Cardiovascular Disease Recognition Using Multiple Machine Learning Algorithms with Feature Scaling Technique

By

Maj Arslan Naseer

NUST-2021-MCS-00000398006

Supervisor

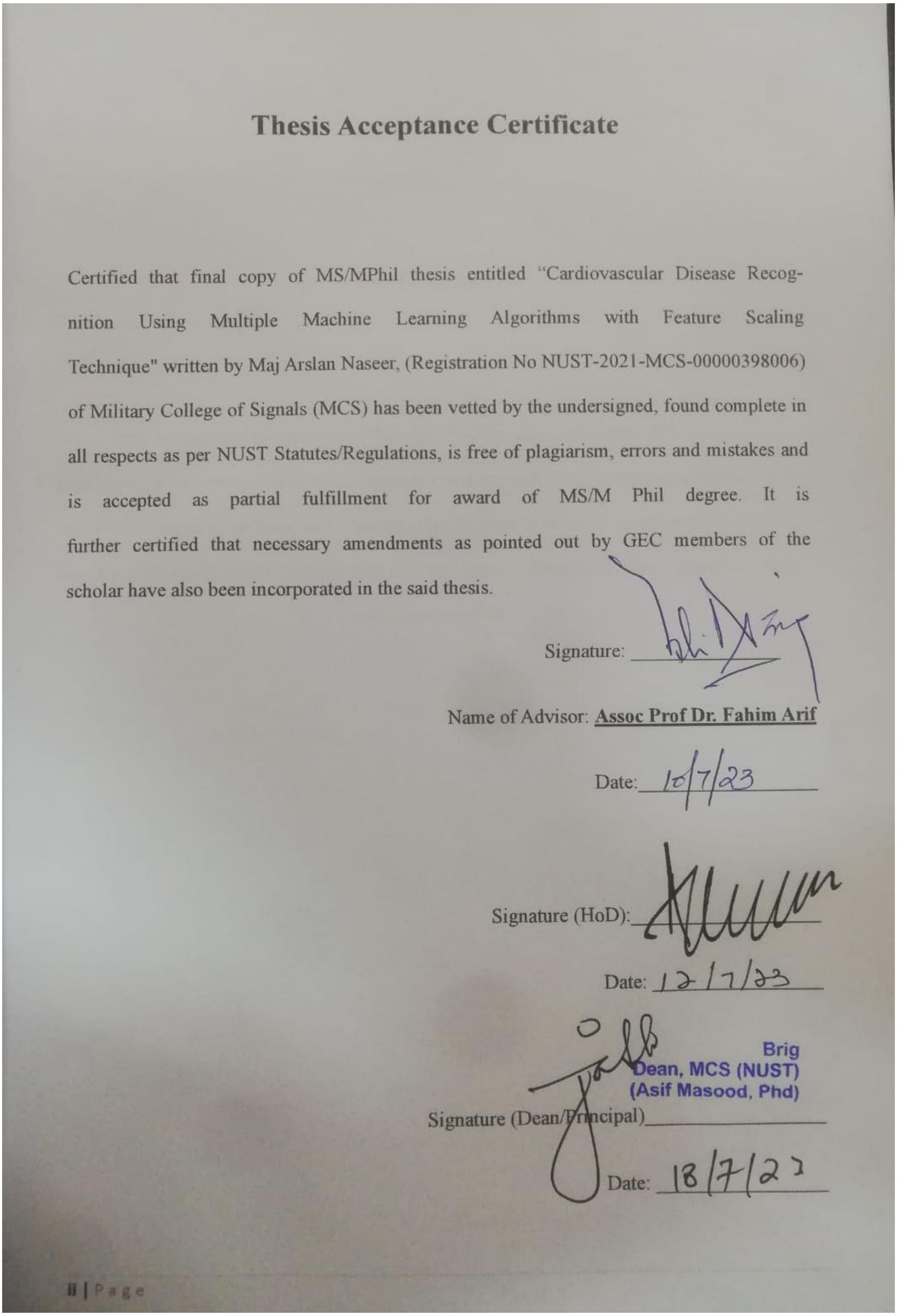
Assoc Prof Dr. Fahim Arif

Department of Software Engineering

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Software Engineering (MSSE)

In

Military College of Signals (MCS),
National University of Science and Technology (NUST),
Islamabad, Pakistan.
(June, 2023)



Thesis Acceptance Certificate

Certified that final copy of MS/MPhil thesis entitled "Cardiovascular Disease Recognition Using Multiple Machine Learning Algorithms with Feature Scaling Technique" written by Maj Arslan Naseer, (Registration No NUST-2021-MCS-00000398006) of Military College of Signals (MCS) has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: __________

Name of Advisor: **Assoc Prof Dr. Fahim Arif**

Date: 10/7/23

Signature (HoD): __________

Date: 12/7/23

Brig
Dean, MCS (NUST)
(Asif Masood, Phd)

Signature (Dean/Principal) __________

Date: 18/7/23

Dedication

This Thesis is dedicated to my beloved Parents, Children and my beloved Wife, who all have been my endless source of love, encouragement, and strength. Your unwavering beliefs in my abilities, countless sacrifices, and relentless support have been the foundation upon which I built my academic pursuits. Without their love and support this research work would not have been made possible.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at Department of Software Engineering at Military College of Signals (MCS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at Military College of Signals (MCS) or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Maj Arslan Naseer**

Signature: _______________

Acknowledgments

In the name of Allah (S.W.A), the Creator and Sustainer of the Universe, to whom belongs all glory and power. He alone has the authority to elevate and humble individuals as He pleases. Truly, nothing can be accomplished without His will. From the moment I stepped foot into NUST until the day of my departure, it was by His divine blessings and guidance that I was able to navigate the path of success. His unwavering support and the opportunities He bestowed upon me were instrumental in completing my research journey.

I humbly acknowledge that no words or actions can fully express my gratitude for the countless blessings He has showered upon me throughout this research period. I am indebted to His boundless bounties and am forever grateful for His divine intervention in my academic pursuits. To Allah (S.W.A), I dedicate this thesis as a humble tribute, recognizing His infinite wisdom and benevolence. It is through His mercy that I have reached this milestone, and I pray that my work may be of benefit to others and serve as a means of pleasing Him.

I would also like to express my heartfelt appreciation to my thesis supervisor, **Brig ® Assoc Prof Dr. Fahim Arif**, for his unwavering support and guidance throughout my thesis. His knowledge, expertise, and dedication to his field have been a source of inspiration to me, and I am grateful for the time and effort he invested in my success. Whenever I encountered any difficulties, he was always available to offer his assistance and provide me with insightful feedback.

In addition, I extend my gratitude to my GEC member, **Brig Adnan Ahmed Khan, PhD** and **Assistant Professor Dr. Yawar Abbas Bangash**, for their continuous availability for assistance and support throughout my degree, both in coursework and thesis. His expertise and knowledge have been invaluable to me, and I am grateful for his unwavering support and guidance.

Lastly, All praises and thanks be to Allah (S.W.A), the Most Merciful and the Most Gracious.

Maj Arslan Naseer

Implication of Research

The findings of this research hold significant potential for Medical Healthcare Centers, particularly in Pakistan, in the field of heart-related issues. The application of the proposed model can assist medical organizations in making prompt decisions regarding heart patients by utilizing various strategies derived from this method. This work aims to provide valuable insights and tools that can aid in the quick assessment and treatment of individuals with heart conditions, ultimately benefiting the healthcare sector in Pakistan.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

ML	--------------------	Machine Learning
CVD	--------------------	Cardiovascular Disease
HLRBM	--------------------	Hybrid Logistic Regression Bagging Model
SVM	--------------------	Support Vector Machine
NB	--------------------	Naïve Bayes
RF	--------------------	Random Forest
KNN	--------------------	K-Nearest Neighbor
DT	--------------------	Decision Tree
LR	--------------------	Logistic Regression
ANN	--------------------	Artificial Neural Network
AB	--------------------	AdaBoost
SMOTE	--------------------	Synthetic Minority Oversampling Technique
GA	--------------------	Genetic Algorithm
TP	--------------------	True Positive
TN	--------------------	True Negative
FP	--------------------	False Positive
FN	--------------------	False Negative
GB	--------------------	Gradient Boosting
LASSO	--------------------	Least absolute shrinkage selection operator

Abstract

Cardiovascular disease (CVD) believes to be a major cause of transience and indisposition worldwide. Early diagnosis and timely intervention are critical in preventing the progression of CVD and improving patient outcomes. Machine learning (ML) algorithms have emerged as powerful tools in CVD recognition, with the potential to assist physicians in making accurate and efficient diagnoses. This research work explores the combination of multiple ML algorithms for CVD recognition, utilizing diverse datasets such as the Cleveland, Hungarian, Switzerland, statlog, and VA Long Beach datasets. Additionally, a CVD dataset comprising 12 attributes and 70,000 records is employed, demonstrating improved results through the proposed and trained model compared to previous prediction techniques for CVD. The performance of various ML techniques, including support vector machines (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), and Logistic Regression (LR), is evaluated and compared. The influence of feature selection and feature scaling on the model's execution is also examined. An ensemble bagging techniques is applied as base classifier which is being embedded with other classifiers. LR classifier embedded with bagging techniques proved to be our proposed model. The findings reveal that the proposed Hybrid Linear Regression Bagging Model (HLRBM) outperforms other models. Furthermore, the study highlights the significance of data preprocessing techniques, such as data normalization and class balancing, which significantly enhance the performance of all models. To this end, standard scalar and ensemble SMOTE techniques are employed. The study emphasizes the importance of selecting an appropriate ensemble technique in conjunction with various ML algorithms and preprocessing methods for CVD prediction. Overall, the research provides valuable insights into the potential of ML in improving CVD risk assessment.

Key Words: Machine Learning, Cardiovascular Disease, Classification Models, E-Healthcare

CHAPTER 1

INTRODUCTION

1.1 Overview

The heart proves to be a vibrant organ that plays a vibrant role in a body's inclusive functionality. Indicators of CVD can fluctuate depending on the explicit situation; however it may include chest tightness or pain, discomfort, shortness of breath, fatigue, dizziness, and swelling in the legs or abdomen [24]. Without a properly functioning of heart, the body cannot survive. These situations can be instigated by various aspects, like drastic change in lipid profile which includes cholesterol, triglyceride and some other factors like diabetes, high blood pressure, smoking, sleeplessness, obesity, lack of physical activity, and family history [16]. Heart disease disturbs around 126 million individuals which is 1.72% of World’s population [11] and it causes one-third of deaths worldwide.

Some people may experience no symptoms at all, particularly in the early stages of the disease [25]. Early detection of CVD is crucial for timely intervention and prevention of adverse outcomes. With the recent advancements in machine learning (ML), there has been a growing interest in using ML algorithms for CVD recognition. ML algorithms have shown promising results in detecting CVD by analyzing various factors, such as demographic data, medical history, clinical examination results, and laboratory test results. In this context, multiple ML algorithms have been proposed, in particular LR, ANN, DT, RF, and SVM, among others. This approach involves training these algorithms with a dataset of CVD patients and non-CVD patients and using them to detect CVD in a new patient. The objective of our research work is to provide an overview of the numerous ML classifiers that have been proposed for CVD recognition and compare their performance based on numerous metrics, such as precision, recall, sensitivity, specificity and accuracy. Ultimately, this research could lead to the development of a reliable and efficient CVD recognition system. This system could aid healthcare professionals in making timely and accurate diagnoses, improving patient outcomes and quality of life [17]. Moreover, the efficiency of these algorithms is extremely reliant on the worth of input data. However, mobile health equipment can also be utilized to implant CVD recognition system using mobile gadgets. These mobile technologies will gather factual data of patients and will deliver proficient health services. Live monitoring of CVD patients can be controlled without making them visit to clinical health centers [18, 19]. Fig.1.1 shows Architecture of CVD Prediction system in E-Healthcare System.

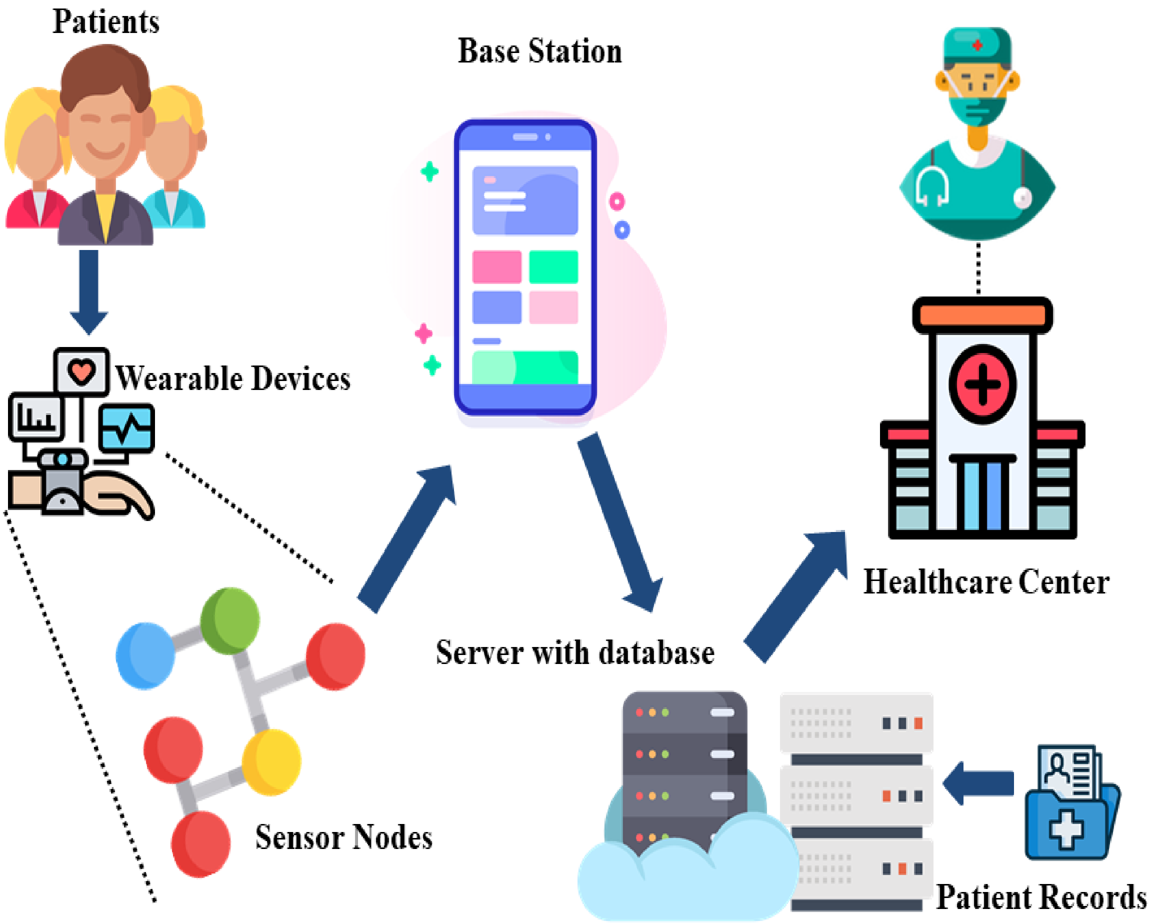


Figure 1. 1: Illustration of overall architecture of CVD prediction model used in E-Healthcare

Feature selection and feature scaling techniques can help to improve the quality of input data for ML classifiers. Feature selection methods aim to identify the most useful attributes that contribute to the accuracy of the model while reducing the dimensionality of the data. Feature scaling techniques, on the other hand, aim to standardize the data to improve the convergence rate of the ML algorithms.

