MACHINE LEARNING BASED FRAMEWORK FOR HEART DISEASE DETECTION

By

RIZWANA YASMEEN

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

2024

Machine Learning Based Framework for Heart Disease Detection

By

RIZWANA YASMEEN

BSSE, National University of Modern Languages, Islamabad, 2021

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

Software Engineering

To

FACULTY OF ENGINEERING & COMPUTING

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computing for acceptance.

Thesis Title: Machine Learning Based Framework for Heart Disease Detection

Submitted by: Rizwana Yasmeen **Registration #:** 47 MS/SE/S21
Master of Science in Software Engineering

Software Engineering
Name of Discipline

Dr. Raheel Zafar _____
Research Supervisor Signature of Research Supervisor

Dr. Sumaira Nazir _____
HOD (SE) Signature of HOD (SE)

Dr. Muhammad Noman Malik _____
Dean (FE&C) Signature of Dean (FE&C)

October, 2024
Date

AUTHOR'S DECLARATION

I Rizwana Yasmin

Daughter of Hakim Khan

Registration # 47 MS/SE/S21

Discipline Software Engineering

Candidate of **Master of Science in Software Engineering (MSSE)** at the National University of Modern Languages do hereby declare that the thesis **Machine Learning Based Framework for Heart Disease Detection** submitted by me in partial fulfillment of MSSE degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work October be cancelled and the degree revoked.

Signature of Candidate

Rizwana Yasmeen

Name of Candidate

October, 2024

Date

ABSTRACT

MACHINE LEARNING BASED FRAMEWORK FOR HEART DISEASE DETECTION

Cardio Vascular Diseases (CVDs), or heart diseases are one of the top-ranking causes of death worldwide. About 1 in every 4 deaths are related to heart diseases, which are broadly classified as various types of abnormal heart conditions. However, diagnosis of CVDs is a time-consuming process in which data obtained from various clinical tests are manually analyzed. Therefore, new approaches for automating the detection of such irregularities in human heart conditions should be developed to provide medical practitioners with faster analysis via reducing the time of obtaining a diagnosis and enhancing results. Electronic Health Records are often utilized to discover useful data patterns that help improve the prediction of machine learning algorithms. Specifically, Machine Learning contributes significantly to solving issues like predictions in various domains, such as healthcare. Considering the abundance of available clinical data, there is a need to leverage such information for the betterment of humankind. In this work, a Stacking model is proposed for heart disease prediction based on the stacking of various classifiers in two levels (Base level and Meta level). Various heterogeneous learners are combined to produce the strong model outcome. The model obtained 98.4% accuracy in prediction with a precision score of 94.56%, recall of 95.6%, and F1-score of 95.89%. The performance of the model was evaluated using various metrics, including accuracy, precision, recall, F1-scores values.

Keywords— Heart disease, Features, Cardiovascular, ML classifiers, Electronic Health Records, Base model, Meta model

TABLE OF CONTENTS

CHAPTER **TITLE** **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CRD	Chronic Respiratory Diseases
CHD	Coronary Heart Disease
LDL	low-density lipoprotein
CABG	Coronary Artery Bypass Graft
AI	Artificial Intelligence
ROC	Receiver Operating Characteristic
PSO	particle optimization
HDFS	Hadoop Distributed File System
WHO	World Health Organization
LR	Linear Regression
ICA	Independent Component Analysis
ICA	Independent Component Analysis
CVD	cardiovascular disease
LR	Linear Regression
KNN	K-Nearest Neighbor
ML	Machine Learning
MLP	Convolutional Neural Network
NB	Naïve Bayes
ADA	AdaBoost Adaptive Boosting
PSD	extreme Gradient Boosting
RF	Random Forest
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt gratitude and deep appreciation to Almighty Allah, whose blessings made this study possible and successful. Without divine support, this achievement would not have been possible.

