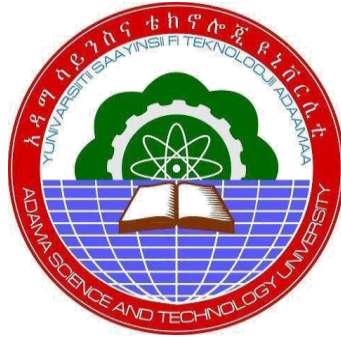


Cardiovascular Disease Prediction Using Machine Learning Approach



Melaku Yilma Arega

A Thesis Submitted to the Department of Computer Science and
Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of
Master of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September, 2023

Adama, Ethiopia

Cardiovascular Disease Prediction Using Machine Learning Approach

Melaku Yilma Arega

Advisor: Dr. Tilahun Melak

A Thesis Submitted to the Department of Computer Science and
Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of
Master of Science in computer science and Engineering

Office of Graduate Studies

Adama Science and Technology University

September, 2023

Adama, Ethiopia

Declaration

I hereby declare that this Master Thesis entitled '**Cardiovascular Disease Prediction Using Machine Learning Approach**' is my original. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation

Name of the student

Signature

Date

Recommendation

I/we, the advisor of this thesis, hereby certify that I/we have read the revised version of the thesis entitled '**Cardiovascular Disease Prediction Using Machine Learning Approach**' prepared under my/our guidance by Melaku Yilma submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering (CSE).

Therefore, I/we recommend the submission of the revised version of the thesis to the department following the applicable procedures.

Major Advisor

Signature

Date

Approval Page of M.Sc. Thesis

I/we, the advisors of the thesis entitled ‘**Cardiovascular Disease Prediction Using Machine Learning Approach**’ and developed by Melaku Yilma, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Major Advisor

Signature

Date

We, the undersigned, members of the board of Examiners of the thesis by **Melaku Yilma** have read and evaluated the thesis entitled “**Cardiovascular Disease Prediction Using Machine Learning Approach**” and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering (CSE).

Chairperson

Signature

Date

Internal Examiner

Signature

Date

External Examiner

Signature

Date

Finally, approval and acceptance of the thesis is contingent upon the submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate committee (SGC).

Department Head

Signature

Date

School Dean

Signature

Date

Office of Postgraduate Studies, Dean

Signature

Date

ACKNOWLEDGMENT

I extend my heartfelt gratitude to the Almighty God for enabling me to overcome challenges and guiding me in all endeavors. His boundless grace and blessings have been instrumental in the completion of this study. Additionally, I offer my gratitude to the Virgin Mary, the mother of God, for her divine support throughout this journey.

I extend my sincere gratitude to my advisor, Dr. Tilahun Melak, for his invaluable guidance and insightful advice. His unwavering support and encouragement have been a driving force behind my research efforts. Dr. Tilahun's dedication to motivating and mentoring me has played a pivotal role in the successful completion of this study. His exceptional professionalism and profound understanding have been instrumental in shaping this research. I am profoundly grateful for his unwavering commitment and expertise, which have been indispensable throughout this journey.

I wish to extend my heartfelt appreciation to the dedicated staff members of St. Paul's Hospital Millennium Medical College for their invaluable cooperation during the data collection phase. I am deeply grateful to Dr. Seifu Bacha (Internist/Cardiologist) for his unwavering interest, guidance, and dedication in sharing his knowledge, which greatly enriched this research endeavor. Furthermore, I am truly honored to convey my special thanks to my mother, Genet Argaw Sima, and to all my family members for their unwavering encouragement, support, and steadfast presence throughout my life journey, culminating in the successful completion of this thesis work.

Lastly, my sincere gratitude extends to all my friends who generously shared their knowledge and insights with me throughout my study.

Table of Contents

LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS AND ACRONYMS	xiv
<i>ABSTRACT</i>	xv
CHAPTER ONE	1
1. INTRODUCTION	1
1.1. Background of the Study.....	1
1.2. Motivation of the study	3
1.3. Statement of the Problem	3
1.4. Research Questions	5
1.5. Objectives of the study	5
1.5.1. General Objective	5
1.5.2. Specific Objectives	5
1.6. Significance of the Study	6
1.7. Scope and limitations	6
1.7.1. Scope of the study.....	6
1.7.2. Limitations of the study	7
1.8. Organization of the study	7
CHAPTER TWO	9
2. LITERATURE REVIEW	9
2.1. Introduction	9
2.2. Cardiovascular disease and its risk factors.....	9
2.3. Prevalence of cardiovascular disease and Economic Burden	9
2.4. Common Cardiovascular disease	10

2.4.1. Coronary heart disease.....	10
2.4.2. peripheral artery disease	11
2.4.3. Rheumatic heart disease	11
2.4.4. Stroke.....	11
2.5. Treatment and management of cardiovascular Disease	12
2.6. Overview of Machine learning.....	12
2.7. Machine learning and health sector.....	14
2.8. Machine learning approach in disease prediction	15
2.9. Feature selection methods	18
2.10. Related works	20
CHAPTER THREE	26
3. RESEARCH METHODOLOGY.....	26
3.1. Introduction	26
3.2. Data collection.....	26
3.3. Cardiovascular diseases prediction modeling	27
3.3.1. Data preprocessing techniques	27
3.3.2. Modelling.....	30
3.3.3. Model evaluation metrics	34
3.4. Performance evaluation metrics for prediction modeling.....	35
CHAPTER FOUR.....	37
4. PROPOSED SOLUTION FOR CARDIOVASCULAR DISEASES PREDICTION	37
4.1. Introduction.....	37
4.2. Proposed architecture for cardiovascular disease prediction	37
4.3. Data preparation and feature description	38
4.4. Data preprocessing phase	42

4.4.1. Data cleaning and handling data values	42
4.4.2. Handling categorical data	43
4.4.3. Feature selection	43
CHAPTER FIVE	45
5. EXPERIMENT AND IMPLEMENTATION.....	45
5.1. Introduction	45
5.2. Implementation environment	45
5.2.1. Tools and Python Libraries.....	45
5.2.2. Package Manager and Environments	46
5.4. Preprocessing implementations.....	47
5.5. Model implementations.....	49
5.4.1. Hyper Parameter Tuning	49
CHAPTER SIX.....	51
6. RESULT AND DISCUSSION	51
6.1. Introduction	51
6.2. Dataset class distribution result.....	51
6.3. Feature selection result.....	54
6.4. model performance evaluation with metrics	55
6.4.1. Models' performance evaluation without feature selection	56
6.4.2. Models' performance evaluation with Lasso regularization	59
6.4.3. Models' performance evaluation with recursive feature elimination.....	59
6.5. Discussion	65
CONCLUSION AND FEATURE WORK.....	67
Conclusion.....	67
Contributions of the study	67

Future Works.....	68
References.....	70
APPENDIX.....	i
APPENDIX A: Sample CVD dataset	i
APPENDIX B: Sample code for stacking model with L1	ii
APPENDIX B: Sample code for stacking model with RFE	iii

LIST OF TABLES

Table 2:1 Selected base classifiers for stacking based ensemble model	17
Table 2:2 Comparison of filter, wrapper, and embedded feature selection models across	19
Table 2:3 Related work summary	24
Table 3:1 Confusion matrix	36
Table 4:1 Dataset features together with descriptions	40
Table 4:2 Missing value imputation	43
Table 4:3 Categorical data	43
Table 5:1 Tools and libraries used implementation.....	45
Table 5:2 Hyperparameter name, values and Optimum value returned by grid search.....	50
Table 6:1 Selected Features using RFE and L1	55
Table 6:2 Models test results without feature selection.....	56
Table 6:3 Overall performance evaluation of models with <i>L1</i>	59
Table 6:4 Overall performance evaluation of models with RFE	59
Table 6:5 Models test results based on stratified 10-fold cross validation	62
Table 6:6 Models test results based on stratified 5-fold cross validation	63
Table 6:7 Performance metrics for Stacking with RFE and Stacking with L1.....	63
Table 6:8 Overall Model Evaluation summery with and without FS techniques	64

LIST OF FIGURES

Figure 2:1 The General Frameworks for Filter, Wrapper, and Embedded model	19
Figure 3:1 The research workflow diagram.....	26
Figure 3:2 Dataset Balancing using SMOTE.....	29
Figure 3:3 Random Forest Classifier	31
Figure 3:4 Extreme Gradient Boosting Trees	32
Figure 3:5 Support Vector Machine	33
Figure 3:6 Stacking classifier architecture.....	34
Figure 3:7 K-fold cross-validation.....	35
Figure 4:1 Proposed Architecture of CVD Prediction.....	38
Figure 4:2 Data preprocessing phase diagram	42
Figure 4:3 Workflow Diagram for Feature Selection	44
Figure 5:1 loading dataset using panda's library	47
Figure 5:2 handling missing values using imputer	48
Figure 5:3 handling categorical values using Label encoder.....	48
Figure 5:4 handling outliers using Z-score	48
Figure 5:5 Normalization using Min-Max scalar.....	48
Figure 5:6 importing python libraries	49
Figure 6:1 Original Dataset Class Distribution.....	51
Figure 6:2 Dataset Class distribution after SMOTE applied	52
Figure 6:3 Numerical feature distribution before Z-score normalization	53
Figure 6:4 Numerical feature distribution after Z-score normalization.....	54
Figure 6:5 Feature importance using Random Forest Classifier	55
Figure 6:6 SVM model's Confusion matrix	56
Figure 6:7 RF model's Confusion matrix	57
Figure 6:8 XGB model's Confusion matrix.....	58
Figure 6:9 Stacking model's Confusion matrix	58
Figure 6:10 Confusion matrix of SVM with RFE	60
Figure 6:11 Confusion matrix of RF with RFE	61
Figure 6:12 Confusion matrix of XGB with RFE.....	61
Figure 6:13 Confusion matrix of Stacking with RFE	62

Figure 6:14 Models accuracy performance comparison	64
Figure 6:15 Learning curve representing training and validation scores.....	65

LIST OF ABBREVIATIONS AND ACRONYMS

A

ACS: acute coronary syndrome

AI: artificial intelligence

C

CVD: Cardiovascular disease

CAD: coronary artery disease

H

HCU: Hypertension and Cardiac unit

L

L1: LASSO regularization

LR: Logistic Regression

M

ML: Machine learning

N

NCD: Non communicable diseases

P

PAD: peripheral artery disease

R

RFE: Recursive feature elimination

RF: Random Forest

RHD: rheumatic heart disease

S

SPHMMC: St. Paul's Hospital Millennium Medical College

SVM: Support vector machine

SMOTE: Synthetic Minority Oversampling Technique

U

UHC: universal health coverage

UCI: university of California Irvine

W

WHO: World Health Organization

RFE: Recursive feature elimination

X

XGB: Extreme Gradient Boosting

ABSTRACT

Cardiovascular diseases (CVDs) are the primary contributors to worldwide mortality, responsible for 17.9 million deaths, constituting 32% of the total global fatalities. Notably, more than 75% of these deaths occurred in low and middle-income countries. In Ethiopia, CVDs kills 170 people every day. It accounts for a significant portion of non-communicable disease (NCD) deaths fatalities and healthcare expenses. Early diagnosis of CVDs is crucial but challenging due to limited access to primary health care program, lack of expertise, shortage of diagnosis and treatment apparatus and the inefficiency of traditional diagnostic methods, which are ineffective, costly, and time-consuming. Recently, machine learning has emerged as a valuable tool to support the diagnosis of CVDs. This study leverages a stacking-based machine learning model to address the challenge of CVD diagnosis. While prior research has focused on general CVD prediction or specific disease types, this study classifies the disease into four common CVDs categories such as CAD, PAD, RHD, and Stroke. The proposed stacking model was implemented using a CVD dataset collected at St. Paul's Hospital Millennium Medical College (SPHMMC) which consists of 2196 instances with 19 features. The obtained dataset was prepared before exposed to the model using different preprocessing techniques, including imputer, z-score, label encoder and min-max normalization. Stratified k-fold cross validation was utilized as dataset splitting methods during model construction process. Performance comparisons were made with three individual machine learning models: SVM, RF, and XGB. Model performance evaluation was carried out with and without applying feature selection techniques (recursive feature elimination (RFE) and lasso regularization (L1) using metrics such as accuracy, precision, recall, and F1-score metrics. Our experiments revealed stacking model with RFE outperformed the others achieving the highest accuracy of 97.55%.

The proposed model helps medical practitioners or cardiologists in classifying CVDs into four common categories effectively, thereby potentially saving lives and reducing the burden of these diseases.

Keywords: Cardiovascular diseases, stacking, machine learning, ensemble learning

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the Study

The cardiovascular system consists of the entire network of blood vessels, such as veins, arteries, and capillaries, which creates an intricate pathway for blood to circulate throughout the body (Hussein, 2017). Any condition that impacts the regular circulation or movement of blood from the heart results in various types of heart diseases and complications, commonly known as cardiovascular diseases (CVDs) (Rabiu, 2020; Yahaya et al., 2020). CVDs are considered non-communicable diseases (NCDs) and are primary contributors to global mortality. As per data from the World Health Organization (WHO, 2021), CVDs were responsible for around 17.9 million fatalities in 2019, constituting 32% of the total worldwide deaths. The proportion of these deaths varies from country to country. For instance, these diseases exhibit a disproportionate impact on low- and middle-income countries, with over 75% of deaths occurring in these regions. The variations in these death figures arise from the limited access to primary healthcare programs in those countries for early identification and treatment of people with risk factors for CVDs. Regarding premature deaths (happens before 70 years of age), the proportion of deaths varies significantly, ranging from 4% in high-income countries to as high as 42% in low-income countries (Angaw et al., 2021). In developing countries like Ethiopia, CVDs contribute significantly to the prevalence and death rates of NCDs, accounting for a significant portion of both mortality and morbidity. In the country, NCDs are responsible for 42% of all deaths and of which 24% by CVDs (Mengistu & Benti, 2022; Shiferaw et al., 2018; Tefera et al., 2017a). Furthermore, CVDs represent 13% of the government's total expenses, which corresponds to 30% of the expenditure allocated to NCDs within the overall health budget (Dombrovskiy et al., 2017).

Generally, CVDs have four main kinds, such as peripheral arterial disease, coronary heart disease, cerebrovascular disease (strokes and transient ischemic attack (TIA)) and Aortic atherosclerosis (Arunachalam, 2020a; Olvera Lopez et al., 2019). Several risk factors that are modifiable, such as high cholesterol and high blood pressure, as well as risk factors that are non-modifiable like age and family history of heart disease, contribute to the development of cardiovascular diseases. It is possible to treat or mitigate the disease if identified early (Uddin & Halder, 2021). However, an increased number of CVDs risk factors has made its early identification

and treatment challenging. (Nadakinamani et al., 2022). Additionally, lack or limitations of medical expertise in this field may lead to false diagnosis. This is often true in developing countries where there is limited trained experts, testing apparatus and other useful resources required to deal with the diseases specially when complex operation is required (such as surgery)(Ghosh et al., 2021). In general, the safety of patients would be improved and serious heart attacks could be avoided with the aid of an accurate assessment of the risk of heart failure. Electrocardiogram (ECG), occultation, and measurements of blood pressure, blood sugar, and cholesterol are the classic methods of diagnosing CVD disease and also help to decide whether the medicine will prescribe or not. However, the diagnosis methods are considered as traditional and are costly and time-consuming. machine learning techniques have assisted CVDs diagnosis process by minimizing resource requirement of traditional methods and has enhanced diagnosis results(Bekkali et al., 2019).

Machine learning techniques supports in building predictive model for health sectors by analyzing and processing medical data in large quantities and provide useful insights. Some of its application areas in this sector are disease diagnosis, drug recommendation, predictive analysis of medical images and drug development. Models build using these techniques like CVD prediction models can assist medical practitioners or cardiovascular expert (cardiologist) in making prompt decisions and saves millions of lives by ensuring timely treatment for patients(Qayyum et al., 2021; Uddin & Halder, 2021).

Recently, there has been a surge in the popularity of various data mining methods and prediction algorithms, including ETC (Extra tree classifier), KNN (K-nearest neighbor), SVM (Support vector machine), LR (Logistic regression), NB (Naïve Baye's) and others. These techniques have been extensively utilized for the purpose of identifying and predicting cardiovascular diseases. Moreover, researchers have shown a growing interest in developing algorithms that incorporate ensemble learning, further enhancing the accuracy and effectiveness of these prediction models(Ashfaq et al., 2022).In an ensemble model, multiple models are constructed and combined to obtain better results(Ashfaq et al., 2022; Rahma & Al-Mousa, 2019; Yuan et al., 2020). This collaboration of multiple models results in a robust and highly effective ensemble model. The fundamental idea behind ensemble learning is that by combining multiple models, the weaknesses of an individual model are likely to be offset by the strengths of other models. This leads an ensemble model's overall performance being superior to that of any individual models (Younis et

al., 2022). Stacking, bagging and boosting are three ensemble learning categories. In this research work we employed stacking ensemble learning to build CVD prediction model which consists of three algorithms as base classifiers and one algorithm as meta classifier. Stacking, which is also named Super Learning, is a kind of ensemble technique where a "meta learner" is trained by combining a number of classification models. The objective of this model is to bring together diverse and robust sets of learners to create a powerful ensemble(Tiwari et al., 2022a). For the proposed model, three algorithms such as SVM (support vector machine), RF (Random Forest), XGB (extreme gradient boosting), were selected as base learner (classifier) and LR (logistic regression) as meta learner (classifier). The stacking classifier we built predicts the diseases into four categories such as peripheral artery disease (PAD), coronary artery disease (CAD), rheumatic heart disease (RHD) and Stroke(WHO, 2021).

1.2. Motivation of the study

In less developed countries, including Ethiopia, cardiovascular diseases (CVD) present a significant healthcare challenge. Resource limitations, lack of expertise, and insufficient diagnostic equipment result in inaccurate diagnoses and suboptimal treatment for CVD patients(Ghosh et al., 2021). This not only endangers lives but also imposes an economic burden on communities in regions such as sub-Saharan Africa. Leveraging current technologies to enhance CVD diagnosis and management is imperative for improving healthcare outcomes and reducing the societal burden of these diseases (Gárate-Escamila et al., 2020). This study seeks to address these critical issues by developing a machine learning model for the prediction of specific CVD types, which has the potential to save lives and reduce the burden of CVDs in resource-constrained healthcare settings.

1.3. Statement of the Problem

Cardiovascular disease (CVD) is a major global health concern which endangers peoples in millions worldwide. The early detection and assessment of the diseases are crucial for reducing the risk of developing CVD complications and providing timely interventions. Preventing and treating this disease is essential, and early identification plays a crucial role in achieving successful outcomes. However, in low- and middle-income countries, this task becomes challenging due to the lack of accessible and beneficial primary healthcare programs. These countries often struggle to provide timely identification and treatment for individuals with known risk factors that can lead to cardiovascular diseases (CVDs), as highlighted in the WHO report from 2021(WHO, 2021).

The report further emphasizes that low- and middle-income countries share 75% of the deaths occurred by CVDs. These dangerous diseases highly impact the society economically by enforcing them to spend an excessive cost for diagnosis and treatment. In general, economies of developing countries are suffered from these diseases.

In Ethiopia, NCDs pose a significant threat to public health, with this CVD being the primary cause of mortality and morbidity among NCDs. RHD, stroke and ischemic heart disease are the most prevalent ones (Ali et al., 2021). In the country CVD accounts for 24% of all deaths given to NCD (Tefera et al., 2017a). These diseases impose a significant burden on health sectors. A study carried out at SPHMMC (St. Paul's Hospital Millennium Medical College in 2021) (Mengistu & Benti, 2022) revealed that from 1,165 yearly medical admissions, 353 cases, or 30%, were related to cardiovascular disease (CVD).

Ethiopia, like many Sub-Saharan African nations, lacks sufficient dataset regarding patterns and clinical characteristics of CVD (Tefera et al., 2017). This knowledge gap is a result of inadequate research on CVD despite the surge of NCDs in developing countries. As a consequence, health systems in Ethiopia have faced difficulties in managing CVD along with other NCDs and infectious diseases. To address the challenges highlighted above, it is essential to leverage current technologies since the risk factors associated with CVD makes the diagnostic process challenging for healthcare providers (Nadakinamani et al., 2022). Researchers have attempted to tackle this problem by creating predictive models through distinct machine learning techniques. However, the focus is on CVD as a general rather than addressing specific disease kinds. Which means if the patient has some kinds of CVD, the developed model predicted as CVD if the patient has the disease with specific kinds. But in real world scenario such kinds of conclusion prevent the patient to get the required treatment because the kind of CVD determines treatment that will be given (Doppala et al., 2021; Tiwari et al., 2022). And some other scholars have tried to address specific disease such as stroke, heart failure and others (Dev et al., 2022; Emon et al., 2020; Tripoliti et al., 2017). So that, to address the mentioned issues we built a model that predicts CVD disease into four kinds. This thesis work aims to address the problem by developing and evaluating machine learning model that predictions multiple kinds of CVD. In this work the CVD classified into four classes, those are CAD, Stroke, PAD, RHD. The study explored set of clinical and demographic features and symptoms to build the model. Additionally, the research identified the key features that are predictive of CVD to enhance the understanding of the comprehension of the

fundamental mechanisms behind these diseases.

The study employed stacking model development approach from ensemble machine learning algorithms. Since this approach requires choosing more than one algorithm as base classifier(learner) and one as meta classifier (learner), three algorithms such as SVM (support vector machine), RF (Random Forest), XGB (extreme gradient boosting), were chosen as base classifier and linear regression (LR) as meta classifier for the stacking implementation. The models' performance was assessed using standard performance evaluation metrics such as recall, precision, accuracy, and F1-score.

The outcomes of this thesis research will contribute to the advancement of a clinical decision support system designed to aid healthcare professionals in identifying and managing cardiovascular diseases at an early stage. Additionally, the study provided insights into the critical clinical and demographic features that are most predictive of CVD. In turn it will improve personalized patient care and population health management.

1.4. Research Questions

The conducted research attempted to answer the following research questions:

Rq1: what are the determinant features to predict cardiovascular diseases?

Rq2: Which feature selection techniques are suitable for predicting CVD based on CVD dataset?

Rq3: Which machine learning model are effective in predicting cardiovascular disease?

1.5. Objectives of the study

1.5.1. General Objective

The general objective of the study is to predict cardiovascular diseases using machine learning approach, which help medical practitioners to diagnose and treat patients.

1.5.2. Specific Objectives

The research follows the following specific objectives to achieve the general objectives of the study.