The combinations of multiple ML algorithms, feature selection techniques and feature scaling methods have the potential to further enhance accuracy and robustness of CVD recognition models. By selecting the most relevant features and normalizing the data, these models can reduce over-fitting and improve generalization to new data. In addition, by combining the outputs of multiple ML algorithms, these models can capture complementary information and improve overall prediction performance.

The main objective and inspiration behind this research is to construct an effective model to detect CVD as accurate and precisely as possible. Fig. 1.2, represents block diagram of proposed hybrid model and required steps followed in this research are summarized as follow:-

- Five Datasets are combined to prepare an effective and mature dataset
- Data Preprocessing techniques min-max scaling and SMOTE are used for normalizing and Class balancing of data
- A Comparison of results is drawn which indicates the difference between without and with preprocessing techniques applied on dataset
- Various ML algorithm like SVM, NB, KNN, RF, LR were applied on UCI
- After preprocessing of data, all five applied algorithm were tempted with bagging method to achieve better results. Proposed Hybrid Model (HLRBM) overtakes in achieving accuracy.

- To endorse efficacy and performance, recommended model is applied to another CVD dataset having 7000 records with 12 attributes.
- A comparison of results is drawn with existing results of former researchers

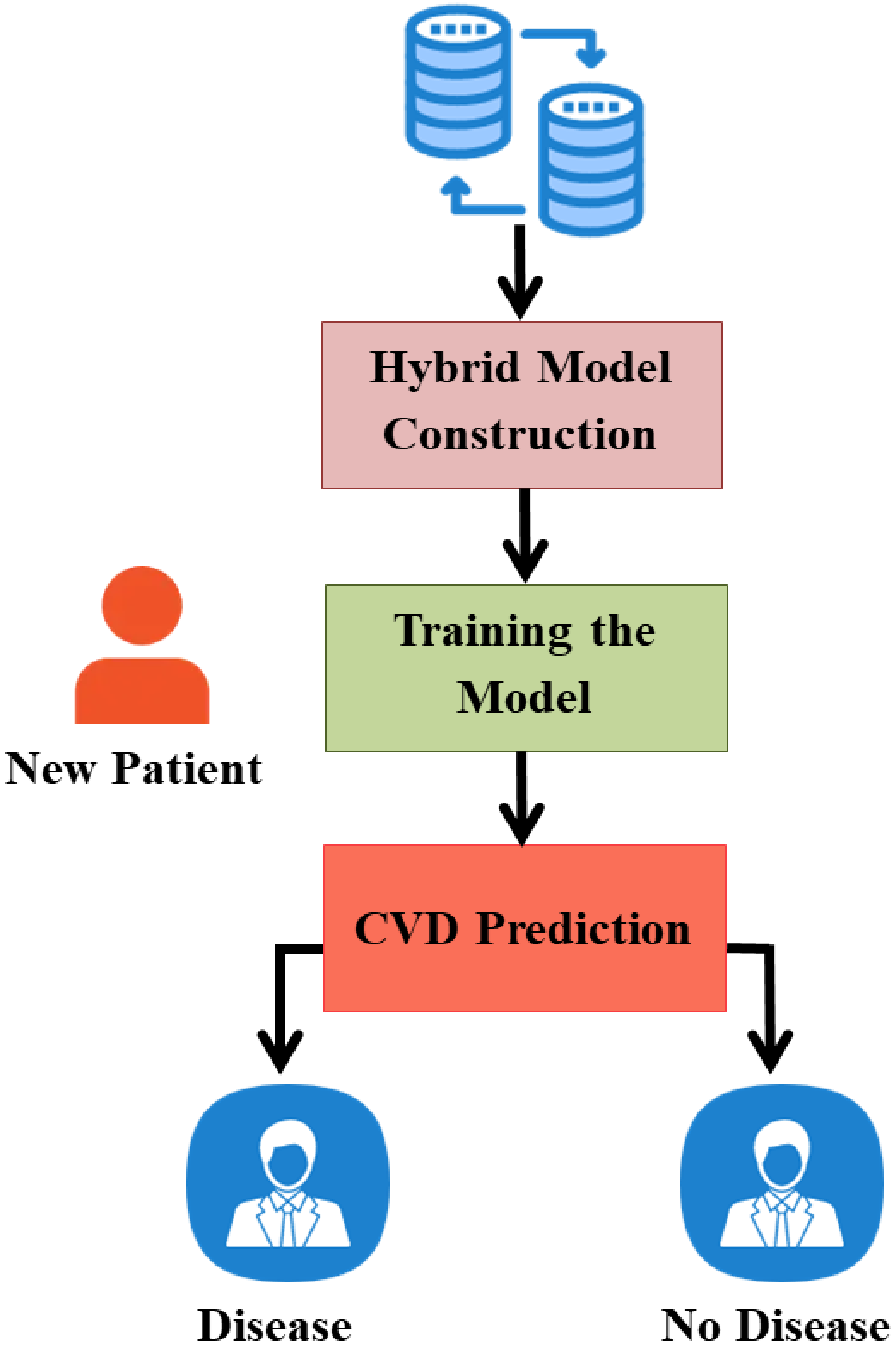


Figure 1. 2: Block Diagram of CVD Prediction Model

1.2 Motivation

Heart disease is a significant global health issue, impacting a large amount of people globally. Common symptoms include shortness of breath, physical weakness, and swollen feet. However, current diagnostic techniques for heart disease are often ineffective in early detection due to challenges related to accuracy and execution time [3]. In the absence of contemporary technology and medical professionals, the task of identifying and addressing heart disease becomes extremely arduous. The European Society of Cardiology reveals that roughly 26 million individuals receive a heart disease diagnosis on a yearly basis, encompassing around 3.6 million new cases each year [24]. The United States has a high prevalence of heart disease among its population.

Traditionally, detecting cardiac disease currently relies on evaluating a patient's medical background, performing physical evaluations, and considering pertinent symptoms under the supervision of a physician. Nevertheless, the effectiveness of this diagnostic approach in identifying individuals with heart disease is restricted. Furthermore, this method is costly and requires significant computational resources for analysis. Consequently, there is a requirement to create a noninvasive diagnostic system utilizing machine learning (ML) classifiers to tackle these obstacles.

Our approach involves analyzing multiple ML classifiers such as SVM, KNN, RF, NB and LR embedded with ensemble techniques to construct a Hybrid model using heart disease datasets to achieve the highest possible accuracy of results.

1.3 Problem Statement

Cardiovascular diseases (CVD) are a significant global health concern, contributing to a high number of deaths worldwide. Timely detection and intervention are crucial for improving patient outcomes and preventing the progression of CVD. The utilization of ML classifiers has become increasingly significant in the field of CVD identification, presenting an opportunity to support healthcare professionals in achieving precise and efficient diagnoses. This research project aims to investigate the efficacy of diverse ML algorithms, encompassing RF, SVM, NB, KNN and LR classifiers, in accurately detecting CVD. The performance of these algorithms will be assessed using various metrics, including accuracy, sensitivity, specificity, and area under the curve (AUC). The outcomes of this study have the potential to contribute towards the development of more dependable and precise tools for the recognition and diagnosis of CVD.

The primary purpose of this research is to construct an effective model that detect CVD with the highest possible accuracy and precision. The proposed Hybrid Model approach surpasses traditional rule-based methods, demonstrating superior performance in the detection of CVD. By leveraging the strengths of various ML algorithms, this approach contributes to more accurate and reliable CVD detection, offering healthcare professionals a powerful tool to aid in diagnosis and patient management.

1.4 Research Objectives

The main objectives of this research work are:-

- A comprehensive examination of ML techniques for accurate detection of heart disease.
- Evaluation of various supervised ML classifiers using heart disease datasets.
- Investigation of feature selection techniques to identify relevant features for analysis.
- Implementation of a proper training and testing methodology for the proposed model, ensuring separation of training and testing data.

- Comparison of different ML classifiers to determine the highest accuracy in predicting heart disease.
- Development of a recommended model that delivers optimal results across diverse datasets.

1.5 Relevance to National needs

The implementation of multiple ML techniques can aid in prediction of heart disease, potentially saving the lives of individuals with cardiovascular conditions. The proposed technique has the potential to identify various cardiac-related issues at an early stage.

1.6 Area of Application

Application of this ML technique can aid to following areas for CVD detection :-

- Hospitals either private or Government.
- Homes in case of emergency.
- Rescue services providers
- Healthcare Centers can also use it.

1.7 Advantages

Followings are the advantages of our research work :-

- Early detection of heart problems to prevent potential risks.
- Minimizing the risk of heart attacks by identifying issues at an early stage.
- Immediate assistance for patients through prompt detection of heart rate abnormalities.

- Prompt monitoring of a patient's heart functionality in response to uncertain readings from the heart and multiple attributes.

1.8 Thesis Organization

The research work has been organized and distributed in the following chapters:-

- **Chapter 1**: A brief introduction is given. Research objectives are listed. Relevance to National need is highlighted followed by area of application, its advantages and justification for selection of the topic is elaborated.
- **Chapter 2**: Describes related works carried out of various CVD databases related to CVD prediction system. A comparison is drawn to observe existing work by various researchers.
- **Chapter 3**: Discuss the overall research methodology including, Overview of Proposed model followed by the application and implementation of proposed model.
- **Chapter 4**: This Chapters presents the results and objective achieved by our proposed model
- **Chapter 5**: This Chapter sums up the research with conclusion drawn and provides direction for future work
- **Chapter 6**: includes References

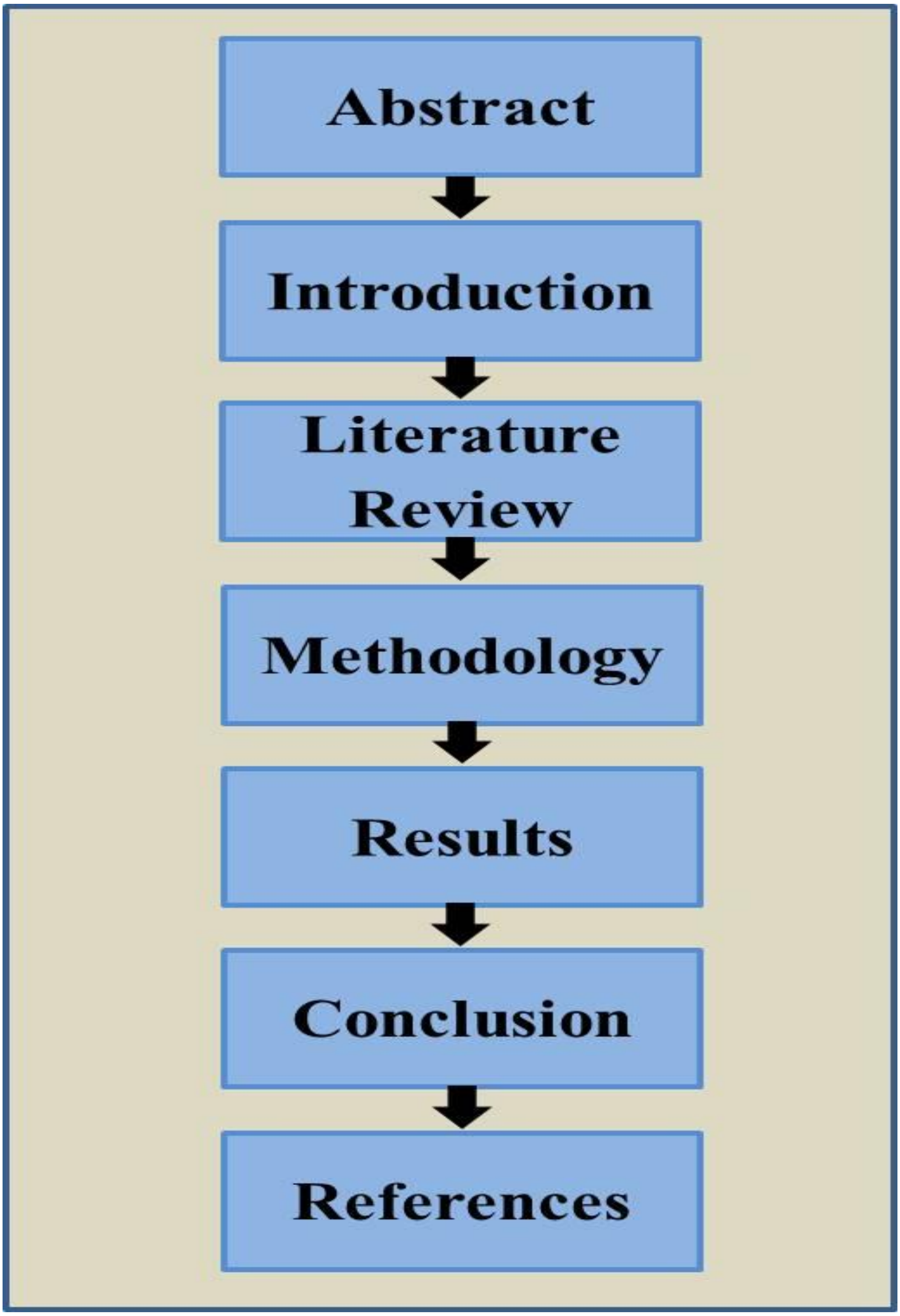


Figure 1. 3: Taxonomy of the thesis

Fig. 1.3, represents the layout of our thesis which is described in detailed in sec 1.8.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In Chapter 1, over all related work is described. It explains about the existing work carried out by numerous former researchers along with their findings and results. Use of traditional classifiers and various data preprocessing techniques are described. The chapter is concluded by giving a detailed comparison of all the work done on CVD dataset by various researchers.

