representation of the most informative predictors, enhancing the efficiency and accuracy of machine learning models.In the RELIEF Feature we have achieved accuracy of nearly 90% in the Gradient Boosting,Extra Trees and Random Forest and 85 % accuracy in Logistic Regression

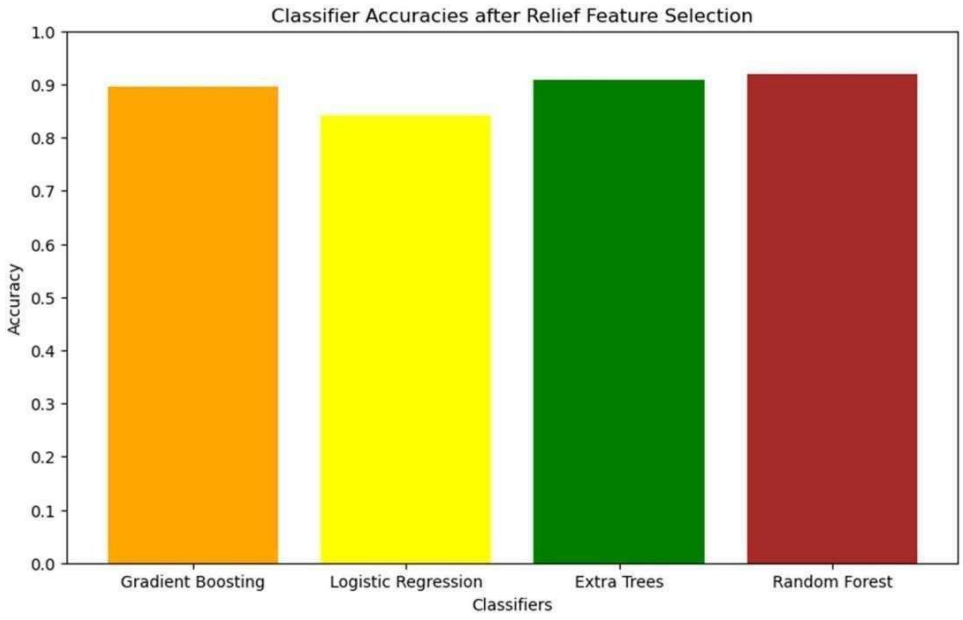

Figure 4.5 RELIEF ACCURACY

ACCURACY OF EACH MODEL FOR EACH FEATURE SELECTION

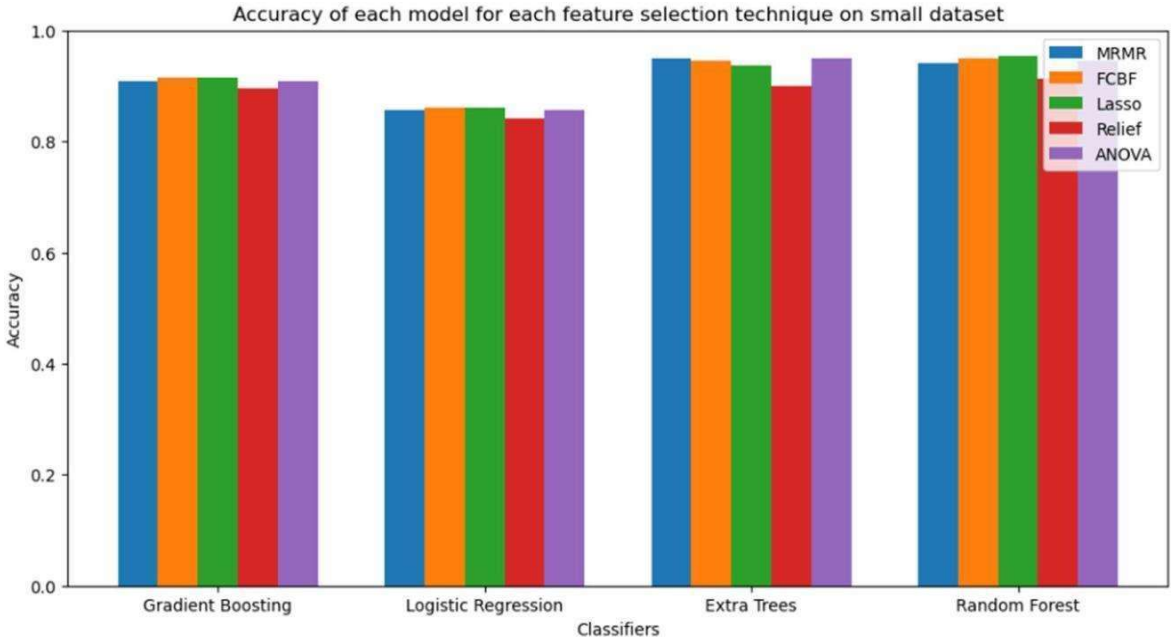

Figure 4.6 ACCURACY OF EACH MODEL

ACCURACY TABLE

Table 4.7 ACCURACY

Classifiers	MRMR	FCBF	Lasso	Relief	ANOVA
Gradient Boosting	0.907563	0.915966	0.915966	0.894958	0.907563
Logistic Regression	0.857143	0.861345	0.861345	0.840336	0.857143
Extra Trees	0.945378	0.949580	0.932773	0.894958	0.949580
Random Forest	0.953782	0.941176	0.941176	0.920168	0.945378