- To conduct a comprehensive analysis of the literature on machine learning techniques for predicting CVDs.
- To collect and preprocess dataset of patient records containing clinical and demographic features

- To Identify key feature selection methods for cardiovascular diseases prediction
- To identify the most significant clinical and demographic features related with CVDs.
- To build predictive model using CVD dataset
- To Evaluate the performance of CVD model through the test set
- To design a prototype for CVD prediction

1.6. Significance of the Study

This research work will benefit several stakeholders, including patients who suffer from cardiac disease due to the limited capability of health workers and the lack of resources required to identify the disease they suffer with. According to a 2017 WHO report (Sharma & Rizvi, 2017), medical practitioners can currently only anticipate 67% of cardiac diseases, highlighting the need for improved diagnosis methods. Additionally, the healthcare sector currently faces a shortage of expert medical professionals, making the need for accurate diagnosis even more pressing. However, the fruit of this research enables medical practitioners to correctly diagnose the disease early and provide the intended treatment, benefiting both patients and the healthcare system. Ultimately, this research will reduce the burden on the healthcare sector, enabling it to focus on mitigating other non-communicable diseases and benefiting the country as a whole. Failing to identify the cardiovascular disease before gating into complications may make the treatment process complex and often leads to surgical procedure which puts the patient in danger.

In addition to the aforementioned benefits, our model will also enhance the ability of medical professionals to accurately identify and treat the disease, empowering them to make effective decisions for their patients. Ultimately, the results of our research will support government policies and strategies aimed at countering the impact of this disease.

1.7. Scope and limitations

1.7.1. Scope of the study

This study introduced cardiovascular diseases prediction model using machine learning approach. The model predicts cardiovascular diseases into four target classes: CAD, Stroke, PAD, and RHD. We employed stacking classifier which contains random forest classifier (RF), extreme gradient descent (XGB), and support vector machine (SVM) as base learner (classifier) and logistic regression as meta learner (classifier). The dataset to train and validate the model was obtained

from SPMMC. The dataset contains medical health records, laboratory test results, and other types of clinical data including symptoms. In the study, data from patients who confirmed a particular type of cardiac disease and patients who have undergone a particular type of diagnostic test were included. Identified variables were demographic information (such as age and sex), clinical data (such as blood pressure, heart rate, fasting blood sugar, cholesterol levels), laboratory test results (such as cardiac biomarkers) and different symptoms. For validation purpose, we used stratified k-fold cross-validation, and for evaluation purpose, we used precision, accuracy, recall, and F1 score.

1.7.2. Limitations of the study

The study has some limitations, such as included patients in the study was aged above 17 years because the study site does not provide the service under the age specified. The prediction was done for four classes of disease (CAD, PAD, RHD and Stroke) due to lack of labelled dataset for other classes which categorized under CVD. This research work does not address the severity of each disease because there is no labelled data. No medical image included to build the model because of labelled data issues.

1.8. Organization of the study

The research is organized into six separate chapters, and their outlines are presented as follows:

Chapter One: In this chapter, we introduced the research topic, outline the research problem, articulate the research questions and objectives. Additionally, we delve into the motivation and significance of the study and specify its scope and limitations.

Chapter Two: In this chapter, a comprehensive review of relevant literature and studies related to the cardiovascular diseases is presented. The diseases prevalence and its economic burdens are elaborated. This section explores existing models, and previous research findings to establish the theoretical framework and gap in knowledge that the current study aims to address.

Chapter Three: This chapter provides a comprehensive overview of the approaches or techniques utilized in the study to design the research methodology. It encompasses various aspects, including the methodology's design, data collection, preparation and analysis techniques. Evaluation Metrics which were employed are included in this chapter.

Chapter Four: This chapter describes the proposed model in detail. The proposed model process workflow, proposed architecture, dataset description and preprocessing

Chapter Five: This chapter explains the experimental setup of the proposed solution to answer

the research question. It also provides the code snippet of useful python library packages data preprocessing and others

Chapter Six: This chapter presents and analyzes the experimental findings in connection with the research questions and objectives.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Introduction

In this chapter, an overview of academic articles, journals, conference papers, and other relevant sources related to cardiac disease and machine learning is presented to provide a brief overview of what has been accomplished thus far and what will be covered in this thesis. The review encompasses CVDs, machine learning techniques, data preprocessing and feature selection methods, selected models and model evaluation approaches.

2.2. Cardiovascular disease and its risk factors

The cardiovascular system is a vital bodily system tasked with the circulation of blood and the distribution of nutrients throughout the entire body. It comprises three main components: the heart, blood vessels, and blood itself. The cardiovascular system has several critical functions, including maintaining blood pressure, regulating body temperature, and transporting oxygen and nutrients to the body's tissues. Cardiovascular diseases are a group of disorders which affects healthy functioning of the system (WHO, 2019). There are several kinds of CVD diseases. Some of them are peripheral arterial disease, rheumatic heart disease, coronary heart disease, stroke. A number of risk factors that causes these diseases are high blood pressure, physical inactivity, poor nutrition, alcohol intake, tobacco usage, high cholesterol, and chewing chat no name a few. (Princy et al., 2020; Tefera et al., 2017). The patients experience distinct symptoms depends on the kinds of cardiac disease they experience even if these diseases have some common shared symptoms.

2.3. Prevalence of cardiovascular disease and Economic Burden

Cardiovascular diseases (CVD) represent a significant global health issue and is associated with comorbid conditions and potentially fatal complications (Tefera et al., 2017). The rapid growth rate of the disease increases the burden on the economies of several countries. According to WHO report in 2019, CVD causes 17.9 million people to die, which nearly 32% of all global deaths (WHO, 2021). Of 17 million premature deaths recorded in 2019 globally by non-communicable diseases, 38% were caused by CVD. This situation has increased the economic burden of the countries by affecting the people who are of working age. Although the occurrence of cardiovascular diseases (CVDs) differs from nation to nation, it is significant to observe that the

overall count of CVD cases almost doubled from 217 million in 1990 to 523 million in 2019. Moreover, the mortality count linked to CVDs rose from 12.1 million in 1990 to 17 million. (Roth et al., 2020). Nearly 80% of the deaths were contributed by LMIC. In Ethiopia, the mortality and morbidity rates are very high. In 2017 alone 2838767 people were affected from CVD and rheumatic heart disease (RHD), stroke, and ischemic heart disease (IHD) took the large number. But in 2021 stroke, IHD, and hypertensive heart disease (HHD) led to the death caused by CVD with about 170 people dying each day as to the journal of Africa reports (Yadeta et al., 2021). Ethiopia spent \$3102527667 on health and 11.11% went to handling NCD which is around \$344.67 million. This figure shows that NCDs such as cardiovascular disease leaves a huge burden on the economy of the country in general and the health sector in particular.

2.4. Common cardiovascular disease

According to WHO cardiovascular disease includes peripheral artery disease, cerebrovascular disease, rheumatic heart disease, deep vein thrombosis pulmonary embolism, and coronary heart disease, stroke.

2.4.1. Coronary heart disease

Coronary artery disease (CAD) is a medical condition marked by an insufficient supply of blood and oxygen to the heart muscle. This insufficiency results from the narrowing or blockage of coronary arteries due to the accumulation of plaque within them. This kind of conditions creates a mismatch between oxygen demand and supply. Coronary artery disease (CAD) is a prevalent global health issue, contributing to 2.2% of the total disease burden and accounting for 32.7% of all cardiovascular diseases. It can be broadly classified into two categories: acute coronary syndrome (ACS) and stable ischemic heart disease (SIHD). If left untreated, CAD can advance to congestive heart failure (CHF), underscoring the importance of timely intervention. The disease prognosis relies on several modifiable and non-modifiable factors. Some of them are age, family history, gender, smoking and also other comorbid situations dyslipidemia, hypertension, diabetes mellitus and chronic kidney disease have also contributed to the outcome. Different tests are used to assess CAD, including electrocardiogram (EKG), echocardiography (echo), CXR (chest X-ray), stress tests, cardiac catheterization, and blood work. The kind of test that will be chosen depend on the patient's presentation and context, helping to diagnose and assess the condition. The specific CAD disease kind decides the treatment that will be given. But typically, medication and non-medical interventions are conducted. Non-pharmacologic interventions are lifestyle changes

include regular exercise, quit smoking and weight reduction etc.(Shahjehan & Bhutta, 2022)

2.4.2. peripheral artery disease

Peripheral arterial disease (PAD) is a situation that leads to decreased blood flow through the arteries, resulting in reduced circulation to the extremities. This can cause pain in the thighs or calves while walking or during physical exertion and we call the pain from walking claudication or to limp. PAD has a significant global impact, affecting more than 200 million adults globally. The PAD's incidence rises significantly in individuals aged 70 and above, with rates reaching as high as 20%. Some of the risk factors contributed to this disease are smoking, diabetes, hypertension and family history. Physical examination, objective test and history of patient results determines the diagnosis result. Patients diagnosed with PAD may or may not exhibit symptoms, but they face an elevated risk of myocardial infarction, and stroke, thrombosis in relation to individuals without arterial disease. Consequently, it is vital for all PAD patients to implement lifestyle adjustments aimed at lowering their risk. Key focal points for these lifestyle changes include reducing cholesterol levels, quitting smoking, and effectively managing diabetes and hypertension.(Michael R. Zemaitis et al., 2023)

2.4.3. Rheumatic heart disease

Rheumatic heart disease (RHD) is a chronic condition that impacts heart valves, impacting approximately 40 million peoples globally, particularly children and young adults living in poverty. RHD arises as a result of a preventable infection initiated by the Strep A (Group A Streptococcus bacterium), which can induce an autoimmune response in the body called acute rheumatic fever (ARF). If not treated, Strep A infection and ARF can increase the likelihood of recurring infections. Although symptoms like fever and other rheumatic manifestations often resolve over time, carditis linked to ARF can lead to enduring harm to one or more of the heart's four valves. RHD widely spread in African and Asian countries. Improving public awareness, promoting reproductive and ani natal service, implementing community-based programs can prevent the disease. (Peters et al., 2020; Watkins et al., 2018)

2.4.4. Stroke

Stroke is a condition affecting the nervous system, where blood vessels become blocked, leading to the formation of clots in the brain. These clots obstruct the normal flow of blood and cause arteries to become clogged, resulting in the rupture of blood vessels and subsequent bleeding. When arteries leading to the brain rupture during a stroke, it causes the rapid death of brain cells

due to insufficient oxygen supply. Stroke rates doubled in low- and middle-income countries from 1990 to 2016, while developed countries saw a 42% decrease during the same period. There are two kinds of stroke: Hemorrhagic and Ischemic occlusions or infections. Both modifiable and non-modifiable risk factors are contributed to stroke disease. Some of those factors are: hypertension, diabetes, lack of physical exercise and drug abuse. Handling both mentioned risk factors can improve effects.(Kuriakose & Xiao, 2020)

2.5. Treatment and management of cardiovascular disease

The treatment and management of the cardiovascular diseases are determined by the specific condition, the severity of the disease, and the individual's overall health. The person who struggles with cardiac disease may originally experience the symptoms such as shortness of breath, chest pain, radiating discomfort in the arms, left shoulders, and elbows, high blood pressure, nausea, dizziness, discomfort while walking, fatigue, and more. (Arunachalam, 2020b). To prevent the disease some general guidelines, need to be followed. The first is lifestyle modification like experiencing regular exercise, reducing alcohol intake, avoid tobacco usage. Secondly, using different kinds of medication such as Cholesterol-lowering drugs (statins) to lower the risk of heart attacks and strokes, Blood pressure medications to reduce the workload on the heart and blood vessels, Antiplatelet and anticoagulant medications to prevent blood clots. Lastly, Medical procedures such as Angioplasty and stenting to open blocked arteries. Another option is coronary artery bypass grafting (CABG), which reroutes blood flow around blocked arteries to enhance circulation and alleviate cardiac issues.

2.6. Overview of Machine learning

The advancement of computers and technology has led to significant breakthroughs in various fields, including medicine, law, business, agriculture, and more. (Rabbi, 2019). In the 21st century, numerous organizations around the globe have used computer systems together with artificial intelligence to meet their customer interest and expectations, maximize their profits, improve production and service, and save other resources. Machine learning is a subset of AI that concentrates on building algorithms and statistical models enabling computers to learn from data and autonomously make decisions without the need for explicit programming. It can work with a huge amount of data to solve complex problems. It can detect patterns from data and give appropriate prediction results or target values. The aim is to empower machines to enhance their performance by scrutinizing and manipulating vast amounts of data. The core idea of machine

learning is to train computers to locate patterns and arrive at intelligent conclusions based on past experiences, rather than relying solely on predetermined instructions. Scholars have categorized machine learning approaches into four main kinds, namely supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Verma & Verma, 2022).

Supervised machine learning

Supervised learning is a machine learning approach where a model is trained using labeled data, which means that the desired or correct output is already known during the training process. The algorithm consists of a target or dependent attribute that is predicted from a set of predictors or independent attributes. The labeled data is utilized to train the algorithm, and the acquired knowledge is subsequently applied to make predictions for new, unseen data. In supervised learning, there are two primary tasks: classification, which forecasts categorical labels, and regression, which predicts continuous numerical values. A range of supervised learning algorithms exists, including DT, NN, LR, SVM, and LR, RF. These algorithms find extensive use in various domains, including healthcare, marketing, finance, and engineering.

Unsupervised machine learning

Unsupervised learning is a machine learning approach method where the data is not labeled or unstructured, and the algorithm discovers patterns or structures without any pre-existing knowledge of the outputs. In unsupervised learning, two common tasks are clustering and dimensionality reduction. Clustering algorithms categorize data points that share similarities in their features, whereas dimensionality reduction methods strive to decrease the number of features while preserving as much of the original information as feasible. Examples of this approach include PCA (Principal Component Analysis), t-SNE (k-means clustering, hierarchical clustering, and t-distributed Stochastic Neighbor Embedding). These techniques are valuable for organizing and summarizing complex data without labeled outcomes, which are applied in customer segmentation, anomaly detection, and data visualization. However, evaluating and interpreting unsupervised learning algorithms can be more difficult than supervised methods.

Semi-supervised machine learning

This is a distinct category of machine learning that incorporates elements of both supervised and unsupervised learning. In this approach, the data used for training the algorithm is partially labeled, meaning that some data points have known outputs or labels, while others remain unlabeled. Semi-supervised learning utilizes the unlabeled data to uncover underlying patterns or structures within

the data, all the while using the labeled data to guide the learning process and enhance the model's accuracy. This approach can be especially valuable when obtaining labeled data is limited or costly. Semi-supervised learning finds applications in fields such as computer vision, speech recognition and natural language processing.

Reinforcement learning

Reinforcement learning falls within the realm of machine learning, involving an agent interacting with an environment and makes decisions based on the feedback or rewards it receives from the environment. The agent learns to identify the actions that lead to the most desirable outcomes based on the rewards or penalties it receives, and the process is iterative over a series of trials. The agent uses this feedback to adjust their behavior and learn to make better decisions in the future. Reinforcement learning is frequently applied in various domains, including autonomous driving, robotics and gaming.

2.7. Machine learning and health sector

The health sector is an essential component of a country to maintain and improve the health of the population. It plays an unprecedented role in preventing, diagnosing, and treating illnesses and injuries, as well as in promoting health and wellness through education and outreach programs. Organizations like the health sector are critical to prevent disease and improving the quality of life in a certain society. However, healthcare sector faces numerous challenges, including a growing burden of illnesses, the presence of comorbidities, rising disability rates, epidemiological transitions, heightened demand for healthcare services, elevated societal expectations, and escalating healthcare expenditures. (Panch et al., 2018). A fundamental overhaul of healthcare systems is essential to tackle the challenges they confront and to realize the goal of universal health coverage (UHC) by the year 2030. ML plays a critical role in transforming this sector by providing medical decision support, automating healthcare billing, and creating medical care regulation just to name a few (Verma & Verma, 2022).

The challenges of health care sectors in low-income countries like Ethiopia are numerous, ranging from the increasing burden of disease, morbidity, mortality, disability, greater demand for health services, delayed diagnosis, and lack of manpower to low productivity of health workers (Atun, 2015). To improve healthcare service, giving more focus on prevention is important but as known every disease is not preventable. However, diagnosing the disease early provide better results and reduces expenses for treatment. Early identification of the disease makes the treatment process

simple and feasible. Machine learning is preferable for such kinds of purposes. It is widely used for disease detection, drug discovery, medical imaging, manufacturing, personalized medicine, and outbreak prediction. Machine learning uses medical imaging, patients' history, and laboratory test results to identify disease presence or absence in a person. This thesis work uses a supervised learning approach to predict CVDs.

2.8. Machine learning approach in disease prediction

In today's world, humans suffer from different kinds of diseases. The effects of diseases on individuals can go beyond physical and also affect their mental well-being. Diseases can arise from various causes but mainly from infection, deficiency, heredity, and organ dysfunction(De & Chakraborty, 2020). In community, medical practitioners have the duty to detect and identify illnesses accurately and offer appropriate treatments to control or cure the diseases. While some diseases can be cured through treatment, chronic diseases are not curable but can be managed with treatment to prevent them from worsening over time. Therefore, it is crucial to detect and treat diseases in their early stages. Recently, Machine learning involvement in healthcare are growing and are used mainly in diagnosis and predictions of so many diseases and assisting medical experts to analyze patient data and driving meaningful insights, and making the right decisions (Dahiwade et al., 2019). Several researchers have tried to apply Machine learning methods in the areas of predicting numerous kinds of diseases. Cardiovascular diseases are one of the diseases that scholars have tried to address through algorithmic approach such as machine learning algorithms (ensemble, linear regression, naïve bayes, decision tree, etc.) and deep learning algorithms (multilayer perceptron, convolutional neural network, etc.). In this study we employed ensemble model development from machine learning algorithmic approach.

Ensemble learning involves aggregating multiple models in supervised machine learning to improve prediction accuracy and robustness(Sagi & Rokach, 2018). It is a broad term that includes various methods for combining models. The technique operates on the principle that errors made by one model can be compensated by other models, resulting in better overall performance compared to using a single model(Sagi & Rokach, 2018b). This approach was selected because it can enhance model's prediction performance by addressing issues such as overfitting, high computational cost, imbalanced data, and the curse of dimensionality. In other words, this method able to improve accuracy and efficiency of the model by minimizing errors and handling complex data. Due to those mentioned and other approaches, we chose the ensemble model development

approach to design a predictive model for CVD disease. The limitation of ensemble model is the demand of extensive computational resources for evaluating prediction unlike single model(Shahid et al., 2015).

Ensemble machine learning basically categorized into three: Bagging, Boosting, and Stacking(Yuan et al., 2020). Bagging entails the generation of multiple datasets through the random sampling of instances from the original training data, with replacement being applied(Ashfaq et al., 2022). These individual datasets are then employed to train separate base classifiers. The ultimate classification decision is reached by combining the results of these models using techniques such as majority voting or averaging. Employing this aggregation method enhances performance and accuracy in comparison to what standalone models can achieve(Ashfaq et al., 2022). Bagging serves the purpose of mitigating overfitting and reducing the variability of a specific classifier by creating multiple training datasets. In contrast, Boosting is an ensemble learning method with the capacity to elevate the performance of weak models to a higher level of strength(Ashfaq et al., 2022). In boosting, data samples are chosen in a sequential manner for model training, with each iteration focused on reducing the errors of the previous one. The core concept behind boosting is its ability to transform a weak classifier into a robust one through a series of iterations(Raza, 2018). Moving to Stacking generalization, commonly known as Stacking, this ensemble framework involves training a distinct machine learning algorithm to integrate predictions from multiple ensemble members (Mienye & Sun, 2022). This technique is particularly powerful as it leverages the strengths of several high-performing predictors to produce predictions that outperform the individual models within the ensemble. Stacking employs diverse base algorithms on the same dataset, providing various perspectives on predictive modeling (Mienye & Sun, 2022). In contrast to bagging, which relies on subsets of input data with decision tree models, stacking employs different algorithms with the same dataset. Unlike boosting, which sequentially corrects predictions, stacking utilizes a single model to learn the optimal way to combine base learners' predictions. For our thesis work, we chose the Stacking method, as it specifically targets individual learners with diverse characteristics. By integrating these distinct models, we aim to extract the strongest aspects from each one, resulting in improved overall performance. Therefore, we believe that the Stacking approach is more suitable for our specific research objectives compared to other ensemble techniques. The algorithms RF, XGB, and SVM was chosen as individual learners for stacking ensemble, and are discussed below:

Random Forest (RF): is an ensemble of decision trees generated from data samples. It aggregates predictions from each individual tree and ultimately determines the best outcome through majority voting. An aggregated trees operate entirely independently of each other. When a test sample is provided to RF, the class label for this sample can be determined by aggregating the voting outcomes from each individual predictions(Dritsas et al., 2022; Parmar et al., 2019).

Extreme Gradient boosting (XGB): is an exceptionally efficient machine learning algorithm that primarily relies on decision tree algorithms as its fundamental components. The ultimate model produced by XGB comprises multiple decision trees, and it employs a sequential dataset training approach to combine weaker predictors (Decision Trees) and attain more robust predictors. To prevent overfitting, it incorporates a regularization term during the modeling process. One of the notable advantages of the XGB algorithm is its optimal utilization of memory resources and its effective handling of missing values during training (Kiangala & Wang, 2021).

Support Vector Machine (SVM): s a widely used and advanced method. It belongs to the category of supervised learning and comes with specific learning algorithms for analyzing data for regression and classification purposes. Besides their ability to perform linear classification, SVMs are proficient at handling non-linear classification tasks through a method called the kernel trick, which covertly transforms their inputs into higher-dimensional feature spaces. In essence, they create boundaries or margins between different classes(Mahesh, 2018).