2.2 Related Work

A literature review is an indispensable module of academic research, involving a critical examination of existing literature pertaining to a specific topic or research question. Its purpose is to comprehensively analyze, synthesize, and evaluate scholarly works to provide a bird eye view of the existing state of knowledge in the field. The primary purpose of conducting a literature review is to recognize significant themes, trends, and findings from previous research that can inform the development of new research questions or hypotheses. It serves as a foundation for further investigation and helps researchers situate their work within the existing body of knowledge.

There are various approaches to conducting a literature review, including deductive, inductive, thematic, and theoretical approaches. A deductive approach involves testing existing theories or hypotheses, while an inductive approach aims to generate new theories or hypotheses based on the literature. A thematic approach involves identifying emerging themes or concepts, while a theoretical approach utilizes existing theories to frame the review. Conducting a literature review requires strong research skills and a critical mindset. Researchers must locate relevant sources, assess the quality of studies, synthesize the findings, and draw meaningful conclusions. It is important to ensure that the review is comprehensive, unbiased, and transparent. A well-executed literature review offers several benefits to academic research, such as identifying research questions, defining the research problem, selecting appropriate research methods, and highlighting the significance of the study. It also helps identify gaps in the existing literature, paving the way for new research questions or hypotheses. For the purpose of conducting the literature survey, the following research papers were extensively examined and analyzed.

2.3 Importance of CVD prediction

CVD is a primary cause of mortality worldwide, proving the development of accurate and reliable predictive models an urgent task. In 2020, it has been reported that approximately 244.1 million people are surviving with heart diseases [12] whereas 19.1 million deaths were caused by CVD globally. ML techniques have shown great potential in predicting the risk of CVD, and several studies have been conducted to compare their performance. ML algorithms have emerged as favorable tools for predicting the risk of CVD based on a wide range of patient data. In this literature review, we summarize the recent research on CVD prediction using multiple ML algorithms. We examine the methodologies, datasets, and performance metrics used in various studies, and identify the strengths and limitations of different ML algorithms for CVD prediction. Some of the research works are addressed below:-

2.4 Use of Traditional Classifiers

In this paper [1], several ML techniques including RF, LR, SVM, NB and Adaboost (AB) were utilized to detect CVD. The Cleveland dataset was used from UCI repository in which missing values were accredited by MICE algorithm. The authors have improved results of classifiers with feature selection technique. Standard Scalar and Synthetic Minority Oversampling Technique (SMOTE) technique were also used for preprocessing of data. The suggested model by the authors has delivered accuracy of 86.6% which was better than all the applied techniques.

The system was established to diagnoses heart diseases [2] based on ML classifiers which includes ANN, LR, KNN, SVM, NB and DT. For removing immaterial and unessential features, the authors have used feature selection technique like Minimal redundancy maximal relevance, Relief, least absolute shrinkage selection operator (LASSO) and Local learning. Outcomes indicated the suggested diagnosis system, fast conditional mutual information feasible with SVM (FCMIM-SVM) attained better performance as related to formerly recommended models.

[3] In this research work, Switzerland, Hungarian, Long Beach VA and Cleveland CVD datasets from UCI database were used. Dataset contained total of 920 records along with 76 features associated with CVD. Preprocessing of data techniques were applied like elimination of noisy data, redundant values, filling of omitted values and sorting of attributes were carried out. The authors have taken only 14 attributes based on which heart diseases were diagnosed through proposed model. Multiple ML classifiers were used including RF, NB, SVM, Gradient Boosting (GB) and LR. The performance of each selected algorithm was obtained which include accuracy, sensitivity and specificity analysis.

2.5 Research Work on Cleveland CVD Dataset

In [4], A unique model was introduced in this paper by uniting five CVD datasets (Cleveland, Switzerland, Long Beach VA, Hungarian and Statlog) to make a larger, mature, and trustworthy dataset. For selection of suitable features LASSO and Relief techniques were utilized which also helped to overcome underfitting and overfitting glitches of machine learning. New hybrid classifiers were introduced by incorporating customary algorithms with bagging and boosting techniques which were used to train dataset.

[5] Authors have formulated a prediction model by using hybrid RF with a linear model (HRFLM) to improve its performance level. UCI Cleveland dataset was used on which an

analytical approach was applied with three connotation rules (apriori, predictive and Tertius) of mining to discover features of CVD. Authors have utilized R studio rattle to achieve CVD catalog. The proposed model HRFLM deliver improved results with an accuracy of 88.7%.

A fusion model was proposed to envisage CVD by utilizing DT and RF algorithm in [6]. Authors have applied both methods individually and later applied combination of these models which produced better results on Cleveland dataset collected from uci.edu. A basic GUI interface was formulated to predict heart disease by giving all required values as input that produced a binary classified calculation. Both models along with hybrid model were applied and mean square error, R-Squared parameter, mean absolute error, root mean square error and accuracy of applied models were calculated and plotted on a graph for a better comparison.

2.6 Research Analysis by former researchers on CVD Dataset

Rahul and Sunit [7] suggested a comparative study and analysis in oct 2020, of all the ML techniques used for prediction of CVD. Authors have given a detailed analysis of each model and advantages / disadvantages were discussed for each machine learning algorithm used for prediction of CVD. Comparison was done to check performance level of each method based on accuracy, precision, sensitivity, recall and error.

[8] Supervised ML classifiers like RF, NB, DT and KNN on UCI, Cleveland dataset were applied in this research paper. In order to attain precise and effective outcomes data preprocessing was done to avoid issues related missing values and noisy data. Performance level of each algorithm was determined and plotted based on precision, recall and accuracy resultantly K-nearest neighbor produced a highest accuracy as compared to former methods used for Cleveland dataset.

2.7 Research Work on CVD Dataset with 12 attributes

In this paper [9], CVD dataset was used which contained 7000 records with 12 attributes where 11 are input features and one output feature. The authors have applied various ML classifiers such as ANN, RF, DT, SVM, KNN and ANN. The optimal results were attained by ANN along with the use of Genetic algorithm (GA). The authors claimed to produce 5.08 percentage improvements of results as compared to other algorithms applied. The results were achieved by using GA-ANN with 3 layers as prominent parameters and 64 neurons. For better results, Softmax function and adagrad optimizer were used which gave 73.3% average accuracy.

In [10], the authors have recommended a hybrid model using KNN and ANN for CVD prediction. Cleveland dataset from UCI was used which have 14 features with 303 records. In this dataset 13 attributes are input features where as one feature is output indicating presence or absence of heart disease against data of each patient. Table 2.1, illustrates the overall existing works described in related work.

Table 2. 1: Comparison of existing work for CVD Prediction by various Former Researchers

Ser. No.	Name of Research Paper	Method Used	Observations	Year of publication	Dataset Used
1.	A decision support system for heart disease prediction based upon machine learning	Naïve Bayes, SVM, Random Forest, Decision Tree, A Hybrid Model by combining GA and RFE algorithms	This paper presents use of various pre processing techniques like feature scaling, class balancing and feature selection methods. Missing values were accredited by MICE algorithm and produced 86.6 % accuracy	2021	UCI, Cleveland Heart Disease Dataset
2.	Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare	ANN, LR, KNN, SVM, NB and DT	In this paper, Feature selection algorithms like Minimal redundancy maximal relevance, Relief, least	2020	UCI, Cleveland Heart Disease Dataset

			absolute shrinkage selection operator and Local learning were used. Performance of used classifiers was tested on particular features designated by feature selection algorithms. Results show that the suggested diagnosis system, fast conditional mutual information feasible with SVM (FCMIM-SVM) attained better accuracy as compared to formerly proposed methods		
3.	Prediction of Cardiovascular Disease Using	SVM, GB, RF, NB and LR	In this paper, Pre-processing of data techniques were	2018	UCI, Cleveland Heart Disease Dataset

	Machine Learning Algorithms+		applied like elimination of noisy data, redundant values, filling of omitted values and sorting of attributes were carried out. The accuracy achieved by the finest algorithm that gave 86.51% accuracy was logistic regression for prediction of CVD as compared to other applied algorithms		
4.	Efficient Prediction of CVD Using Machine Learning Algorithms With Relief and LASSO Feature Selection	DTBM, RFBM, KNNBM, ABBM, GBBM	In this case study, New hybrid classifiers were introduced by incorporating customary classifiers with	2021	Combination of five datasets (Cleveland, Switzerland, Long Beach VA, Hungarian and Stat log) were

	Techniques		bagging and boosting techniques which were used to train dataset Least Absolute Shrinkage and Selection Operator (LASSO) and Relief techniques ware used for feature selection		used
5.	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	NB, LR, DT, RF, SVM and HRFLM (Hybrid)	In this paper, Authors have formulated a prediction model by using hybrid random forest with a linear model (HRFLM) to improve its performance level. Authors have utilized R studio rattle to achieve	2019	UCI, Cleveland Heart Disease Dataset

			CVD catalog of Cleveland dataset from UCI repository. The proposed model HRFLM produced better results with an accuracy of 88.4%.		
6.	Heart Disease Prediction using Hybrid machine Learning Model	DT, RF and Hybrid (DT+RF) algorithms were used	In this research work, A basic GUI interface was formulated to predict heart disease by giving all required values as input that produced a binary classified calculation that means 0 for negative and 1 for positive in case of heart disease. Both models along with	2021	UCI, Cleveland Heart Disease Dataset

			hybrid model were applied and mean square error, R-Squared parameter, mean absolute error, root mean square error and accuracy of applied models were calculated and plotted on a graph for a better comparison. Outcome of this research resulted into achieving accuracy of 88.7% with proposed Model.		
7.	Machine Learning Techniques for Heart Disease Prediction: A Comparative Study	Review paper	Authors suggested a comparative study and analysis in oct 2020, of all the ML techniques used for prediction	2020	Comparative study

	and Analysis		of CVD. Authors have given a detailed analysis of each model and advantages / disadvantages were discussed for each machine learning algorithm used for prediction of CVD. Comparison was done to check performance level of each method based on accuracy, root mean square error, sensitivity, precision and recall		
8.	Heart Disease Prediction using Machine Learning Techniques	RF, NB, DT and KNN	This research work illustrated that In order to attain precise and effective outcomes data pre-processing was done to avoid	2020	UCI, Cleveland Heart Disease Dataset

			issues related missing values and noisy data. Performance level of each algorithm was determined and plotted based on precision, recall and accuracy resultantly K-nearest neighbor produced a highest accuracy as compared to former methods used for Cleveland dataset		
9.	An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction	ANN, RF, DT, SVM, KNN and ANN	The authors claimed to produce 5.08 percentage improvements of results as compared to other algorithms applied. The results were achieved by using GA-ANN	2022	CVD Dataset having 70000 records

			with 3 layers as prominent parameters and 64 neurons. For better results, Softmax function and adagrad optimizer were used which gave 73.3% average accuracy		
10.	PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY	KNN and ANN	In this Paper, the authors have proposed a hybrid model using K-nearest neighbor and ANN for prediction of heart disease. UCI Cleveland dataset was used which have 14 attributes with 303 records. In this dataset 13 attributes are input features where as	2017	UCI, Cleveland Heart Disease Dataset

			one feature is output indicating presence or absence of heart disease against data of each patient		

2.8 Summary

The studies reviewed in this literature cover various types of CVD datasets used for Heart Disease prediction. Various ML algorithms are utilized by numerous researchers to achieve better accuracy of Heart prediction. Multiple data preprocessing techniques were used to normalized data and class balancing of data and furthermore, various feature selection techniques were used mentioned in literature review. Various classifiers like SVM, KNN, RF, DT, GB, ANN, LR and Linear models were also used.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the overall functioning and methodology of proposed model. An overview of proposed model is described with diagram containing workflow of proposed model to be followed for CVD detection. Fourteen Attributes of CVD dataset used in this work are properly described along with table 3.1, containing description of each feature, values ranges and data types. Application of proposed model is explained with the help of fig 3.2. Next subpart of this chapter, describes the justification of proposed model with co-relation of attributes of Cleveland dataset being utilized. Pseudo code of proposed algorithm is written which easily indicates the flow construction of Hybrid model. A detailed diagram is drawn which tells about how a CVD will be detected by using proposed Hybrid model.

3.2 Overview of Proposed Model

In this study, a CVD prediction system has been developed by the authors. The three phases of this hybrid system are data gathering, data preprocessing and the model creation. Missing data is imputed, features are chosen, features are scaled, and class balancing is carried out during the preprocessing stage. Before applying classifiers data must be standardized or normalized. The standard scalar is used to standardized the data, guaranteeing that each feature has a mean of 0 ($\boldsymbol{\mu}$) and a standard deviation ($\sum$) of 1. Transformation formula from [21] is appended below:-

$$\boldsymbol{Standardization, X} = \frac{\boldsymbol{X} - \mu}{\boldsymbol{\sigma}} \qquad (1)$$

The SMOTE technique is also used for class balancing to handle imbalance data. ML traditional classifiers like SVM, NB, KNN, RF and LR were applied on selected features. After data preprocessing, the data was divided into 80% of training and remaining 20% of data into test data. Various ensemble methods with traditional classifiers are imputed to create a combination over same dataset. Finally, the classifier determines if the person is CVD positive or negative. A remarkable difference is observed in results while applying these classifiers when used without data preprocessing and after preprocessing techniques. Different training methods are applied to check the performance of each model so that we can choose finest hybrid model for our trustworthy dataset. However, our proposed model HLRBM resulted in providing more accurate results than other models. Fig. 3.1, depicts the framework and application technique of recommended hybrid model for CVD prediction.