I am immensely thankful to all the individuals and sources whose unwavering supportand encouragement played a pivotal role in the completion of this study. Their honest espousal has been invaluable, and I am sincerely grateful for their contributions. I owe a special debt of gratitude to my research supervisor Dr. Raheel Zafar, whose dedication and guidance were instrumental in shaping my research journey. Their commitment and relentlessefforts left no stone unturned, and I am truly grateful for their mentorship.

To every person who has contributed to my success in ways both seen and unseen, I extend my heartfelt thanks. Your support has been an indispensable part of this endeavor, andI am deeply appreciative for everything you have done.

DEDICATION

I wholeheartedly dedicate this thesis work to the pillars of my life: my beloved parents, my team lead the remarkable teachers, and my cherished friends, all of whom have been unwavering in their support throughout my entire educational journey. Their boundless love, encouragement, and belief in my potential have been the driving force behind my pursuit of knowledge and academic excellence.

To my parents, whose unwavering love and sacrifices have been a constant source of strength, I owe my deepest gratitude. They have stood by me during the challenging and triumphant moments, providing unwavering support and guiding me with their wisdom.

I am equally indebted to my extraordinary teachers, who have not only imparted knowledge but also instilled in me the virtues of discipline, perseverance, and integrity. Theirpassions for teaching and dedication to their students have inspired me to strive for greatness in every endeavor I undertake.

CHAPTER 1
INTRODUCTION

Predicting heart disease using machine learning involves collecting and preprocessing patient data, including medical history and lifestyle factors. Supervised learning algorithms like logistic regression and neural networks are used to develop predictive models. These models are trained and validated to ensure accuracy and robustness. The models are then integrated into clinical systems to provide real-time decision support for healthcare providers. Continuous updates and monitoring are essential to maintain and improve model performance.

1.1 Context

Nowadays everyone is so busy in their life and working so much that they do not have time to take care of themselves. Due to their hectic lives, people regularly suffer stress, anxiety, sadness, and many other problems. Considering everything, they become unhealthy and contract major diseases. A variety of diseases, including cancer, heart disease, and others take the lives of millions of people every year, heart disease, also known as cardiovascular disease (CVD) is the leading cause of death in the medical field.

Many people feel stressed, anxious, and depressed because of their busy lives. These feelings can lead to health problems. Heart disease is a big problem, causing a lot of deaths around the world. As per to the World Health Organization (WHO), about 31% of all deaths are because of heart issues. In one year, around 15.2 million people die from heart diseases. Stress and a fast-paced lifestyle can contribute to these heart problems. Unhealthy habits like poor eating, lack of exercise, smoking, and too much drinking can also make things worse. It's essential to prevent these issues by promoting a healthy lifestyle with good food, regular exercise, and managing stress [2].

The WHO suggests that early detection and proper medical care are crucial in dealing with heart diseases. Regular health’s check-ups help catch problems early. So, taking care of mental and physical health, along with advancements in medical research, can help reduce the number of people getting sick from heart-related issues [3].

The heart is like the boss of our circulation system. It's a strong muscle that pumps blood throughout the full body. It forms a difficult system to collaborate with veins, arteries, and blood, which control the body's blood flow. Any flaw or injury that prevents something from happening normally would result from the heart blood flow [4]. These are usually referred to as cardiovascular diseases (CVD) according to reports by the "World Health Organization (WHO)" are some of the world's deadliest diseases [2].

Heart and blood vessel disorders are the main cause of high blood pressure, which also includes brain diseases, coronary heart diseases (heart attacks), heart defects, and artery diseases [3]. According to the 2007 WHO report, there were an estimated 58 million deaths worldwide in 2005 due to the above-mentioned causes. An estimated 30% of total deaths between 2006 and 2015 were because of non-communicable illnesses including cardiovascular illnesses; these causes were anticipated to rise by 17% [5].