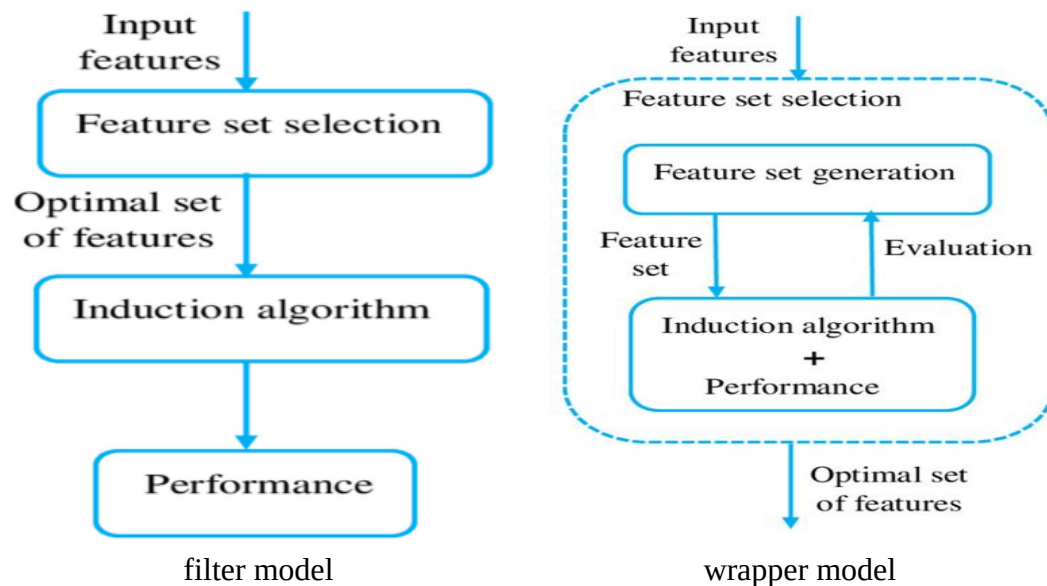
Table 2:1 Selected base classifiers for stacking based ensemble model

Algorithms	Advantage	Disadvantage
Random forest (RF)	handles both linear and nonlinear relationships, Simple to implement, high classification performance, and efficient, can deal with base-variance trade-off, often scaling of the variable is no required.	Computationally expensive for high dimensional data, suffers interpretability and unable to determine importance of each variable
Extreme gradient boosting (XGB)	Scalable, flexible, handle sparse data, support distributed platform, improved classification accuracy, uses memory resources optimally(Kiangala & Wang, 2021)	Relatively inefficient in terms of time, less interpretable, its performance is affected by class imbalance in dataset(Kiangala & Wang, 2021)

Support Vector Machine (SVM)	Resistant to overfitting, effective for high dimensional dataset, relatively efficient in terms of memory(Ray, 2019).	Requires relatively more computational power, not much suitable for data with much noise and very high dimensional space, does not give probability estimates(Ray, 2019).
------------------------------	---	---

2.9. Feature selection methods

Feature selection is the process of choosing relevant attributes for certain kinds of algorithms to improve performance and reduce storage and computational time requirements (Dhal & Azad, 2022). It is used to deduct the number of attributes in a dataset by selecting the most relevant ones for creating a model. This helps to eliminate contradictory and noisy data and to increase performance of the model. In CVD prediction, feature selection techniques have used to retrieve significant features from patients' medical histories, laboratory test results, and demographic information.



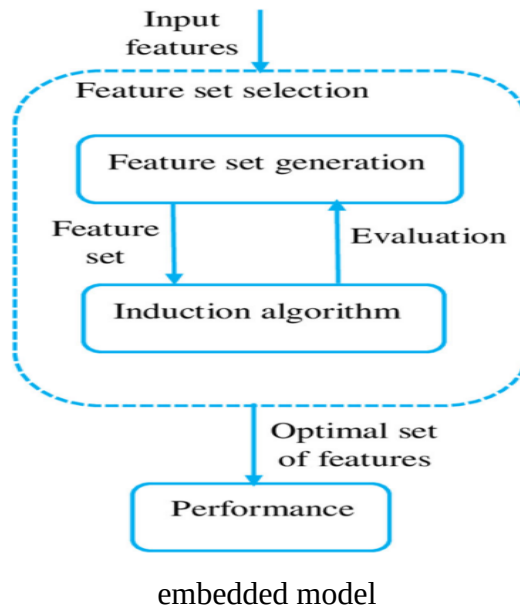


Figure 2:1 The General Frameworks for Filter, Wrapper, and Embedded model

In general, feature selection models can be divided into three primary types: filter, wrapper, and embedded, and their summarized working nature is going to be presented below(Dhal & Azad, 2022):

Table 2:2 Comparison of filter, wrapper, and embedded feature selection models across several key aspects: (Dhal & Azad, 2022)

Parameters	Filter	Wrapper	Embedded
Procedure	Statistical test	Cross-validation	Cross-validation
Assessment	Feature subset relevance	Feature subset	Feature subset
Criteria		usefulness	usefulness
Searching	involves evaluating features either through nested feature subsets or by assessing the individual rankings of features.	All possible feature subsets.	Is controlled by the learning process
Results	Robust against overfitting. There is a probability to fail on selecting useful features.	closest to overfitting. Selects the most important attributes although it has massive time complexity.	Less prone to overfitting. relatively smaller time complexity.

For the prediction of CVDs, we selected feature selection methods from the two categories by considering their ability to improve model performance. We chose recursive feature elimination from the wrapper method, and lasso regularization (L1) from the embedded method. The models were trained using those methods and the one which outperform the other was selected for model development. Some descriptions of the methods are given below.

Recursive feature elimination (RFE): a feature selection method that iteratively creates models, removes bad performing feature during each iteration and ranks the features based on their elimination order. RFE is required to identify the most significant features of a given task. Due to its simplicity in application and deployment, along with its effectiveness in choosing crucial features from training datasets for deciding target variables while removing less impactful features, the RFE algorithm has gained widespread popularity (Senan et al., 2021)

Least Absolute Shrinkage and Selection Operator (LASSO) regularization: this method combines feature selection and regularization by introducing an additional penalty term into the cost function. This helps to reduce the magnitude of less important feature coefficients to zero, effectively removing them from the model and selecting the most relevant features for the task at hand. LASSO improves model interpretability by removing variables that have no meaningful relationship to the response variable. This procedure not only enhances interpretability but also reduces the possibility of overfitting (Fonti & Belitser, 2017).

2.10. Related works

Machine learning algorithms have found extensive application in predicting various types of illnesses, including cardiovascular diseases. scholars tried to address the issue by following different kinds of problem-solving approaches. In this section, the works of the researcher related to cardiovascular disease prediction using machine learning approaches have been reviewed.

(Arunachalam, 2020b) proposed a prediction algorithm to predict cardiovascular disease using 300 instances with 13 attributes. The employed algorithms were Gradient Boosting, Random Forest, Extremely Randomized Trees (Extra Trees Classifier), Multi-Layer Perceptron (MLP) Classifier, Support Vector Machine, and Logistic Regression, and got the maximum accuracy of 91.7% by SVM and MLP. Data cleaning and feature engineering such as categorical encoding was used during pre-processing phase and specificity and sensitivity evaluation metrics were used to obtain accuracy. The drawback identified in this work is that the researcher only handled noise and missing values in preprocessing phase. But feature selection plays a significant role in improving

performance by minimizing execution time. It supports as to choose determinant features which has an impact on the result. Another flaw of this work is that size of the dataset used to build the model was limited.

(Alqahtani et al., 2022) proposed a model based on ensemble through machine learning and deep learning models to predict CVD. A dataset of 7000 from Kaggle with eleven attributes was used for the research and RF feature selection method was used to extract important features. From the machine learning classifier, KNN, RF, XGB, DT, and from deep learning, deep neural network (DNN) and Keras deep neural network (KDNN) were used. To identify the correlation between variables and eliminating extraneous, duplicate, and repetitive features from the dataset, The PCC (Pearson correlation coefficient) was utilized. The machine learning ensemble uses the output of classifiers and utilizes weighted majority method to provide the results. The model attains an accuracy of 88.7% but employed one feature selection method

In 2019, (Bekkali et al., 2019) proposed a SVM with linear kernel model to predict CVD after exploring different machine learning algorithms. They utilized the dataset from UCI online database which contains 303 instances with 13 features. they tried to improve the performance of the model by filling up the missing cells with the mean values of each column during a preprocessing stage. Finally, they found an accuracy of 86.8%. Limited dataset and no indication of whether features selection methods utilized or no during model construction in the article is identified as a limitation.

Aseel Alfaidi(Alfaidi et al., 2022), suggests a method of using Machine Learning (ML) models to diagnose the likelihood of an individual having CVD. The study conducted experiments utilizing seven different algorithms and a publicly available dataset to build the models. A Chi-square test was utilized to identify the most crucial features for predicting CVD. The results of the experiment indicated that the Multi-Layer Perceptron model had the most accurate diseases predictor, achieving an accuracy of 87.23%. the authors suggested applying other feature selection methods, model development techniques (such as deep learning) and datasets as the future work to obtain better results. failing to try other feature selection method seen as drawbacks.

Arroyo & Delima, 2022 introduced a genetic algorithm artificial neural network (GA-ANN) with the aim of improving the performance of artificial neural networks (ANNs). They utilized the SoftMax function as the activation function, PSO (Particle Swarm Optimization) as feature selection, and employed the Adagrad optimizer for optimization purposes. To evaluate the efficacy

of their proposed model, they conducted a comparative assessment against alternative forecasting algorithms, including RF, LR, DT, and SVM. Remarkably, GA-ANN achieved an accuracy rate of 73.43%, surpassing the performance of the other six algorithms. limitation regarding this research is using only accuracy as evaluation metrics and unable to try other method for identifying important features.

(V. D. A. Kumar et al., 2022) proposed a prediction model based on SVM for identification of cardiovascular disease at early stages. Symptoms and risk factors associated with the patients were utilized as parameters to build the model. Sensitivity, Accuracy, Specificity, and were used as evaluation metrics and 81.2187% accuracy was achieved. The source of collected data and amount utilized is not indicated in this paper and whether feature selection techniques are utilized or not is unknown. This research work lacks diversity of machine learning model

(Islam et al., 2020) proposes the use of machine learning techniques to forecast CVD and identify the risk factors linked with it. The authors have used a publicly available UCI dataset and they carried out preprocessing and feature selection tasks to train and test four machine learning algorithms including DT, LR, SVM, and naïve bayes. The algorithms were compared against their accuracy and LR was chosen by performing better than the other with 86.25% accuracy. The limitations that were recognized pertain to the size of the dataset, and the authors also noted that while they employed technique for feature selection, they did not explicitly specify this method.

(Geetha et al., 2021) conducted research work that uses K-nearest neighbor to identify the likelihood of a person having cardiac disease at an early stage based on several risk factors. Google Co lab was utilized for experiment purposes. At the time of model development, the dataset (Obtained from UCI and has 301 instances) was divided into a 67%:33% ratio where 67% for training and 33% for testing. Euclidean distance was applied for calculating the distance between data points. The tried to build the model by varying the value of K to 1 to 21 and obtained highest accuracy of 87%, recall of 0.437, precision of 0.8428, and F-score of 0.617 when the valued of k set to 8. The con of this work is the dataset size and limited diversity in machine learning model utilization.

(Tiwari et al., 2022b) proposed stacking based ensemble framework for the prediction cardiovascular disease. In this approach, a careful selection of algorithms, including the random forest classifier, extra tree classifier, extreme gradient boosting, and gradient boosting classifier, was made based on accuracy comparisons against six other classifiers. The dataset used for this

analysis comprised 11 features and 1190 instances obtained from the IEEE data port. The model's performance was assessed using various evaluation metrics, including specificity, ROC, F1-score, AUC, sensitivity, MCC, and accuracy. The proposed framework attained an accuracy of 92.34%. The authors suggested dataset size improvement as their future work. In this paper, whether feature selection method is utilized or not was not mentioned.

(Ahmad, 2020) utilized four machine learning techniques (SVM, KNN, RF, DT) to detect whether a person has a CVD disease or not, with the most accurate algorithms selected for prediction. The dataset was obtained from Kaggle and has 600 instances with 14 attributes. The model was implemented on a web platform using Flask Python framework with 13 input fields for user prediction. The results obtained using the K-Nearest Neighbors Classifier algorithm have an accuracy of approximately 85.83%, indicating a good performance for the model. In this paper, Standard Scaler was used for scaling purposes. Utilizing only accuracy as evaluation metrics and nothing was written about preprocessing and feature selections. Additionally, dataset size needs an improvement.

Table 2:3 Related work summary

Authors and year	Title	Methods and findings	Evaluation metrics	Research Gaps of related papers
(Arunachalam, 2020)	Cardiovascular prediction using machine learning algorithms	RF, SVM, ETC, LR, and MLP with maximum accuracy of 91.7% by SVM and MLP.	Accuracy, sensitivity, and specificity	➤ Small number of parameters used ➤ Limited to binary classification (CVD or Not) ➤ Only 300 instances of dataset with 13 attributes used ➤ Limited diversity of machine leaning model utilized
(Alqahtani et al., 2022)	Cardiovascular Disease Detection using Ensemble Learning	ML Ensemble of RF, KNN, DT, XGB, and DL Ensemble of DNN and KDNN. ML ensemble attain the highest accuracy of 88.7%	ROC-AUC, Precision, Accuracy, F1 score, and Recall	➤ The model classifies the disease into binary classes (CVD or Not) ➤ Limited feature selection methods (RF)
(Bekkali et al., 2019)	Identification of cardiovascular disease using machine learning	SVM-RBF kernel, SVM-Linear kernel and NB. SVM-Linear kernel performs well with 86.8%	sensitivity, specificity, Accuracy, Precision, and F-measure	➤ small dataset of 303 instances with 13 features ➤ Limited to binary classification (CVD or Not) ➤ No feature selection method ➤ Limited diversity of machine leaning model utilized
(Alfaidi et al., 2022)	Machine Learning: Assisted Cardiovascular Diseases Diagnosis	LR, DT, NB, SVM, RF, KNN, and MLP. MLP provides 87.23% accuracy which is the highest.	Accuracy, sensitivity, specificity, Precision, and F-measure	➤ Limited to binary classification (CVD or Not) ➤ limited feature selection method (Chi-square test)

(Arroyo & Delima, 2022)	An Optimized Neural Network Using Genetic Algorithm for CVD Prediction	GA-ANN, LR, ANN, SVM, DT, KNN, and RF. GA-ANN attained the highest accuracy of 73.43.	Accuracy.	➤ Used only one evaluation metrics (Accuracy) ➤ Limited to binary classification ➤ No feature selection
(Islam et al., 2020)	Cardiovascular Disease Forecast using Machine Learning Paradigms	LR, DT, NB, and SVM. LR has beating accuracy of 86.5%	F1 score, Recall, Precision, and Accuracy	➤ Limited to binary classification (CVD or Not) ➤ Small dataset to train the model (301 instances with 12 attributes) ➤
(Geetha et al., 2021)	A Method of Cardiovascular Disease Prediction using Machine Learning	KNN with 87% accuracy	Precision, Accuracy, Recall, and F1 measure	➤ Limited diversity of machine leaning model utilized ➤ Limited to binary classification (CVD or Not) ➤ Dataset with 303 instances with 14 features used, which is limited
(Tiwari et al., 2022b)	Ensemble Framework for Cardiovascular Disease Prediction	Stacking based ensemble with four algorithm such as RFC, ETC, GBC and XGB. The attained accuracy result is 92.34%	Specificity, ROC, Sensitivity, MCC, F1-score, and Accuracy	➤ Limited to binary classification (CVD or Not) ➤ No feature selection was employed
(Ahmad, 2020)	Cardiovascular Diseases (CVDs) Detection using Machine Learning Algorithms	SVM, KNN, RF, and DT. With the highest accuracy of 85.83% KNN	Precision and Accuracy	➤ Limited to binary classification (CVD or Not) ➤ Dataset size needs an improvement (600 instances with 14 features)

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1. Introduction

This section outlines various strategies, procedures, and methodologies that aim to utilize machine learning methods for predicting cardiovascular diseases. This encompasses the utilization of techniques such as data preprocessing and feature engineering. It contains the description of models and performance evaluation metrics utilized in the study. Additionally, the section defines systematic approaches, and procedures to be followed. Experimental scientific research was used to design the research.

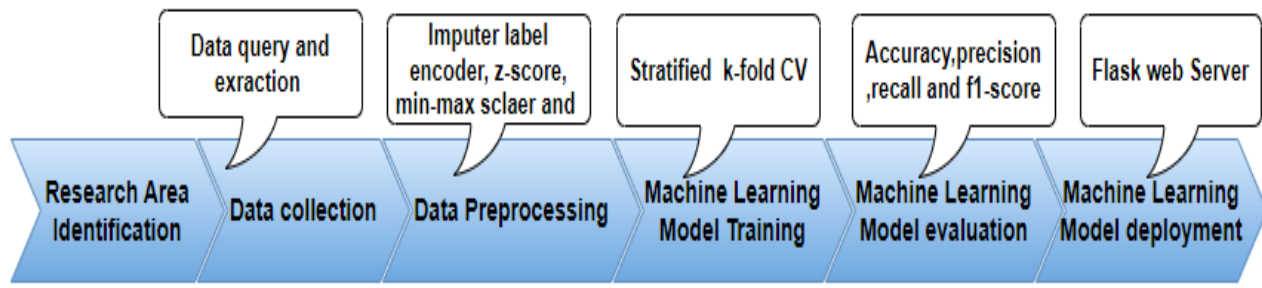


Figure 3:1 The research workflow diagram

3.2. Data collection

The data required for the study was obtained from electronic record of cardiac patients admitted to the hypertension and cardiology unit that attends their treatment during the period 2019 to 2023 at SPHMMC (St. Paul's Hospital millennium medical college) Addis Ababa, Ethiopia. Moreover, distinct parameters to diagnose cardiovascular disease are identified by: consulting cardiologist, assessing different local and global references on cardiovascular diseases, and considering online available features of the dataset regarding the diseases. The dataset was obtained by reviewing the patient history in the HCU. The collected dataset has 19 features, 18 independent and 1 dependent feature with a 4-target class values. SPHMMC is the largest tertiary teaching public hospital in Addis Ababa, Ethiopia (Mengistu & Benti, 2022). The Hospital plays a vital role as a crucial healthcare institution, offering services to patients requiring follow-up care, emergency cases, and referrals from other medical facilities and healthcare providers throughout the country. The hospital is governed by a board under federal ministry of health. According to a different source of information. According to a different source of information, the hospital provides service to

1200 emergency and outpatient individuals daily. In addition, it offers an opportunity for researchers and policymakers to monitor and evaluate programs.

3.3. Cardiovascular diseases prediction modeling

3.3.1. Data preprocessing techniques

Data preprocessing encompasses a set of techniques and procedures used to convert raw data into a format that is well-suited for analysis and machine learning applications. This stage holds immense importance in machine learning because it ensures quality of the data which directly influences the accuracy and efficiency of the models (Goyal, 2021). The collected dataset was available in a format where not suitable for the models to train since it has missing values, noise, irrelevant information, and inconsistencies. By working on the raw dataset, we made it suitable to build the model. It is known that in order to develop a usable model, it is necessary to go through several steps on the gathered data to ensure data quality in such a way that the model can learn from and perform as expected (Shan et al., 2019). Furthermore, proper preprocessing of the datasets is essential to achieve better accuracy and efficiency in proposed model during training and testing.

Data cleaning: involves addressing several issues we observed in the prepared dataset, such as missing values, irrelevant data, outliers, and misspelled categorical data values. We obtained those mentioned information by employing various exploratory techniques, with Python serving as a tool for implementation. The problem was handled by applying different techniques. Irrelevant information was removed, missing values were filled using Simple imputation technique such as mean, median and mode. this technique involves replacing missing values in each individual value by utilizing all the non-missing values of either a quantitative or qualitative attribute (Emmanuel et al., 2021). It is widely used because of its simplicity. During imputation process, mean and median imputation were chosen for numerical features and mode was for categorical features for implementations (Lin & Tsai, 2020). In the mean-based method, missing values are substituted by using the calculated average value of that attribute derived from the available dataset. Conversely, in the mode approach, the missing attribute values are filled in using the attribute value from the observed data that occurs most frequently. Regarding outliers, we manage them through Z-score outlier detection techniques. Outliers are data points that significantly deviate from the typical distribution of the provided data. They may result in detrimental consequences, such as model misspecification, skewed estimates, and exaggerated standard errors. (Saleem et al., 2021). Z-score uses mean and standard deviation values to discover outliers in a dataset and computed as

follows(Nwodo et al., 2021).

$$\mathbf{Z} - \text{score}(i) = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where **Z-score**= outlier detector, x_i = each object in the distribution, μ = distribution mean, and σ = standard deviation. An absolute **Z-score** $|\mathbf{Z}|$ value greater than 3 is usually considered an outlier(Saleem et al., 2021). Based on this threshold values, we detected and removed outliers from CVD dataset.

Handling categorical data: since machine learning only be able to deal with numerical data, categorical (nominal or ordinal) data should be transformed or encoded into numerical values(Potdar et al., 2017). To encode categorical variables in CVD dataset, we used one of widely used encoding techniques which is label encoding (D McGinnis et al., 2018). Label encoding involves assigning integers to categories within a categorical variable based on the total number of categories. In this approach, additional columns are not introduced to the dataset. Instead, it depends on an implied order for the variable, which may not always convey significant information. We performed label encoding on categorical variables (Potdar et al., 2017).

Handling imbalanced data: Imbalanced data arises when there is a significant imbalance or disproportionate ratio of observations among different classes. To address this issue, there are generally two levels of solutions available such as data-level solutions and algorithmic-level solutions (Rout et al., 2018). In our case there are four classes in the collected CVD dataset and we observed dataset imbalance between them. The two algorithms (RF and XGB) we selected incorporates the mechanism of handling class imbalance internally which operates by assigning class weight to each majority and minority class to make them balanced. But for other algorithms, we employed powerful and commonly used class imbalance handling techniques from data-level solutions called CC oversampling(Rout et al., 2018; Spelman & Porkodi, 2018).

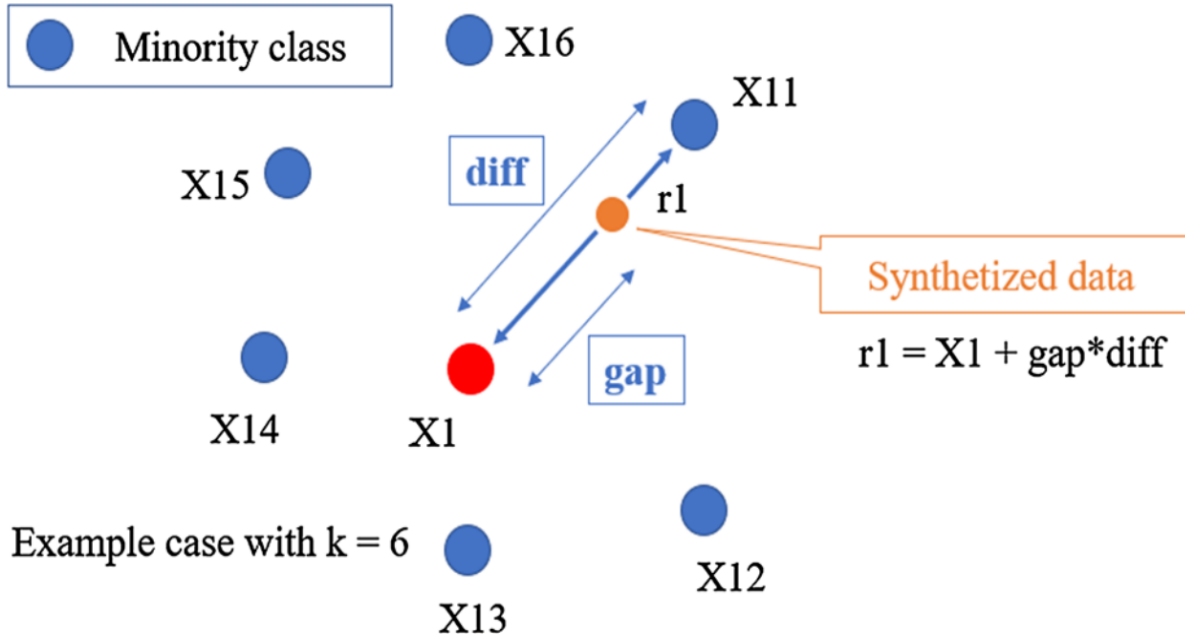


Figure 3:2 Dataset Balancing using SMOTE(Lerra et al., 2022)

In Synthetic Minority Over-Sampling Technique (SMOTE) systematic samples are generated for the minority class. These synthetic samples are not simply duplicated randomly; instead, they're generated by utilizing information from the existing data. SMOTE generates new data points by creating a linear combination of two similar samples from the minority class. It selects one sample randomly from the nearest neighbors of another minority class sample. Notably, this generation process takes place within the feature space, not the data space.(Kumari et al., 2020).

