CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION

K. NEHAS REDDY - 21951A04B6

N. MANISH - 21951A0492

B.SNEHA – 21951A04K4

CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION

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

BACHELOR OF TECHNOLOGY IN
ELECTRONICS AND COMMUNICATION ENGINEERING

by

K. NEHAS REDDY	**21951A04B6**
N. MANISHKUMAR	**21951A0492**
B. SNEHA	**21951A04K4**

Under the guidance of

Dr. G MARY SWARNALATHA
Assistant Professor

Department of Electronics and Communication Engineering

INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)

Dundigal, Hyderabad – 500 043, Telangana

May 2025

DECLARATION

We certify that

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

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

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

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

e) Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, wherever necessary.

Signature of the Student

Place: HYDERABAD

21951A04B6

Date:

21951A04K4

21951A0492

CERTIFICATE

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

Supervisor	**Head of the Department**
Dr. G. Mary Swarnalatha	**Dr. P Munaswamy**
Assistant Professor	**Professor**

Date:

APPROVAL SHEET

This project report done **CARDIOVASCULAR DISEASE DETECTION USING OPTIMAL FEATURE SELECTION** by **K. NEHAS REDDY (21951A04B6), N.MANISH KUMAR (21951A0492), B. SNEHA (21951A04K4)** 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 V Narasimha Prasad

Date:

Place: HYDERABAD

ACKNOWLEDGEMENT

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

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

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

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

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

ABSTRACT

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

Keywords: Cardiovascular Disease(CVD), Decision trees, random forest**.**

CONTENTS

Table of Contents **Page No**

LIST OF FIGURES

LIST OF TABLES

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,

1.2 OBJECTIVES

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

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

- **Distribution of Numerical Features**

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

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

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

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

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

- **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.

- **Accuracy of Each Model on Each Selection 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