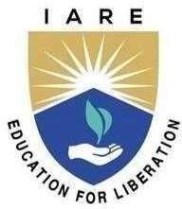
**CARDIOVASCULARDISEASEDETECTIONUSING**

**OPTIMAL FEATURE SELECTION**

***K. NEHAS REDDY 21951A04B6***

**CARDIOVASCULARDISEASEDETECTIONUSING**

**OPTIMAL FEATURE SELECTION**

***A Project Report***

***submitted in partial fulfillment of the requirements fortheaward ofthe degree of***

**BACHELOROF TECHNOLOGY**

**IN ELECTRONICSANDCOMMUNICATIONENGINEERING**

***by***

**K. NEHAS REDDY – 21951A04B6**

***Under the guidance of***

**Dr.GMARYSWARNALATHA AssistantProfessor**

**Departmentof Electronics and Communication Engineering**

**INSTITUTE OF AERONAUTICAL ENGINEERING (Autonomous)**

**Dundigal,Hyderabad – 500 043,Telangana**

**May, 2025**

**©2024, K. Nehas Reddy, All Rights Reserved**

I

**DECLARATION**

Icertifythat

a) The work contained in this report is original and has been done by me under the guidance of my supervisor(s).

b) The work has not been submitted to any other Institute for any degree or diploma.

c) I have followed the guidelines provided by the Institute in preparing the report.

d) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e) Whenever I have used materials (data, theoretical analysis, figures, and text) from othersources, Ihave given due credit tothemby citing them in thetext ofthereport andgiving theirdetails in thereferences. Further, I have taken permission from the copyright owners of the sources, wherever necessary.

**Place: HYDERABAD**

**Signatureofthe Student**

**21951A04B6**

**Date:**

II

**CERTIFICATE**

This is to certify that the project report entitled **CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE** **SELECTION** submitted by **K.NEHAS REDDY** to the Institute of Aeronautical Engineering, Hyderabad in partial fulfilment of the requirements for the award of the Degree Bachelor of Technology in **ECE** is a bonafide record of work carried out by him under my 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:**

III

**APPROVAL SHEET**

This project report entitled **CARDIOVASCULAR DISEASE DETECTION USING**

**OPTIMAL FEATURE SELECTION** by **K.NEHAS REDDY (21951A04B6),** is approved for the award of the Degree **Bachelor** **of** **Technology** in **ELECTRONICS AND COMMUNICATION ENGINEERING.**

**Examiner(s)** **Supervisor**

**Dr.G. Mary Swarnalatha Assistant Professor**

**Principal**

**Dr.L VNarasimha Prasad**

**Date:**

**Place:HYDERABAD**

IV

**ACKNOWLEDGEMENT**

I 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 my pleasure to acknowledge the help of all those individuals

and my family support responsible for foreseeing the successful completion of my project.

I would like to thank my project guide **Dr. G MARY SWARNALATHA, Assistant Professor of ECE** 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.

Iexpress my sinceregratitude to **Dr.L.V.Narasimha Prasad,Professor**and **Principal**

who has been a great source of information for my work.

I thank our college management and respected **Sri M. Rajashekar Reddy, Chairman,**

**IARE, Dundigal** for providing me with the necessary infrastructure to conduct the project work.

I take this opportunity to express my deepest gratitude to one and all who directly or

indirectly helped me in bringing this effort to present form.

V

**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: CardiovascularDisease(CVD), Decision trees, random forest.**

VI

**CONTENTS**

**Table of Contents** **Page** TitlePage I Declaration II Certificate III Approval Sheet IV Acknowledgement V Abstract VI Contents VII List of Figures VIII List of Tables IX Chapter 1 - Introduction 1

1.1Introduction 1 1.2Objectives 2

1.3Feasibility 3 1.4Existing Methodologies 4 1.5System Requirements 6

Chapter 2 - Review of Relevant Literature 8 Chapter 3 – Methodology 15

3.1Project Structure 15 3.2 Statistical Features 21 3.3 Numerical Distribution 23 3.4 ROC curves 26

Chapter 4 - Results and Discussions 28 4.1MRMR Model Results 28 4.2 LASSO Model Results 29 4.3 FCBF Model Results 30

4.4 ANOVA Model Results 31 4.5RELIEF Model Results 32 4.6Accuracy of Each Model 33 4.7Accuracy Table 33 4.8Confusion Matrix Table 34 4.9Pearson Correlation and Confusion Matrix 35

Chapter 5 - Conclusions andFuture Scope 36

5.1 Conclusion 36 5.2Future Scope 36 References 37

VII

**LISTOF FIGURES**

**FigureNo.** **FigureName**

3.1 METHODOLOGY

**PageNo.**

14

3.2 PROJECT WORKFLOW 15

3.3 STATISTICAL PROPERTIES 23

3.4 DENSITY DISTRIBUTION 24-25

3.5 ROC CURVES 27

4.1 MRMR 28

4.2 ANOVA 29 4.3 FCBF 30 4.4 LASSO 31 4.5 RELIEF 32 4.6 ACCURACY OFEACH MODEL 33 4.7 ACCURACY TABLE 33

4.8 CONFUSIONMATRICES 34 4.9 PEARSON CORRELATION COEFFICIENT MATRIX 35

VIII

**LISTOF TABLES**

**TableNo.** 1

2

3

**TableName** STATISTICSTABLE

ACCURACY TABLE

CONFUSION MATRIX TABLE

**PageNo.** 23

33

34

IX

**CHAPTER-1**

**INTRODUCTION**

**1.1 INTRODUCTION**

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

**1.2 OBJECTIVES**

• **StatisticalProperty ofEachFeature ofSmallData**

➢ 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.

•**DistributionofNumerical Features**

➢ Analyze the distribution of numerical features to identify patterns, skewness,

and potential outliers. This can bevisualized through histograms, box plots, and

density plots.

•**Accuracy ofAll Models onSmall 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.

•**ROCCurves forMrMr, 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.

•**Overall Results of All Classifiers with Confusion Matrix**

➢ Compile and present the overall results of all classifiers using confusion

matrices. This will help in understanding the true positives, false positives, true

negatives, and false negatives for each model.

2

•**Accuracy ofEach Model on EachSelection Technique**

➢ Compare the accuracy of each model when different feature selection techniques

are applied. This involves analyzing how each method impacts the model’s

predictive performance.

•**Pearson Correlation Between All the Features**

➢ Calculate the Pearson correlation coefficients between all pairs of features to

assess the degree of linear correlation. This can help in identifying redundant

features and understanding feature interdependencies.

**1.3 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 3

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 theirability to discriminatebetween instances that areclose to eachother. 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.

**1.4 EXISITINGMETHODOLOGIES**

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 various features.

Random Forest identifies the most important features contributing to machine states or

faults. This information aids in understanding the critical factors leading to specific

predictions, providing insights into the health of industrial equipment. Random Forest is 4

➢**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

5

1.5 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.

1.5.1 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.

1.5.2 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.

6

Each technique is applied separately to understand its individual impact on classifier performance. 1.5.3 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.

1.5.4 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.

1.6 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 on the small dataset, followed by visualization and comparative analysis.

7

1.7 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.

1.8 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.

1.8.1 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, improving

8

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.

1.8.2 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.

 Ultimately, the enhancement of predictive accuracy and reliability directly contributes to improved healthcare delivery, as it supports evidence-based decisions, early interventions,

9

and targeted treatment plans, all of which are essential in managing and preventing cardiovascular events.

1.8.3 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.

1.8.4 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 specialists. In contrast, ML-based systems leverage readily available patient data—such as

10

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.

1.8.5 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.

 Furthermore, the democratization of ML tools through open-source platforms and cloud integration makes it possible to scale diagnostic capabilities across remote and rural

11

regions. Community health programs can leverage these tools to perform mass screenings and prioritize high-risk individuals for further medical attention.

 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.

1.8.6 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.

1.8.7 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

12

erroneous or redundant entries. Missing values can be imputed using statistical techniques such as mean, median, or more sophisticated methods like K-Nearest Neighbors (KNN) imputation. Standardization (zero mean and unit variance) or normalization (scaling to a range) is applied to ensure uniform feature contributions, especially important for algorithms sensitive to feature scales like SVM and KNN. 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.

1.8.8 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

13

supporting more informed clinical decision-making. By minimizing irrelevant or redundant data, this step also reduces model complexity and training time, making real-time or embedded deployment in medical devices more feasible.

1.9 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.

1.9.1 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-

14

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.

1.9.2 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.

15

**1.9.3 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, stableand high-speed internet connectivity is necessary, especially foraccessing 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**.**

16

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.

**Chapter4**, 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.

17

**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

18

**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

19

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**

Themethodology employed in this study is asystematicmapping of existing content

Which is the main following the framework proposed by Petersen et al. The proc 20

-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 forin-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,Kalali Vanajad, 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.

21

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 as

22

dataset values. The process includes splitting, preprocessing, training, and testing the

input data. After submitting the inputs, the machine learning model processes the data

internally, and the final prediction is displayed, indicating whether the user is suffering

from heart disease. This application’s future scope includes extending its capability to

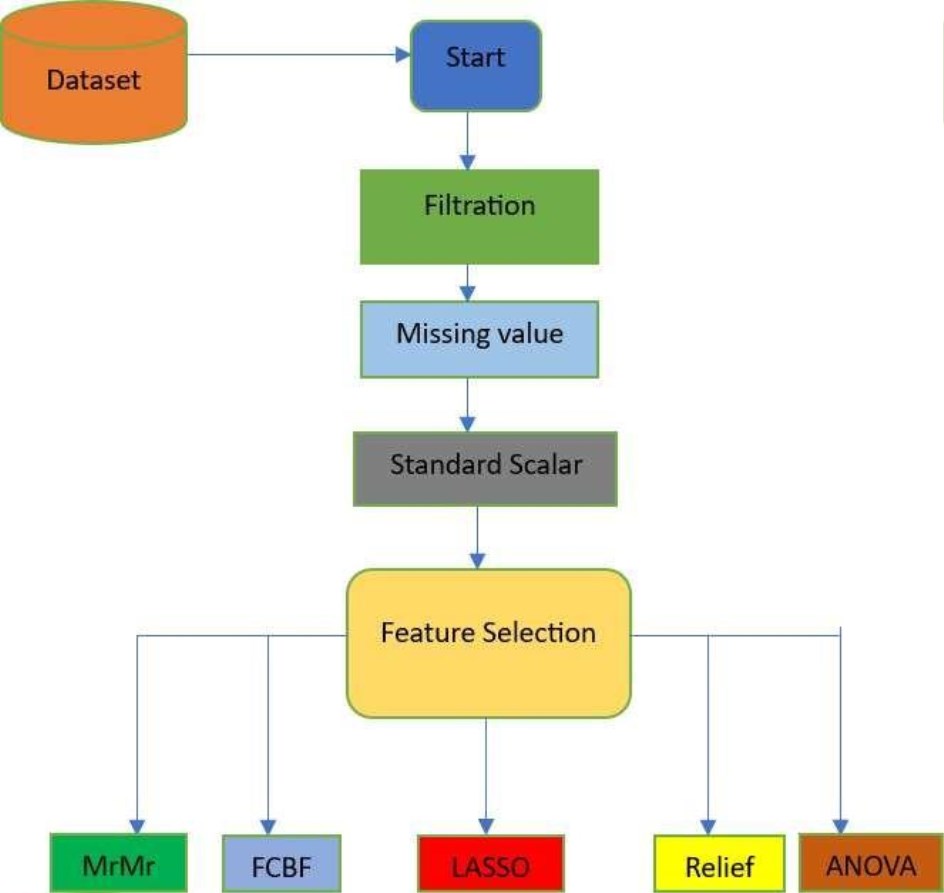
predict other diseaseslike kidney and lung-related diseases by incorporating additional

attributes. The aim is to evolve this application into a common platform for

predicting various diseases, making it a comprehensive disease prediction

tool.