Feature Scaling: feature scaling techniques enable us to normalize our dataset to the required standards. This is another task we did before training our proposed ML model. By Conducting a feature scaling task in the dataset, we improved its performance. Two of algorithms we selected (RF and XGB) are fairly invariant to feature scaling but others are greatly affected. In the study, we used Min-Max normalization methods. Min-Max Normalization stands as one of the most commonly employed methods for data normalization. This technique involves a linear transformation of variables, utilizing the min (minimum) and max (maximum) values within the dataset (x). Through this process, the data is scaled or mapped, ensuring that the lowest value of x corresponds to 0 and the highest value of x corresponds to 1(Kappal, 2019).

The formula of Min Max normalization(scaling):

$$M_{scaled} = \frac{x - \min(X)}{\max(X) - \min(X)} \quad (2)$$

Where:

M_{scaled} = new min-max value

x = old value

$\min(X)$ = minimum value

$\max(X)$ = maximum value

3.3.2. Modelling

The model selection phase requires evaluating and comparing different model based on their performance using techniques and choose one with best performance with respect to the others. To address the problem of cardiovascular disease prediction, an ensemble-based stacking model was selected. Stacking is an ensemble method that entails training a "meta learner" by combining multiple classification models(Tiwari et al., 2022b). This process includes instructing a secondary-level meta learner to determine the optimal combination of the underlying base learners. In stacking a variety of strong and diverse learners can merged and provide a powerful ensemble model. The property poses by this model enable us to exploit the advantage of the individual models chosen to incorporate. The performance of stacking model is compared with three better performing models which is chosen based on their advantages they pose. The description of proposed and the selected models are given below.

Random Forest (RF): We opted for this algorithm because of its simplicity in implementation and its high predictive capabilities. Moreover, when it comes to risk prediction, the random forest algorithm has the potential to increase the precision of risk assessments by uncovering previously unknown risk factors and detecting intricate relationships among them(Dritsas et al., 2022; Sagi & Rokach, 2018b; Yang et al., 2020).

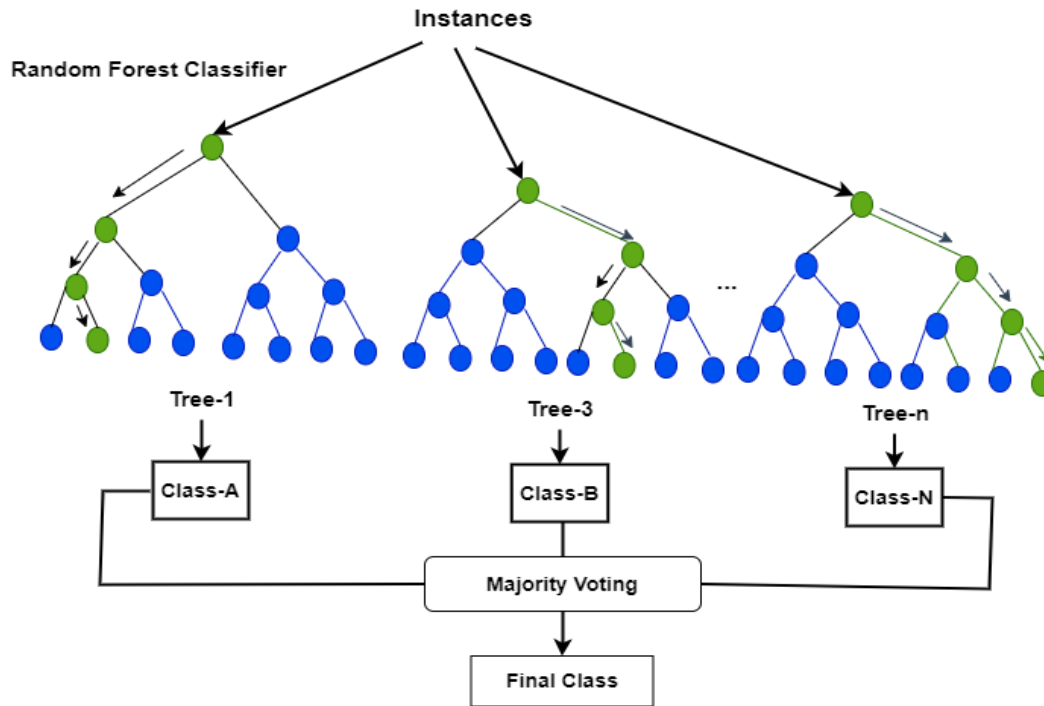


Figure 3:3 Random Forest Classifier(Azhari et al., 2019)

Random forest is applicable for regression and classification problems(Kiangala & Wang, 2021). The diagram shown above is for a classification problem since we proposed a solution for a problem that is classification in nature. In this problem the target class has discrete values, the random forest takes the majority vote to classify the final class but in a regression problem where the output values are continuous, the algorithms take the average of the whole decision tree classification to provide the final output.

Extreme Gradient boosting (XGB): is an algorithm that uses tree boosting and has become very popular among practitioners in the field. It has achieved success in many machine learning tasks, as evidenced by the fact that in 2015, 17 out of 29 winning solutions on Kaggle used XGB. To improve scalability, XGB has incorporated various algorithmic optimizations and refinements compared to GBM. One of these refinements is the addition of a regularization component to the loss function that aims to build ensembles that are simpler and more generalizable. Additionally, XGB can run faster than other models as it supports distributed platforms and can be distributed across multiple machines(Sagi & Rokach, 2018b).

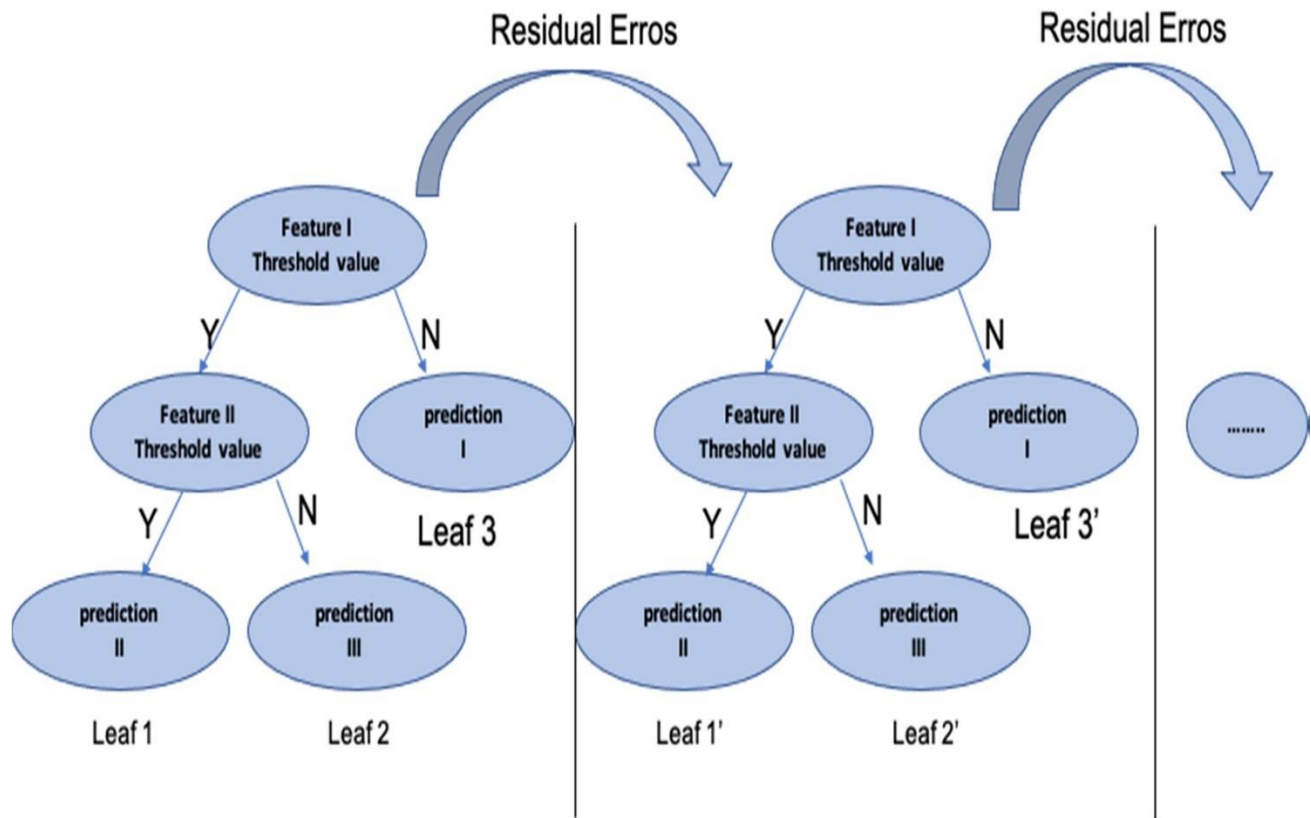


Figure 3:4 Extreme Gradient Boosting Trees (Ibrahim Ahmed Osman et al., 2021)

XG is an application of gradient boosting decision trees. In this algorithm, decision trees are employed sequentially, and weights are allocated to all independent variables, which are then utilized in the trees to make predictions. Variables that were predicted incorrectly are given higher weights and subsequently used in the second decision tree. These individual predictors are combined to construct a robust and precise model.

Support Vector Machine (SVM): is a supervised machine learning algorithm commonly used for both regression and classification tasks. It's commonly used for classification because it can achieve high accuracy with low computation power.

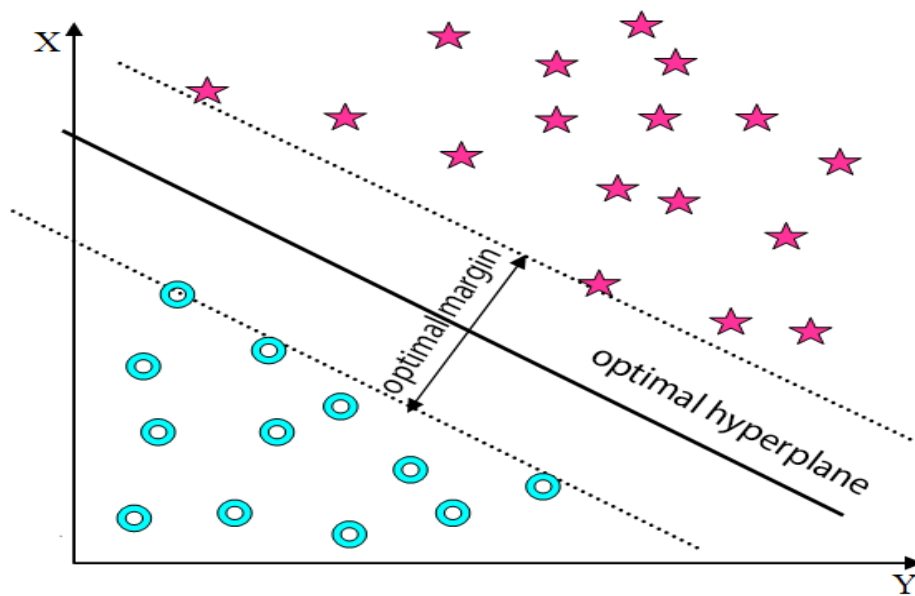


Figure 3:5 Support Vector Machine(S. Kumar et al., 2019)

The main objective of this classifier is to establish a hyperplane that can optimally separate the classes by adjusting the distance between data points and the hyperplane. Each attribute in the dataset is represented as a coordinate and a hyperplane is plotted for each one. The classifier identifies the hyperplane that separates one class from another to perform classification. distinct kernels are used to determine the hyperplane (Arunachalam, 2020b).

Stacking models

The cardiovascular disease prediction using machine learning approach was proposed in this study and particularly machine learning based stacking model is employed in this work. Since the stacking model development approach requires more than one heterogeneous base classifiers(learners) and one meta classifier. We selected three algorithms such as SVM, XGB and RF as base learners and one algorithm which is LR as meta classifier based on better predictive performance they pose.

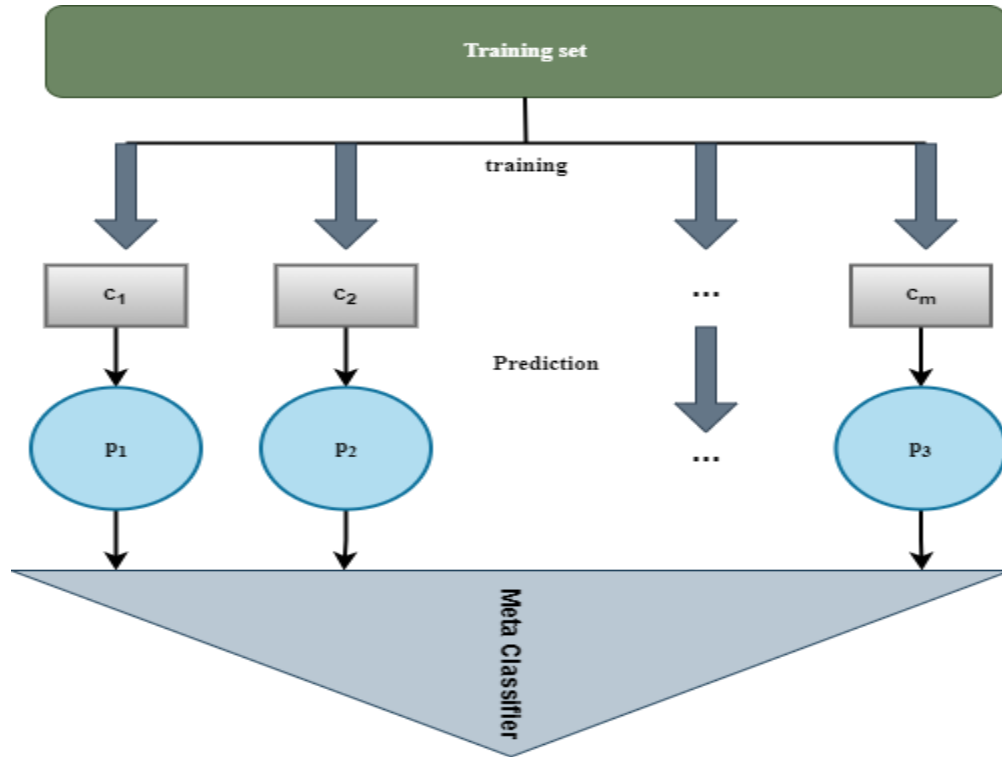


Figure 3:6 Stacking classifier architecture(Shan et al., 2019)

The selected algorithms enhance the model's performance and offer several advantages. For instance, SVM has the ability to prevent overfitting and efficiently utilize memory resources. XGB is both flexible and scalable, while RF excels in handling the bias-variance tradeoff.

3.3.3. Model evaluation metrics

Precise prediction is expected from the designed model to be applicable in the intended area. Different dataset splitting methods are available to build the model. we used stratified K-fold cross validation as splitting to construct the proposed model. The prepared dataset served as training and testing during splitting. The training dataset is used to fit the model whereas testing dataset is used during model's performance assessment. To evaluate the performance thoroughly, evaluation metrics were used. The objective of evaluation model is to establish the model's generalization ability to new data never the model exposed before.

Stratified K-fold cross-validation

Cross-validation is a technique that involves the successive alternation of training and validation stages in multiple rounds. This method is used to assess a model's performance and its capacity to generalize to new and unseen data. The main components of cross-validation include partitioning

the dataset into folds and sequentially rotating the training and validation sets among them(Raschka, 2018). Stratified K-fold cross-validation utilized is this work, which makes sure that each fold better represents the whole dataset.

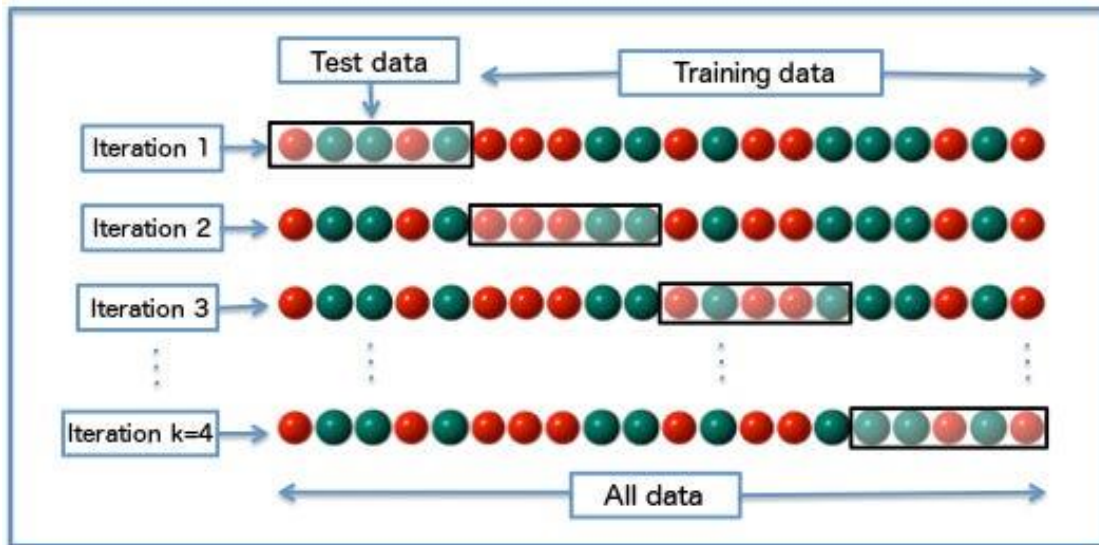


Figure 3:7 K-fold cross-validation(Doppala et al., 2021)

We opted for Stratified K-fold cross-validation due to its ability to address potential issues in standard K-fold cross-validation. In the traditional approach, the dataset is randomly divided into folds, which can lead to imbalanced fold distributions and potential bias in model evaluation, especially in the presence of imbalanced class distributions. This can result in some folds lacking samples from minority classes, significantly impacting model assessment. Stratified K-fold cross-validation was specifically designed to mitigate these concerns by rearranging datasets to ensure that each fold maintains a more representative sample of the entire dataset, particularly with regard to class distribution. This choice was made to provide a more robust and unbiased assessment of the model's performance.

3.4. Performance evaluation metrics for prediction modeling

Model evaluations are utilized to appraise the effectiveness of a model, explicate its efficient functioning, and assess the relationships it establishes(Raschka, 2018). Selecting appropriate evaluation metrics is essential because different machine learning algorithms use different metrics to obtain their performance. To assess the capability of the model, we used the performance evaluation metrics described below.

Accuracy: we used this evaluation metric to know the number of observations the model predicted

correctly for the disease. it is a measure of how closely predicted values align with actual values, indicating the model's performance across all classes and is denoted by:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision: is determined by dividing the total count of true positives by the sum of true positives and the count of false positives detected by the proposed model among all the expected positive cases in the datasets.

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Recall: characterizes the model's capability to identify all relevant instances of a specific class within a dataset. It is quantified as the ratio of true positives to the sum of true positives and false negatives.

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (5)$$

F1 score is a metric that calculates the harmonic mean of both recall and precision values. It considers both false negatives and positive instances in its calculation.

$$\textbf{F1 - score} = 2 * \left(\frac{\textbf{Precision} * \textbf{Recall}}{\textbf{precision} + \textbf{Recall}} \right) \quad (6)$$

Confusion matrix: is a table employed to assess the performance of a classification model by contrasting the predicted labels to the actual labels. We used a confusion matrix to visualize the report of four classes (CAD, PAD, RHD and Stroke). The tables below depict the confusion matrix of four classes.

Table 3:1 Confusion matrix

predicted Values	Actual Values				
		CAD	PAD	RHD	STROKE
	CAD	True CAD	False PAD	False RHD	False STROKE
	PAD	False CAD	True PAD	False RHD	False STROKE
	RHD	False CAD	False PAD	True RHD	False STROKE
	STROKE	False CAD	False PAD	False RHD	True STROKE

CHAPTER FOUR

4. PROPOSED SOLUTION FOR CARDIOVASCULAR DISEASES PREDICTION

4.1. Introduction

This section delves into proposed model architecture, descriptions of features, and other essential components necessary to realize the proposed model. Furthermore, the model's development phases and the steps undertaken during the study are explained. The model is defined to address the research questions and to fill identified gap identified in literature review.

4.2. Proposed architecture for cardiovascular disease prediction

This section outlines the architecture proposed to achieve accurate prediction of cardiovascular diseases using stacking ensemble-based machine learning model development approach. The primary goal of the model is to effectively categorize diseases, specifically CAD, PAD, RHD, and Stroke, leveraging the available collected data. This study investigates the efficacy of machine learning algorithms in predicting cardiovascular disease (CVD). To construct the model different tasks were carried out on the dataset including data preprocessing and feature selection and then preprocessed dataset utilized as input for selected the models including RF, SVM, and XGB. The results of these models are combined using stacking ensemble to create a new proposed model for prediction. Performance evaluation is conducted on the model to assess its performance.

Various performance evaluation metrics are utilized to determine the model with better performance, ensuring the best future predictions on the test dataset. Once all these stages are concluded, the subsequent and final step is model deployment. Deployment involves integrating the machine learning prediction model into the existing production environment, providing real-world decision-making based on data to assist domain experts, cardiologists, and patients. By enabling early disease identification, mortality and morbidity risks associated with the disease can be reduced, benefiting both individuals and the overall health system.