To decrease the number of deaths from heart diseases and chronic respiratory illnesses, medical experts as well as researchers are doing all the work on this topic. A survey done by the WHO in 2016 highlights the seriousness of the problem. According to this study, about 19 million individuals died in 2016 as a result of these two diseases. According to studies, a million people die from chronic respiratory disorders each year, compared to 17.5 million who pass away from cardiovascular disease (CVD) [6]

According to another most recent research based on WHO study from 2017, About 17.9 million people die from heart-related diseases every year. Estimated 31% of all fatalities globally. "Every year an estimated 17.9 million people die from cardiovascular disease. The risk that can increase the chances of heart disease includes diabetes, smoking, or too much drinking, high cholesterol, and high blood pressure [7].

The research showed that one of the major causes or drawbacks is the lack of information about the main signs and symptoms of this disease. Generally, fewer people's deaths occur as a result of High blood sugar, high blood pressure, high cholesterol, and other problems. Smoking increases the possibility of heart and chronic respiratory diseases. The reason that there are so many deaths because of Chronic Respiratory Diseases (CRD) [1] and Cardiovascular Disease (CVD) [2] is the lack of prediction related to the disease's vital signs [8].

The World Health Organization has shown in many papers that the number of deaths from cardiovascular cases has increased, mainly due to insufficient preventive measures, insufficient medical equipment, and a lower number of doctors, especially in low-income countries. Some risk elements enhance the chances of heart disease in a person. The most serious health problem is heart disease (HD), which has impacted a large number of people worldwide [9]. Breathlessness and weakness in the body, such as swollen feet are important signs of heart disease. A better knowledge of Coronary Heart Disease (CHD) and its complex factors like stroke, high low-density cholesterol, high blood pressure, smoking, diabetes, unhealthy diet, and physical fitness is urgently required to help in prevention, early detection, and advancements in the medical field[10].

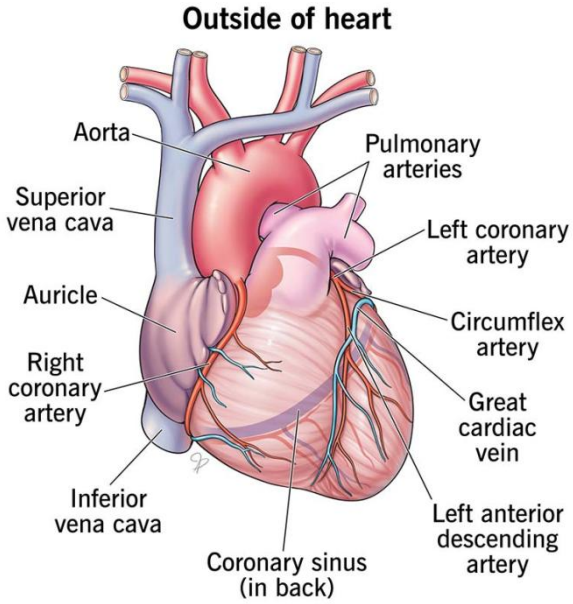

Figure 1.1: Function of Heart Representation [2]

Due to many factors, including execution time, accuracy and present heart disease diagnosis methods are not very effective in early identification. Therefore, researchers are working to find an effective method for the detection of heart disease. When new tools and qualified medical skill personnel are unavailable the treatment and diagnosis of heart disease are particularly complex [11]. According to projections from the European Society of Cardiology, there are currently around 26 million heart disease patients worldwide, with 3.6 million new cases reported annually [12]. In the US, the majority of people suffer from heart disease. A healthcare professional would usually diagnose heart disease following a checkup of the patient's diagnostic results, medical history, and any warning signs. However, this diagnosis result cannot be used to accurately identify those suffering from cardiac disease.