23

**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

24

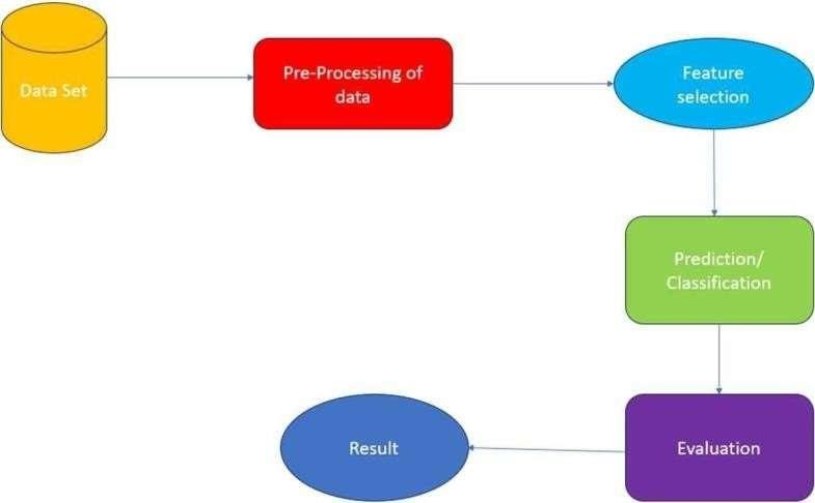


Figure 3.2:ProjectWorkflow

**3.1** **PROJECT STRUCTURE**

The project structure 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 thehistorical Cardiovascular disease data necessary fortraining 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(MinimumRedundancyMaximum 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).

25

➢**LASSO.py (Least AbsoluteShrinkage 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.

26

➢**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. ThePandas library is used extensively fordata 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 likeGradientBoostingClassifier 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.

27

➢**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 features, potentially improving the

model’s predictive accuracy**.**

28

➢**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.

29

**3.2 Statistical Feature**

Thetableprovides acomprehensive statistical summary foreach featurein thedataset,

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.

30

• **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.

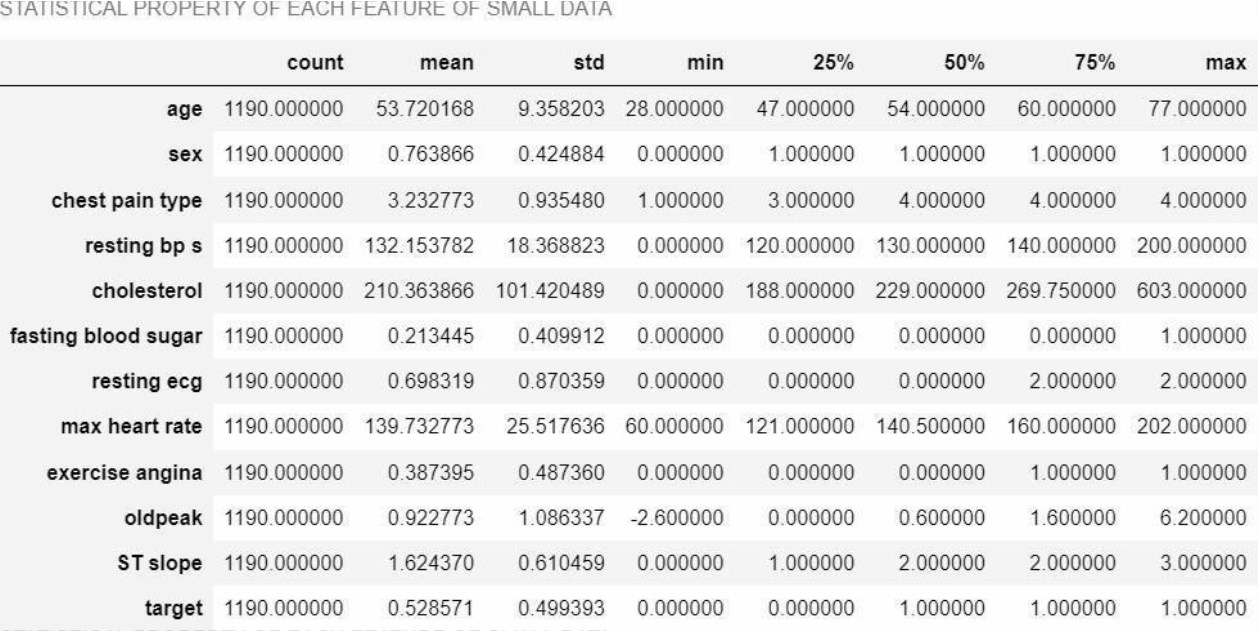
• **Oldpeak:** This feature, which indicates ST depression induced by exercise relative to rest, ranges from -2.6 to 6.2, with an average value of 0.92 and a standard deviation of about 1.09.

• **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.

31

Fig. 3.2STATISTICAL PROPERTY TABLE



**3.3** **DISTRIBUTIONOF NUMERICAL FEATURESINTHEDATASET**

Understanding the distribution of numerical featuresiscrucial for data analysisand 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.

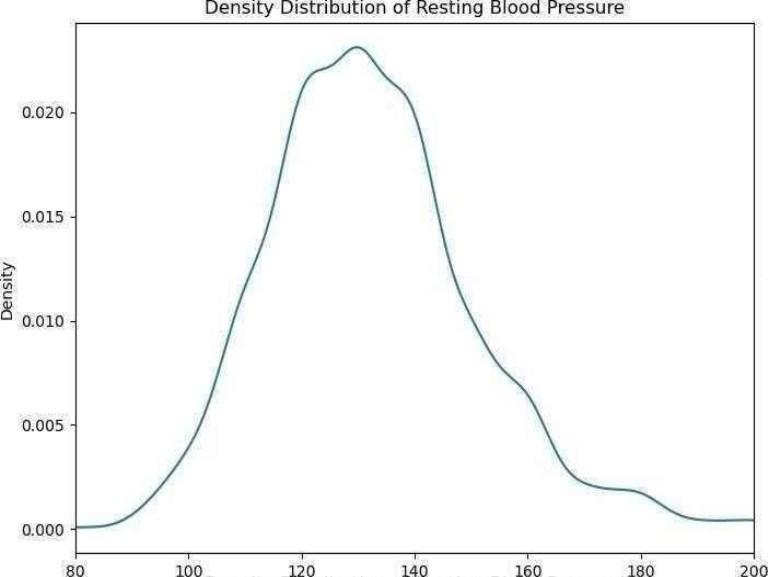
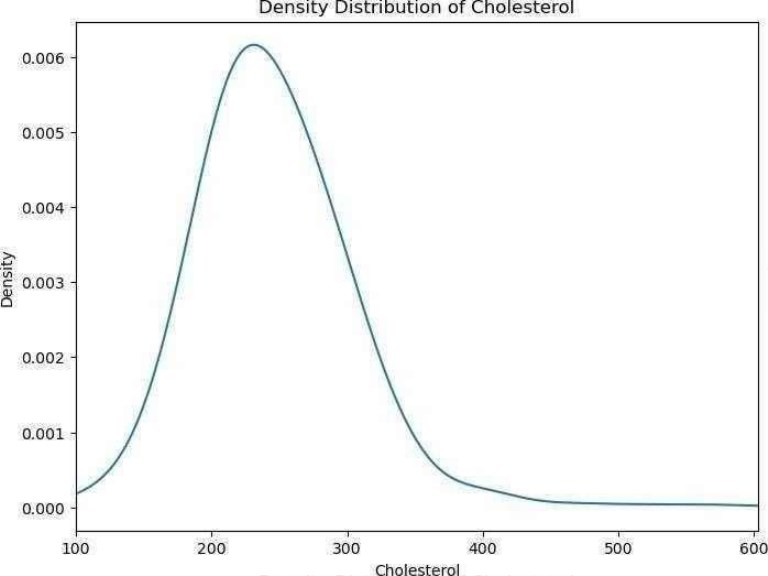
**RestingBloodPressure(resting bp s)**

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

32

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 vary widely from 0 to 603 mg/dL. The density distribution plot for

cholesterol shows a skewed distribution with a longtail towardsthe higher values. Most individuals

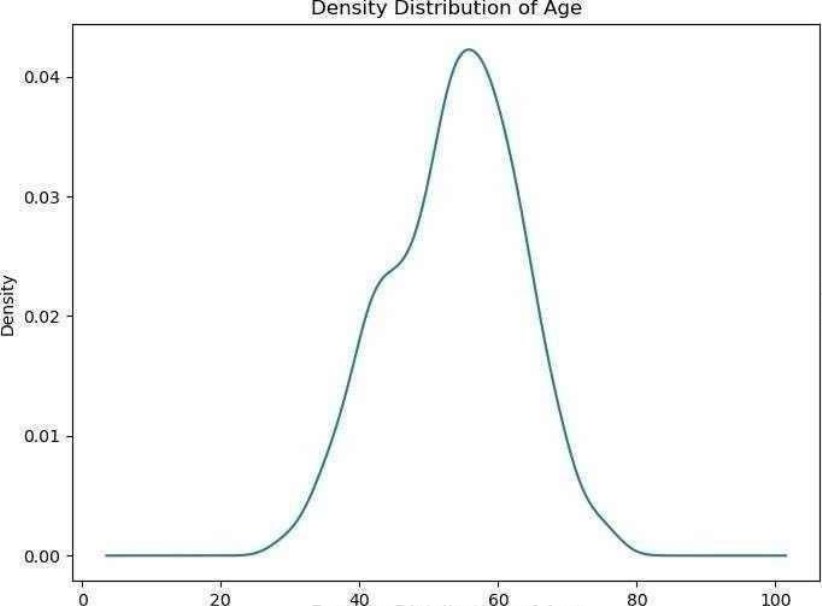
have cholesterol levels between 150 and 300 mg/dL, with the mean value being around 210 mg/dL.