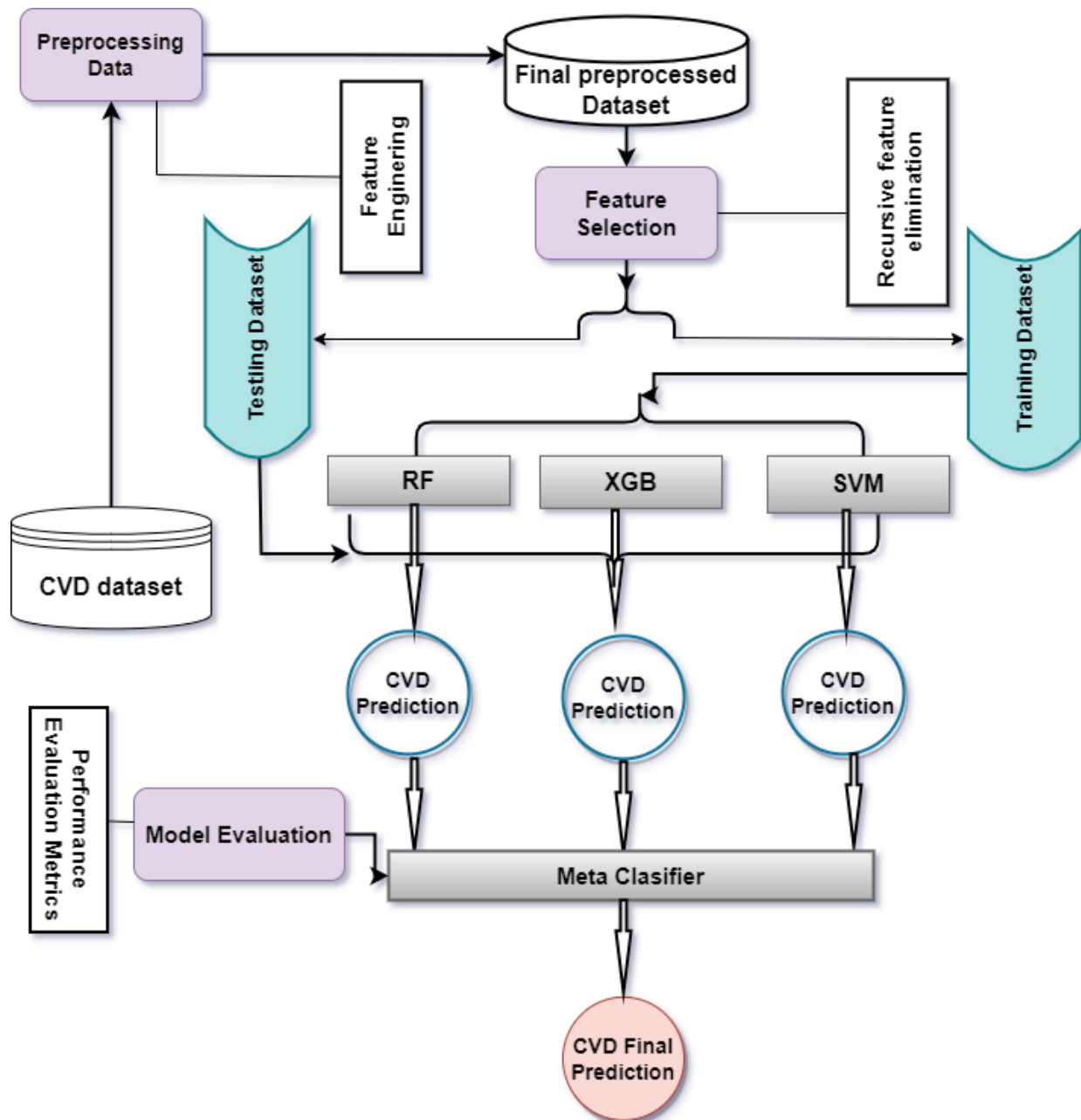


Figure 4:1 Proposed Architecture of CVD Prediction

4.3. Data preparation and feature description

The aim of this study is to develop a machine learning model for predicting cardiovascular disease. To accomplish this, it is necessary to prepare dataset specifically for the risk factors and symptoms associated with cardiovascular diseases. Since there is no published or annotated dataset on cardiovascular diseases in Ethiopia. We managed to collect the dataset which annotated by the help of two cardiologists who works in SPHMMC. The process of constructing this Dataset

involves gathering Dataset from various sources, such as patient history, laboratory test results, and demographic information. The collected Dataset is then carefully prepared to ensure its suitability for building predictive model. A crucial aspect of machine learning research is to format the data appropriately for the selected machine learning tools, techniques, and tasks. Consequently, the researcher took the necessary steps to prepare the dataset in a manner that aligns with the requirements of the chosen machine learning tasks and the specific tools utilized.

The predictor utilized a Dataset sourced from St. Paulos Millennium Medical College. This Dataset encompasses records spanning from 2019 to 2023, and its initial size in excel format is 178 kilobytes (KB). It comprises 2196 instances, each representing the information of a single patient, and incorporates 19 distinct features. Out of which, 18 of them are considered independent features and 1 as dependent feature.

Initially, the collected dataset was not in the required format, which is not conducive to effective model learning. To address this limitation, the dataset is prepared using Microsoft Excel as CSV file format. The prepared dataset encompasses various essential components, including demographic information of patients (such as gender, age, family history of CVD), behavioral risk factors (smoking and physical inactivity), health-related factors (body mass index (BMI), diabetes, kidney disease, and hypertension), laboratory test results (cholesterol level, heart rate, fasting blood sugar, systolic and diastolic blood pressure, and) and disease specific symptoms. Additionally, the dataset includes a target class of four possible value such as CAD, PAD, RHD and Stroke. This transformation ensures that the dataset is suitable for the intended machine learning tasks and it was effectively utilized in the selected machine learning processing, modeling and evaluation tools. The description of the features with respect to the context in the study is showcases in the table below:

Table 4:1 Dataset features together with descriptions

Features	Data type	Descriptions and possible values
Age	Numerical	cardiac patients starting from age 17 is included in the study: Age of patients in years (17 to 120)
Gender	Nominal	both male and female patients were considered for the study: Male:0, Female:1
Kidney disease History (HKD)	Nominal	the patients who identified as both cardiac and kidney disease (different level of stages) were incorporated in addition to the patients who diagnosed as cardiac disease and no kidney disease: No:0, Yes:1
Diabetic History (HD)	Nominal	the records of cardiac patients with and with no diabetic disease (two stages) were included during data gathering: No:0, Yes:1
Total Cholesterol level (TCHL)	Numerical	A metric that indicates the overall quantity of cholesterol present in your bloodstream. Measured in milligrams per deciliter (mg/dL): Serum cholesterol in mg/dl (126 to 564)
Systolic blood pressure (SBP)	Numerical	The elevated number measured during blood pressure assessment, indicating the force applied to artery walls as the heart contracts and propels blood throughout the circulatory system: Measured in millimeters of mercury (90 to 200 mmHg)
Diastolic blood pressure (DBP)	Numerical	The lower value measured during blood pressure evaluation, indicating the pressure applied to artery walls when the heart is in a relaxed state between contractions: Measured in millimeters of mercury (60 to 140 mmHg)
BMI	Numerical	Computed from wight and height of a person and used to assess the wight based on the computed result. Measured in kilogram per height square(kg/h ²)
Smoking status (SS)	Nominal	Indicates previous and current smoking status of the patients: No:0, Yes:1

Heart rate (HR)	Numerical	Indicates the frequency of your heart beats per minute: measured in beat per minute (50-170 BPM)
Physical activity (PHA)	Nominal	Indicates whether the patients have the trends to do physical exercise: No:0, Yes:1
Chest pain type (CPT)	Nominal	Shows kinds of the pain the patients experience that identified during disease diagnosis: NAP:0, TA:1, ATA:2, and ASY:3
Fasting blood sugar (FBS)	Numerical	A measurement of blood glucose level after a period of fasting and often used to diagnose diabetes. Measured in milligram per deciliter (70 to 150 mg/dL)
Hypertension History (HHP)	Nominal	The conditions occurs when the blood pressure is elevated in the arteries and this is happened when the wall of blood stream is narrowed due to plaque buildup in the arteries. Blood pressure is measured using electronic sensor. Basically, hypertension has two stages (stage I and II): No: 0, Stage I:1, Stage II:2
Plaque buildup (PLB)	Nominal	Represents the kinds of plaque developed in blood stream of different parts of your body.
Rheumatic fever (RF)	Nominal	Shows whether the patients experience rheumatic fever before: No:0, Yes:1
Symptoms (SYM)	Nominal	The values of this feature show the symptoms specific to a kind of the cardiovascular disease the patient diagnosed to.
Family history of heart disease (FHD)	Nominal	This feature represents whether the patient's parents or children has the history of heart disease: No:0, Yes:1
CVD diagnosis result (Target)	Nominal	Represents the particular kind of cardiovascular disease the person may diagnosed with. The target classes of the disease are CAD, PAD, Stroke and RHD: CAD:0, PAD:1, RHD:2, and Stroke:3

4.4. Data preprocessing phase

Data preprocessing is critical because it assesses the data quality and extracts valuable information that influences the learning model's performance(Alqahtani et al., 2022). Data cleaning, handling missing, categorical and numerical values were done at this stage. Other tasks such as normalization and features engineering were done on the dataset before proceeding to train the model.

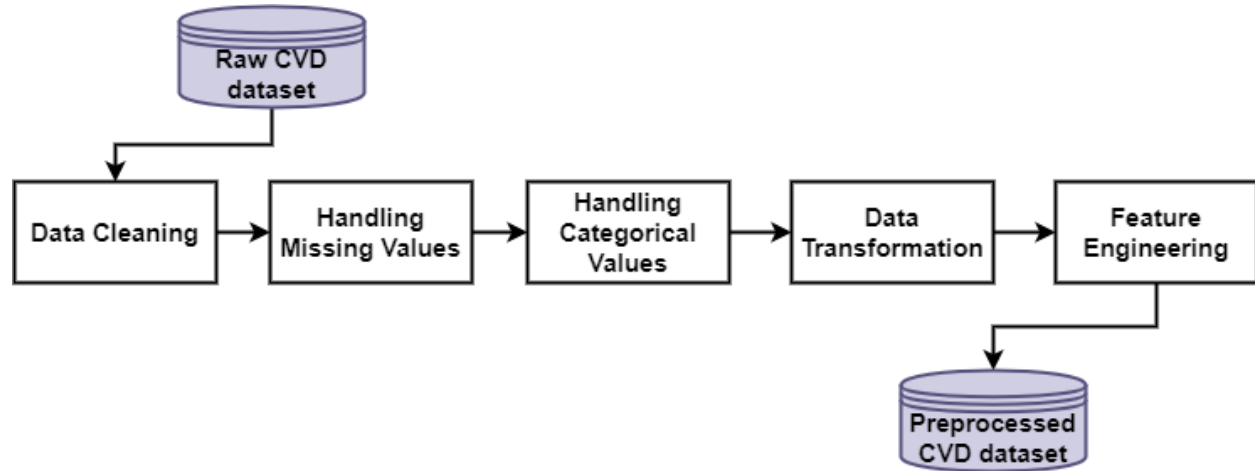


Figure 4:2 Data preprocessing phase diagram

4.4.1. Data cleaning and handling data values

Data cleaning is an important step in a preprocessing that encompasses identifying and resolving issues with incorrectly formatted or incomplete data in a dataset. In this study, the collected dataset contains some instances of improperly formatted data or inappropriate data that need to be eliminated. Additionally, the dataset also has missing values, which is a problem that can impact the accuracy of data analysis. Missing values can hinder the prediction accuracy and affect the performance of models generated from classification algorithms. The dataset was cleaned using different data cleaning techniques and the missing values are handled using mean, median and mode imputation techniques. We applied mean and median imputation on numerical features, mode imputation on categorical features. The following table shows information about the feature that has missing values.

Table 4:2 Missing value imputation

Feature Name	Data Type	Number of missing values	Missing rate %
Cholesterol total	Numeric	2	0.046%
Systolic blood pressure	Numeric	5	0.228%
diastolic blood pressure	Numeric	6	0.273%
BMI	Numeric	15	0.683%
Smoking	Nominal	5	0.228%
Heart Rate	Numeric	3	0.137%
Physical activity	Nominal	4	0.182%
Fasting blood sugar	Numeric	17	0.773%
Hypertension	Nominal	3	0.137%
Family History	Nominal	3	0.137%
Symptoms	Nominal	1	0.046%

4.4.2. Handling categorical data

Categorical data in the dataset was changed to numerical data to make appropriate for proposed machine learning model to learn from it. Two of feature encoding techniques were employed for the conversion (label encoding is for the feature that has ordinal categorical features and one-hot encoding for nominal categorical features).

Table 4:3 Categorical data

Feature Name	Data type
Hypertension	Nominal
Smoking	Nominal
Physical activity	Nominal
Symptoms	Nominal
Kidney disease	Nominal
Chest pain	Nominal
Diabetes	Nominal
plaque buildup	Nominal

4.4.3. Feature selection

Two different features selection methods such as LASSO regularization (L1) and recursive feature elimination (RFE) were employed to select relevant subset of features and improve performance

of the model we proposed.

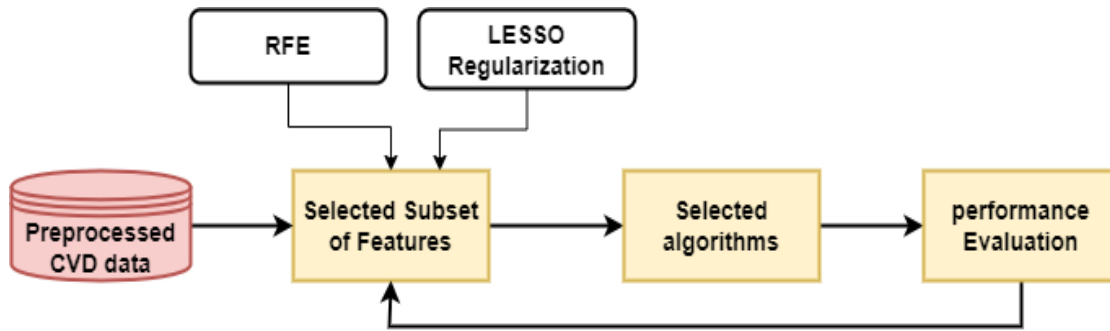


Figure 4:3 Workflow Diagram for Feature Selection

Recursive feature elimination (RFE)

This technique is included under wrapper method which aimed to improve the performance of a model by removing redundant and weak features while keeping the independent and strong ones. It follows an iterative process known as backward feature elimination, where the model is initially built using all the features and ranks them based on their importance. The least important feature is then removed, and the model is rebuilt, recalculating the feature importance. This process is repeated until a sequence of feature rankings, denoted as T , is obtained.

During each round of backward feature elimination, T_i denotes the highest-ranked attributes on which the model is retrained and evaluated for its performance. The T_i value that leads to the most optimal performance is determined, and the associated top-performing attributes are utilized to create the ultimate model.(Misra & Singh Yadav, 2020).

CHAPTER FIVE

5. EXPERIMENT AND IMPLEMENTATION

5.1. Introduction

In this chapter, the experimentation and implementation of the solution we proposed was demonstrated based on the methodologies discussed in chapter three and four. Development tools and frameworks used in this study are explained. The aim is to build the model with better performance after going through different experiments.

5.2. Implementation environment

The implementation of the CVD prediction model involved the use of various tools and packages. Python was selected as the preferred programming language due to its powerful machine learning processing capabilities and user-friendly nature. The implementation process encompassed data preprocessing, model construction, and deployment on a Flask server. Throughout the study, Python was employed for the implementation and testing of the suggested model, as well as for assessing the performance of the proposed predictor.

5.2.1. Tools and Python Libraries

Different modules and libraries used for implementing the proposed solution models are mentioned together with its version and a bit explanation in the table below.

Table 5:1 Tools and libraries used implementation

Tools and packages	version	description
Python	3.9.13	Python stands out as a favored platform due to its user-friendly and potent programming language, particularly when it comes to developing machine learning applications.
Anaconda navigator	2.4.1	It is a graphical user interface (GUI) application designed for desktop use. It simplifies tasks such as running programs and handling conda packages, and channels, and environments making it unnecessary to rely on command-line interface (CLI) commands.
Scikit-learn ¹	1.0.2	Model training and testing are carried out using this library

¹ [Getting Started — scikit-learn 1.2.2 documentation](#)

Numpy	1.23.5	For textual data to numeric data conversion
Pandas	1.4.4	Used for handling and manipulating data frame
Seaborn	0.11.2	To depict mathematical and statistical graphs
Jupyter notebook	6.4.12	A web-based interactive development environment used for document creation and sharing purposes. Different stages of machine learning tasks are carried out here.
Matplotlib	3.5.2	A visualization tool in python
Flask server		A python framework for availing services in the web. Used for developing prototype in this study.
Microsoft word	2019	To prepare the document
Microsoft Excel	2019	To prepare the dataset

5.2.2. Package Manager and Environments

Google Colab², developed by Google, is a browser-based platform that enables developers to write and execute Python code using Google Cloud infrastructure. We used this convenient environment for running code and installing libraries necessary for building the model. By utilizing google Colab, developers can enhance the efficiency of Python programming when developing machine learning applications, leveraging popular libraries like Sklearn, and other relevant tools.

Anaconda Navigator³ serves as a user-friendly graphical interface for Anaconda Individual Edition, eliminating the need for command-line tools by providing a convenient way to launch applications and manage packages and environments. To implement of our proposed approach, we opted for the 64-bit supported version of Anaconda, which facilitated the installation of the latest Python version along with a comprehensive range of modules and integrated development environments (IDEs).

Deployment environment

We used both Google Colab and Anaconda to implement the model. Personal computer with the configuration of Intel(R) Core (TM) i3-6100U CPU @ 2.30GHz 2.30 GHz, 8.00 random access memory (RAM) with x64-based processor are used to deploy the tools discussed.

² [Welcome To Colaboratory - Colaboratory \(google.com\)](https://colab.research.google.com/)

³ [Anaconda Navigator — Anaconda documentation](https://anaconda.com/anaconda-navigator/)

5.4. Preprocessing implementations

Tasks related to preprocessing are missing and categorical value handling, dealing with outliers, normalization, implementing techniques to handle class imbalance and others were carried out under this phase. To prepare the CVD dataset we started by loading dataset file before proceed to data preparation phases.

Python programming language is used throughout implementation, and different important library packages were imported. The below figure shows the loading of the dataset from personal computer drive into google colab and assign into a variable as a data frame using read_csv file methods and access the data as column name of datasets.



```
from google.colab import files
uploaded = files.upload()
from google.colab import files

#read dataset as dataframe using python library
import pandas as pd
df=pd.read_csv('compiled cardiac datasets final prepro.csv')
```

Figure 5:1 loading dataset using panda's library

Handling missing value implementation

We handled missing values using imputation techniques of mean and median. We applied mean and median imputation for numerical features and mode for nominal features. Simple Imputer module is used for this purpose.

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='most_frequent')
imputer.fit(df[['Hypertension']])
df['Hypertension'] = imputer.transform(df[['Hypertension']])

imputer = SimpleImputer(strategy='mean')
imputer.fit(df[['Fasting blood sugar']])
df['Fasting blood sugar'] = imputer.transform(df[['Fasting blood sugar']])
```

```

imputer = SimpleImputer(strategy='median')
imputer.fit(df[['Heart rate']])
df['Heart rate'] = imputer.transform(df[['Heart rate']])

```

Figure 5:2 handling missing values using imputer

Handling categorical value implementations

Machine learning models can only learn from numerical data. So that categorical data values in the dataset need to be converted into numerical data values. Since we have categorical features in our dataset, we used one of encoding techniques which is Label Encoder as shown below.

```

from sklearn.preprocessing import OneHotEncoder, LabelEncoder
label_encoder = LabelEncoder()
for column in categorical_columns:
    df[column] = label_encoder.fit_transform(df[column])

```

Figure 5:3 handling categorical values using Label encoder

Handling outliers' implementations

we used z-score to detect outliers and to remove them based on specified threshold value. The calculated z-score results of any data point is greater or less than the specified threshold value, that data point considered as outlier and removed from the dataset.

```

import numpy as np
threshold=3
z_scores = pd.DataFrame(np.abs((numerical_df - numerical_df.mean()) / numerical_df.std()), columns=numerical_df.columns)
numerical_df_clean = numerical_df[(z_scores <= threshold).all(axis=1)]
df=numerical_df_clean

```

Figure 5:4 handling outliers using Z-score

Normalizing numerical data values

We implement Min-Max normalization technique on numerical data values. Min-Max technique makes all values of numerical on the intervals of 0 and 1. 0 as minimum and 1 as maximum value in the interval.

```

from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df[numerical_features] = scaler.fit_transform(df[numerical_features])

```

Figure 5:5 Normalization using Min-Max scalar

5.5. Model implementations

The model that predicts CVD based on ensemble-based stacking machine learning approach, was evaluated using stratified 10-fold cross-validation and four performance measurement metrics. The proposed model trained using previously preprocessed dataset. To accomplish the tasks, necessary python libraries were imported as shown below.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, cross_val_predict, StratifiedKFold
from sklearn.metrics import accuracy_score, classification_report, precision_score, recall_score, f1_score
from imblearn.over_sampling import RandomOverSampler
import numpy as np
from sklearn.metrics import mean_squared_error, confusion_matrix
from sklearn.linear_model import Lasso
from sklearn.model_selection import train_test_split, learning_curve, GridSearchCV, cross_val_score
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import RFE
import pandas as pd
from sklearn.svm import SVC
import matplotlib.pyplot as plt
from sklearn import svm, metrics
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import StackingClassifier
from xgboost import XGBClassifier
from sklearn.linear_model import LogisticRegression
from imblearn.over_sampling import SMOTE
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import itertools
from sklearn.model_selection import cross_validate
from sklearn.metrics import make_scorer, precision_score, recall_score, f1_score
```

Figure 5:6 importing python libraries

5.4.1. Hyper Parameter Tuning

Optimizing hyperparameters is a crucial step prior to deploying a model for a specific task, as it directly influences algorithm's performance. By employing automatic hyperparameter tuning, researchers can alleviate the fatigue associated with manual fine-tuning. This approach not only saves time and effort but also enhances the overall effectiveness of the model. After performing automatic hyperparameter tuning using grid search, best hyperparameters are identified and used in the model development. Grid search systematically evaluates different combinations of

hyperparameters within predefined ranges to identify the best set⁴. Since all machine learning algorithms has hyperparameters(Hutter et al., 2019), automatic setting is required to optimize model performance. Algorithms included in stacking model we proposed has hyperparameters we set during optimizations. SVM has C (Regularization Parameter): which controls the trade-off between maximizing the margin and minimizing prediction errors, Kernel: enables the model to deal with non-linear problems (Wu et al., 2012) and degree. RF has n_estimators (number of trees in the forest), max-depth (maximum depth of the tree) and min_samples_split (The minimum number of samples required to split an internal node) and XGB has n_estimators, max-depth and learning_rate (The step size shrinkage to prevent overfitting). The hyperparameter together with values provided to grid_search algorithm is given in the table below. Grid search provides the best combinations of hyperparameter values which improved performance of the model proposed.