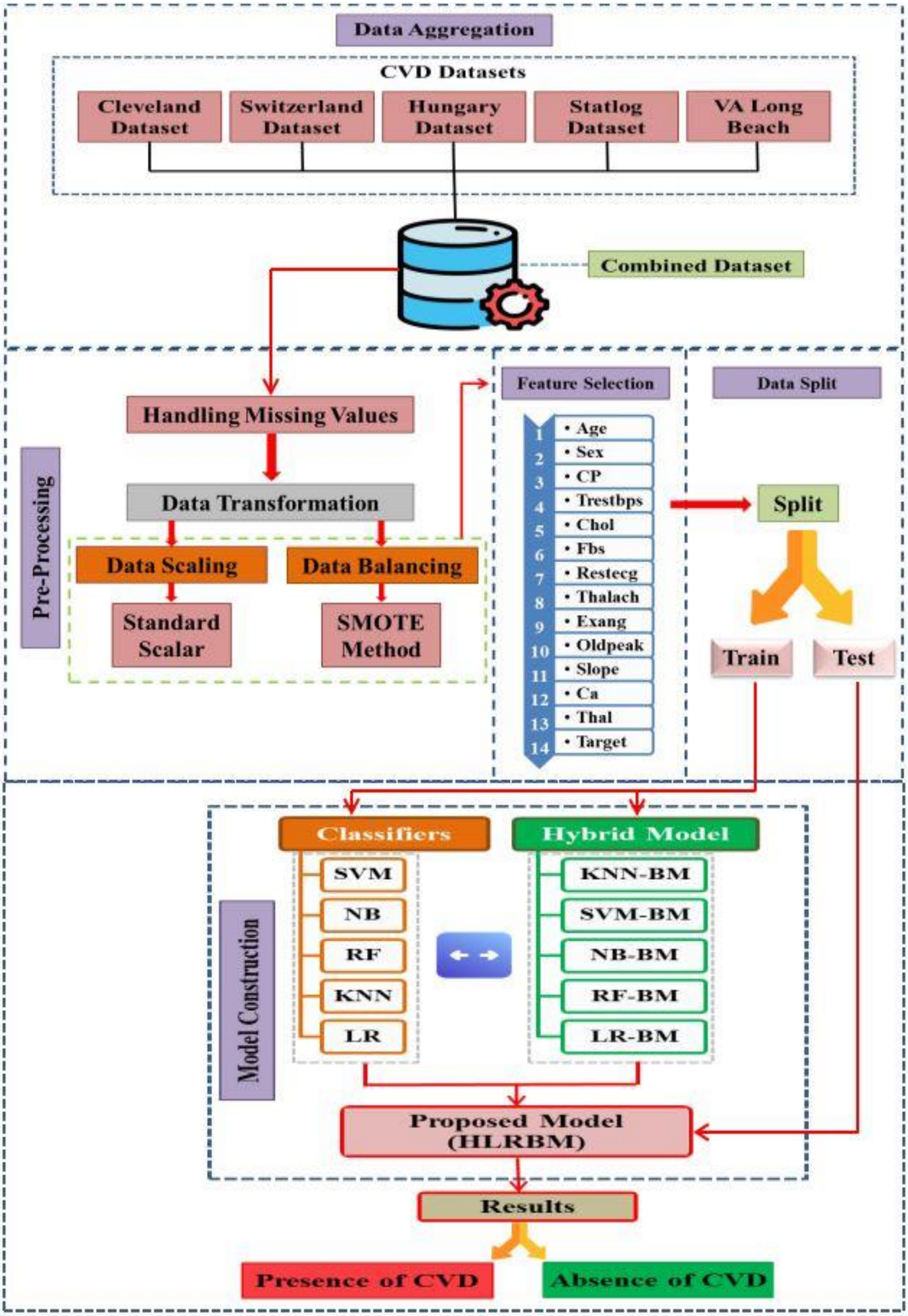


Figure 3. 1: Suggested Architecture of proposed Model for CVD detection

3.3 Dataset Description

The datasets available on UCI repository [14] [15] were utilized to construct an effective dataset by merging five CVD datasets (Switzerland, VA long Beach, Cleveland, Statlog and Hungary). Table 3.2, describes the dataset's attributes along with their value ranges. The dataset used in this research work contains 303 records having 14 attributes out of which 13 are input features along with 1 output feature indicating presence or absence of CVD. Following are the comprehensive depiction of each attribute:-

- **Age**: The age of a person in completed years. This feature signifies the patient's age at the time of data collection.
- **Sex**: This attribute represents the patient's gender and is encoded as 0 indicating female and 1 indicating male.
- **Chest Pain Type**: This attribute provides information about the chest pain encountered by the patient. It is classified into four categories: 1 denoting typical angina, 2 indicating atypical angina, 3 for non-angina pain and 4 indicating asymptomatic.
- **Resting Blood Pressure**: This feature symbolizes the patient's blood pressure while at rest, measured in mm Hg (millimeters of mercury) upon admission in hospital.
- **Serum Cholesterol**: The cholesterol level is measured in milligrams per deciliter (mg/dl) indicates the amount of serum cholesterol present in the patient's blood.
- **Blood Sugar in fasting**: This attribute provides information about the patient's fasting blood sugar level, measured in mg/dl. A value of 1 indicates high blood

sugar, defined as greater than 120 mg/dl, while a value of 0 indicates normal blood sugar, equal to or less than 120 mg/dl.

- **Resting Electrocardiographic Results**: This attribute represents the results of the patient's resting electrocardiogram (ECG) test. It is encoded as follows: 0 denoting normal, 1 indicating ST-T wave abnormality, or 2 presenting apparent or certain left ventricular hypertrophy.
- **Maximum Heart Rate**: This attribute indicates the peak heart rate attained by the patient during exercise.
- **Exercise-Induced Angina**: This attribute indicates whether the patient experienced angina (chest pain) while exercising. A value of 1 indicates the presence of induced angina, while a value of 0 signifies the absence of angina.
- **ST Depression Induced by Exercise Relative to Rest**: This feature measures the amount of ST depression observed during exercise compared to the patient’s resting stage.
- **Slope of the Peak Exercise ST Segment**: This attribute represents the shape of the ST segment during the peak exercise, classified as 1 denotes upsloping, 2 indicating flat or 3 for downsloping.
- **Number of Major Vessels Colored by Fluoroscopy:** This attribute denotes the count of major blood vessels that have been visualized and colored through fluoroscopy. The possible values range from 0 to 3.
- **Thallium Stress Test Results**: The attribute shows the results of the thallium stress test, which measures blood flow to the heart. It is encoded as 3 indicating normal, 6 denoting fixed defect, and 7 represents reversible defect.

- **Target**: The target variable indicates the presence of cardiovascular disease. A 0 value means no presence of disease, while reading greater than 0 represents the presence of disease.

Table 3. 1: Dataset Description in details including value ranges

No.	Feature Name	Feature Code	Description	Values Range	Data Type
1	Age	Age	Age in years completed	between 29 and 77	Numeric
2	Sex	Sex	Male: 1, female: 0	0 or 1	Nominal
3	Type of chest pain	CP	Typical angina: 1, atypical angina: 2 non-angina pain: 3, asymptomatic: 4	1 to 4	Nominal
4	Resting blood pressure	Trestbps	Patient's Resting Blood Pressure Range	94 to 200 mm Hg	Numeric
5	Serum cholesterol	Chol	Cholesterol level in mg/dl	126 to 564 mg/dl	Numeric
6	Fasting blood sugar	Fbs	Fasting Blood Sugar >120 mg/dl (true:1, false: 0)	0 or 1	Nominal
7	Resting electrocardiographic results	Restecg	Normal: 0, ST-T wave abnormality:1, Hypertrophy: 2)	0, 1 and 2	Nominal

8	Maximum heart rate achieved	Thalach	Heart Rate of Patients	71 to 202 HR	Numeric
9	Exercise-induced angina	Exang	Patient experienced angina during exercise (Yes=1, No=0)	0 or 1	Nominal
10	ST depression induced by exercise relative to rest	Oldpeak	Depression caused by exercise, Up sloping: 1, Flat: 2, down sloping: 3	1 to 3	Numeric
11	The slope of the peak exercise ST segment	Slope	Slope of peak exercise	1, 2, 3	Nominal
12	Number of major vessels (0–3) colored by fluoroscopy	Ca	Major Vessels colored by fluoroscopy with range 0 to 3	0 to 3	Numeric
13	Thallium	Thal	Represents thallium stress test, Normal:3, fixed defect: 6, reversible defect: 7	3, 6, 7	Nominal
14	Target	Target	Output, Heart disease present: 1, heart disease absent: 0	0 or 1	Nominal

3.4 Application of Proposed Hybrid Model

The suggested model is prove to be an efficient model once its suitable application is justified and it also aid to deal with real world challenges. Fig. 3.2, illustrates the workflow of our proposed model. This intelligent devised model can be utilized in various health centers to predict CVD in an effective way. The following procedure can be followed to attain prediction of CVD.

- Data of each patient is collected and put together in a database
- The main features of patient's data will be selected as an input to our proposed model HLRBM to perform prediction
- Selected features will be handled in our trained model
- As a result, binary output will be generated either 0 or 1. Result 0 identifies as negative (absence of CVD). 1 in case of CVD results are positive
- In case of 1 patient will be guided to visit Heart Specialists for further investigations
- Data of each patient will be preserved in a database to aid better prediction in future

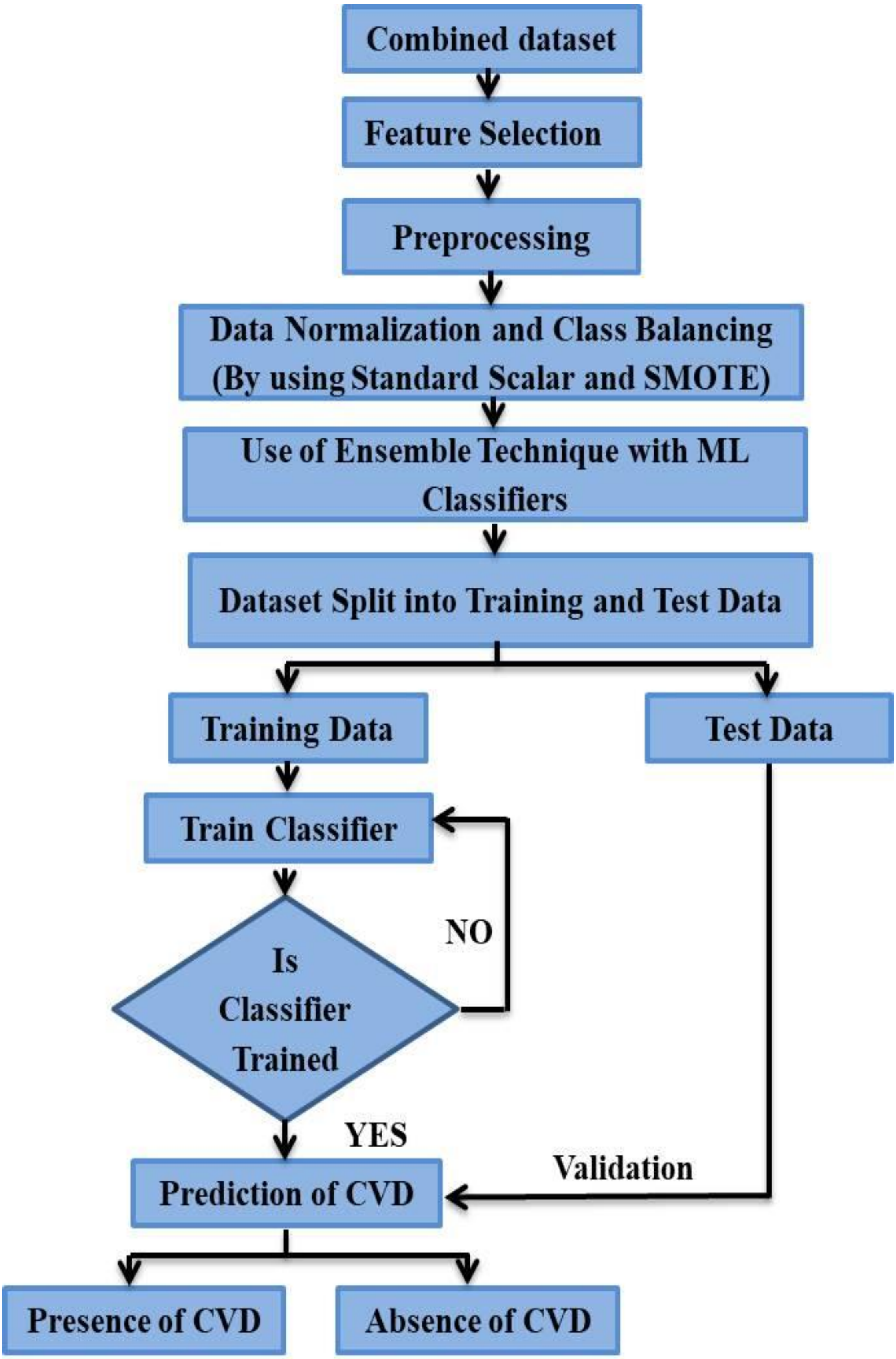


Figure 3. 2: Proposed CVD Prediction System

3.5 Justification of Proposed Hybrid Model

The unique hybrid model has been devised by using five various ML classifiers. Consequently, an ensemble Bagging technique is used to make this model efficient and persistent in achieving results. data preprocessing techniques used for data normalization and class balancing remains the hallmark difference of former results. After achieving improvements in result in data preprocessing phase, an ensemble bagging techniques is applied as base classifier which is being embedded with other classifiers. Logistic regression classifier embedded with Bagging techniques as based classifier proved to be our proposed model. All the five models produced better results where as our proposed model HLRBM overtakes in achieving accuracy. Our paper highlights the former researchers results which have used ensemble technique with various ML classifiers and their results are shown in our comparison to our achieved results.

Various studies have already been conducted by using various ML algorithm that contracts with the same dataset don’t produced better results as expected. After an exhausted study we came to a conclusion that some models don’t perform well as those systems don’t identify most important and highly co-related features. Fig. 3.3, represents the co-relation of feature of CVD dataset. For that matter we tried to split highly co-related features and grouped together. Some features are having numeric values and some features are using nominal values according to their nature of readings.