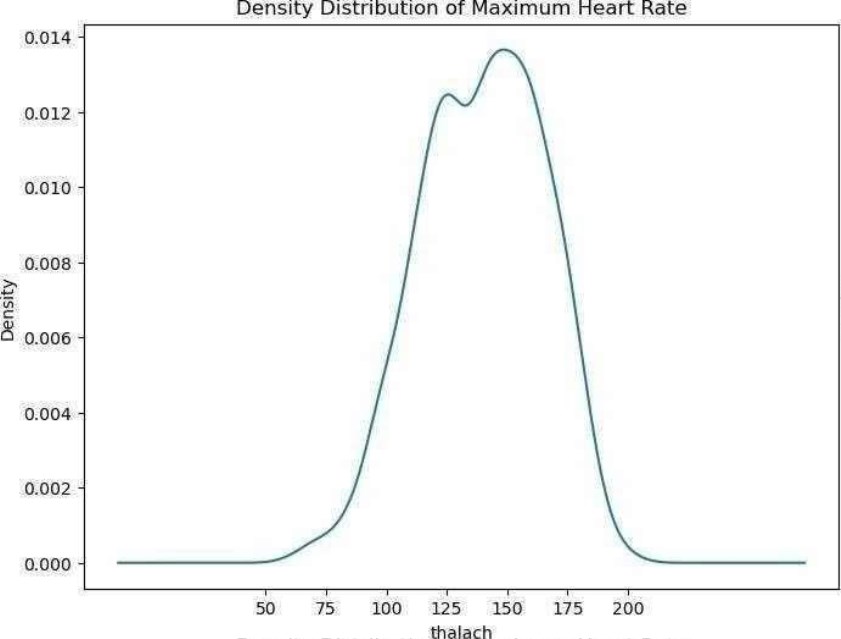
The distribution indicates that while high cholesterol levels are present, they are less common

compared to moderate levels.

Fig. 3.3&3.4Density Distribution of Cholestrol and Blood Pressure

33

**MaximumHeart Rate(maxheart rate)**



The maximum heart rate 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 typically displays a bell-shaped curve, suggesting a normal distribution with

most values clustering around the mean.

Fig. 3.5 & 3.6Density Distribution of Age and Maximum Heart Rate

34

**3.4** **ROC CURVES**

➢**Definition:**

• ROC (Receiver Operating Characteristic) curves are graphical plots used to

evaluate the performance of binary classification models. ➢**Axes:**

• The X-axis represents the False Positive Rate (FPR), which is the proportion of negative instances incorrectly classified as positive.

• The Y-axis represents the True Positive Rate (TPR), also known as sensitivity or recall, which is the proportion of positive instances correctly classified.

➢**IdealPoint:**

• 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 that follows the diagonal line from (0,0) to (1,1) represents a

modelwith no discriminatory power, equivalent to random guessing.

➢**AUC - Area Underthe 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 performanceno 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).

➢**Interpretation:**

• ROC curves are particularly useful when the classes are imbalanced

35

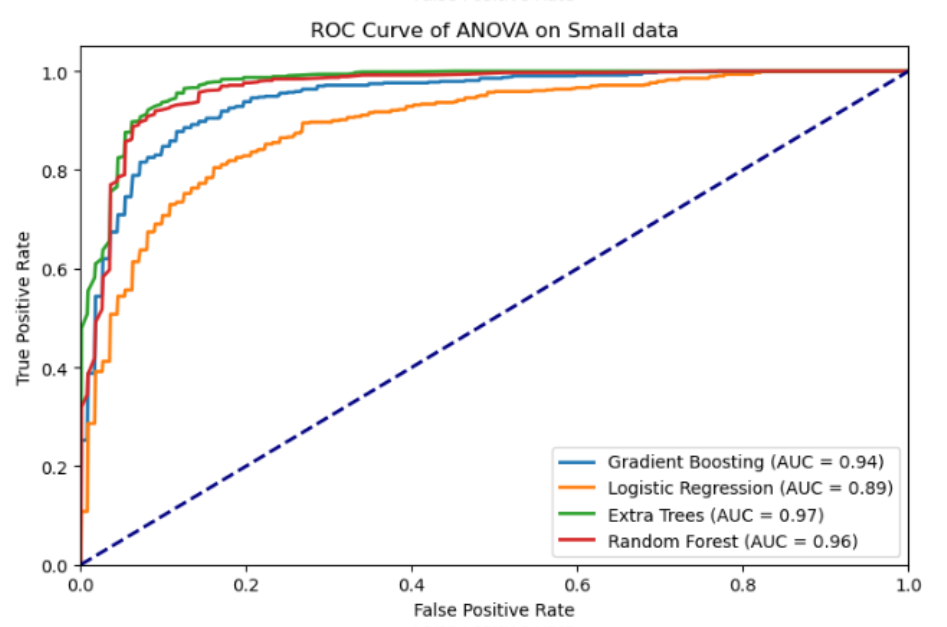
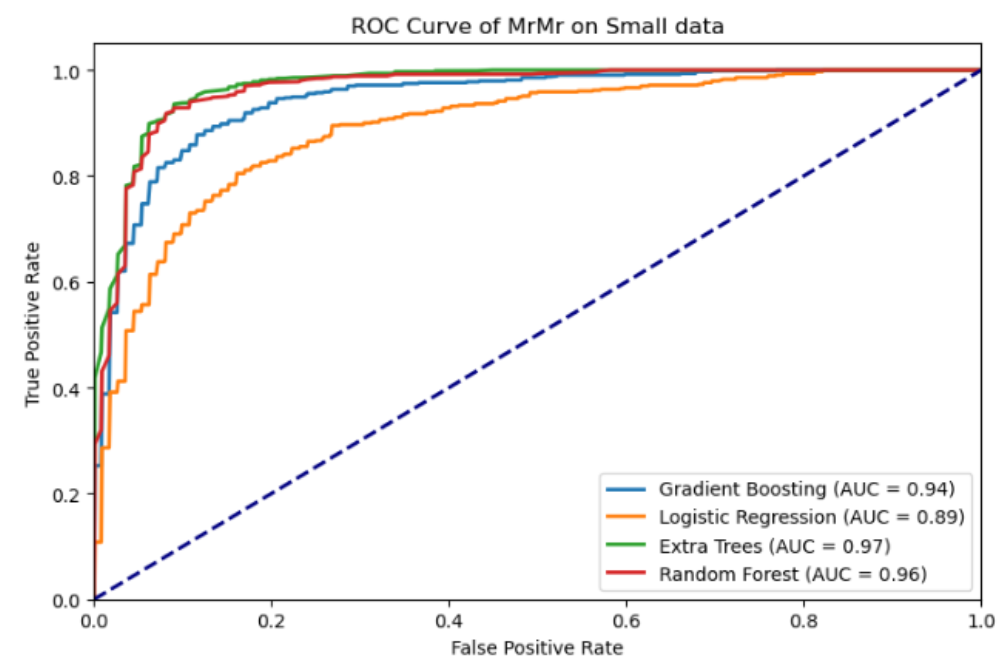
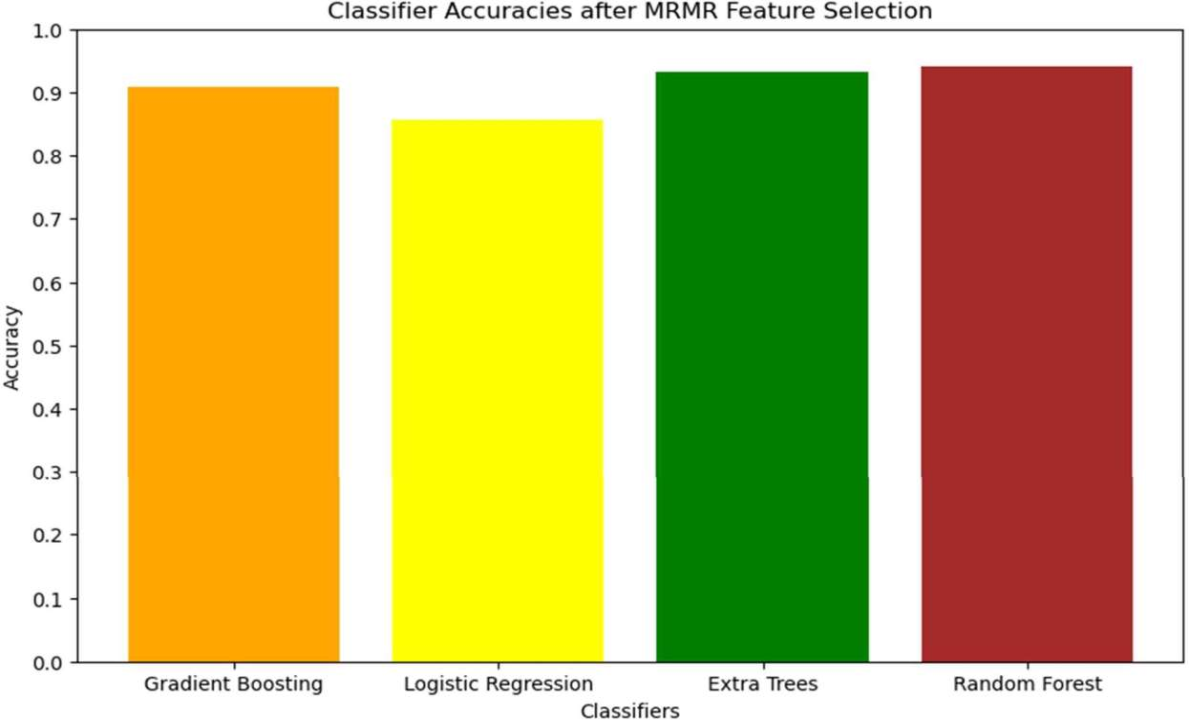