Table 5:2 Hyperparameter name, values and Optimum value returned by grid search

Models	Hyperparameter name and values	Optimum value
SVM	Kernel => linear, RBF, poly, sigmoid	RBF
	Degree => 2,3,4,5	2
	C => 0.1,1,2,5,10	10
RF	max_depth => 3,6,9,10,12	6
	min_samples_split => 5,10,15,20	10
	n_estimators => 40,50,100,200	50
XGB	learning_rate => 0.02,0.05,0.1,0.2	0.2
	max_depth => 3,6,9,10,12	1
	n_estimators => 40,50,100,200	40

⁴ <https://www.section.io/engineering-education/grid-search/>

CHAPTER SIX

6. RESULT AND DISCUSSION

6.1. Introduction

In this chapter, we elucidate the findings derived from various experimental results obtained through machine learning algorithms required for predicting cardiovascular disease. Utilized algorithms are SVM, RF, XGB, LR and stacking of those algorithms. Sklearn and other important python libraries were used for implementations.

6.2. Dataset class distribution result

The first step in applying preprocessing to the cardiovascular dataset is to load the dataset into the Python environment. To make the dataset available for preprocessing, we use the pandas library, which supports loading dataset as data frames. We utilize the `read_csv` function in pandas to read the file in comma-separated value format (CSV) and load it as a data frame. This allows us to access the instances of the dataset using column names. The original dataset has a total of rows 2196 and 19 columns and has the following class distribution:

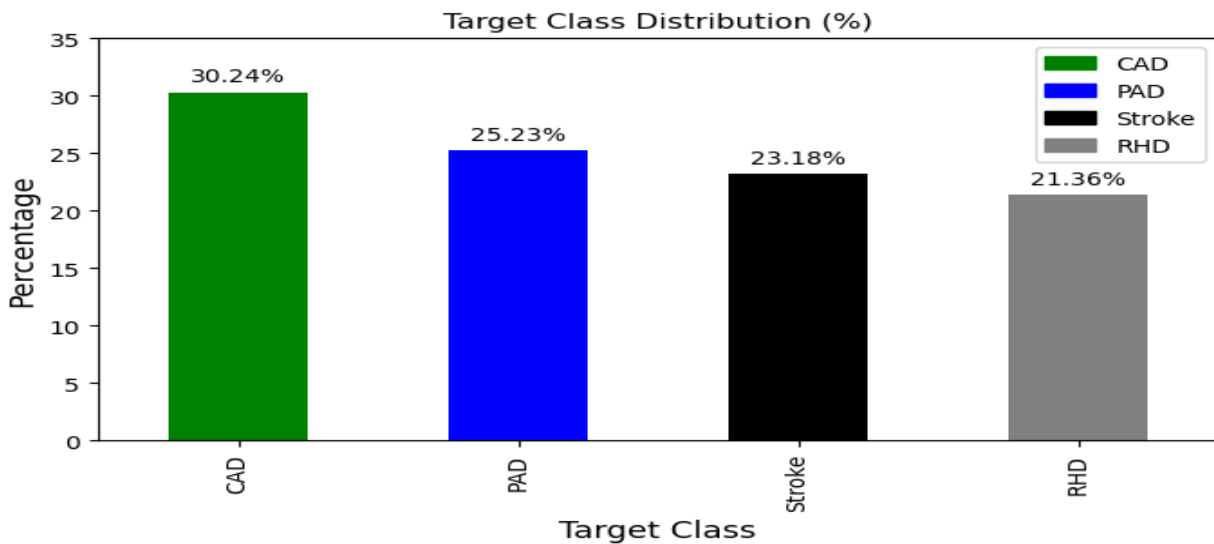


Figure 6:1 Original Dataset Class Distribution

The presented dataset illustrates four distinct classes of CVD: peripheral artery disease (PAD), stroke, coronary artery disease (CAD), and rheumatic heart disease (RHD). Within the dataset's total of 2196 patient records, the distribution is as follows: CAD comprises 664 records, PAD encompasses 554 records, stroke represents 509 records, and RHD is associated with 469 records.

These numbers provide insight into the dataset's composition, revealing the frequencies of these cardiovascular conditions among the collected patient records.

The depicted dataset class distribution shows imbalance between classes. The figure below shows the balanced instances among target classes after applying Synthetic Minority Oversampling Technique (SMOTE) on the dataset.

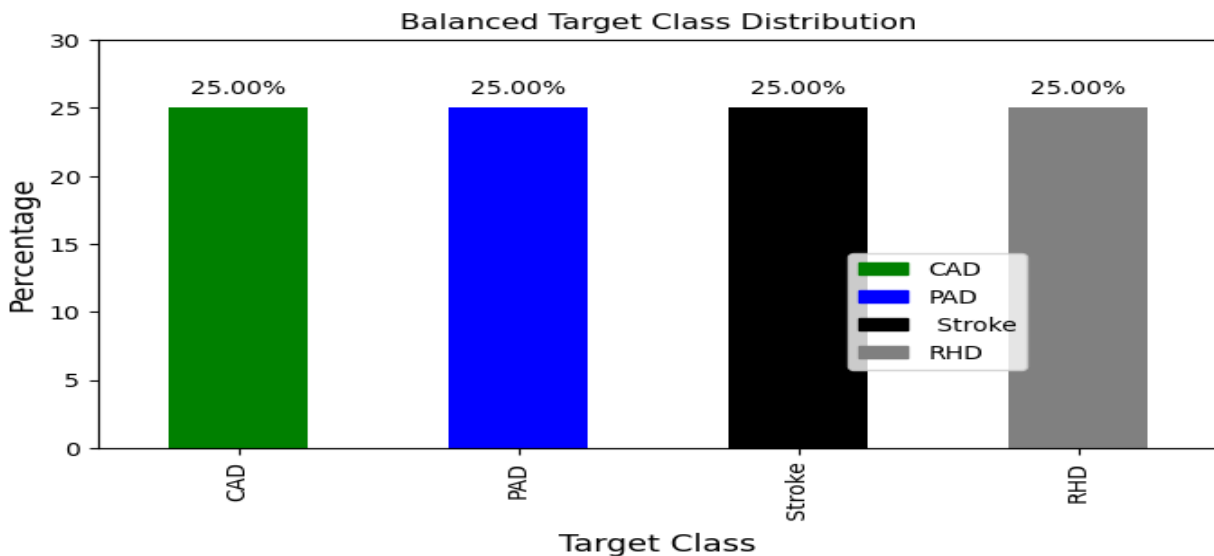


Figure 6:2 Dataset Class distribution after SMOTE applied

The SMOTE class imbalance handling method handles the problem by oversampling the minority classes. Now each class has an equal number of instances, which means 664 instances each.

Feature distribution results

The distribution values of numerical features in the CVD dataset were identified using a histogram. Outliers were detected in each feature except Age using the z-score outlier detection method. Z-score uses mean and standard deviation values. The distribution value of the features before applying Z-score normalization is shown below in a histogram.

Feature distribution before Z-score Normalization

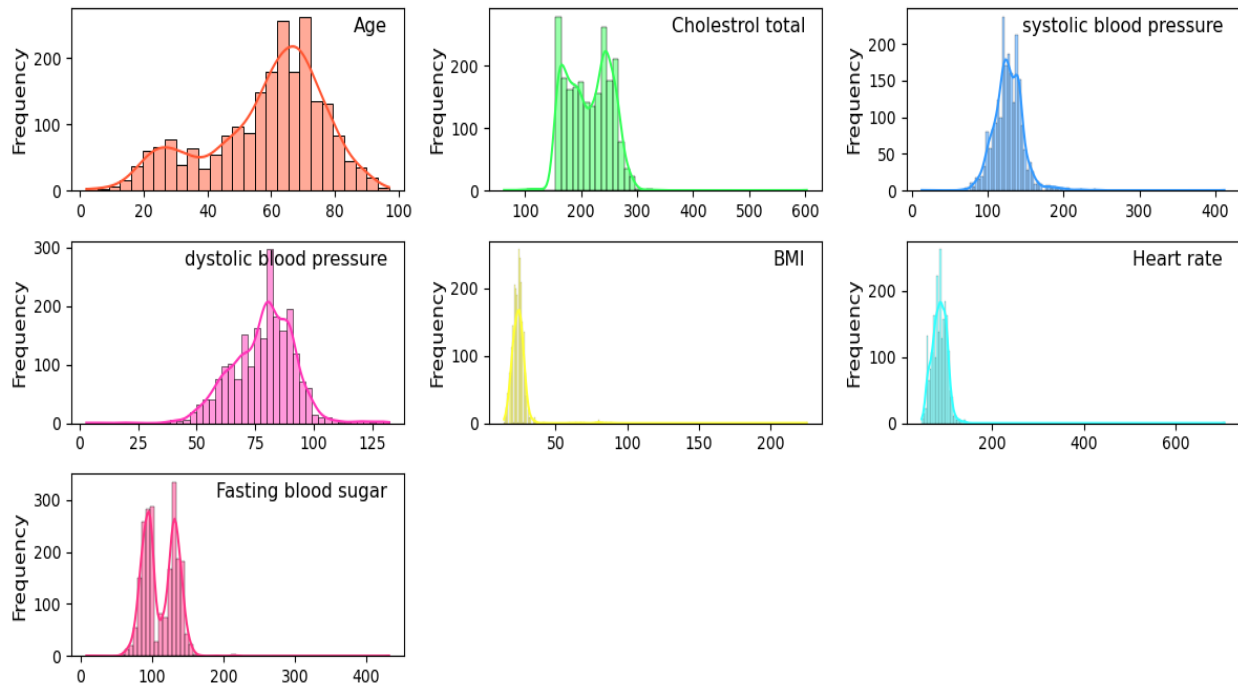


Figure 6:3 Numerical feature distribution before Z-score normalization

As shown in Figure 6:3, even if the range of outliers is different from feature to feature, all six numerical features such as cholesterol, systolic and diastolic blood pressure, BMI, heart rate and fasting blood sugar except age have outliers. For instance, most of the values of data points in cholesterol lie between 100 and 300, and some others lie below and above the indicated range. The data points that significantly deviate from the mean are considered as outliers. Therefore, outliers are observed from the visualized diagram using histograms before applying Z-score, which is one of the outlier handling techniques.

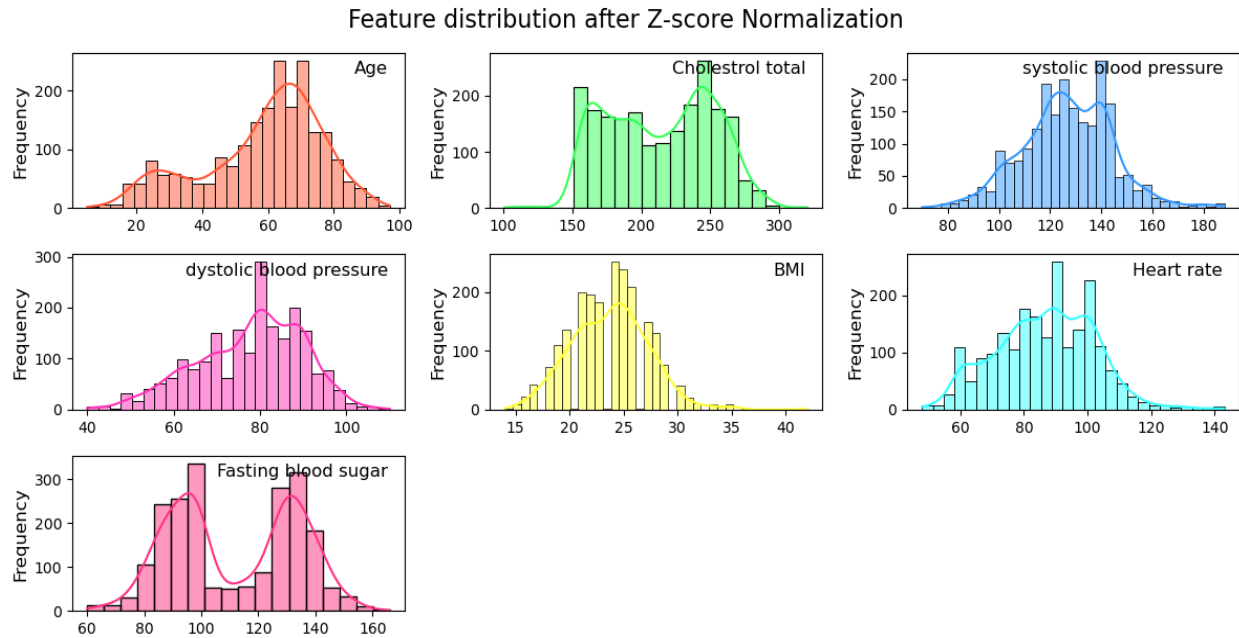


Figure 6:4 Numerical feature distribution after Z-score normalization

Figure 6:4 shows the distribution of the data points after applying Z-score normalization techniques. The detected data points as outliers were removed based on specified threshold value of 3. The data points that deviate from Z-score of 3 in either side often considered as outliers (Saleem et al., 2021).

6.3. Feature selection result

Importance of features based on their ranking or contribution on the prediction outcomes are quantified using random forest classifier. The ranked list features are depicted in the figure below:

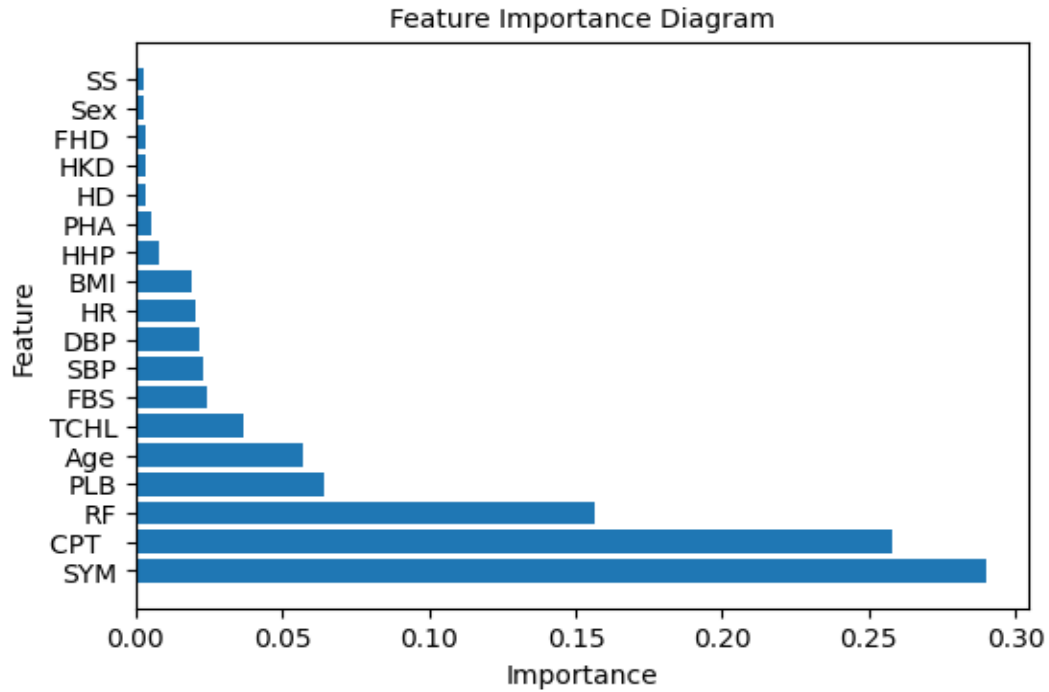


Figure 6:5 Feature importance using Random Forest Classifier

As shown in the figure, the significance of the features with respect to their contribution in deciding the model prediction results are different. To improve CVD model predictive performance, feature selection methods are employed to choose more important feature for training the selected models such as XGB, RF, SVM, and stacking.

Table 6:1 Selected Features using RFE and L1

Feature selection method	Number of original features	Number of Selected Features
RFE	18	12
L1	18	9

6.4. model performance evaluation with metrices

To assess and evaluate four of machine learning models we built, we utilized distinct performance evaluation metrices such as precision, accuracy, recall and f1_score. An assessment of those models was carried out on different circumstances based on Stratified K-fold cross validation: using dataset before and after employing feature selection techniques. each given experimental results are based on 10-fold stratified cross validation. Detailed assessment results were given by considering models confusion metrics outcomes.

6.4.1. Models' performance evaluation without feature selection

The performance outcomes of SVM, XGB, RF, and Stacking without applying any feature selection method on dataset are given in the table given below:

Experiment I:

Table 6:2 Models test results without feature selection

Techniques	Performance Evaluation metrics			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
SVM	92.77%	91.68%	91.59%	91.69%
RF	93.17%	92.88%	92.83%	92.83%
XGB	95.20%	94.31%	94.29%	94.33%
Stacking	96.47%	95.83%	95.83%	95.84%

The models' performance was evaluated using the CVD dataset in this experiment, without employing any feature selection methods. Four evaluation metrics such as Precision, Recall, F1-score, and Accuracy were utilized to assess the models. The experimental results presented in Table 6.4 highlight the stacking model as the top-performing model among the others, achieving evaluation metrics: 96.90% Precision, 97.01% Recall, 96.88% F1-score, and 96.80% Accuracy. Furthermore, an assessment of each model was conducted based on a confusion matrix. The discussion of these findings is provided below.

Support vector machine (SVM): is one of the models employed in the study. It provides results for each cardiovascular disease classes of CAD, PAD, RHD and Stroke.

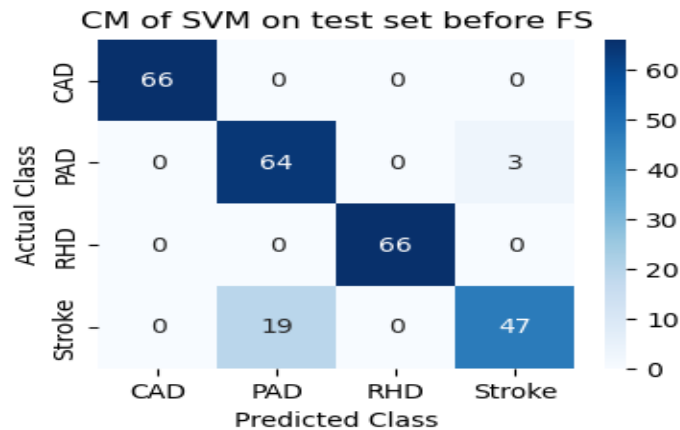


Figure 6:6 SVM model's Confusion matrix

Confusion metrics depicted in figure 6.1, the performance of the SVM algorithm is assessed on cardiovascular disease dataset. The result shows that the algorithm accurately classifies 66 out of 66 instances as CAD, 64 out of 67 instances as PAD, 66 out of 66 instances as RHD, and 47 out of 66 instances as Stroke. As observed in the figure the algorithm mistakenly misclassifies 3 instances of PAD as Stroke, 19 instances of Stroke as PAD. but all instances are correctly classified for CAD and RHD.

Random Forest Classifier (ETC): is one of the other models we employed to compare its resulting performance with the stacking model we proposed. This model provides results for cardiovascular classes of CAD, PAD, RHD and Stroke.

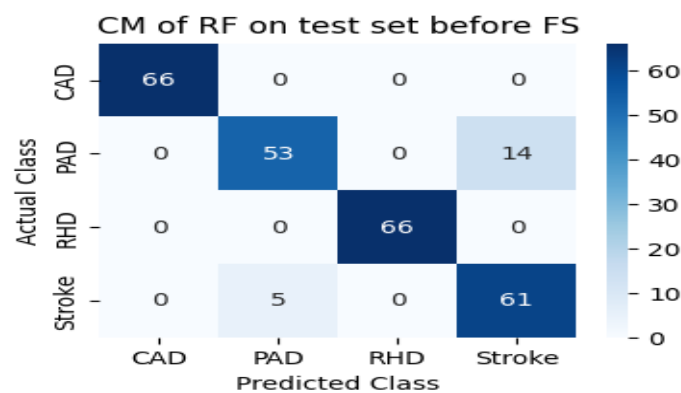


Figure 6:7 RF model's Confusion matrix

Confusion matrix which illustrated in figure 6:7 shows the evaluation results of Ransom Forest Classifier on the test dataset. The classifier correctly classifies 66 out of 66 as CAD, 53 instances out of 66 as PAD, 66 instances out of 66 as RHD, and 61 instances out of 66 as Stroke respectively. Random Forest Classifier algorithm misclassifies 14 instances of PAD as Stroke, 5 instances of stroke as PAD respectively.

Extreme gradient boosting (XGB): is other technique we utilized in this work to construct the model using cardiovascular disease dataset and compared its performance with our proposed model. The results attained by this model with the preprocessed dataset before applying any feature selection method has described below based on confusion matrix result.

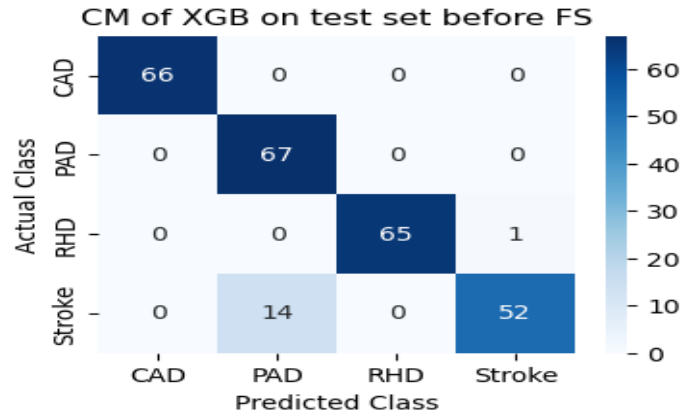


Figure 6:8 XGB model's Confusion matrix

Confusion metrics in figure 6:8 below indicates that extreme gradient boosting algorithms classifies 66 out of 66 instances as CAD, 67 out of 67 instances as PAD, 65 out of 66 instances as RHD and 52 out of instances 66 as Stroke respectively. And it wrongly misclassifies 1 instance of RHD as Stroke and 14 instances of Stroke as PAD.

Stacking model: is one of an ensemble-based techniques we used to develop a model to classify cardiovascular diseases. This model was the one we proposed to address the problem. The performance of this model on the dataset before applying feature selection techniques is explained based on confusion metrics results.

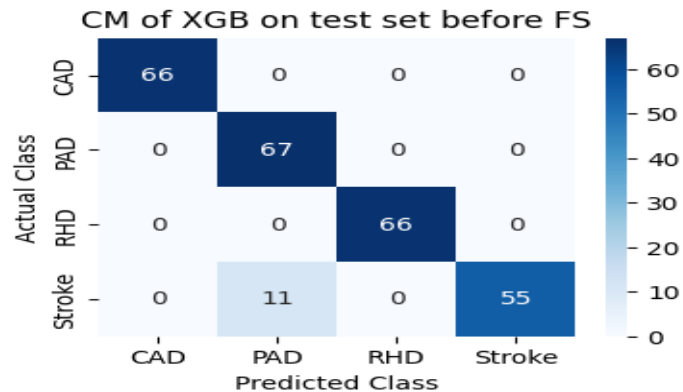


Figure 6:9 Stacking model's Confusion matrix

Figure:9 showcases the confusion matrix for the stacking model, offering an assessment of its predictive performance. The metrics reveals classification of instances cross four classes. As such the model classifies instances 66 out of 66 as CAD, 67 out of 67 as PAD, 66 out of 66 as RHD and 55 out of 66 as Stroke respectively. Misclassified instances by the model are 11 instances of Stroke as PAD.