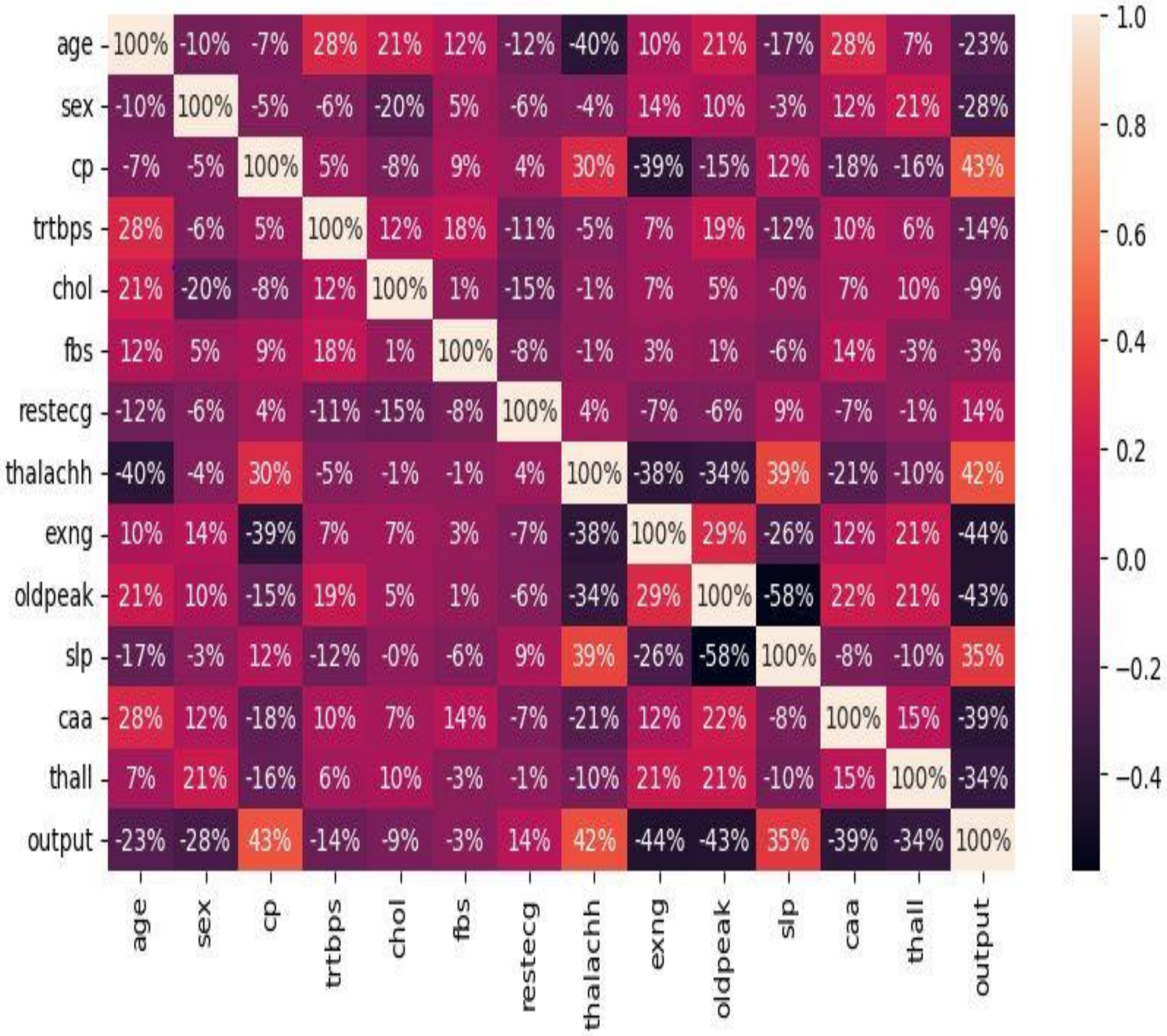


Figure 3. 3: Representation of Co-related features of CVD Dataset

We tried to make it unique while using our proposed model to another CVD dataset publically available having 70000 records with 12 attributes. 11 attributes are serves as input features where the 12th feature is output feature giving result in binary form either 0 or 1 indicating presence or absence of CVD. A comparison is carried out of former outputs with improved results achieved by our proposed model on same dataset having 70000 rows (record of patients). Pseudo code of proposed model is illustrated below:-

Pseudo code 1: Pseudo code of proposed Model

```
Input: CVD Datasets, ML Classifiers
Results: CVD Accuracy, precision, recall, F1score and ROC-AUC
BEGIN
Datasets ←{UCI (Cleveland, Hungarian, Switzerland, statlog, VA long Beach)};
Scalars ←{Standard scalar (min-max), SMOTE};
Classifiers ←{SVM, NB, KNN, RF, LR}
SMOTE-SVM, SMOTE-NB,SMOTE-KNN, SMOTE-RF, SMOTE-LR
For model ∈ Classifiers do
    Step2: model←{TrainClassifier{ScaledX_train, X_trainTarget};
    Step3: model.fit← {X_train, Y_train};
    Step4: y_predict← {X_test};
    Step5: Accuracy←ComputeAccuracy{y_predict, X_TestTarget};
    Step6: Precision←ComputeAccuracy{ y_predict, X_TestTarget};
    Step7: Recall←ComputeAccuracy{ y_predict, X_TestTarget};
    Step8: F1score←ComputeAccuracy{ y_predict, X_TestTarget};
    Step9: ROC-AUC score←ComputeAccuracy{ y_predict, X_TestTarget};
End for
Return {Accuracy, Precision, Recall, F1 score, ROC-AUC}
END
```

Fig. 3.4, represents overall working and the steps taken in selecting proposed model from where we selected our proposed hybrid Model (HLRBM)

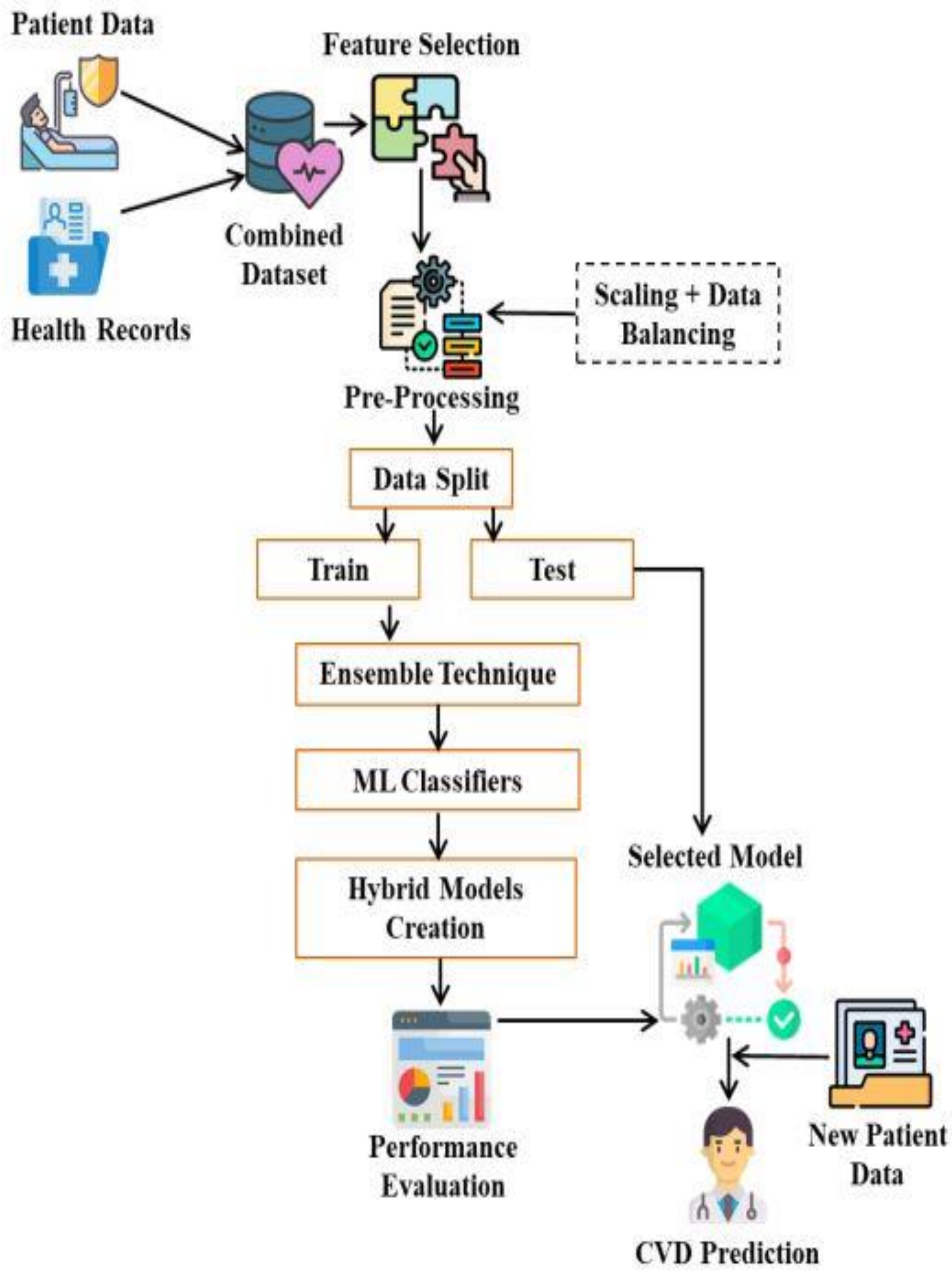


Figure 3. 4: Workflow of Proposed CVD Prediction System

3.6 Summary

In this chapter, an overview of proposed model is given followed by a detailed diagram of suggested Architecture of proposed model. The dataset is described along with a comprehensive discussion on dataset is discussed in table 3.1. Application of proposed model in real world is discussed in points. A complete workflow diagram is also appended for better understanding. Furthermore, Justification of proposed model also elaborated supported by co-relation matrix of all 14 features used in this research work followed by pseudo code of proposed model.

CHAPTER 4

IMPLEMENTATION OF PROPOSED MODEL

4.1 Introduction

In this chapter, implementation of proposed hybrid model HLRBM is described in details which includes environment used for this research work, data preprocessing and ensemble techniques applied. All the ML classifiers utilized for this research are described in details followed by evaluation of parameters.

4.2 Environment

All the experiments are executed using language Python 3 in Google Colab. 8GB RAM with windows 10 on Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz is used in this research work.

4.3 Data Pre-processing Techniques

In this modern era, a large amount of data can be collected via the internet, different valuable experiments and surveys etc. Most of the time data collected for research is noisy data and contain missing or null values. In this case, some popular techniques are applied like deletion and imputation that can deal with missing values. Furthermore, before applying any kind of ML algorithm data must be normalized or class balancing is executed to make an efficient dataset. We use standard scalar min-max scaling that normalized the data and SMOTE technique was used to balance the dataset. The techniques used for preprocessing of data are described as:-

4.3.1 Standard Scalar

Standard Scalar is a widely used scaling technique in ML. This normalization process guarantees that all features are uniformly scaled and prevents certain features from exerting excessive influence on the learning procedure. By standardizing the features, convergence of optimization algorithms is improved, facilitating faster and more stable training. It also enables fair comparison and interpretation of feature effects on the model. Standard Scalar is particularly beneficial for algorithms relying on distance metrics and is commonly applied in SVM, linear regression, LR, and neural networks.

4.3.2 SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a method used in ML to address imbalanced datasets. By interpolating between existing instances of the minority class, this technique generates mockups to augment the minority class. SMOTE helps in class balancing, reducing over fitting and improving model performance. It retains the inherent characteristics of the minority class while increasing the number of samples. However, it is important to be cautious of potential noise or overgeneralization introduced by the synthetic samples. Evaluating the performance with and without SMOTE is recommended to determine its suitability in a given scenario.

After data preprocessing phase, the data is distributed into 80% of training and remaining 20% into Test data. The algorithms were trained and on training data followed by achievement of results from Test data. Five algorithms which include SVM, NB, KNN, RF and LR were applied on dataset for achieving results. A comparison is drawn of attained results of with and without using data preprocessing techniques.

4.4 Ensemble Techniques of ML

Ensemble methods are assorted with classifiers to achieve better results. The main purpose behind this is that weak learners combined can work with strong learners to become an efficient combination [20]. Fig. 5 illustrates the ensemble process [20]. Mainly Bagging and Boosting techniques are used to produced more accurate and efficient results. In our research work we used Bagging method which is described below:-

4.4.1 Bagging Method

Bagging (Bootstrap Aggregating) is a ML ensemble technique where multiple models are trained on various subsections of training data, created through bootstrapping (sampling with replacement). The main purpose is to generate numerous divisions of data out of training models. Arbitrarily elected pools of subset data are recycled to train their DT. Resultantly, we acquire an ensemble of various models [22]. The individual models, typically decision trees, are then combined through voting or averaging to make predictions, resulting in improved accuracy and reduced variance. Bagging helps to reduce over fitting and improve generalization by leveraging the diversity among the models. It also helps resolving of missing values problems and preserves accuracy.

4.4.2 Boosting Method

Boosting is a ML ensemble technique which sequentially trains a series of weak models to create a robust model. Every fragile model are trained on a modified version of the training data, where the misclassified samples are given higher weights. The absolute prediction is made by joining the estimation of all the fragile models. Boosting iteratively focuses on the trials that are hard to categorize, continuously improving the model's performance. It helps to reduce bias and increase accuracy by emphasizing the learning from previously misclassified examples. Usually Boosting builds better predictive models [23].

In our research work we utilized Bagging technique with five classifiers which include SVM, NB, KNN, RF and LR to formulate hybrid models. The hybrid Models: SVM-BM, NB-BM, KNN-BM, RF-BM and HLRBM are constructed and applied on training and Test data of our dataset. Resultantly, our proposed model HLRBM outperformed and produced better results among other models.

4.5 Classification Modeling

Multiple classifiers used in our research work which are embedded with ensemble technique. Each model has its own impact on dataset. Specific description of used algorithms are given below:-

4.5.1 SVM

A Support Vector Machine (SVM) is a supervised ML classifier that splits data into various modules by discovering the best hyper plane in a high-dimensional attribute space. It exploits the edges, the distance among the hyper plane and the nearest data points from each class, to improve generalization. SVM can handle non-linear boundaries using the "kernel trick" that transforms the feature space. They are trained by minimizing classification errors while maximizing the margin. SVMs are used for classification and regression tasks, and their applications include text categorization, image classification, and financial forecasting. They are powerful, versatile algorithms that excel in handling high-dimensional data and complex patterns.