Fig.3.7 ROC CURVE ANOVA

Fig 3.8 ROC CURVE MRMR

36

**CHAPTER-4**

**RESULTS AND DISCUSSION**

**MRMRFeatureSelection 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

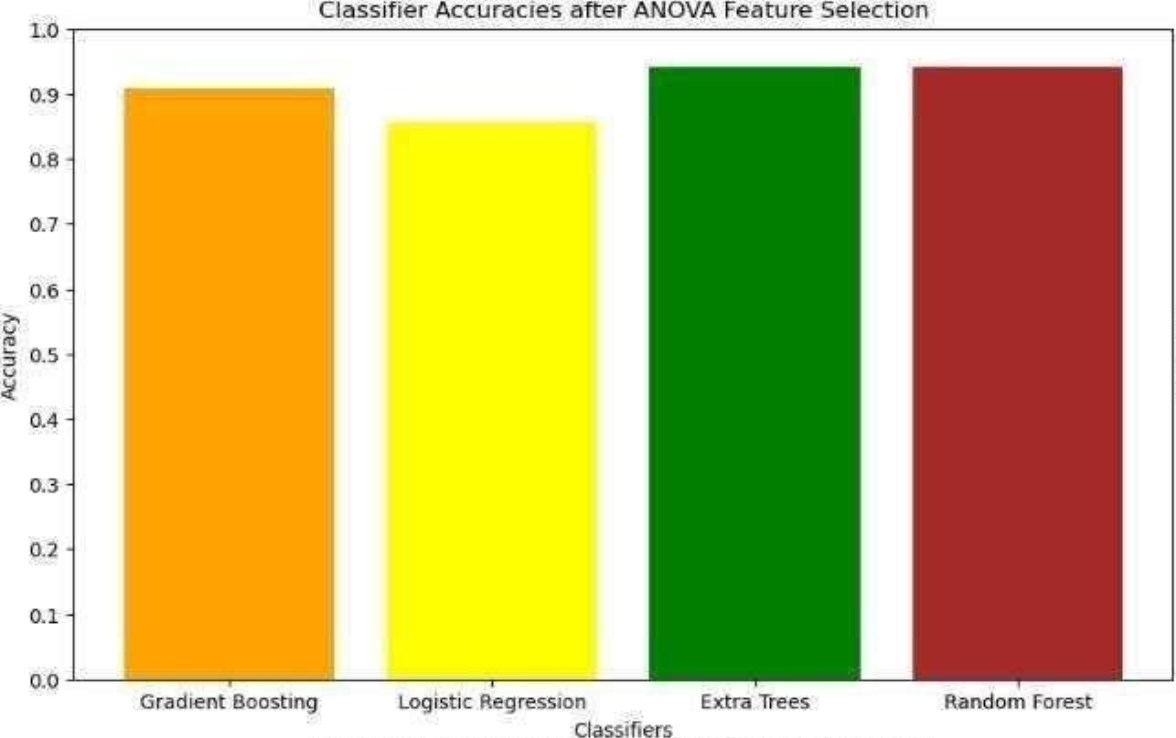
to enhance the predictive performance of machine learning models in this we have achieved

an accuracy of 85-90% using the gradient boosting, Logistic Regression, Extra Trees,

Random Forest.

Fig 4.1 MRMR ACCURACY

37

**ANOVAFeatureSelection 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

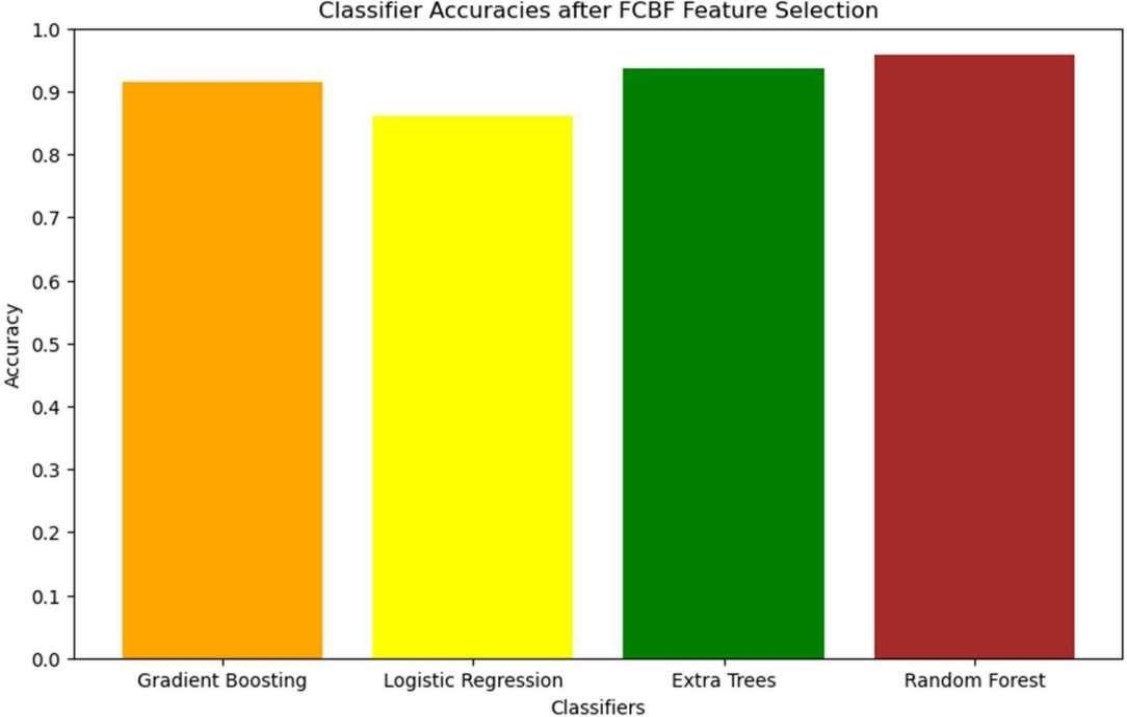
variance in the outcome variable, identifying those with the most significant impact. In the

ANOVA feature selection we have achieved an accuracy of 85-90% and above 90% in

Extra Trees and Random Forest

Fig 4.2 ANOVA ACCURACY

38

**FCBFFeatureSelection 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

redundancyamong thefeatures. This method is particularly advantageousfor 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

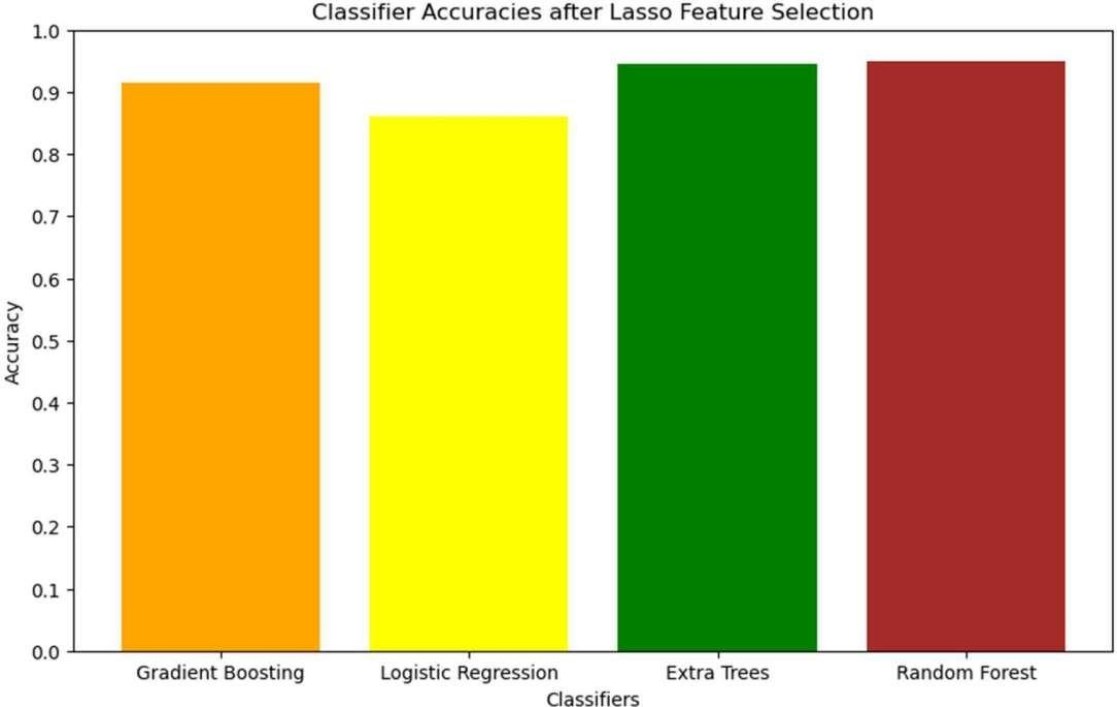
most critical predictors, enhancing the efficiency and accuracy of machine learning models. In

the Gradrient and Extra Trees and Random Forest we have scored more than 90% accuracy

successfully.

Fig 4.3 FCBF ACCURACY

39

**LASSOFeatureSelection 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

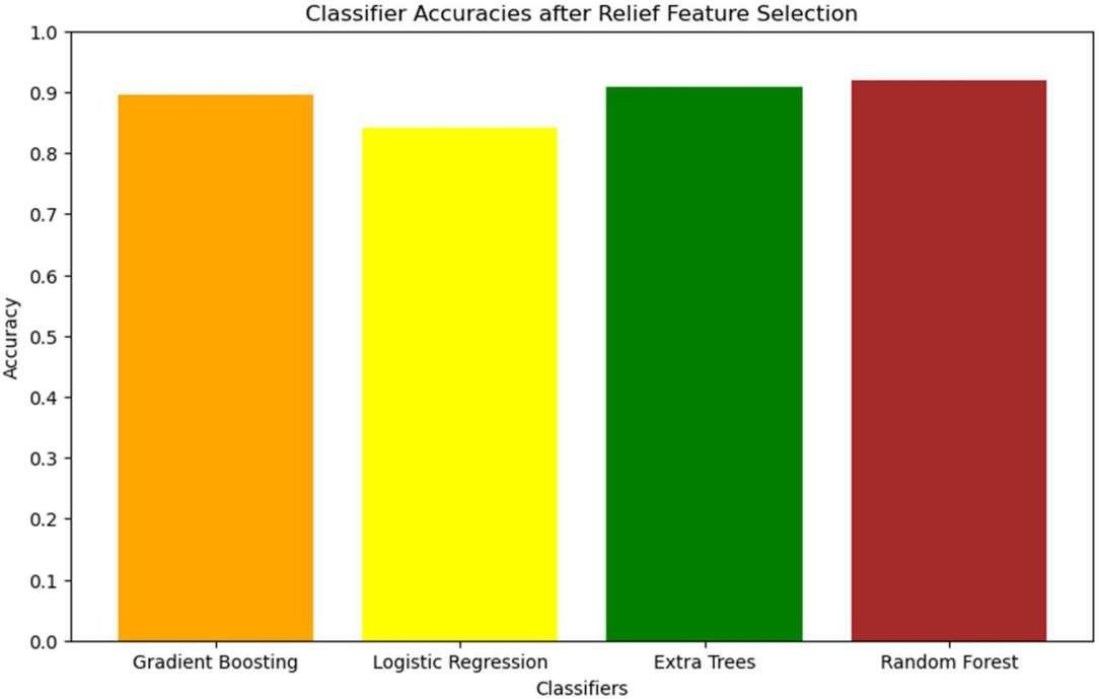
for high-dimensional datasets where the goal is to enhance model interpretability and reduce

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.