Furthermore, it is very costly and computationally complex to analyze. Heart attack disease ispresently the most dangerous disease in all countries. The best strategy to reduce the rates of death carried by any disease is by early identification of heart disease and correct medication. To more accurately detect and predict diseases, machine learning algorithms/methods are widely used in the medical field[13]. The rise of heart disease is now a major problem on a global level. As a result, the healthcare department needs to enhance and modify the diseases that are managed to reduce their impact on the population. In the proposed study, we have analyzed all these aspects and tried to provide a better prediction of heart disease through various techniques which include deep learning techniques as well as machine learning. In the study, the researchers explored different factors related to heart disease and attempted to make predictions using different kinds of methods, including both machine learning and some advanced deep learning techniques. The goal of predicting heart-related diseases is significant, as it can have a substantial impact on both the medical field and the lives of individuals [14]. The ratio of deaths reduces as a result of the effective analysis and system's use of artificial fuzzy systems and machine learning classifiers to diagnose heart disease. Several researchers and used the Cleveland heart disease data set to tackle the heart disease identification issue [15]. The machine learning models for prediction require particular data for testing and training. In this research work, we have introduced a machine learning-based diagnosis technique for identifying heart diseases. Different machine learning algorithms and data processing tools are used to accurately predict heart diseases. The most utilizing machine learning technique is the Nearest Neighbor (KNN), Artificial Neural Network (ANN), Logistic Regression (LR), Random Forest (RF), Naive Bayes, Decision Tree (DT), and SVM have been executed to provide the good results.

1.2 Machine Learning

Machine learning is a branch of computer science that allows a computer to learn without the assistance of external applications. These machine learning approaches can be used to forecast the outcome of specific inputs. Machine Learning is a set of techniques that enable computers to learn without the need for human interaction. Medical diagnostics, stock market analysis, DNA sequence classification, games, robots, predictive analysis, and other applications have all benefited from machine learning. We're particularly interested in predictive analysis, where machine learning allows us to create complicated models that October be used to make predictions [16]. People benefit significantly from these models

since they provide relevant data that helps them make better decisions. Designing and developing algorithms for computers to predict behavior based on a dataset collected that's what Machine Learning is all about. It's an artificial intelligence sub-discipline. In recent years, there has been a tremendous increase in the use of these algorithms in the field of education. Pattern recognition and decision making are the major goals of these machine learning systems [17]. First, patterns are recognized, and then rules are formed based on the supplied data. The behavior is expected and decisions are made based on those guidelines.

Machine learning is a branch of computer science that is distinct from the basic computing methods that are used to solve problems. The algorithms used in machine learning are developed in such a way that the system or computer can evaluate data inputs, create training sets, and produce the required range determined output using statistical estimation [18].

Machine learning is a branch of science that uses patterns and inferences to make decisions. It builds a system that requires less human intervention using a statistical model and algorithms. Supervised learning and unsupervised learning are the two primary kinds of machine learning algorithms. This study focuses on supervised learning, which employs a classification algorithm to predict the heart disease. The algorithm and data are used to determine the outcome of machine learning in the field of health. It's important to set the appropriate statistical methods for predicting the heart disease. The machine learning algorithm determines the efficiency of the prediction result [19].

Most machine learning classifiers, such as the), K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Naive Bayes, allow multiclass classification by default . It's important to keep in mind that some research studies have introduced and compared various machine learning and data mining techniques. In 2014, several machine learning algorithms were used to predict the heart diseases [20].

Previously, manual machine learning techniques were used to review data and providebusiness projections. Currently, the rapid growth in educational site, medical site, and another etc. Its utilization to improve the quality of managerial decisions are the biggest challenges faced medical filed [21].

Machine Learning (ML) application is concerned with the extraction the unique features and recognition of the relationship among the parameters in a large amount of data.

M. Dayalanet.al machine Learning is subjected to two main objectives; prediction and data possible outcomes [22].

Machine learning aids various fields by providing tools to analyze data, recognize patterns, and make informed predictions. In healthcare, it predicts disease outcomes and personalizes treatment plans. In finance, it detects fraud and manages risk. In marketing, it helps in customer segmentation and targeted advertising. In manufacturing, it optimizes supply chains and predicts equipment failures. In transportation, it enhances route planning and enables autonomous vehicles. Machine learning's ability to process vast amounts of data and improve decision-making processes makes it invaluable across multiple domains.

Machine learning optimizes manufacturing processes by analyzing data from sensors on machines. This helps detect defects, reduce downtime, and improve overall production efficiency.