4.5.2 Naïve Bayes

The NB algorithm is a simple yet powerful ML classifier based on Bayes' hypothesis. It accepts that all attributes in a dataset are independent of each other, hence the "naive" assumption. The probability of a particular instance belonging to a specific class is calculated by multiplying the conditional probabilities of each feature, given that class the algorithm works well with large datasets and is computationally efficient. Despite its simplifications, NB often produces competitive results, making it a popular choice in machine learning applications.

$$\mathrm{P(A|B) = P(B|A) * \frac{P(A)}{P(B)}} \tag{2}$$

Here, the probability we aim to calculate, P(A|B), is referred to as the posterior probability, while the prior probability of the event, P(A), is known as the marginal probability [26].

4.5.3 Logistic Regression

The LR algorithm is supervised ML technique utilized for binary classification related problems. It simulates the association among input features and the probability of belonging to a particular class by the logistic function. LR estimates the parameters by minimizing the logistic loss function, typically using gradient descent optimization. During training, for a given input, LR computes the possibility for each class and assigns the predicted class as the one with the highest probability. LR is popular due to its simplicity, interpretability, and efficiency. It is widely used in numerous applications, such as disease diagnosis, sentiment analysis and spam detection where the aim is to predict binary outcomes based on input features.

4.5.4 Random Forest

The RF algorithm is a ML technique that utilizes an ensemble approach by combining the predictions of multiple decision trees. This combination allows for more accurate and reliable predictions. By randomly selecting subsets of the training data and features for each tree, the algorithm creates a "forest" of DT. During training, each tree learns patterns and makes predictions independently. When making predictions, the algorithm aggregates the predictions of all the trees to determine the final outcome through voting or averaging. They are used for classification and regression tasks, and their applications include areas such as finance, healthcare, and image recognition, where accurate predictions and interpretability are essential.

4.5.5 KNN

This algorithm is most versatile and intuitive classification algorithm. To classify new instances, this approach identifies the K nearest neighbors in the training set and assigns the majority class label among them [27]. KNN measures the similarity between instances using a distance metric, typically Euclidean distance. KNN relies on the supposition that instances in the similar class are adjacent to each other in attribute space. Being a non-parametric algorithm, it does not rely on explicit assumptions regarding the underlying data distribution. KNN has the ability to handle classification problems with multiple classes and is also applicable in regression tasks. While simple and easy to implement, KNN's performance can be sensitive to the choice of K and the feature scaling.

4.6 Evaluation Parameters

On a scale of precision, recall, F1 score, accuracy and ROC-AUC classifiers performance were assessed. If a patient with the condition is anticipated the system determines the person to have cardiac disease, then the result is a true positive; in other case it's a false negative. Similar to the last example, a prediction that a healthy person will remain disease-free is said to be a true negative; and false positive in other case. These terms are precisely defined below [28]:-

- **True Positive (TP):** Instances correctly identified as positive when they are truly positive.
- **True Negative (TN):** Instances correctly identified as negative when they are truly negative
- **False Positive (FP):** Instances incorrectly identified as positive when they are actually negative.

- **False Negative (FN):** Instances incorrectly identified as negative when they are actually positive
- **Accuracy**: It is a performance parameter that gauges the system's propensity for accurate prediction.

$$\boldsymbol{Accuracy = \frac{TN+TP}{TN+FP+TP+FN}} \tag{3}$$

- **Precision**: Precision measures the capability of a system to produce only relevant results.

$$\boldsymbol{Presision = \frac{TP}{TP+FP}} \tag{4}$$

- **Recall:** Recall is the measure of the model's ability to identify all positive instances correctly, indicating the proportion of true positives out of all actual positives.

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \tag{5}$$

- **F-Measure**: F-Measure combines results of precision and sensitivity using harmonic mean.

$$\boldsymbol{F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall}} \tag{6}$$

4.7 Experimental Work with proposed Hybrid Model

The technique that is used to improve accuracy of results are utilizing more time on data preprocessing techniques. Resultantly, better accuracies are achieved in each applied model. These 12 attributes provide a range of information about an individual's demographics, physical characteristics, lifestyle factors, and medical history. By analyzing and modeling these attributes, researchers and healthcare professionals can have perceptions of the risk factors linked with CVD and make strategies for anticipation and treatment.

To check the efficacy of our recommended model we used another Dataset [13] from kaggle ML repository. The dataset contains 70000 records with 12 attributes. Out of 12 attributes 1 to 11 attributes are considered to be input features and 12th attribute is output feature.

The dataset [13] includes attributes such as age in completed years, sex (male or female), height, weight, systolic blood pressure, diastolic blood pressure, cholesterol levels, glucose levels, smoking status, alcohol consumption, physical activity level and the presence or absence of CVD. A comparison of results is drawn in results and analysis section which indicates that result achieved by our recommended model are better than results achieved at [9].

4.8 Summary

Implementation of complete proposed model is discussed in this chapter. Data preprocessing techniques are discussed and how these techniques enhanced the results are discussed in details. Whereas all the five traditional ML classifiers are explained followed by evaluation parameters are highlighted based on which all the models were examined efficiently.

CHAPTER 5

RESULTS AND ANALYSIS

5.1 Introduction

The results section of a thesis is a critical component that presents the findings of the research study in a comprehensive and meaningful manner. It is an opportunity for the researcher to analyse and interpret the data gathered from various research methods and present them in an organized and structured way. The results section is crucial in demonstrating the validity of the research and the extent to which the research question or hypothesis has been answered.

In conclusion, the results section of a thesis is a critical component that presents the findings of the research study. The researcher should pay attention to the organization of the data, the language used, and the critical evaluation of the results presented. A well-crafted results section will provide insights into the validity of the research and the implications for the field of study. The methodology that we adopted here for our thesis gave us very promising results.

A range of classification methods is employed to diagnose patients with heart disease, including SVM, NB, KNN, RF and LR. The UCI Cleveland dataset is used for the studies. The diagnosis of heart disease was made using several medical parameters extracted from the dataset. The classification was performed by utilizing these factors, where class 1 indicated the presence of an illness and class 0 indicated the absence of a disease.

5.2 Performance of classifiers without preprocessing

Firstly, experiments are performed on all selected 14 features without applying any kind of data preprocessing technique. System performance is evaluated using metrics such as accuracy, precision, recall, ROC-AUC, and F-measure. Table 5.1, Indicates the results of five classifiers applied on all 14 features without using any data preprocessing techniques :-

Table 5. 1: Performance of classifiers without applying preprocessing techniques

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	68.85	67	85	75	68
NB	81.96	81	88	85	82
KNN	63.93	71	59	65	64
Random Forest	83.64	83	88	86	84
LR	82.01	81	88	85	82

5.3 Performance of Classifiers with Scaling and Class Balancing technique

Results achieved indicate that scaling and class balancing have positive influence on each classifier. Standard Scalar and SMOTE technique used in research work proved to be a best suitable on our dataset. Table 5.2, indicates improvements of results achieved by data preprocessing techniques are better than the former ones. Overall, accuracy of each classifier is improved where specifically accuracy increased by SVM is 18.03%, NB increased by 3.56%, KNN improvement is 18.03%, RF increased by 1.6% and LR improved by 3.23%

Table 5. 2: Performance improvement with pre-processing techniques:

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	86.88	84	94	89	87
NB	85.52	91	88	90	87
KNN	81.96	81	88	85	82
RF	85.24	88	85	87	85
LR	85.24	86	88	87	85

Fig.[5.1- -5.5] illustrate that with the use of scaling and classing balancing technique, results of each evaluation parameters in term of accuracy, precision, recall and F1 score are improved. Graphical representation of each parameter is indicated below:-

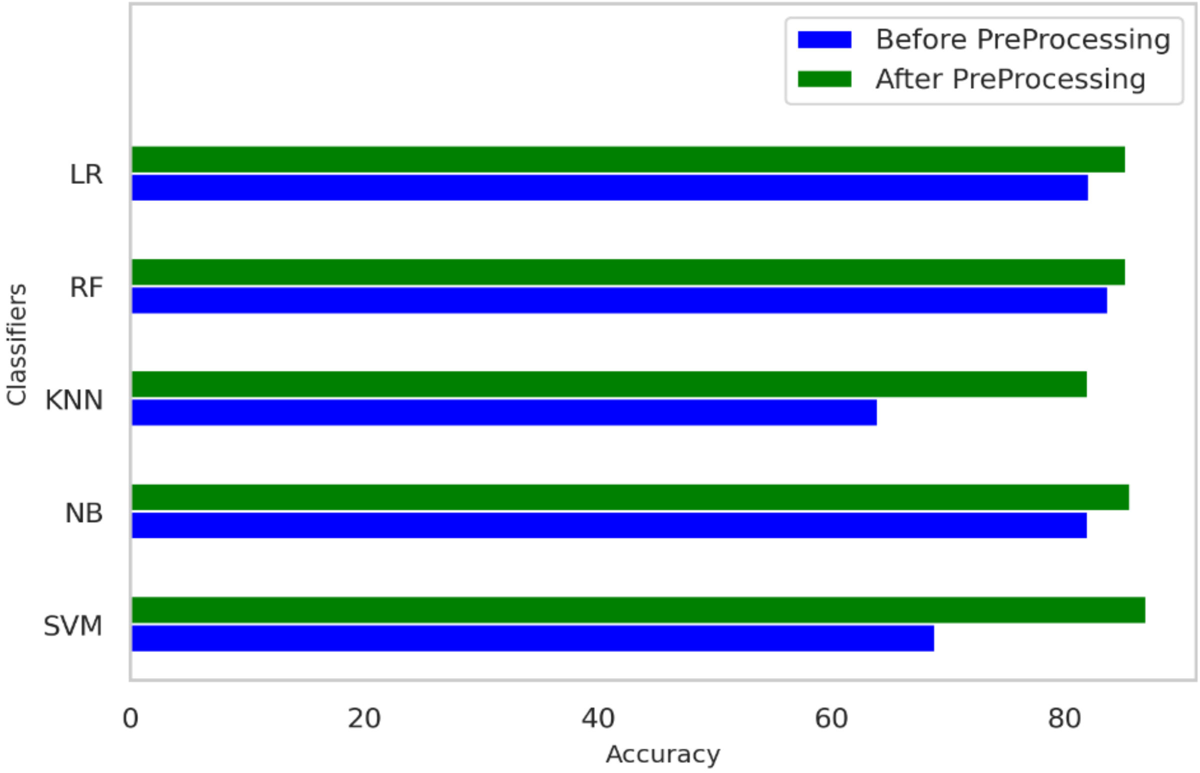


Figure 5. 1: Accuracy Improvement of classifiers using Standard Scalar and SMOTE Technique

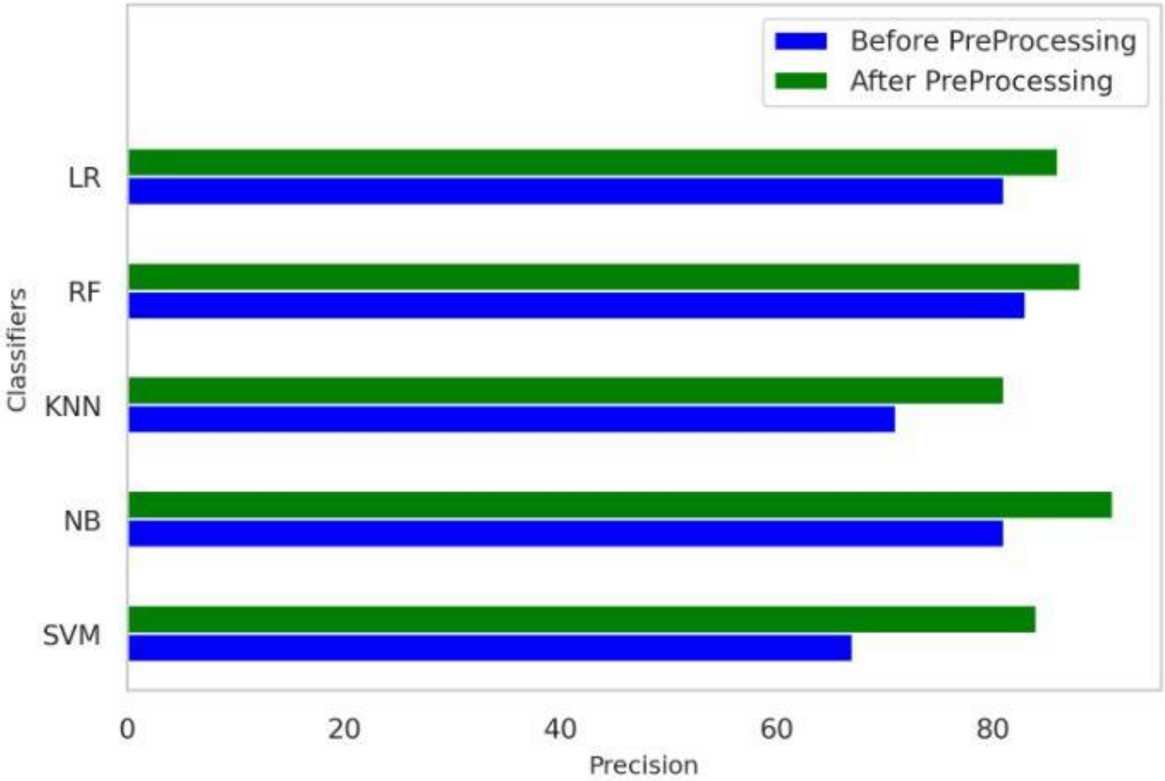


Figure 5. 2: Precision Improvement of classifiers using Standard Scalar and SMOTE Technique