6.4.2. Models' performance evaluation with Lasso regularization

The performance evaluation results of models using Lasso regularization (L1) method in terms of evaluation metrics such as precision, accuracy, recall and f-score are indicated in the table below.

Experiment II:

Table 6:3 Overall performance evaluation of models with *L1*

Techniques	Performance Evaluation metrics			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
SVM	92.51%	91.67%	91.55%	92.51%
RF	95.64%	95.63%	95.63%	95.37%
XGB	95.61%	95.21%	95.19%	95.22%
Stacking	95.83%	95.67%	95.66%	96.76%

In this experiment, the lasso regularization (L1) feature selection technique was employed to identify significant features from the dataset, having a more pronounced impact on the model's performance. L1 selected 9 out of 18 features such 'Sex', 'HKD', 'HD', 'SS', 'SPT', 'HHP', 'PL', 'SYM', and 'RF' as determinant and these features were used to build the model. The results of Precision, Accuracy, Recall, and F1-score, based on L1 is presented in Table 6:5. Notably, the proposed stacking model emerged as the superior performer compared to other models. It achieved best results: 95.83% Precision, 95.67% Recall, 95.66% F1-score, and 96.76% Accuracy.

6.4.3. Models' performance evaluation with recursive feature elimination

The performance evaluation results of models using recursive feature elimination method in terms of evaluation metrics such as precision, recall, accuracy and f-score are discussed as follows.

Experiment II:

Table 6:4 Overall performance evaluation of models with RFE

Techniques	Performance Evaluation metrics			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
SVM	94.52%	93.42%	93.35%	93.45%
RF	95.84%	95.71%	95.71%	95.71%
XGB	96.76%	95.92%	95.76%	95.90%
Stacking	98.04%	97.56%	97.50%	97.55%

In this thesis work, another feature selection technique employed was Recursive Feature Elimination. From the CVD dataset, 12 crucial features were identified using these techniques and utilized for model construction. Through a conducted experiment, the models yielded varying results, as depicted in Table 6.5, based on evaluation metrics: Precision, Accuracy, Recall and F1-score, and among these models, the stacking model outperforms the others, achieving the highest scores in all four metrics. Specifically, the stacking model attained remarkable results of 98.04% Precision, 97.56% Recall, 97.50% F1-score, and 97.55% Accuracy.

Moreover, an assessment of each model is carried out based on a confusion matrix. The discussion of these findings is provided below.

The SVM model's confusion matrix, as illustrated in Figure 6:10, reveals that out of the 266 testing samples within the dataset, the SVM model correctly classifies 259 instances. There were, however, 16 stroke cases that were mistakenly categorized as PAD, and 1 PAD case was incorrectly labeled as stroke.

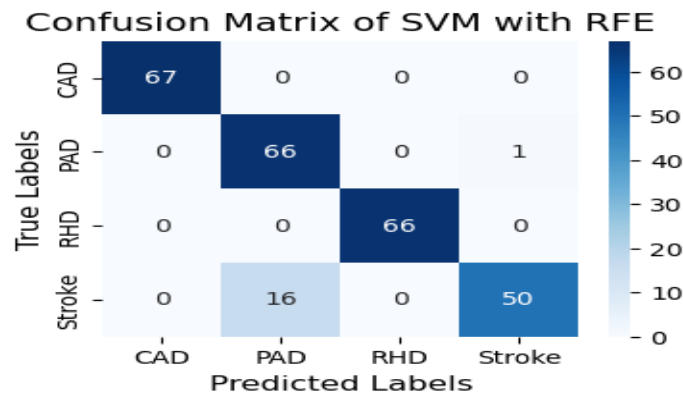


Figure 6:10 Confusion matrix of SVM with RFE

RF model's confusion matrix, as depicted in Figure 6.11, indicates that out of the 262 testing samples in the dataset, the RF model accurately classifies 252 instances. However, there were 6 cases of PAD that were incorrectly classified as stroke, and 4 cases of stroke that were mistakenly categorized as PAD.

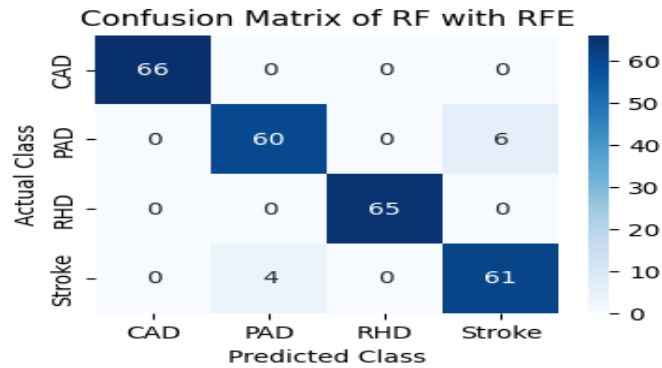


Figure 6:11 Confusion matrix of RF with RFE

XGB model's confusion matrix, as depicted in Figure 6:12, reveals that among the 263 testing samples in the dataset, the XGB model correctly classifies 213 instances. However, there were 6 cases of PAD that were erroneously classified as stroke, and 4 instances of stroke that were mistakenly categorized as PAD.

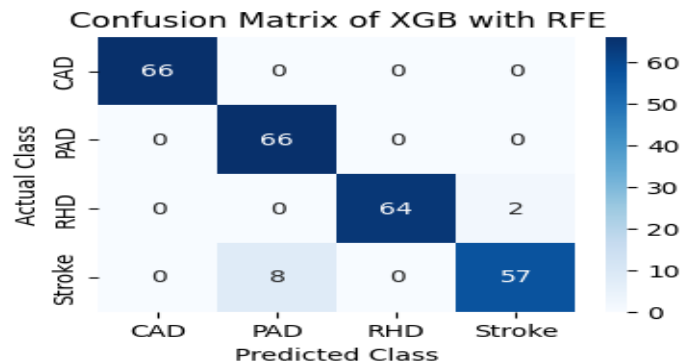


Figure 6:12 Confusion matrix of XGB with RFE

In Figure 6:13, the confusion matrix for the Stacking model following the application of RFE indicates that the model accurately classifies all instances of CAD and RHD. However, it misclassifies 5 out of 67 PAD instances as Stroke and 2 out of 67 Stroke instances as PAD.

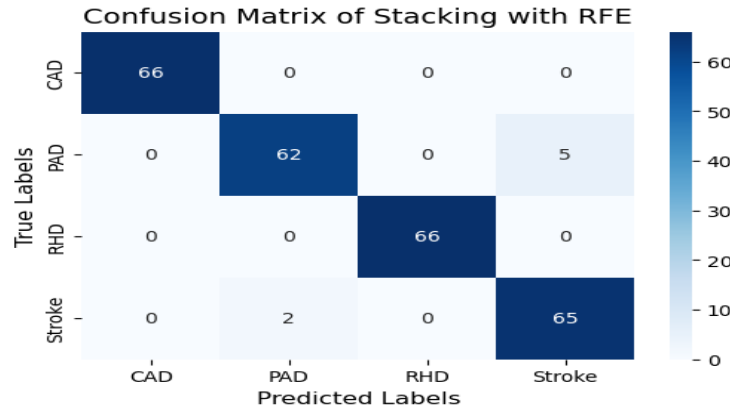


Figure 6:13 Confusion matrix of Stacking with RFE

Stratified K-Fold Cross Validation Results with 5-Values and 10-values of the models

The 5-fold and 10-fold stratified cross validation results in each fold for models with RFE are provided in table 6.2 and 6.3 and they showcase accuracy of the models results in each iteration together with overall mean accuracy of the models.

Table 6:5 Models test results based on stratified 10-fold cross validation

Folds	Techniques			
	SVM	RF	XGB	Stacking
1 st fold	89.84%	96.24%	98.87%	99.89%
2 nd fold	92.48%	95.49%	98.12%	95.11%
3 rd fold	94.73%	98.12%	99.62%	96.99%
4 th fold	93.23%	96.24%	98.87%	90.22%
5 th fold	93.60%	96.24%	88.85%	99.62%
6 th fold	93.23%	93.61%	88.35%	93.93%
7 th -fold	94.73%	93.21%	98.49%	99.95%
8 th fold	94.73%	93.96%	89.43%	99.75%
9 th fold	93.23%	96.23%	99.58%	99.58%
10 th fold	94.73%	97.74%	99.25%	99.62%
Accuracy(mean)	93.45%	95.71%	95.90%	97.55%

Stratified 10-fold cross validation involved 10 iterations times across CVDs dataset to partition it into training and testing sets. The performance of the models based on these 10-fold evaluations is presented in terms of accuracy, along with the corresponding mean results. The mean accuracy of **97.50%** for Stacking

with RFE indicates best performance across classes of CAD, PAD, stroke and RHD.

Table 6:6 Models test results based on stratified 5-fold cross validation

Folds	Techniques			
	SVM	RF	XGB	Stacking
1 st fold	86.84%	96.99%	98.50%	99.80%
2 nd fold	89.84%	96.42%	99.25%	96.04%
3 rd fold	93.23%	93.79%	88.51%	93.40%
4 th fold	94.36%	93.97%	90.96%	99.75%
5 th fold	93.23.%	94.54%	99.44%	94.34%
Accuracy(mean)	91.50%	95.14%	95.33%	96.76%

When the value of stratified cross validation is 5, the test set is switched across each fold in each elevation loop, similar to how it's done in 10-fold stratified cross-validation. In an experiment involving a stacking with RFE model on 5-fold stratified cross-validation, the obtained accuracy was **96.76%%**. The result shows the proposed model performs well in 10-fold stratified cross validation than 5-fold stratified cross validation with an accuracy of **97.55%** in our case.

The experimental results of stacking model based on recursive feature elimination and lasso regularization are summarized in the table below:

Table 6:7 Performance metrics for Stacking with RFE and Stacking with L1

Techniques	Performance Evaluation Metrics			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Stacking with L1	95.83%	95.67%	95.66%	96.76%
Stacking with RFE	98.04%	97.56%	97.50%	97.55%

According to the data in Table 6.7, the Stacking ensemble model with Recursive Feature Elimination (RFE) demonstrates superior performance compared to the Stacking ensemble model with L1 regularization across all the performance evaluation metrics, even when considering the original features.

The overall evaluation summary of the chosen model with and without feature selection techniques are indicated in Table 6.8.

Table 6:8 Overall Model Evaluation summary with and without FS techniques

Techniques	Selected Features	Performance evaluation metrics			
		Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
SVM	18	92.77%	91.68%	91.59%	91.69%
SVM with LR	9	92.51%	91.67%	91.55%	92.51%
SVM with RFE	12	94.52%	93.42%	93.35%	93.45%
RF	18	93.17%	92.88%	92.83%	92.83%
RF with LR	9	95.64%	95.63%	95.63%	95.37%
RF with RFE	12	95.84%	95.71%	95.71%	95.71%
XGB	18	95.20%	94.31%	94.29%	94.33%
XGB with LR	9	95.61%	95.21%	95.19%	95.22%
XGB with RFE	12	96.76%	95.92%	95.76%	95.90%
Stacking	18	96.90%	97.01%	96.88%	95.84%
Stacking with LR	9	95.83%	95.67%	95.66%	96.76%
Stacking with RFE	12	98.04%	97.56%	97.50%	97.55%

The summary of the Table 6:6 shows the performance evaluation results of selected models based on conducted experiments before and after applying features selection techniques on the CVD dataset.

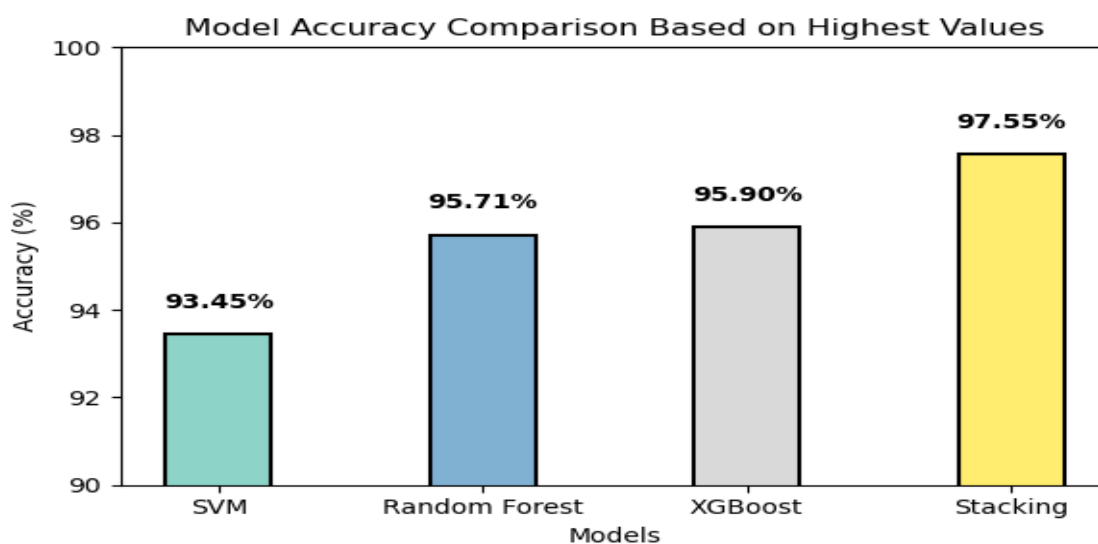


Figure 6:14 Models accuracy performance comparison

As observed in the table indicates the stacking model, we proposed with recursive feature elimination as feature selection techniques outperforms the other by attaining **98.04% precision, 97.56% recall, 97.50% f1-score, and 97.55% accuracy.**

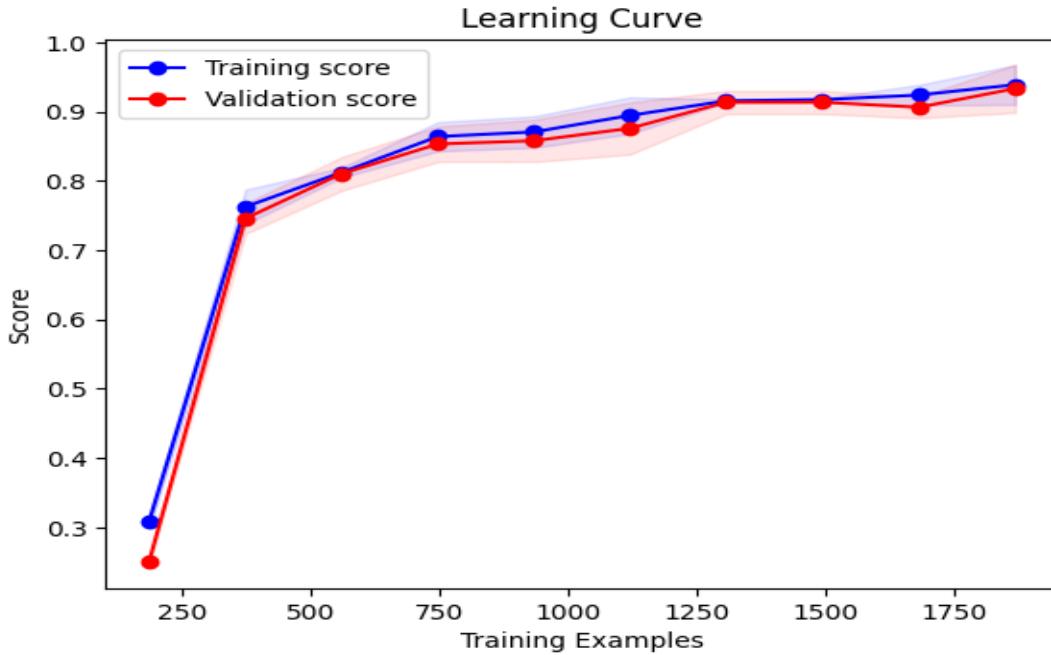


Figure 6:15 Learning curve representing training and validation scores

6.5. Discussion

This study aimed to predict cardiovascular diseases into CAD, PAD, RHD and Stroke. We conducted various experiments to address the research question raised in section 1.4. We selected the algorithms SVM, XGB, RF, and stacking to carry out these experiments. To split the CVD dataset during model development, we employed stratified 5-fold and 10-fold cross-validation, and the results of the models in both cases are summarized in Tables 6:5 and 6:6. However, stratified 10-fold cross-validation performed better, so all subsequent experiments were conducted based on 10-fold cross-validation. We utilized two distinct feature selection methods: recursive feature elimination (RFE) and lasso regularization (L1), to identify key determinant features for building the models. As shown in Table 6:1, out of the 18 features in the cardiovascular disease dataset, L1 identified 9 features as determinants, while RFE identified 12 features as determinants. The performance summary of the models based on these two cases is provided in Table 6:7. It indicates that the models' performance, based on the features identified by RFE, performed better. The identified features include Age, HD (History of Diabetes), TCHL (Total cholesterol level),

SBP (Systolic blood pressure), DBP (Diastolic blood pressure), PHA (Physical Activity), CPT (Chest pain type), FSB (Fasting blood sugar), PLB (Plaque buildup in the arteries), SYM (symptoms), FHD (Family history of heart disease) and RF (history of rheumatic fever). Therefore, the stacking model we proposed performed best when using RFE as the feature selection method with 12 features identified as determinants. **This answers both Rq1 and Rq2.**

To evaluate the performance of the selected models, we employed four performance evaluation techniques. Table 6:8 provides a performance summary for SVM, XGB, RF, and Stacking based on these metrics. In the model proposed for predicting cardiovascular diseases, SVM, XGB, and RF were combined as base learners, with LR serving as the meta learner to form a stacking ensemble for predicting cardiovascular diseases. We compared the performance of the proposed stacking ensemble with that of three individual models: SVM, RF, and XGB, both with and without feature selection techniques. The results of the experiment in the summary Table demonstrate that SVM with RFE attained a precision of 94.52%, recall of 93.42%, F1-score of 93.35%, and accuracy of 93.45%, RF with RFE attained a precision of 95.84%, recall of 95.71% an F1-score of 95.71%, and accuracy of 95.71%, XGB with RFE attained a precision of 96.76% recall of 95.92%, F1-score of 95.76%, and accuracy of 95.90%, **Stacking with RFE attained a precision of 98.04%, recall of 97.56%, F1-score of 97.50%, and accuracy of 97.55% respectively. The Stacking with RFE, the model we proposed outperforms the others. This answers Rq3.**

CONCLUSION AND FEATURE WORK

Conclusion

In conclusion, this thesis addressed the critical issue of cardiovascular diseases (CVDs), which are the leading cause of death globally. CVDs account for a significant portion of mortality, particularly in low and middle-income countries like Ethiopia. Challenges such as limited access to healthcare, shortages in expertise, and diagnostic resources, along with the costly and time-consuming nature of traditional diagnostic methods, hinder early identification and treatment. Our research focused on leveraging machine learning techniques to tackle this problem, with a specific emphasis on predicting CVDs into four distinct categories: CAD, PAD, Stroke, and RHD. We collected and prepared a dataset from St. Paulo's Hospital Millennium Medical College (SPHMMC) in consultation with experienced cardiologists, addressing issues like missing data, outliers, and class imbalance. To enhance our model's performance, we employed feature selection techniques, including lasso regularization (L1) and recursive feature elimination, and created a stacking ensemble consisting of SVM, RF, and XGB as base classifiers, with logistic regression as the meta-classifier. Through rigorous experimentation, employing 10-fold cross-validation, our results demonstrated that the stacking model, particularly when using the recursive feature elimination method for feature selection, outperformed other approaches. It achieved remarkable performance with a precision of 98.04%, recall of 97.56%, F1-score of 97.50%, and an accuracy of 97.55%.

In the fight against cardiovascular diseases, early identifications and accurate predictions are crucial for timely and effective treatment. Our research contributes to this effort by providing a robust machine learning model tailored to categorize CVDs effectively, thereby potentially saving lives and reducing the burden of these diseases.

Contributions of the study

- ✚ The dataset utilized in this study was gathered from St. Paul's Hospital Millennium Medical College (SPMMC) which is a valuable contribution that can benefit other researchers to make further improvement in CVD diagnosis and prediction.
- ✚ This study introduced a novel approach to the prediction of CVDs by categorizing them into four common categories: CADPAD), RHD, and Stroke. This categorization represents a significant contribution as it allows for more specific and targeted diagnosis.

- ✚ This thesis work contributes to the field of medical diagnostics by leveraging machine learning techniques to improve the accuracy and efficiency of CVD diagnosis. This shift from traditional diagnostic methods to advanced technology has the potential to improve CVD diagnosis, particularly in regions with limited access to healthcare services.
- ✚ The application of the model we built has significant clinical implications. It can assist medical practitioners and cardiologists in effectively predicting CVDs into common categories. This not only enhances the accuracy of diagnosis but also has the potential to save lives and reduce the burden of CVDs, especially in regions like Ethiopia where the prevalence of CVDs is a major healthcare concern.

Future Works

The model we proposed is performed well in predicting the common classes of CVD based on CAD dataset provided to the model. We recommend interested researches specifically in the domain of health sector to further investigate its implantability specially to the disease that requires immediate response.

As a future works, the following bulleted tasks are recommended:

- ✚ Future research should consider to incorporate the prediction of the CVD disease with their respective severity level and also better to include other diseases in this category.
- ✚ Future researchers can use diversified dataset including images of different kinds and can increase the applicability of the model together with its performance and effectiveness.
- ✚ Consider to increase the dataset size and other method to increase the performance of this model further.

Case Study: Applying CVD Prediction Model at SPMMC

The case study presented here exemplifies the practical applicability of our research within the healthcare sector, focusing on St. Paul's Hospital Millennium Medical College (SPMMC) in Addis Ababa, Ethiopia. SPMMC), a bustling healthcare facility serving a diverse patient population. With its extensive cardiology department, the hospital frequently encounters challenges related to CVD diagnosis and prediction due to resource constraints and high patient volumes. SPMMC has been grappling with the timely and accurate diagnosis of CVDs, which is crucial for effective patient care. Traditional diagnostic methods often result in delays, misclassifications, and increased patient burden. To address this challenge, we will collaborate with SPMMC to integrate

our machine learning model into their existing diagnostic workflow. The model, developed based on a dataset collected from the hospital's records, will be fine-tuned for the hospital's specific needs. The implementation involved training the medical staff on using the model, ensuring seamless integration with their diagnostic processes, and regular monitoring and feedback collection. Our developed model was shown to Dr. Seifu Bacha, the head of cardiology unit at SPMMC. He remarked “The incorporation of this model will transform our ability to diagnose CVDs accurately and swiftly. It will become an invaluable tool in our daily practice, saving lives and reducing the burden on our medical staff”.