1.3 Prediction

Heart-related diseases, also known as heart diseases (HDs), have become the main cause of death, not only in Pakistan but globally in recent years. Predicting these diseases early is crucial for effective treatment and accurate and efficient methods are needed for this purpose [24]. Researchers are using machine learning methods to analyze large and complex medical data. In recent years, several studies have explored various machine-learning approaches to help healthcare professionals identify heart-related issues. This article gives an overview of different models created using these algorithms and methods [25][26].

1.4 Prediction of Heart Disease

The heart is a vital organ in human body since responsible for pumping blood and supplying oxygen to various organs. If the heart doesn't work well, it can affect other organs. Due to busy lifestyles and unhealthy eating habits, there's an increased risk of heart attack-related diseases. By analyzing health information using machine learning and deep learning,

we can understand and predict the risk and symptoms of heart diseases [26]. Heart disease risk factors include high blood pressure, high cholesterol, diabetes, smoking, drinking too much alcohol, and obesity. For example, when a person has high blood pressure, the heart has to work harder. By taking steps to manage these factors, one can reduce the risk of developing heart-related issues [27]. We've studied these factors and used various methods, including machine learning and deep learning techniques, to estimate the likelihood of getting heart disease. Predicting heart-related diseases is crucial for both the healthcare system and individuals. Additionally, other heart problems can be explained as follows [28].

1.5 Heart Treatment and Risk Factor

Heart treatment involves various medical approaches and interventions to manage heart-related conditions and diseases. The type of treatment prescribed depends on the specific heart condition, its severity, the patient's overall health, and other factors. Here are some common treatments for heart conditions.

1.5.1 Cholesterol-Lowering Medications

Cholesterol-lowering medications, commonly called statins, are drugs used to reduce cholesterol levels in the blood. Cholesterol is a fatty substance that can build up in the arteries, causing a condition called atherosclerosis and increasing the risk of heart disease. Statins work by blocking an enzyme in the liver that produces cholesterol. This leads to a decrease in low-density lipoprotein (LDL) cholesterol, also known as "bad" cholesterol. These medications not only help lower LDL cholesterol levels but October also offer other benefits, such as stabilizing existing plaque, reducing inflammation, and improving the overall health of blood vessels. It's important to lower cholesterol by following a heart-healthy diet, staying active with regular exercise, and avoiding smoking.

1.5.2 Blood Pressure Medications

Medications for blood pressure control are given to treat high blood pressure (hypertension), which is a major risk factor for heart disease, stroke, and other cardiovascular issues. There are several classes of blood pressure medications, each targeting different mechanisms that contribute to high blood pressure:

1.5.3 Heart Failure Medications

Heart failure medications are special drugs used to help people with a condition called heart failure. In heart failure, the heart struggles to pump blood effectively. These medications are used to improve the heart's pumping ability, reduce the buildup of excess fluid in the body, and relieve symptoms like shortness of breath and fatigue. They work by making the heart pump stronger, getting rid of extra water in the body, and making people feel better overall.

1.5.4 Heart Transplant

A heart transplant is a difficult medical procedure where a healthy donor heart is used to replace a damaged or failing heart. When other treatments are unsuccessful for a patient with end-stage heart failure, the typical procedure involves extracting the recipient's injured heart and delicately transplanting the donor's heart into its place. Connecting it to the recipient's major blood vessels and atria. Immunosuppressive medications are administered to prevent rejection of the transplanted heart. This procedure aims to improve the recipient's quality of life and overall survival by restoring proper cardiac function.

1.5.5 Coronary Artery Bypass Surgery

Coronary Artery Bypass surgery is a medical procedure used to help with a heart problem called coronary artery disease. This happens when the blood vessels supplying the heart don't work well because they are narrowed or blocked by fatty deposits. This blockage can lead to chest pain or heart attacks. In CABG surgery, doctors create new paths for blood to flow to the heart using healthy blood vessels from other parts of the body. This helps the blood reach the heart muscle more easily. These healthy blood vessels are called grafts, and they are typically taken from the patient's leg, arm, or chest.