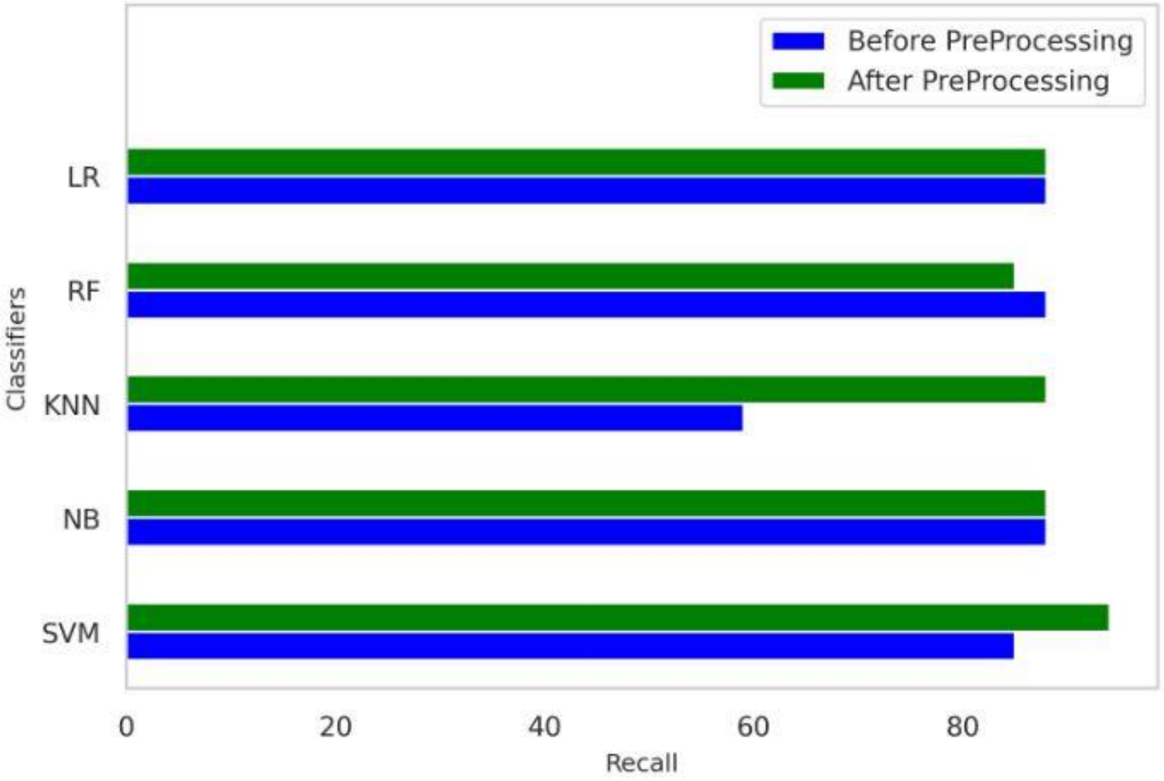


Figure 5. 3: Recall Improvement of classifiers using Standard Scalar and SMOTE Technique

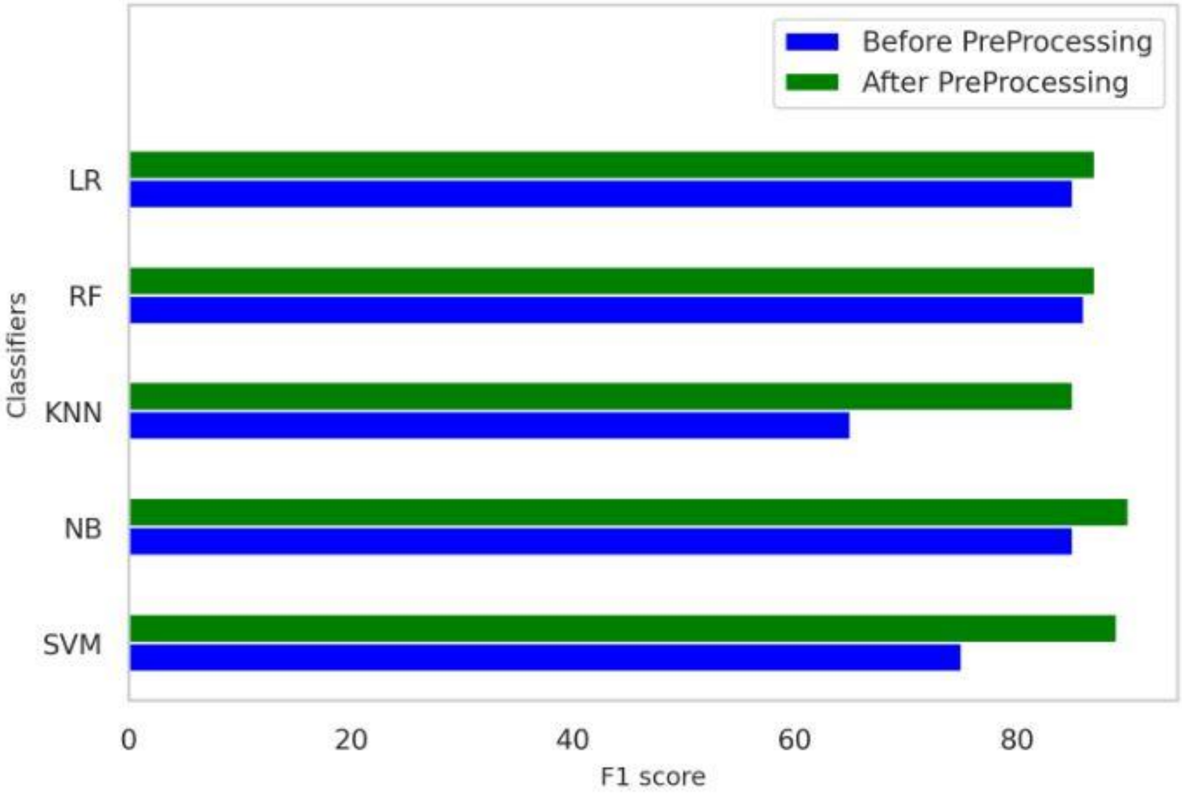


Figure 5. 4: F1 Score Improvement of classifiers using Standard Scalar and SMOTE Technique

5.4 Performance improvement using Ensemble Technique

This section describes that Ensemble Bagging technique is used as based classifier and all five classifiers are tested which produced further improvements in results. The all five hybrid model embedded with ensemble bagging techniques proved to be fruitful on account of results achieved as precision, accuracy, recall and F1 score. Table 5.3, illustrates the improvements of results in each case where as proposed model HLRBM produced better accuracy. A graphical presentation of achieved results is visualized in fig. 5.5.

Table 5. 3: Results of numerous Models with Proposed Hybrid Model

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM-BM	88.52	91	88	90	89
NB-BM	86.88	88	88	88	87
KNN-BM	86.93	84	94	89	87
RF-BM	86.88	86	91	89	87
HLRBM	90.16	93	88	90	90

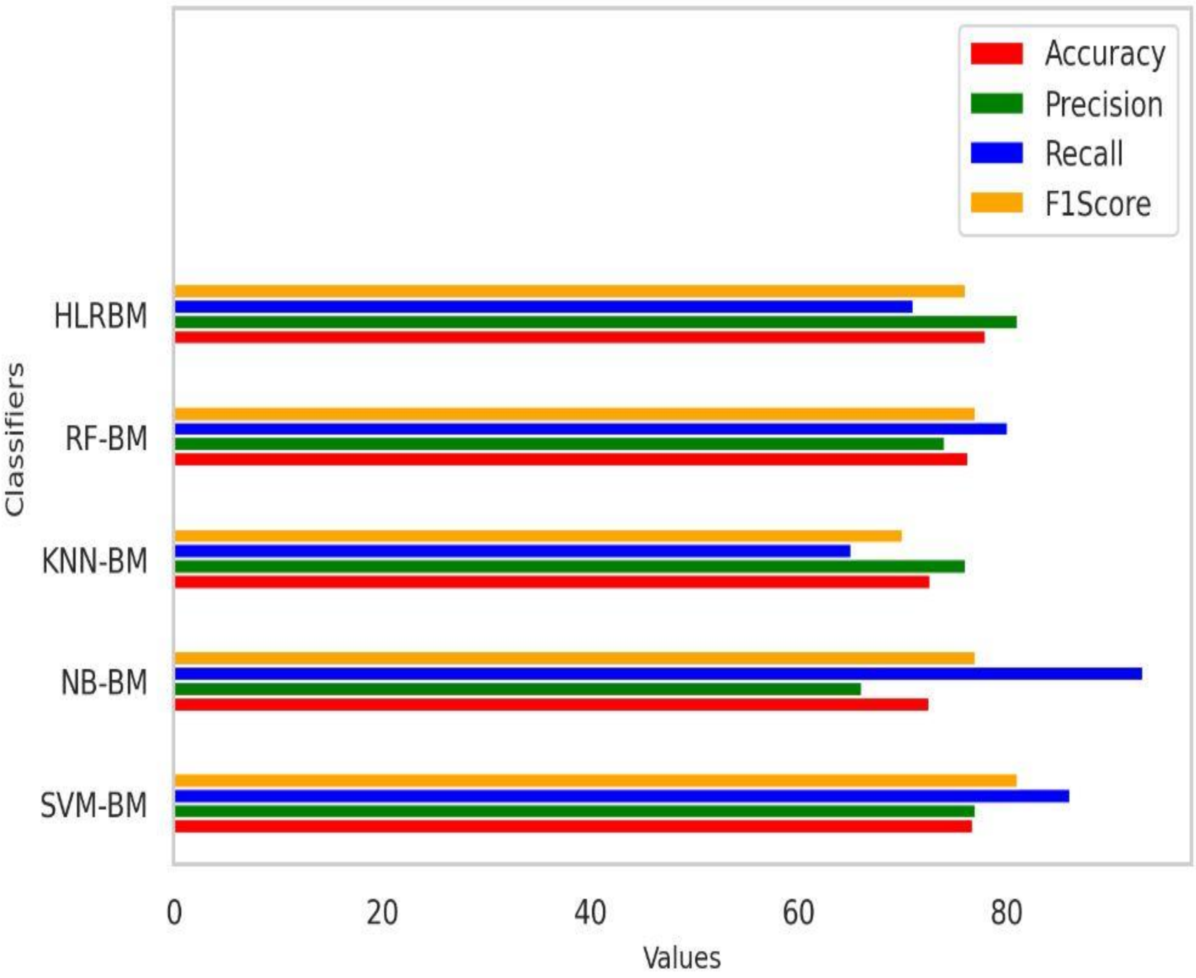


Figure 5. 5: Graphical representation of each classifier with proposed Model

5.5 Comparison of Results with former Researcher Results

A comparison of results achieved is drawn with existing researcher's results. Our research work proves that with the use of better data preprocessing techniques and formulation of unique hybrid model overall satisfactory results can be achieved for prediction of CVD. Table 5.4, illustrates the overall comparison of application of numerous models on same dataset by various researchers.

Table 5. 4: Comparison of proposed Model with various existing research work

Reference	Classifiers	Accuracy %	Precision	Recall /(Sen%)	F1 Score
J.P Li et al: [2]	LR K-NN ANN SVM NB DT	83 69 60 85 75 70	95 70 100 95 90 72	75 64 0 75 78 83	-
P.Gosh et al:[4]	DT RF K-NN AB GB	86.97 88.65 83.61 89.07 86.97	-	-	-
Pooja Rani et al: [1]	NB SVM LR RF AdaBoost Hybrid Model	85.07 84.16 83.24 83.85 82.34 86.60	84.37 86.92 86.62 88.46 88.96	82.31 81.09 82.92 84.14 83.53	83.33 83.91 84.73 86.25 86.16
Dinesh Kumar et al: [2]	LR RF NB GB	86.51 80.89 84.26 84.26	-	-	-

	SVM	79.77			
S. Mohan et al: [5]	NB	75.8	90.5	79.8	84.5
	LR	82.9	89.6	91.1	90.2
	DT	85	86.0	98.8	91.8
	RF	86.1	87.1	98.8	92.4
	SVM	86.1	86.1	100	92.5
	HRFLM (Hybrid)	88.4	90.1	92.8	90
M. Kavitha et al: [6]	DT	79			
	RF	81	-	-	-
	Hybrid (DT+RF)	88			
Our Work	SVM-BM	86.88	84	94	89
	NB-BM	88.52	91	98	90
	KNN-BM	86.93	89	91	90
	RF-BM	88.52	89	91	90
	HLRBM (Proposed)	90.16	93	88	90

In fig. 5.6, A detailed comparison is visualized in graphical form to have better idea of results in term of accuracy of each classifiers applied on same dataset.

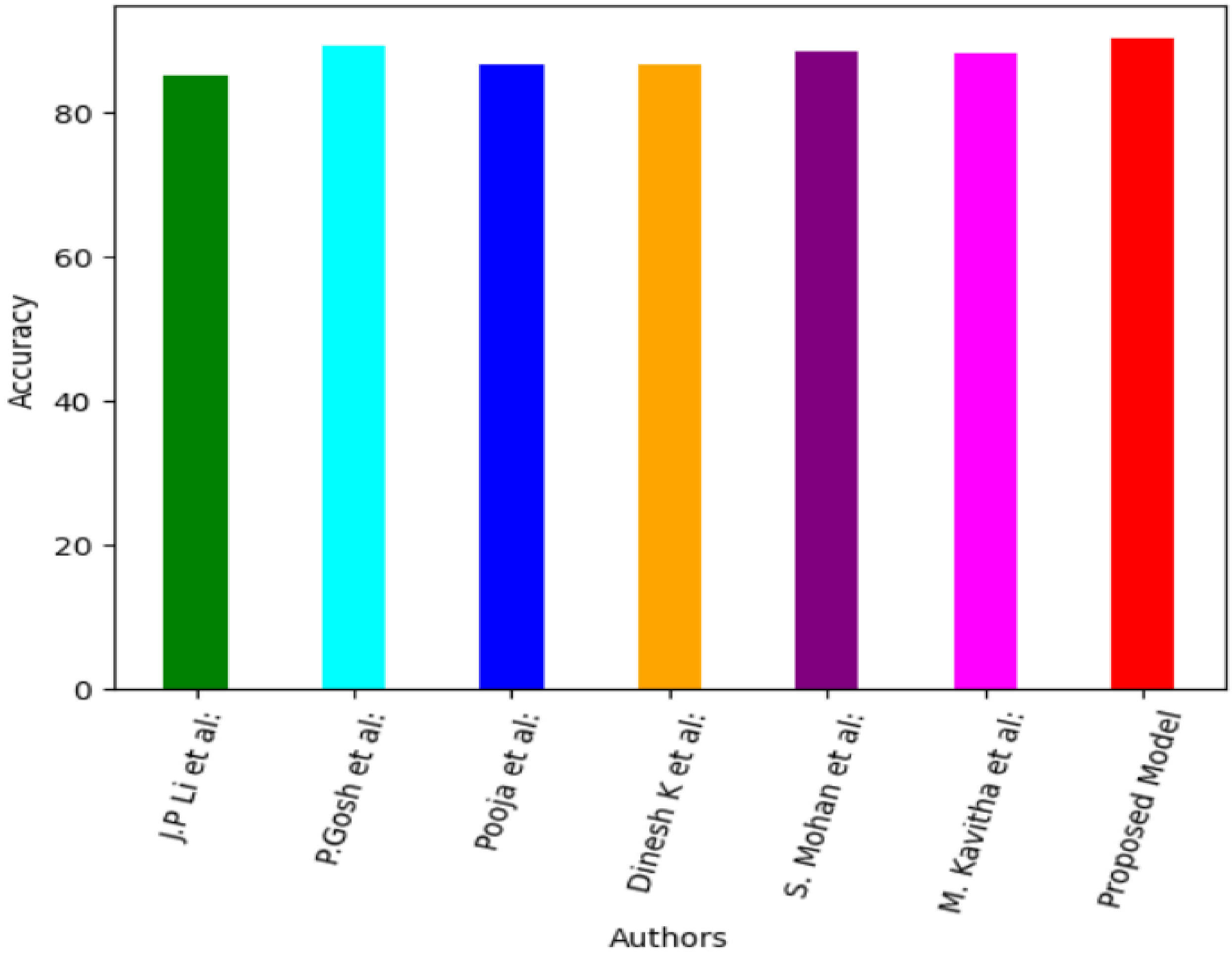


Figure 5. 6: Improvement in accuracy of Proposed Model with existing work

5.6 Results of additional experimental work

All the preprocessing techniques are applied to another dataset which is publically available having 70000 records [13] with 12 attributes. Five algorithms were applied including recommended Model that gave the maximum results of the former results. The efficacy and efficiency of our research work is proved to be productive after having comparison of results on existing dataset. Table 5.5, shows the results achieved on datasets utilized to check performance of recommended model on other than Cleveland dataset used in earlier part. A graphical visualization is illustrated in fig. 5.7, to compare evaluation parameters against proposed models.