Special Acknowledgment

This research was sponsored by Adama Science and Technology University with a grant number

ASTU/SM-R/852/23

Adama, Ethiopia

References

- Ahmad, E. (2020). Cardiovascular Diseases (CVDs) Detection using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 8(6), 2341–2346. <https://doi.org/10.22214/ijraset.2020.6376>
- Alfaidi, A., Aljuhani, R., Alshehri, B., Alwadei, H., & Sabbeh, S. (2022). Machine Learning: Assisted Cardiovascular Diseases Diagnosis. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 2). www.ijacsa.thesai.org
- Ali, S., Misganaw, A., Worku, A., Destaw, Z., Negash, L., Bekele, A., Briant, P. S., Johnson, C. O., Alam, T., Odell, C., Roth, G. A., Naghavi, M., Abate, E., & Mirkuzie, A. H. (2021). The burden of cardiovascular diseases in Ethiopia from 1990 to 2017: Evidence from the Global Burden of Disease Study. In *International Health* (Vol. 13, Issue 4, pp. 318–326). Oxford University Press. <https://doi.org/10.1093/inthealth/ihaa069>
- Alqahtani, A., Alsubai, S., Sha, M., Vilcekova, L., & Javed, T. (2022). Cardiovascular Disease Detection using Ensemble Learning. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/5267498>
- Angaw, D. A., Ali, R., Tadele, A., & Shumet, S. (2021). The prevalence of cardiovascular disease in Ethiopia: a systematic review and meta-analysis of institutional and community-based studies. *BMC Cardiovascular Disorders*, 21(1). <https://doi.org/10.1186/s12872-020-01828-z>
- Arroyo, J. C. T., & Delima, A. J. P. (2022). An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction. *Journal of Advances in Information Technology*, 13(1), 95–99. <https://doi.org/10.12720/jait.13.1.95-99>
- Arunachalam, S. (2020a). Cardiovascular Disease Prediction Model using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 8(6), 1006–1019. <https://doi.org/10.22214/ijraset.2020.6164>
- Arunachalam, S. (2020b). Cardiovascular Disease Prediction Model using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 8(6), 1006–1019. <https://doi.org/10.22214/ijraset.2020.6164>
- Ashfaq, A., Imran, A., Ullah, I., Alzahrani, A., Alheeti, K. M. A., & Yasin, A. (2022). Multi-model Ensemble Based Approach for Heart Disease Diagnosis. 2022 International Conference on Recent Advances in Electrical Engineering and Computer Sciences, RAEE and CS 2022. <https://doi.org/10.1109/RAEECS56511.2022.9954490>

- Atun, R. (2015). Transitioning health systems for multimorbidity. In *The Lancet* (Vol. 386, Issue 9995, pp. 721–722). Lancet Publishing Group. [https://doi.org/10.1016/S0140-6736\(14\)62254-6](https://doi.org/10.1016/S0140-6736(14)62254-6)
- Azhari, M., Alaoui, A., Achraoui, Z., Ettaki, B., & Zerouaoui, J. (2019, October 2). Adaptation of the random forest method: Solving the problem of Pulsar Search. *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3368756.3369004>
- Bekkali, M., Institute of Electrical and Electronics Engineers. Morocco Section, École nationale des sciences appliquées, F., Jāmi‘at Sīdī Muḥammad ibn ‘Abd Allāh, & Institute of Electrical and Electronics Engineers. (2019). Identification Of Cardiovascular Diseases Using Machine Learning.
- D McGinnis, W., Siu, C., S, A., & Huang, H. (2018). Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data. *The Journal of Open Source Software*, 3(21), 501. <https://doi.org/10.21105/joss.00501>
- Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 1211–1215. <https://doi.org/10.1109/ICCMC.2019.8819782>
- De, S., & Chakraborty, B. (2020). Disease Detection System (DDS) Using Machine Learning Technique. In V. Jain & J. M. Chatterjee (Eds.), *Machine Learning with Health Care Perspective: Machine Learning and Healthcare* (pp. 107–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-40850-3_6
- Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, 2. <https://doi.org/10.1016/j.health.2022.100032>
- Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, 52(4), 4543–4581. <https://doi.org/10.1007/s10489-021-02550-9>
- Dombrovskiy, V., Workneh, A., Shiferaw, F., Small, R., & Banatvala, N. (2017). Prevention and control of noncommunicable diseases in Ethiopia The case for investment Prepared for the Ministry of Health of Ethiopia by The World Health Organization United Nations Development Programme.

- Doppala, B. P., Midhunchakkravarthy, & Bhattacharyya, D. (2021). A Novel Approach to Predict Cardiovascular Diseases Using Machine Learning. *Advances in Intelligent Systems and Computing*, 1280, 71–80. https://doi.org/10.1007/978-981-15-9516-5_6
- Dritsas, E., Alexiou, S., & Moustakas, K. (2022). Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00516-9>
- Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Mamun, M. S. Al, & Kaiser, M. S. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 1464–1469. <https://doi.org/10.1109/ICECA49313.2020.9297525>
- Fonti, V., & Belitser, E. (2017). Feature Selection using LASSO.
- Gárate-Escamila, A. K., Hajjam El Hassani, A., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, 19. <https://doi.org/10.1016/j.imu.2020.100330>
- Geetha, A., Ece, D., Potluri, P. V, Prasada, S., Borra, R., Prasad, E., Potluri, V., & Vidya Sagar, K. (2021). A Method of Cardiovascular Disease Prediction using Machine Learning. www.ijert.org
- Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, 9, 19304–19326. <https://doi.org/10.1109/ACCESS.2021.3053759>
- Hussein, M. U. (2017). Physics and The Cardiovascular System The Effect of the Non-Thermal Plasma Needle on Pseudomonas Aeruginosa Bacteria View project medical physics View project. <https://www.researchgate.net/publication/320702503>
- Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). The Springer Series on Challenges in Machine Learning Automated Machine Learning Methods, Systems, Challenges. <http://www.springer.com/series/15602>
- Ibrahim Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A.

- (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, 12(2), 1545–1556. <https://doi.org/10.1016/j.asej.2020.11.011>
- Islam, S., Jahan, N., & Khatun, M. E. (2020). Cardiovascular Disease Forecast using Machine Learning Paradigms. *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, 487–490. <https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00091>
- Kappal, S. (2019). Data Normalization using Median & Median Absolute Deviation (MMAD) based Z-Score for Robust Predictions vs. Min-Max Normalization.
- Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications*, 4, 100024. <https://doi.org/10.1016/j.mlwa.2021.100024>
- Kumar, S., Koolwal, V., Mohbey, K. K., Kumar, S., Koolwal, V., & Mohbey, K. K. (2019). SENTIMENT ANALYSIS OF ELECTRONIC PRODUCT TWEETS USING BIG DATA FRAMEWORK. In *Jordanian Journal of Computers and Information Technology (JJCIT)* (Vol. 05, Issue 01). <http://www.tweepy.org/>
- Kumar, V. D. A., Swarup, C., Murugan, I., Kumar, A., Singh, K. U., Singh, T., & Dubey, R. (2022). Prediction of cardiovascular disease using machine learning technique-a modern approach. *Computers, Materials and Continua*, 71(1), 855–869. <https://doi.org/10.32604/cmc.2022.021582>
- Kumari, C., Abulaish, M., & Subbarao, N. (2020). Using SMOTE to Deal with Class-Imbalance Problem in Bioactivity Data to Predict mTOR Inhibitors. *SN Computer Science*, 1(3). <https://doi.org/10.1007/s42979-020-00156-5>
- Kuriakose, D., & Xiao, Z. (2020). Pathophysiology and treatment of stroke: Present status and future perspectives. In *International Journal of Molecular Sciences* (Vol. 21, Issue 20, pp. 1–24). MDPI AG. <https://doi.org/10.3390/ijms21207609>
- Lerra, F., Candido, A., Liverani, E., & Fortunato, A. (2022). Prediction of Micro-scale Forces in Dry Grinding Process Through a FEM—ML Hybrid Approach. *International Journal of Precision Engineering and Manufacturing*, 23(1), 15–29. <https://doi.org/10.1007/s12541-021-00601-2>

- Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2), 1487–1509. <https://doi.org/10.1007/s10462-019-09709-4>
- Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research*. <https://doi.org/10.21275/ART20203995>
- Mengistu, M. D., & Benti, H. (2022). Assessment of magnitude and spectrum of cardiovascular disease admissions and outcomes in Saint Paul Hospital Millennium Medical College, Addis Ababa: A retrospective study. *PLoS ONE*, 17(12 December). <https://doi.org/10.1371/journal.pone.0267527>
- Michael R. Zemaitis, Julia M. Boll, & Mark A. Dreyer. (2023). peripheral artery disease. <https://www.ncbi.nlm.nih.gov/books/NBK430745/>
- Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. In *IEEE Access* (Vol. 10, pp. 99129–99149). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ACCESS.2022.3207287>
- Misra, P., & Singh Yadav, A. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. *International Journal on Emerging Technologies*, 11(3), 659–665. www.researchtrend.net
- Nadakinamani, R. G., Reyana, A., Kautish, S., Vibith, A. S., Gupta, Y., Abdelwahab, S. F., & Mohamed, A. W. (2022). Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/2973324>
- Nwodo, B. C., Abisoye, O. A., Mohammed, A. D., & Sulaimon A. Bashir. (2021). Outlier Detection in Multivariate Time Series Data using a Fusion of K Medoid, Standardized Euclidean Distance and Z-Score. *Communications in Computer and Information Science*, 1350, 340–354. https://link.springer.com/chapter/10.1007/978-3-030-69143-1_21
- Olvera Lopez, E., Arif, ;, & Affiliations, J. (2019). Cardiovascular Disease White Coat wearing and laundry patterns among Internal Medicine Department Physicians View project Prevalence of folic acid and vitamin-B12 deficiency in the inpatient and outpatient setting among community-dwelling adult population in New York View project Cardiovascular Disease. <https://www.ncbi.nlm.nih.gov/books/NBK535419/?report=printable>
- Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health

- systems. *Journal of Global Health*, 8(2). <https://doi.org/10.7189/jogh.08.020303>
- Parmar, A., Katariya, R., & Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 26, pp. 758–763). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-03146-6_86
- Peters, F., Karthikeyan, G., Abrams, J., Muhwava, L., & Zühlke, L. (2020). Rheumatic heart disease: Current status of diagnosis and therapy. In *Cardiovascular Diagnosis and Therapy* (Vol. 10, Issue 2, pp. 305–315). AME Publishing Company. <https://doi.org/10.21037/CDT.2019.10.07>
- Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175(4), 7–9. <https://doi.org/10.5120/ijca2017915495>
- Princy, R. J. P., Parthasarathy, S., Hency Jose, P. S., Raj Lakshminarayanan, A., & Jeganathan, S. (2020). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 570–575. <https://doi.org/10.1109/ICICCS48265.2020.9121169>
- Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and Robust Machine Learning for Healthcare: A Survey. In *IEEE Reviews in Biomedical Engineering* (Vol. 14, pp. 156–180). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/RBME.2020.3013489>
- Rabbi, F. (2019). A review of the recent trends in the use of machine learning in business. www.ijaiml.com
- Rabiu, A. (2020). A Survey of Performances of Some Selected Machine Learning Algorithms for Cardiovascular Disease Prediction.
- Rahma, A., & Al-Mousa. (2019). Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method.
- Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. <http://arxiv.org/abs/1811.12808>
- Ray, S. (2019). A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 35–39. <https://doi.org/10.1109/COMITCon.2019.8862451>

- Raza, K. (2018). Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. In *U-Healthcare Monitoring Systems: Volume 1: Design and Applications* (pp. 179–196). Elsevier. <https://doi.org/10.1016/B978-0-12-815370-3.00008-6>
- Rout, N., Mishra, D., & Mallick, M. K. (2018). Handling imbalanced data: A survey. *Advances in Intelligent Systems and Computing*, 628, 431–443. https://doi.org/10.1007/978-981-10-5272-9_39
- Sagi, O., & Rokach, L. (2018a). Ensemble learning: A survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 4). Wiley-Blackwell. <https://doi.org/10.1002/widm.1249>
- Sagi, O., & Rokach, L. (2018b). Ensemble learning: A survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 4). Wiley-Blackwell. <https://doi.org/10.1002/widm.1249>
- Saleem, S., Aslam, M., & Rukh Shaukat, M. (2021). A REVIEW AND EMPIRICAL COMPARISON OF UNIVARIATE OUTLIER DETECTION METHODS (Vol. 37, Issue 4).
- Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H. H., Alqarni, A. A., Alsharif, N., Uddin, M. I., Alahmadi, A. H., Jadhav, M. E., & Alzahrani, M. Y. (2021). Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. *Journal of Healthcare Engineering*, 2021. <https://doi.org/10.1155/2021/1004767>
- Shahid, A., Sreenivas, S., & Abdolhossein, S. (2015). Ensemble Learning methods for Decision Making.
- Shahjehan, R. D., & Bhutta, B. S. (2022). *Coronary Artery Disease*. StatPearls Publishing, Treasure Island (FL). <http://europepmc.org/books/NBK564304>
- Shan, L., Yang, Z., Zhang, H., Shi, R., & Kuang, L. (2019). Predicting duration of traffic accidents based on ensemble learning. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 268, 252–266. https://doi.org/10.1007/978-3-030-12981-1_18
- Sharma, H., & Rizvi, M. A. (n.d.). International Journal on Recent and Innovation Trends in Computing and Communication Prediction of Heart Disease using Machine Learning Algorithms: A Survey. <http://www.ijritcc.org>
- Shiferaw, F., Letebo, M., Misganaw, A., Feleke, Y., Gelibo, T., Getachew, T., Defar, A., Assefa,

- A., Bekele, A., Amenu, K., Teklie, H., Tadele, T., Taye, G., Getnet, M., Gonfa, G., Bekele, A., Kebede, T., Yadeta, D., Gebremichael, M., ... Tadesse, Y. (2018). Non-communicable Diseases in Ethiopia: Disease burden, gaps in health care delivery and strategic directions. <http://vizhub.healthdata.org/gbd->
- Spelman, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1–11. <https://doi.org/10.1109/ICCTCT.2018.8551020>
- Tefera, Y. G., Abegaz, T. M., Abebe, T. B., & Mekuria, A. B. (2017a). The changing trend of cardiovascular disease and its clinical characteristics in Ethiopia: Hospital-based observational study. *Vascular Health and Risk Management*, 13, 143–151. <https://doi.org/10.2147/VHRM.S131259>
- Tefera, Y. G., Abegaz, T. M., Abebe, T. B., & Mekuria, A. B. (2017b). The changing trend of cardiovascular disease and its clinical characteristics in Ethiopia: Hospital-based observational study. *Vascular Health and Risk Management*, 13, 143–151. <https://doi.org/10.2147/VHRM.S131259>
- Tiwari, A., Chugh, A., & Sharma, A. (2022a). Ensemble Framework for Cardiovascular Disease Prediction.
- Tiwari, A., Chugh, A., & Sharma, A. (2022b). Ensemble Framework for Cardiovascular Disease Prediction. <https://www.sciencedirect.com/science/article/abs/pii/S0010482522004164>
- Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. In *Computational and Structural Biotechnology Journal* (Vol. 15, pp. 26–47). Elsevier B.V. <https://doi.org/10.1016/j.csbj.2016.11.001>
- Uddin, M. N., & Halder, R. K. (2021). An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach. *Informatics in Medicine Unlocked*, 24. <https://doi.org/10.1016/j.imu.2021.100584>
- Verma, V. K., & Verma, S. (2022). Machine learning applications in healthcare sector: An overview. *Materials Today: Proceedings*, 57, 2144–2147. <https://doi.org/10.1016/j.matpr.2021.12.101>
- Watkins, D. A., Beaton, A. Z., Carapetis, J. R., Karthikeyan, G., Mayosi, B. M., Wyber, R., Yacoub, M. H., & Zühlke, L. J. (2018). Rheumatic Heart Disease Worldwide: JACC

- Scientific Expert Panel. In *Journal of the American College of Cardiology* (Vol. 72, Issue 12, pp. 1397–1416). Elsevier USA. <https://doi.org/10.1016/j.jacc.2018.06.063>
- WHO. (2021). Cardiovascular diseases (CVDs). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)#](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#)
- Wu, X., Tang, W., & Wu, X. (2012). Support vector machine based on hybrid kernel function. *Lecture Notes in Electrical Engineering*, 154 LNEE, 127–133. https://doi.org/10.1007/978-1-4471-2386-6_17
- Yadeta, D., Walelgne, W., Fourie, J. M., Scholtz, W., Nel, G., Scarlatescu, O., & Gebremichael, M. (2021). PASCAR and WHF Cardiovascular Diseases Scorecard project. *Cardiovascular Journal of Africa*, 32(1), 37–46. <https://doi.org/10.5830/CVJA-2021-001>
- Yahaya, L., David Oye, N., & Joshua Garba, E. (2020). A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. *American Journal of Artificial Intelligence*, 4(1), 20. <https://doi.org/10.11648/j.ajai.20200401.12>
- Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., & Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-62133-5>
- Younis, E. M. G., Zaki, S. M., Kanjo, E., & Houssein, E. H. (2022). Evaluating Ensemble Learning Methods for Multi-Modal Emotion Recognition Using Sensor Data Fusion. *Sensors*, 22(15). <https://doi.org/10.3390/s22155611>
- Yuan, K., Yang, L., Huang, Y., & Li, Z. (2020). Heart Disease Prediction Algorithm Based on Ensemble Learning. *Proceedings - 2020 7th International Conference on Dependable Systems and Their Applications, DSA 2020*, 293–298. <https://doi.org/10.1109/DSA51864.2020.00052>

APPENDIX

APPENDIX A: Sample CVD dataset

	Age	Sex	HKD	HD	TCHL	SBP	DBP	BMI	SS	HR	PHY	CPT	FBS	HHP	PB	SYM	FHD	RF	Target
0	55	Female	no	no	159	117	61	24	no	62	no	NAP	88	no	no	no	yes	no	CAD
1	56	Female	no	no	196	99	62	25	no	61	no	NAP	99	no	no	no	yes	no	CAD
2	60	Male	no	no	190	91	55	24	no	60	no	NAP	100	no	no	no	no	no	CAD
3	58	Male	no	no	175	87	38	16	yes	61	no	NAP	80	no	no	no	no	no	CAD
4	59	Male	no	no	193	101	59	16	no	58	yes	NAP	85	no	no	no	no	no	CAD
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2191	54	Male	yes	yes	265	140	91	25	no	93	no	no	142	stage II	no	joint pain	no	yes	RHD
2192	56	female	no	no	245	132	90	24	yes	105	no	no	115	stage I	no	joint pain	no	yes	RHD
2193	68	Male	yes	yes	254	138	92	25	no	95	no	no	130	stage I	no	joint pain	no	yes	RHD
2194	29	Female	no	no	245	121	75	24	yes	113	no	no	81	no	no	joint pain	no	yes	RHD
2195	74	Male	yes	yes	222	132	81	26	no	115	no	no	142	stage I	no	joint pain	yes	yes	RHD

2196 rows × 19 columns

APPENDIX B: Sample code for stacking model with L1

```
lasso = Lasso(alpha=0.01) |
stacking_model = StackingClassifier(estimators=base_models, final_estimator=LogisticRegression(max_iter=1000, random_state=42))
for fold_idx, (train_index, test_index) in enumerate(cv.split(X_resampled, y_resampled), 1):
    X_train, X_test = X_resampled.iloc[train_index], X_resampled.iloc[test_index]
    y_train, y_test = y_resampled.iloc[train_index], y_resampled.iloc[test_index]

    lasso.fit(X_train, y_train)
    sfm = SelectFromModel(lasso, prefit=True)
    X_train_selected = sfm.transform(X_train)
    X_test_selected = sfm.transform(X_test)
    stacking_model.fit(X_train_selected, y_train)
    y_pred_fold = stacking_model.predict(X_test_selected)
    accuracy = accuracy_score(y_test, y_pred_fold)
    fold_accuracies.append(accuracy)
avg_accuracy = np.mean(fold_accuracies)
print(f"Average Accuracy Across Folds: {avg_accuracy:.4f}")

y_pred_cv = cross_val_predict(stacking_model, X_resampled, y_resampled, cv=cv)

# Calculate precision, recall, and F1-score
acc = accuracy_score(y_resampled, y_pred_cv)
precision = precision_score(y_resampled, y_pred_cv, average='macro')
recall = recall_score(y_resampled, y_pred_cv, average='macro')
f1 = f1_score(y_resampled, y_pred_cv, average='macro')

# Print precision, recall, and F1-score
print("Macro Precision:", acc)
print("Macro Precision:", precision)
print("Macro Recall:", recall)
```

APPENDIX B: Sample code for stacking model with RFE

```
base_models = [
    ('svm', SVC(kernel='poly', random_state=42)),
    ('RF', RandomForestClassifier(random_state=42)),
    ('xgboost', XGBClassifier(random_state=42))
]
base_model_param_grids = {
    'svm': {'C': [0.1, 1, 10], 'kernel': ['poly', 'rbf', 'linear'], 'degree': [2, 3, 4]},
    'RF': {'n_estimators': [50], 'max_depth': [6], 'min_samples_split': [2, 5, 10]},
    'xgboost': {'n_estimators': [40], 'max_depth': [1], 'learning_rate': [0.01, 0.1, 0.2]}
}
stacking_model = StackingClassifier(estimators=base_models, final_estimator=LogisticRegression(max_iter=1000))
estimator = LogisticRegression(max_iter=2000, random_state=42)
rfe = RFE(estimator=estimator, n_features_to_select=num_features_to_select)
X_train_rfe = rfe.fit_transform(X_resampled, y_resampled)
sum_conf_matrix = None

best_base_models = {}
for model_name, model in base_models:
    param_grid = base_model_param_grids[model_name]
    grid_search = GridSearchCV(model, param_grid, cv=StratifiedKFold(n_splits=10), scoring='accuracy')
    grid_search.fit(X_train_rfe, y_resampled)

    best_model = grid_search.best_estimator_
    best_base_models[model_name] = best_model
    print(f"Best hyperparameters for {model_name}: {grid_search.best_params_}")
stacking_model.set_params(**best_base_models)

skf = StratifiedKFold(n_splits=10)
for train_index, test_index in skf.split(X_train_rfe, y_resampled):
    X_train_fold, X_test_fold = X_train_rfe[train_index], X_train_rfe[test_index]
    y_train_fold, y_test_fold = y_resampled[train_index], y_resampled[test_index]

    stacking_model.fit(X_train_fold, y_train_fold)

    y_pred_fold = stacking_model.predict(X_test_fold)
    accuracy = accuracy_score(y_test_fold, y_pred_fold)
    precision = precision_score(y_test_fold, y_pred_fold, average='macro')
```