Fig 4.4 ANOVA ACCURACY

40

**RELIEFFeatureSelection 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 effectivefordatasets with amix ofcontinuousandcategorical

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

Fig 4.5 RELIEF ACCURACY

41

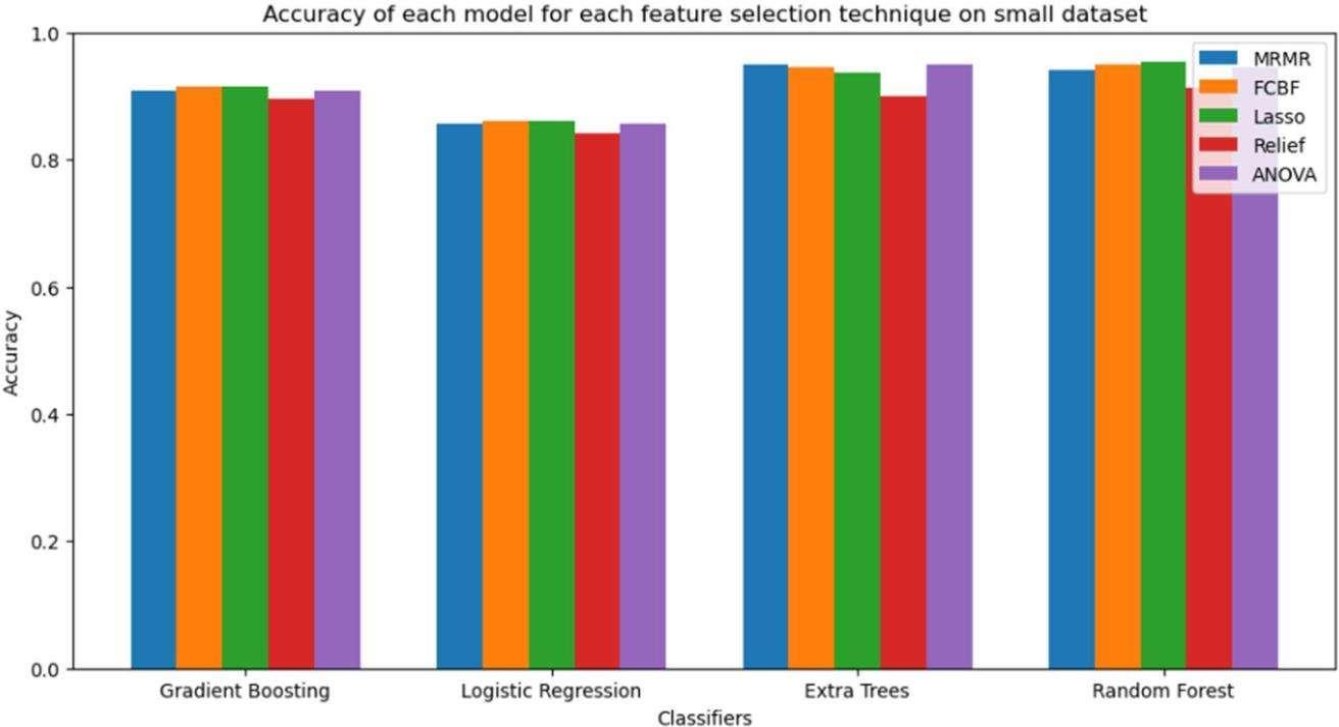
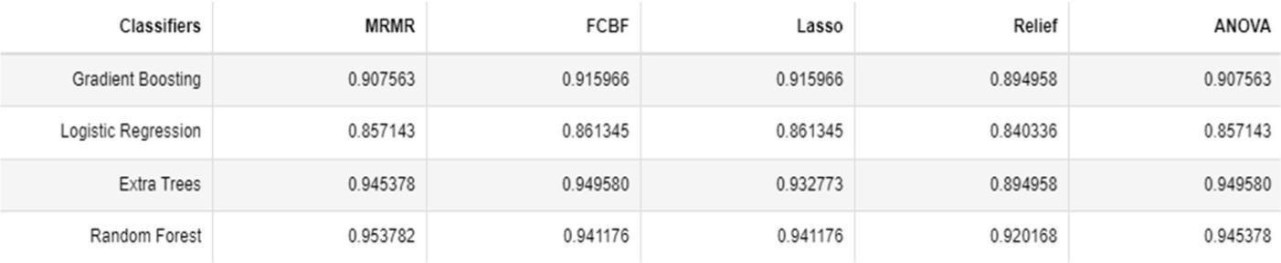
**ACCURACYOFEACHMODEL FOREACHFEATURE SELECTION**

Fig 4.6 ACCURACY OF EACH MODEL

**ACCURACY TABLE**

Fig 4.7 ACCURACY TABLE

42

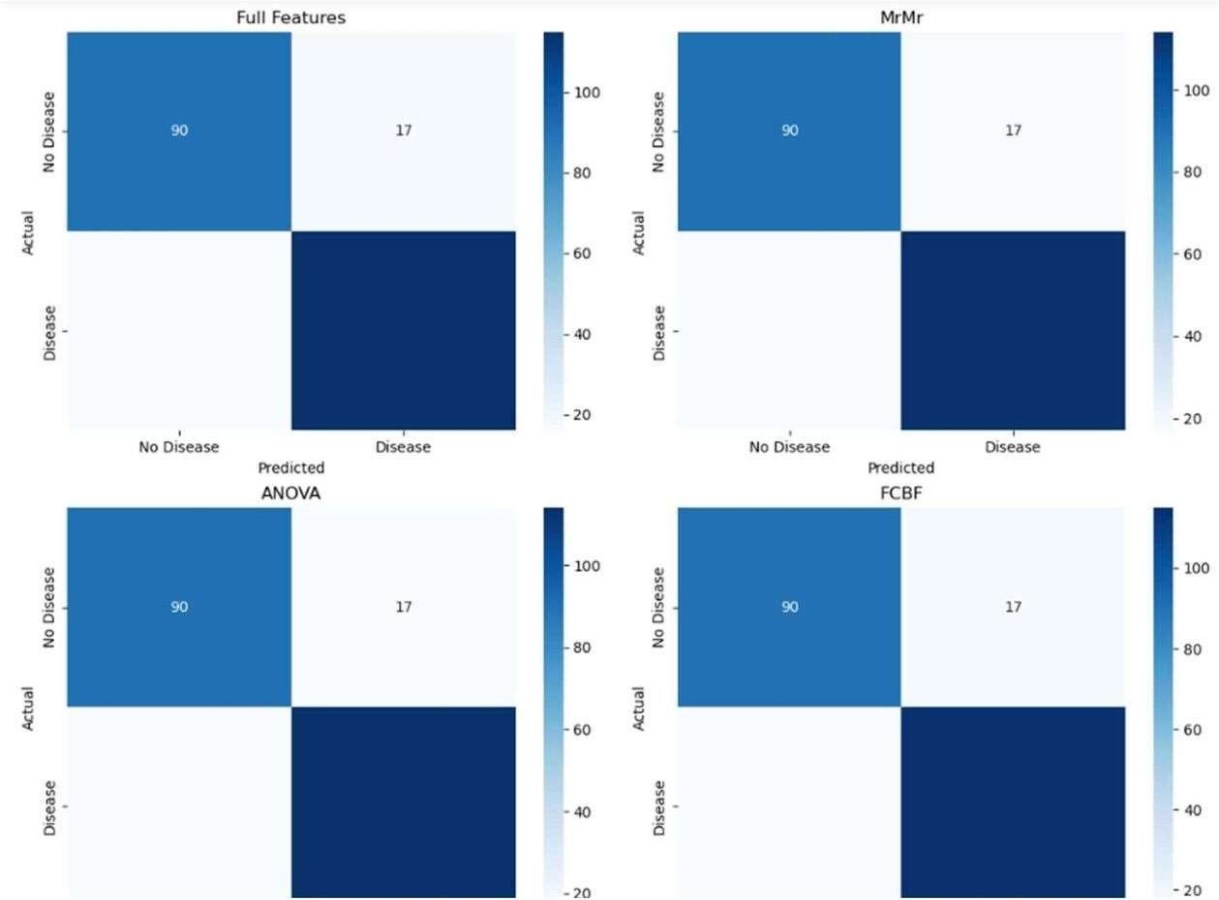
**CONFUSION MATRICES**

Fig 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

43

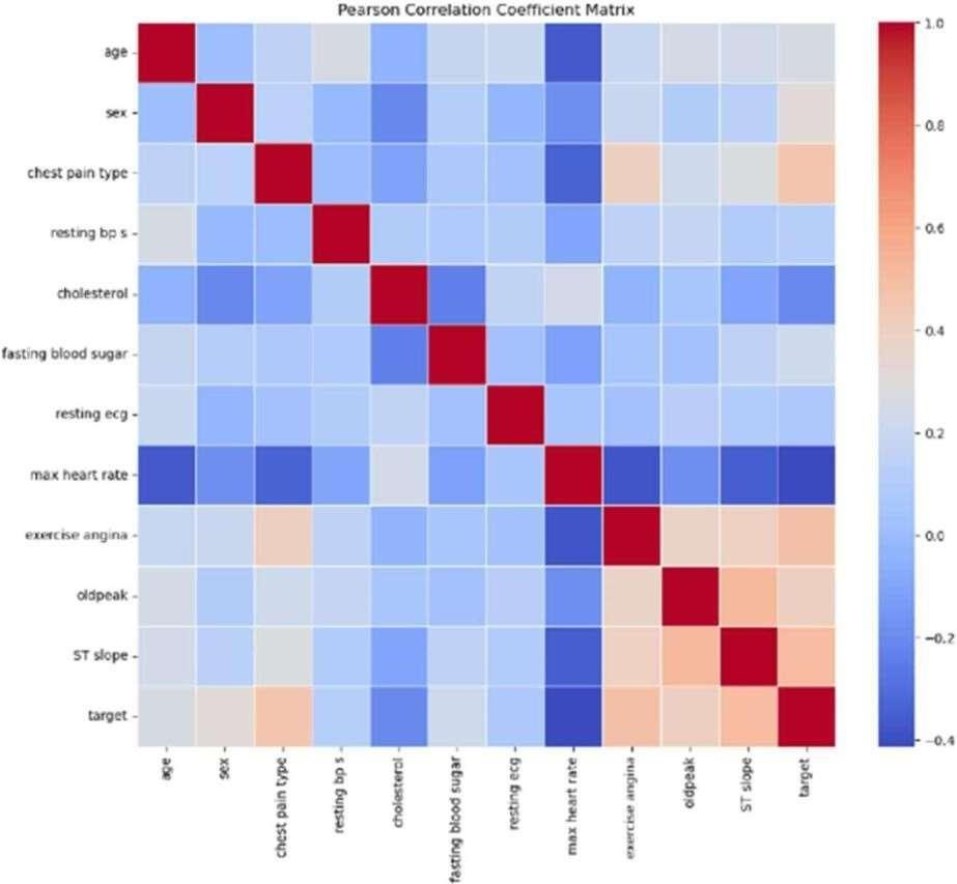
**PEARSONCORRELATIONCOEFFICIENTMATRIX**

Fig 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.