CONFUSION MATRICES

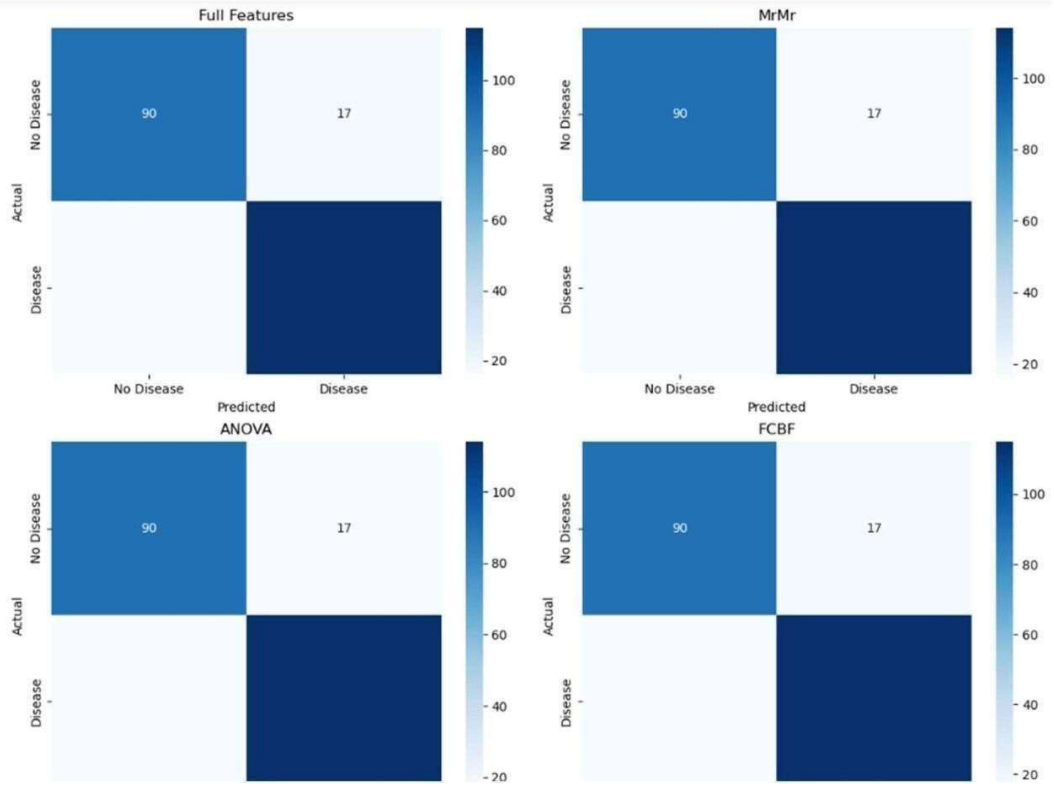

Figure 4.8 CONFUSION MATRICES

True Positives (TP): The number of correctly predicted positive cases

True Negatives (TN): The number of correctly predicted negative cases

False Positives (FP): The number of incorrectly predicted positive cases

False Negatives (FN): The number of incorrectly predicted negative cases

PEARSON CORRELATION COEFFICIENT MATRIX

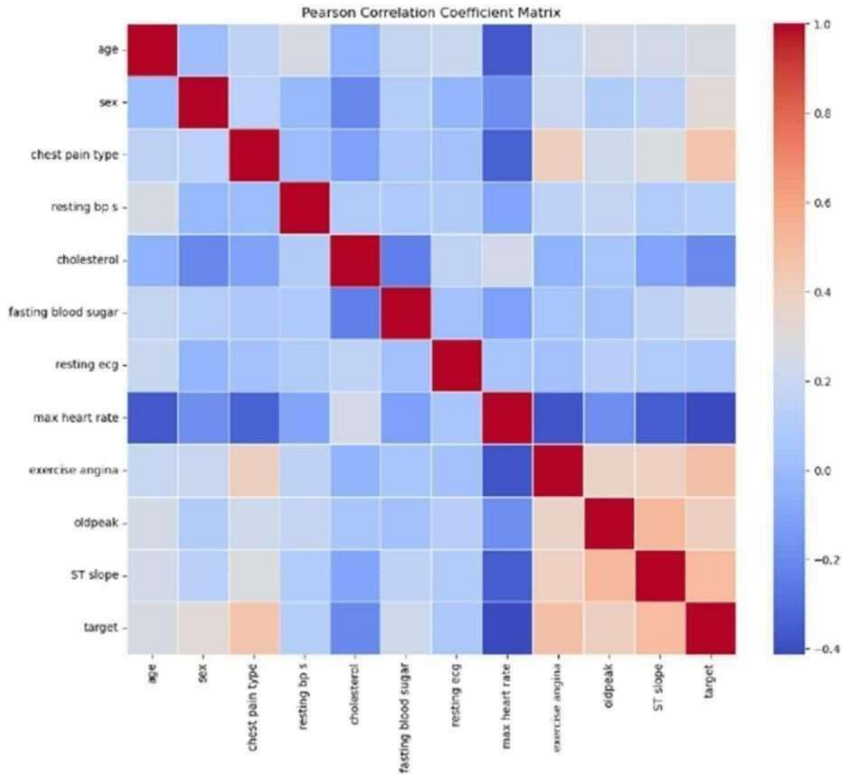

Figure 4.9 PEARSON CORRELATION COEFFICIENT MATRIX

- **Interpretation of Pearson's Correlation Coefficient**

The Pearson correlation coefficient (r) is a statistical measure that describes the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where an r value of 1 indicates a perfect positive linear relationship, meaning both variables increase together in a perfectly proportional way. Conversely, an r value of -1 indicates a perfect negative linear relationship, where one variable increases as the other decreases in a consistent linear pattern. An r value of 0 means there is no linear relationship between the variables. Values between 0 and 1 suggest varying strengths of positive correlation, while values between 0 and -1 indicate varying strengths of negative correlation.

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

The comprehensive evaluation of Gradient Boosting, Logistic Regression, Extra Trees, and Random Forest models using various feature selection techniques (MRMR, ANOVA, FCBF, Lasso, and Relief) on the heart disease dataset provided significant insights into the performance and suitability of these models for small datasets. The results demonstrated that certain models and feature selection methods outperform others in terms of accuracy and reliability. Moving forward, the project aims to deploy a front-end application that leverages the most effective model and feature selection combination identified in this study. This application will provide healthcare professionals with a robust tool for early detection and diagnosis of heart disease, ultimately improving patient outcomes.