Table 5. 5: Results achieved by recommended models on different dataset

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM-BM	76.66	77	86	81	77
NB-BM	72.55	66	93	77	73
KNN-BM	72.62	76	65	70	73
RF-BM	76.25	74	80	77	76
HLRBM	77.90	81	71	76	78

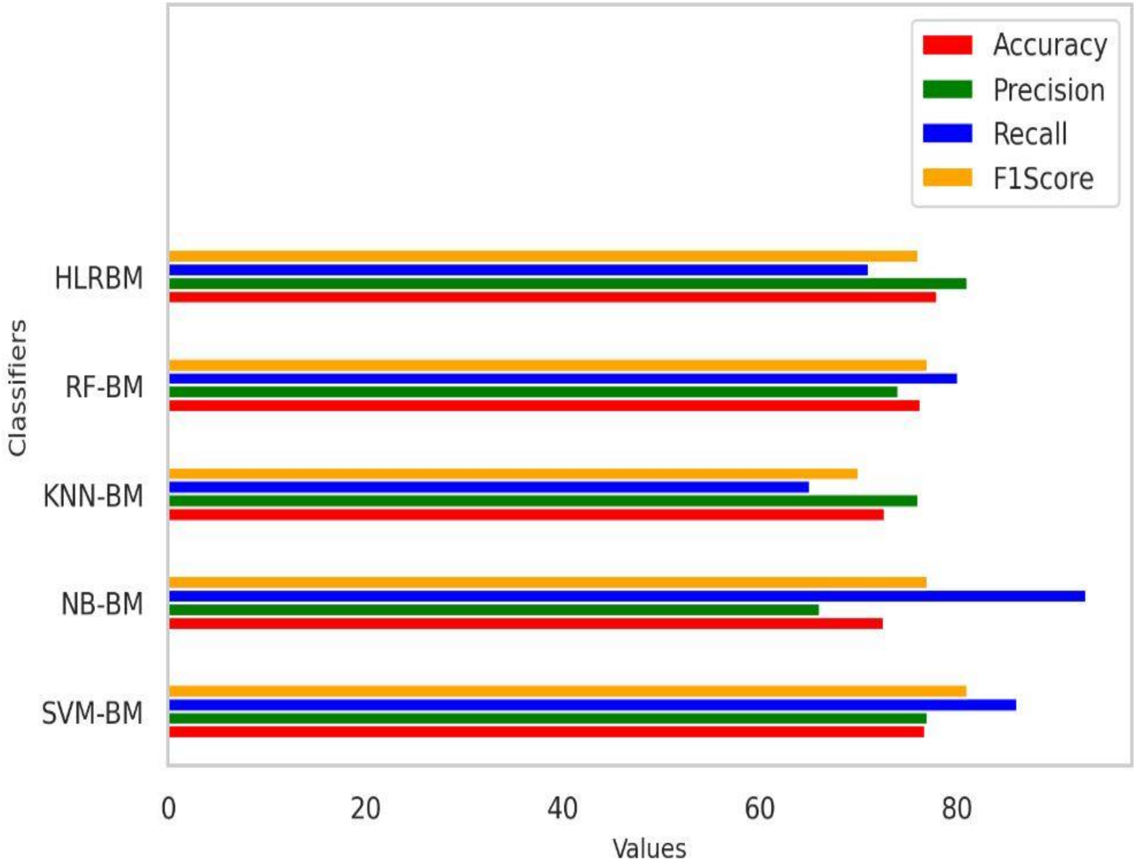


Figure 5. 7: Graphical representation of results achieved by proposed model on different dataset

Experimental work is carried out to check efficacy of our proposed model. A comparison is drawn to show its performance on different dataset having 70000 records publically available. Table 5.6, Shows its performance against existing work on this dataset.

Table 5. 6: Comparison of results of different dataset with existing work

Reference	Algorithms	Accuracy %	Precision	Recall	F1 Score
Jan Carlo et al: [9]	ANN LR DT RF SVM KNN GA-ANN	68.35 72.35 61.72 68.94 72.16 68.34 73.43	-	-	-
Own Work	SVM-BM NB-BM KNN-BM RF-BM HLRBM	76.66 72.55 72.62 76.25 77.90	77 66 76 74 81	81 77 70 77 76	77 73 73 76 78

5.7 Summary

Results and analysis are highlighted in this chapter, where impact of preprocessing data techniques and overall performance of proposed model is discussed. Results revealed that proposed hybrid model HLRBM outperforms other model. Experimental work done with HLRBM proved to be a remarkable achievement as trained model beats the accuracy of former researchers carried out on CVD dataset comprising on 12 attributes along with 70000 records.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion and Objective achieved

The development of an improved hybrid model (HLRBM) for CVD detection using ML shows significant promising results. The model combines multiple ML classifiers and leverages relevant medical parameters to achieve enhanced accuracy and efficiency in identifying CVD. The results obtained from this study demonstrate the potential of ML techniques in improving the accuracy of CVD detection, which can aid in early diagnosis and intervention, leading to better patient outcomes. Before constructing a hybrid model, Data preprocessing techniques were utilized to enhance accuracy of CVD detection making it more efficient and reliable model. The hybrid model was fabricated using ensemble techniques, taking Bagging method as base classifiers and merging it with five traditional classifiers SVM, NB, KNN, RF and LR used in this research work. Furthermore, Hybrid model HLRBM proved to be our recommended model which outperformed other models.

Experimental work is carried out with proposed hybrid model on different CVD Dataset having 12 attributes and 70000 records. HLRBM produced promising results which beats accuracy achieved by former researcher on this CVD dataset.

6.2 Limitations

However, there are several avenues for future work in this area. Firstly, the model could benefit from incorporating additional patient data from diverse populations to improve its generalizability and robustness. This would help ensure that the model performs effectively across different demographic groups and can be widely applicable in real-world healthcare settings. Furthermore, the integration of advanced deep learning methods, such as recurrent neural networks or attention mechanisms, could potentially enhance the model's performance by capturing complex temporal dependencies and extracting more informative features from the data.

6.3 Future Work

Additionally, the model's interpretability can be further explored by incorporating explainable AI techniques, allowing healthcare professionals to understand the underlying reasons behind the model's predictions. This would increase trust and acceptance of the model within the medical community. Lastly, conducting extensive validation studies using large-scale clinical datasets and comparing the performance of the hybrid model against existing diagnostic methods would be crucial for assessing its clinical utility and effectiveness in real-world scenarios. Overall, the improved hybrid model for CVD detection shows promise in revolutionizing e-healthcare systems. Future research and development in this field can greatly contribute to advancing early detection and proactive management of CVD, ultimately improving patient care and reducing the burden on healthcare systems.

CHAPTER 7

REFERENCES

[1] Rani, P., Kumar, R., Ahmed, N.M.O.S. *et al.* A decision support system for heart disease prediction based upon machine learning. *J Reliable Intell Environ* **7**, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6

[2] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[3] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.

[4] P. Ghosh *et al.*, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in *IEEE Access*, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[5] Mohan, S. and Chandrasegar, T., Srivastava Gautam Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, pp.81542-81554.

[6] Kavitha, M., G. Gnaneswar, R. Dinesh, Y. Rohith Sai, and R. Sai Suraj. "Heart disease prediction using hybrid machine learning model." In *2021 6th international conference on inventive computation technologies (ICICT)*, pp. 1329-1333. IEEE, 2021.

[7] Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health Technol.* **11**, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

[8] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* **1**, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[9] Jan Carlo T. Arroyo and Allemar Jhone P. Delima, "An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction," Journal of Advances in Information Technology, Vol. 13, No. 1, pp. 95-99, February 2022.

[10] Malav, Amita & Kadam, Kalyani & Kamat, Pooja. (2017). PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY. International Journal of Engineering and Technology. 9. 3081-3085. 10.21817/ijet/2017/v9i4/170904101.

[11] Khan M, Hashim M, Mustafa H, et al. (July 23, 2020) Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. Cureus 12(7): e9349. doi:10.7759/cureus.9349

[12] Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, Boehme AK, Buxton AE, Carson AP, CommodoreMensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Ferguson JF, Generoso G, Ho JE, Kalani R, Khan SS, Kissela BM, Knutson KL, Levine DA, Lewis TT, Liu J, Loop MS, Ma J, Mussolino ME, Navaneethan SD, Perak AM, Poudel R, Rezk-Hanna M, Roth GA, Schroeder EB, Shah SH, Thacker EL, VanWagner LB, Virani SS, Voecks JH, Wang N-Y, Yaffe K, Martin SS; on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2022 update: a report from the American Heart Association [published online ahead of print Wednesday, January 26, 2022]. Circulation. doi: 10.1161/CIR.0000000000001052

[13] S. Ulianova. (2019). Cardiovascular Disease dataset. [Online]. Available: https://www.kaggle.com/sulianova/cardiovasculardisease-dataset

[14] Heart Disease Datasets From UCI Machine Learning Repository. Accessed: Apr 14, 2023. [Online]. Available: https://archive.ics.uci. edu/ml/datasets/Heart+Disease

[15] Heart Disease Statlog Dataset of UCI Machine Learning Repository. Accessed: May 14 Apr, 2023. [Online]. Available: http://archive. ics.uci.edu/ml/datasets/statlog+(heart)

[16] Subhadra K, Vikas B (2019) Neural network based intelligent system for predicting heart disease. Int J Innov Technol Explor Eng 8(5):484–487

[17] Jain A, Tiwari S, Sapra V (2019) Two-phase heart disease diagnosis system using deep learning. Int J Control Autom 12(5):558– 573.
http://sersc.org/journals/index.php/IJCA/article/view/2690

[18] Vithanwattana N, Mapp G, George C (2017) Developing a comprehensive information security framework for mHealth: a detailed analysis. J Reliab Intell Environ 3(1):21–39. https://doi. org/10.1007/s40860-017-0038-x

[19] Jusob FR, George C, Mapp G (2017) Exploring the need for a suitable privacy framework for mHealth when managing chronic diseases. J Reliab Intell Environ 3(4):243–256. https://doi. org/10.1007/s40860-017-0049-7

[20] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informat. Med. Unlocked, vol. 16, no. 2, 2019, Art. no. 100203

[21] A. Acharya, "Comparative study of machine learning algorithms for heart disease prediction," M.S. thesis, Helsinki Metropolia Univ. Appl. Sci., Helsinki, Finland, Apr. 2017. [Online]. Available: https://
www.theseus.fi/bitstream/handle/10024/124622/Final%20Thesis.pdf?
sequence=1&isAllowed=y

[22] Ensemble Techniques of Bagging. Accessed: Apr. 23, 2023. [Online]. Available:
https://quantdare.com/what-is-the-difference-betweenBagging-and-Boosting/

[23] An Explanation of Ensemble Bagging Techniques. Accessed: Apr. 23, 2023. [Online]. Available: https://towardsdatascience. com/ensemble-methods-Bagging-Boosting-and-stacking-c9214a10a205/

[24] Heart disease - Symptoms and causes - Mayo Clinic. Accessed: Jun 8, 2023. [online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#SnippetTab

[25] Coronary Heart Disease - Symptoms | NHLBI, NIH. Accessed: Jun 8, 2023. [online]. Available: https://www.nhlbi.nih.gov/health/coronary-heart-disease/symptoms#:~:text=Sometimes%20coronary%20heart%20disease%20is%20%E2%80%9Csilent%2C%E2%80%9D%20meaning%20you,is%20so%20important.%20Symptoms%20of%20a%20heart%20attack

[26] Posterior Probability: Definition + Example - Statology. Accessed: Jun 8, 2023. [online]. Available: https://www.statology.org/posterior-probability/

[27] Lee, W. "Supervised learning-classification using K-nearest neighbors (KNN)." *Python Machine Learning*. Wiley, 2019. 205-220

[28] Taamneh, Madhar. "Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm." *Journal of safety research* 66 (2018): 121-129

[29] Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. Journal of Big Data, 6(1), 1-21.

[30] Alneamy, J. S. M., & Alnaish, R. A. H. (2014). Heart disease diagnosis utilizing hybrid fuzzy wavelet neural network and teaching learning based optimization algorithm. Advances in Artificial Neural Systems, 2014, 6-6.

[31] Morita, K., Tsuka, H., Kuremoto, K. I., Kimura, H., Kawano, H., Yokoi, M., ... & Tsuga, K. (2019). Association between buccal mucosa ridging and oral feature/symptom and its effects on occlusal function among dentate young adults in a cross-sectional study of Japan. CRANIO®.

[32] Majumder, S., & Pratihar, D. K. (2018). Multi-sensors data fusion through fuzzy clustering and predictive tools. Expert Systems with Applications, 107, 165-172.

[33] Marreiros, G., Martins, B., Paiva, A., Ribeiro, B., & Sardinha, A. (Eds.). (2022). Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings (Vol. 13566). Springer Nature.2