44

**CHAPTER-5 CONCLUSIONANDFUTURESCOPE**

**5.1 CONCLUSION**

The comprehensive evaluation of Gradient Boosting, Logistic Regression, Extra Trees, and Random Forest models usingvarious featureselection 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 FUTURESCOPE**

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.

45

**REFERENCES**

[1]A.K.Gárate-Escamila,A.Hajjam El Hassani,andE.Andrès,“Classificationmodelsfor 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,”.

[8]S. Anitha and N. Sridevi, “HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES”.

[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,”.

46

Similarity Report ID: oid:3618:67451721



PAPER NAME

Base paper - final - final.pdf

WORD COUNT

2827 Words

PAGE COUNT

6 Pages

SUBMISSION DATE

Sep 24, 2024 9:52 AM GMT+5:30

CHARACTER COUNT

16370 Characters

FILE SIZE

1.0MB

REPORT DATE

Sep 24, 2024 9:53 AM GMT+5:30

15% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

7% Internet database

Crossref database

0% Submitted Works database

14% Publications database

Crossref Posted Content database

Excluded from Similarity Report

Bibliographic material

Summary

2

Cardiovascular Disease Analysis Using Python

G. Mary Swarnalatha E

I

H

Nehas Reddy Kyatham E

I

H

N. Manish Kumar E

I Aeronautical Engineering Hyderabad, I

B. Sneha E

I Aeronautical Engineering Hyderabad, I

A — Worldwide, cardiovascular diseases are still among the

major health issues due to their prolonged healing process.9 paper, present machine g technique f r early CVD d n. Our system utilizes electro-cardiogram (ECG) data and focuses on feature selection optimization as a means of improving prediction accuracy. We achieved outstanding accuracy rates of 100 on both small and large datasets by using sophisticated classifiers. This approach has the potential to transform patient management practices and decrease CVD-related mortality rates. Cardiovascular disorders (CVDs) pose a major global health challenge accounting for high numbers of deaths across the globe(Siontis et al.,2011) (Lyon et al., 2011) .In this regard, we propose an innovative machine learning technique that is based on electrocardiograms (ECGs) to achieve remarkable accuracy in early CVD detection. Our system concentrates on feature selection optimization so as to enhance prediction scores. Using available state-of-the-art classifiers enabled us achieving 100% predictively paling small databases including lessthan20,000 records as well as extensive ones containing millions of observations which is promising for transforming patients’ supervision systems.

Index Terms—CVD,CardioVascular,ECG

I. INTRODUCTION

Public health is indeed a critical global concern, impacting millions of lives. Let’s delve into some key issues related to public health:

Long COVID: Long COVID is a significant health issue in 2023. It affects individuals for months, disrupting their ability to engage in daily activities, work, and relationships. Research is urgently needed to find effective treatments and preventive measures for long COVID 1. Mental Health: Mental disorders remain a leading cause of disability worldwide. The COVID-19 pandemic, war, and violence have further exacer- bated mental health challenges. Understanding risk factors and offering prevention strategies at the population level are crucial

Impact of Climate Change: Climate change directly affects health, from extreme heat to indirect effects like flooding, droughts,andairpollution.These environmental changes impact mental well-being, food security, and water availability

Chronic Diseases (CDs): CDs contribute significantly to overall

1

1

mortality. They h com- pared to

other diseases.1U e choices ( poor diet,

g, and ) play a major role

in CDs 1. Global Burden1

s a substantial b s.

allocating a significant portion of its GDP for treatment.

II. TRADITIONAL SYSTEMS

8 e will talk about different g that can be used for predicting cardiovascular disease

(CVD). This is because traditional systems have problems with their accuracy in assessing the risk of CVD (CardiovascularDisease).Machine learning algorithms (MLA) have been seen to give a better prognosis for CVD. MLAs contribute to improved clinical decision making, and enhance personalized care. Research Gap and Need for Exploration: However, there is little knowledge and empirical data available in this area despite the advances made. It is important that more accurate and efficient models are generated by filling these gaps. Optimization Approach: Particle Swarm Optimization (PSO): In an earlier study undertaken using MrMr and Relief identification of characteristics but results were not satisfactory. The present research optimizes its model findings through PSO approach that makes use of least effort hence produces empirical data for CVD prediction. Multiple Machine Learning Technis: We employed four different machine learning technologies to predict CVD. We intend to merge these methods into a single model so as to improve patient outcomes while also lowering healthcare expenses. To sum up, our research supports robust CVD prediction models which lead to better patient care and health outcomesoverall. TheRisk assessment and strategyto determine are the key components and tools like Framingham risk score and ASCVD Risk Calculator helping estimate the likelihood of cardiovascular events based the factors like age, cholesterol levels and the smoking status of the person is determined of the many factors that are considered of the heart related issues.The world healthorganization also issuedmany different guidelinesbasedon this. The people should follow their diet and nutrition in many ways and also we need to be careful about their life style in many ways .The risk assessment tools help us to calculate the features of the heart in many different ways and also the tools or the features give us the more or less accurate features of the particular referred cardiovascular diseases which more or less trouble the person in nature and the Framingham risk score also helps us in evaluating many different parameters of the characteristics heart disease which we also need to consider in evaluating the nature of the heart disease which is taken into consideration of many different aspects which can be more useful while determining .

13

The Study: Unleashing Dimensionality Reduction (DR) Methods Objective: The study aimed to optimize arrhythmia classification using unsupervised DR methods. Researchers explored five DR techniques: PCA, fastICA, KPCA, hNLPCA, and PPA.Methodology: Probabilistic n-grams: These were used to extract relevant features from electrocardiogram (ECG)

signals. DR Algorithms: PCA: A classic method for reducing dimensionality. fastICA: Equip 1

A: Utilized p kernels. hNLPCA: H . PPA:

. Classifier: A probabilistic neural network (PNN). Key Findings: FastICA Triumphs: Using fastICA w

1

led to an impressive Fscoreof99.83Time Trade-Off:While hNLPCA and KPCA are time-consuming for low-dimensional mapping, they offer potential benefits. PPA Superiority: PPA outperformed PCA by 10Dataset Limitations: The study ana- lyzed a

1

s. Gener- alizability to larger

populations requires further investigation. Beyond Arrhythmias: NYHA Rating and CRT NYHA Rating:1

g assesses heart failure severity. Regularly tracking1 provides

1

valuable insights into treatment re- sponse. C c T): CRT t

. patients. Monitoring NYHA class in electronic health records (EHR) helps gauge CRT effectiveness. Conclusion: A Heartfelt Journey Automated arrhythmia classification, coupled with NYHA assessment and CRT, empowers clinicians to make informed decisions. As research continues, we move closer to personalized cardiac care, where algorithms and human expertise harmonize for healthier hearts.Introduction: Heart disease, a leading global health concern, demands accurate prediction methods. Re- searchers explore machine learning algorithms to enhance diagnosis and prognosis. Let’s delve into recent findings and challenges.

This work proposes a nursing assistant framework that leverages m e risk . The framework utilizes readily available parameters like age, gender, andheart rate to assess risk. Additionally,the model incorporates

neural codes to enhance accuracy and robustness, enabling

1 l CVD r .

A recognized l s approach is its dependence on a limited set of parameters, potentially overlooking other relevant factors. To address this, the integration of data from wearable sensor devices holds promise. These devices can provide continuous streams of health data, enabling cost-effective detection of early cardiac issues through big data analytics and machine learning. Apache Spark, a distributed computing platform, can be employed for real-time analysis, optimizing machine learning for CVD prediction.

Advancements in Machine Learning for CVD Risk Assess-esearchers have actively explored machine learning tech- niques forCVDriskprediction.One study achieved high accuracy(over 90 percent) using ensemble learning methods, demonstrating its effectiveness compared to traditional tech- niques. Another approach implemented a achieving promising results. However, limitations exist, such as the need for multi-class classification for various heart disease stages. While machine learning offers significant potential for CVD risk prediction, challenges remain. Studies have highlighted the importance of ensuring data privacy and security, theneedforbroadervalidationacross diverse populations, and the interpretability of the models for clinical application. Machine learning presents a powerful tool for developing accessible and reliable heart disease risk assessment tools. Integration of previous research findings and taking care of limitations within this technology can lead to early detection of heart diseases revolutionalized by them as well as better outcomes for patients.