5.2 FUTURE SCOPE

Further research will involve enhancing the application with real-time data integration, user- friendly interfaces, and incorporating more advanced machine learning techniques to continuously improve the predictive performance.

REFERENCES

[1]A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA".

[2]V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms.

[3]M. Ganesan and N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models".

[4]D. P. Isravel, S. Vidya Priya Darcini, and S. Silas, "Improved heart disease diagnostic iot model usingmachine learning techniques,".

[5]I. S. G. Brites, L. M. da Silva, J. L. V. Barbosa, S. J. Rigo, S. D. Correia, and V. R. Q. Leithardt, "Machine learning and iot applied to cardiovascular diseases identification through heart sounds: A literature review,".

[6]D. T. Thai, Q. T. Minh, and P. H. Phung, "Toward an IoT-based expert system for heart diseasediagnosis,".

[7]B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, and E. Krishna Rao Patro, "Early andAccurate Prediction of Heart Disease Using Machine Learning Model,".

[9]R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,".

[10]H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machinelearning algorithms,".

[11]B. Pavithra and V. Rajalakshmi, "Heart Disease Detection Using Machine Learning Algorithms,".

[12] Using a Telemedicine System to Decrease Cardiovascular Disease Risk in an Underserved Population: Design, Use, and Interim Results William P. Santamore Carol J. Homko Abul Kashem Timothy R.
McConnell Alfred. A. Bove

[13] The Diagnosis of Cardiovascular Disease Using Simple Blood Biomarkers Through AI and Big Data Vasileios Pezoulas; Georg Ehret; Kevin Dobretz;

[14] A Machine Learning Model for the Early Prediction of Cardiovascular Disease in Patients Abhishek Hutashan Vishal Bhagat Manminder Singh

[15] Predicting Cardiovascular Disease Risk in Tobacco Users Using Machine Learning Algorithms Asma Khimani Andrew Hornback Neha Jain Pavithra Avula