III. METHODOLOGY

Noise Reduction and Classification: Pre-processing tech-

niques improve ECG signal accuracy by removing noise. In Figure 1, it is represented a detailed schematic representation of the

Decision Tree outperforms KNN and Naive Bayes for de- tecting The chart explains the framework’s structure as well as its components.

abnormal heart rhythms. Accurate diagnosis of heart- related

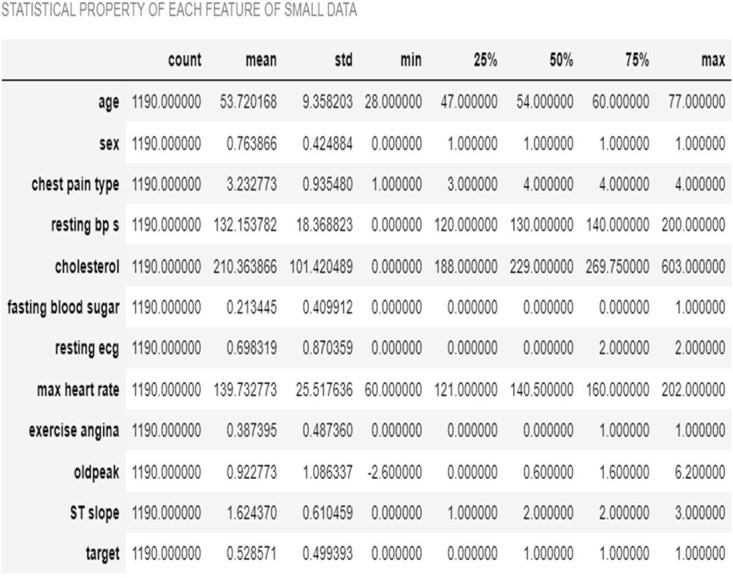
diseases is achievable. Feature Selection and Dimen- sionality 12. is the a Reduction: Machine learning predicts heart disease symptoms. . is Filtration of unwanted data CHI-PCA with RF achieves high accuracies, iden- tifying 3. Step-3 is Feature Selection

relevant features. Ethical considerations and state- of-the-art 4. Step-4 is Results and Discussions comparisons remain unexplored. Data Mining and Healthcare 5. Step-5 is Conclusion and Future Scope Impact: Heart disease research receives significant funding. Data

mining facilitates medical record interpretation. Supervised algorithms (SVM, k-NN, Naive Bayes) play a pivotal role. In the quest for healthier hearts, machine learning emerges as a powerful ally C a e h globally, highlighting e need for early detection strategies. While traditional methods like ECG tests offer valuable insights, their daily use for everyone might not be

7

feasible. This article explores the potential of using

Data Filtration

Feature Selection

Properly scaled features are crucial for algorithm performance and convergence speed.

E. Attribute Selection

MrMr

Classifier

Min Redundancy Max Relevance (MRMR):

MrMr is a feature selection method based on filter.. Its goal is to identify relevant features while minimizing redundancy.

1

It achieves this by iteratively removing f t

.

Gradie nt Boosti

Extra Tree

Rando m Forest

Logistic Regressi on

Figure-1

1 s heavily Datasets’ quality greatly affects the accuracy of classification metrics. Therefore, we have selected the

following datasets to show the importance of data and evaluate its generalizability in our research.1 t

) t) which was e

n old but d created . Various d s include

V among others. This d t has s with l number f

5 i as follows: to 1 (showing more disease)The 2nd1 s

, ) that

is s s done yearly on o .

concerning c service utilization are collected

during the survey period. Specifically, this dataset

emphasizes on the 2015 BRFSS which includes 253680 1 cleansed n

heart disease presence or absence.

B. Data Pre-processing: Data Pre-processing: Preprocessing is necessary for accurate representation of data and appropriate training and testing of classification algorithms since it switches raw data to meaningful integrations..

1 : Dealing with m a common challenge in data analysis. Factors like data collection errors,incompletesurveys, oromissions can lead t

D. Standard Scaling: Standard Scaler from machine learning literally sucks all the life out ofyou whenyou talk aboutdata processing methods as it modifies continuous variables devoid of normal distribution into a standard distribution.

F. STATISTICAL PROPERTY OF EACH DATA

Figures and Tables

a d t

4

4

4

i s.. Insert f after they are cited in the text.

G. Numerical Distribution

H. Age: This is how many years old you are. It’s like counting how many birthdays you’ve had! I. Resting Blood Pressure: Imagine your blood is like a little river flowing through your body. Blood pressure tells us how hard that river pushes against the walls of your blood vessels. We want it to be just right, not too high or too low! J. Cholesterol: Think of cholesterol as tiny helpers in your blood. Some are good (HDL) and some are not-so- good (LDL). We want more of the good ones and less of the not-so-good ones. K. Maximum Heart Rate: Your heart is like a superhero! The maximum heart rate is the fastest your heart can beat. It’s like when you run really fast or play tag – your heart races!

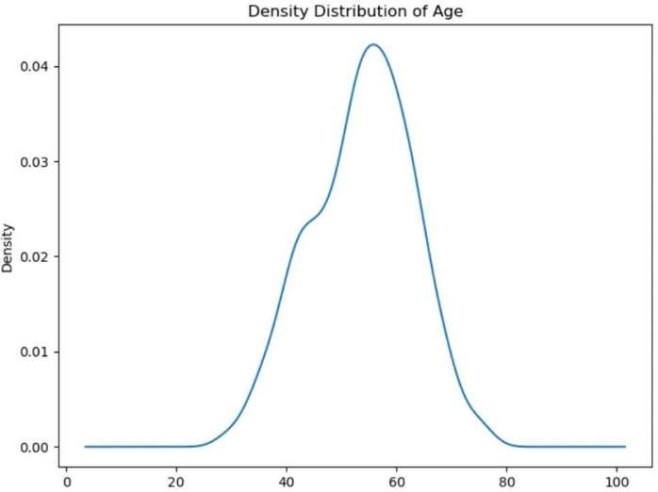
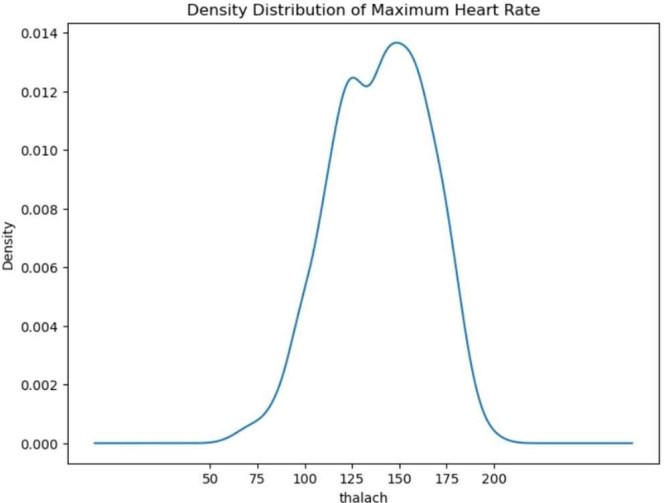
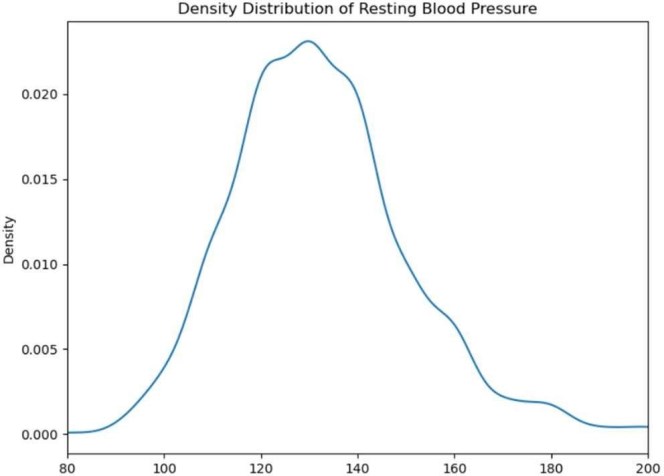
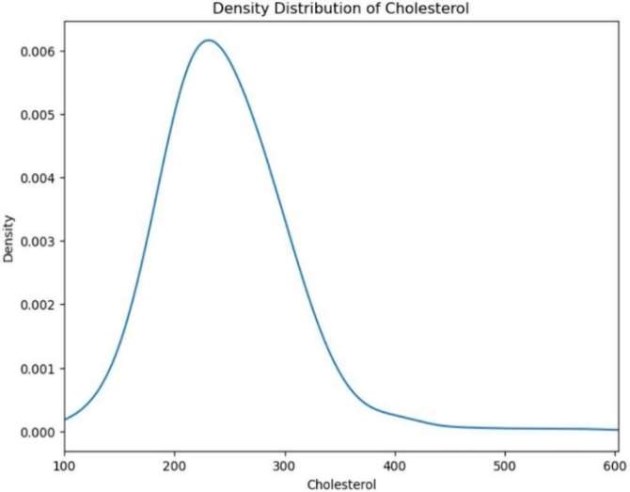


Figure-2

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

Figure-4

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.

Figure-3

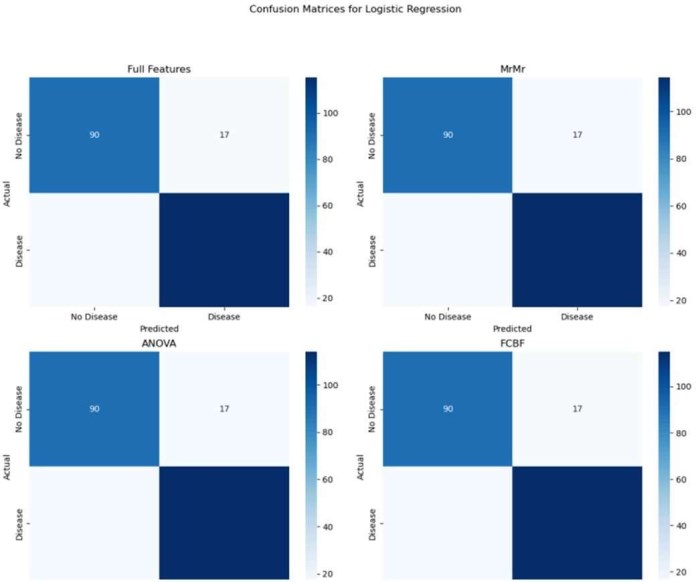
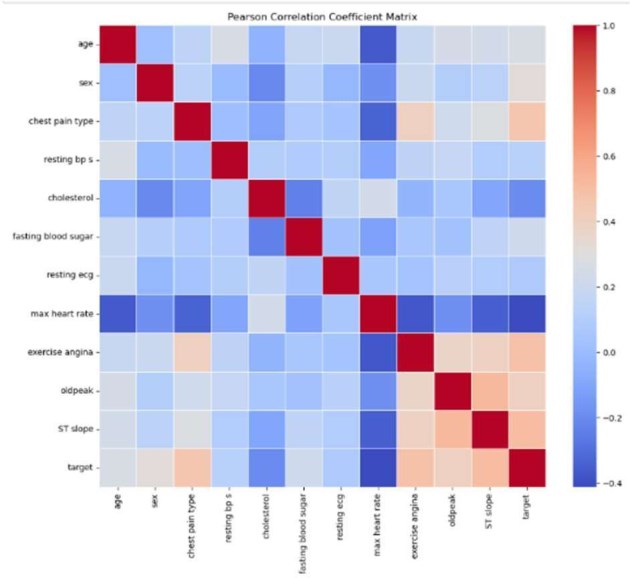
Cholestrol levels in the dataset vary widely from 0 to 603 mg/dL.

The density distribution plot for cholesterol shows a skewed distribution with a long tail towards higher values

Figure-5

The maximum heart rate 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



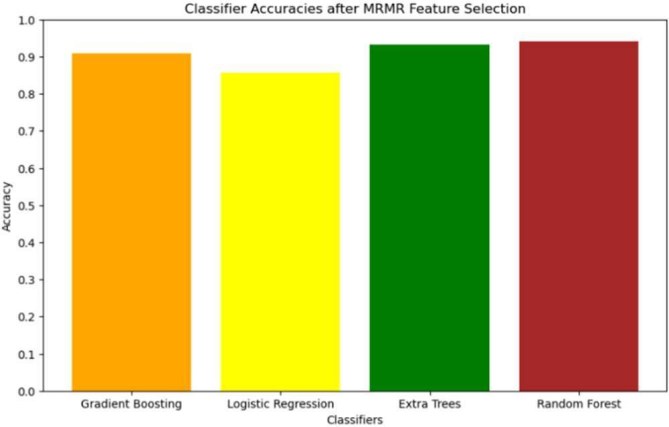
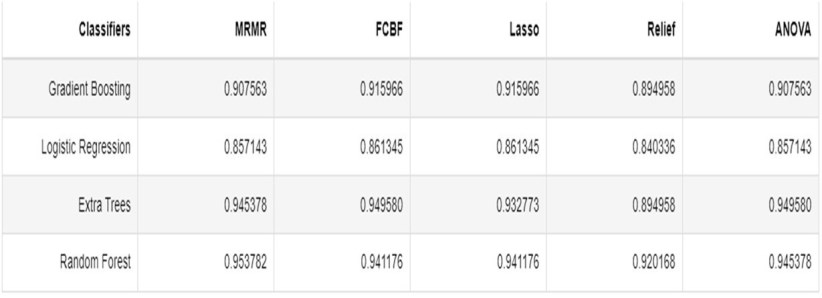
IV.RESULTS AND DISCUSSIONS

Figure-6

The M e

6

0

s e aims 1s s that have

r with t while maintaining minimal redundancy among them. This method is particularly effective in scenarios where the dataset contains a large number of features.

Figure-7

All the models have achieved an average of nearly 87.38 % accuracy using the feature selection technique.

Figure-8

The scattering parameters contain5 T

s namely e

Figure-9

e

The Pearson coefficient, also known as3 n

c a statistical m n o

t

V.Conclusion and Future Scope

The study focused on evaluating the accuracy of various machine learning models on a small dataset, specifically the heart disease dataset. The models considered were Gradient Boosting, Logistic Regression, Extra Trees, and Random Forest. The evaluation was performed using several feature selection techniques: MRMR, ANOVA, FCBF, Lasso, and Relief. The findings from these evaluations provide valuable insights into the performance and suitability of these models for small datasets.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. ClerkMaxwell,ATreatise onElectricity and Magnetism, 3rded., vol. [3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange

anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: Univer-sity Science, 1989.

[8] S. Rehman, E. Rehman, M. Ikram and Z. Jianglin, "Cardiovascular disease (CVD): assessment prediction and policy implications", BMC Public Health, vol. 21, no. 1, pp. 1299, 2021.

[9] O. Atef, A. B. Nassif, M. A. Talib and Q. Nassir, "Death/Recovery Prediction for Covid-19 Patients using Machine Learning", 2020.

[10] A. B. Nassif, I. Shahin, M. Bader, A. Hassan and N. Werghi, "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data", Mathematics, 2022.

[11] H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed and Q. Nasir, "Wearable Devices Smartphones and Interpretable Artificial Intelligence in Combating COVID-19", Sensors, vol. 21, no. 24, 2021.

[12] O. T. Ali, A. B. Nassif and L. F. Capretz, "Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution", 2013 3rd International Conference on Communications and Information

[13] A. Nassif, O. Mahdi, Q. Nasir, M. Abu Talib and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease", Jan. 2018.

[14] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye and M. Ashour, "Effective diagnosis and monitoring of heart disease", Int. J. Softw. Eng. its Appl, vol. 9, no. 1, pp. 143-156, 2015.

[15] K. Vembandasamyp, R. R. Sasipriyap and E. Deepap, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International J. Innov. Sci. Eng. Technol, vol. 2, no. 9, 2015

Similarity Report ID: oid:3618:67451721



15% Overall Similarity

Top sources found in the following databases: 7% Internet database

Crossref database

0% Submitted Works database

14% Publications database

Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

**1**

**2**

**3**

**4**

[Tahseen Ullah, Syed Irfan Ullah, Khalil Ullah, Muhammad Ishaq, Ahmad...](https://doi.org/10.1109/ACCESS.2024.3359910) Crossref

[Samineni Peddakrishna, Taimoor Khan. "Design of UWB Monopole Ant...](https://doi.org/10.1016/j.aeue.2018.09.014) Crossref

[en.hocmarketing.org](https://en.hocmarketing.org/univariate-analysis-basic-statistical-techniques-for-data-analys) Internet

[cs.overleaf.com](https://cs.overleaf.com/latex/templates/acii-2022-template/gqhfwdctrpby.pdf)

Internet

9%

1%

1%

1%

**5**

**6**

**7**

[weasychair.org](https://wwww.easychair.org/publications/preprint_download/2V4r) Internet

[peerj.com](https://peerj.com/articles/cs-1997v0.1/submission) Internet

[scholarscompass.vcu.edu](https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5691&context=etd)

Internet

<1%

<1%

<1%

**8**

[Arunpradeep. N, Dr. G. Niranjana. "Different Machine Learning Models ...](https://doi.org/10.35940/ijrte.F7310.038620)

Crossref

<1%

Sources overview

Similarity Report ID: oid:3618:67451721



<1%

**9**

**10**

[Huilin Ge, Zhiyu Zhu, Kang Lou, Wei Wei, Runbang Liu, Robertas Damaš...](https://doi.org/10.3390/sym12030434) Crossref

[Jingxin Ren, XianChao Zhou, Ke Huang, Lei Chen, Wei Guo, KaiYan Fen...](https://doi.org/10.1016/j.compbiomed.2023.107883)

<1%

Crossref

**11**

**12**

**13**

[koreascience.kr](https://koreascience.kr/article/JAKO202230654206154.pdf) Internet

[coursehero.com](https://www.coursehero.com/file/58582463/Chp3pdf/) Internet

[Dhafer G. Honi, Laszlo Szathmary. "A one-dimensional convolutional n...](https://doi.org/10.1016/j.imu.2024.101535)

Crossref

<1%

<1%

<1%

Sources overview

02/11/2024, 11:57 Gmail - ICICI-2024 Conference Paper Provisional Acceptance -Reg

**Nehas Reddy Kyatham <nehasreddyk@gmail.com>**

**ICICI-2024 Conference Paper Provisional Acceptance -Reg**

**ICICI ECE** <icici@vjit.ac.in> Mon, Oct 28, 2024 at 6:03 AM To: Nehas Reddy Kyatham <nehasreddyk@gmail.com>

DearAuthor,

It's a pleasure to inform you that your paper entitled "Cardiovascular Disease Analysis Using Python with Paper ID: ICICI-158381" has been provisionally accepted for presentation in the **Scopus IndexedAIPConference ICICI-2024 to be held during 28 -29, November 2024, being organized by the Department of Electronics and Communication Engineering, VJIT, Hyderabad, India.** Herewith we are attaching Similarity Report and Reviewers comments also.Please ensure the following points before submitting your Camera Ready Paperon or before **01-11-2024 to** [**https://forms.gle/pGPZ5xEXbShtgU2V7**](https://forms.gle/pGPZ5xEXbShtgU2V7)

1. Your Paper will be published underAIP Scopus indexed conference proceedings (Rs.10000/-) category after successful presentation.

2. Similarity index from single source should be less than 5% 3. Paper must be revised as per Reviewers comments

4. Paper must be strictly as per theAIP template format given link<https://icici2024.com/images/AIPCP_Article_Template.docx> 5. Equations must be written by the equation editor.

6. Graphs, figures and diagrams must have good resolution.

7. Copy and paste of diagrams must be avoided and diagrams must be drawn by using standard software.

8. Please revise the paper thoroughly to avoid any mistakes in your resubmitted paper. Spelling mistakes and grammatical errors must be eliminated. Equation, table and figure numbers in the text and captions of the manuscript must be correct.

9. Carefully review the article for typographical errors and make necessary corrections.

10. Authors should remove self-citations and citations that are unrelated to the content 11. No need to include headers, footers in the document.

12. Don't allot any page numbers and maximum number of pages should not exceed 10 13. Please check the information of all the authors.

14. Proofread your manuscript thoroughly to confirm that it will require no revision.

**Reviewer Comments**

https://mail.google.com/mail/u/0/?ik=411c4e1918&view=pt&search=all&permmsgid=msg-f:1814115803679968903&simpl=msg-f:1814115803679968903 1/2

02/11/2024, 11:57

1.

2.

3. 4.

5.

6. 7.

Gmail - ICICI-2024 Conference Paper Provisional Acceptance -Reg

The paper is in IEEE double column format.

Present info on the database used for experiment

References should be men oned in the running text too. Results to be explained elaborately and clearly

Mention uniqueness of the proposed method? Conclusion to be support the methodology and title Figures (images) should be of high quality.

**Kindly Note:**

· The Author(s) is (are) sole responsible for any work quality issues/citation/self-citation issues/mismatched content issues/ review & revision process delays/ not responding for review comments/copyright issues/ copyright and proofread response delays during submission/review/revision/proofreading related issues and rejection from the organizers/ publishers.

· After submitting your papers if there is any discrepancy with respect to above mentioned points their papers will not be processed to proceed further.

Best Regards,

ICICI-2024 Conveners Dr. S. Thulasi Prasad, Dr. V.G. Sivakumar, Department of ECE,

VidyaJyothi Institute of Technology (A), Hyderabad – 500075.

Contact No: 9700668133, 7305278513

**4 attachments**

**ICICI-158381.pdf** 1447K

**AIP\_Conference\_Proceedings\_License\_Agreement to Author.pdf** 229K

**AIPCP\_Article\_Template (5).docx** 288K

**Registration Guidlines.pdf** 194K

https://mail.google.com/mail/u/0/?ik=411c4e1918&view=pt&search=all&permmsgid=msg-f:1814115803679968903&simpl=msg-f:1814115803679968903 2/2