1.5.6 Angioplasty and Stent Placement

Angioplasty and stent placement is a specialized medical procedure designed to treat “coronary artery disease, a condition characterized by the narrowing or blockage of the coronary arteries that supply” oxygen-rich blood to the heart muscle. This procedure is

intended to restore proper blood flow to the heart, alleviate symptoms like chest pain (angina), and reduce the risk of heart attacks.

1.6 Different Machine Learning Process

Machine learning is a field within artificial intelligence that focuses on teaching computers to learn from data without being explicitly programmed. It involves developing algorithms and models that can analyze large datasets, identify patterns within them, and make predictions or decisions based on those patterns [23].

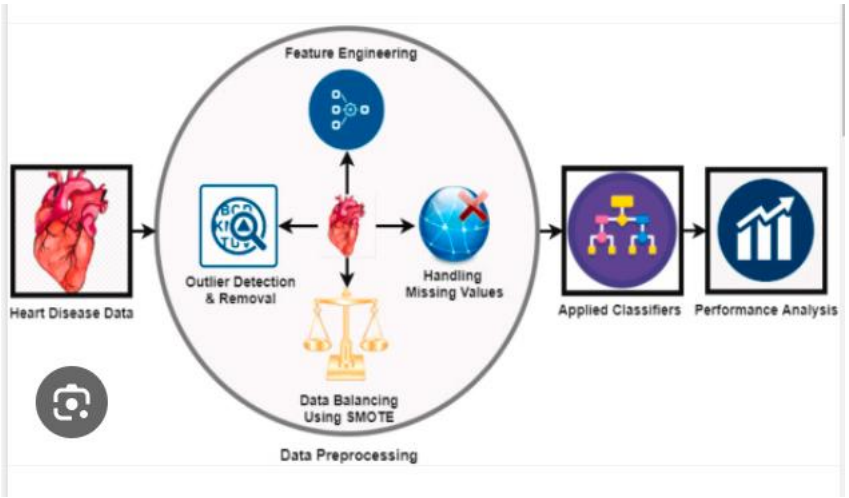

Figure 1.2: Machine Learning Process to Prediction of Heart Disease [25]

Data

Machine learning depends heavily on data. This data can come in various forms, such as text, images, audio, or numerical values. The better the variety and importance of the information, the more effectively the machine learning program can understand and apply its knowledge in different situations.

Features

Features are specific characteristics or attributes within the data that the algorithm uses to make predictions. Identifying relevant features and extracting meaningful information

from collecting the data and properly arranging the data for developing machine learning models that are accurate.

Labels

In supervised learning, labeled data is used for training. Labels are the correct answers or outcomes associated with the input data. The algorithm learns to map features to labels through training.

Algorithm/Model

Machine learning algorithms are mathematical formulas that are able to recognize and understand patterns and relationships from data. Various algorithms are better suitable to specific kinds of problems, like sorting things into groups, making predictions, finding patterns, and so on.

Training

During the training process, the algorithm analyzes the input data and adapts its internal settings to understand the patterns present. The ultimate aim is to minimize discrepancies between the predicted outputs and the actual labels.

Testing and Validation

Testing involves checking a system or model for errors or flaws, ensuring it work correctly. Validation confirms that the system or model meets its intended requirements or objectives and is reliable.

Deployment

After training and validation, the model is ready to make predictions on new data. This real-world application of the model is referred to as inference.

1.7 Supervised Learning

The most common application of supervised learning is in classifying issues. The primary purpose of this learning is to develop a classification model that allows the computer to learn about the input and forecast the outcome. Collecting data, finding an appropriate method, creating the model, and applying the model for prediction are the main steps in supervised learning. To develop a model in supervised learning, you need a set of inputs and known responses. New data is mapped to the required answers or outputs using the model that was created. The probability of all the given inputs is also provided by the supervised learning algorithms. In the dataset, there should be no missing values. It is impossible to predict the outcome if any values are missing. Nave Bayes, Logistic Regression, Decision Trees, and Neural Networks are examples of algorithms used in this form of learning.

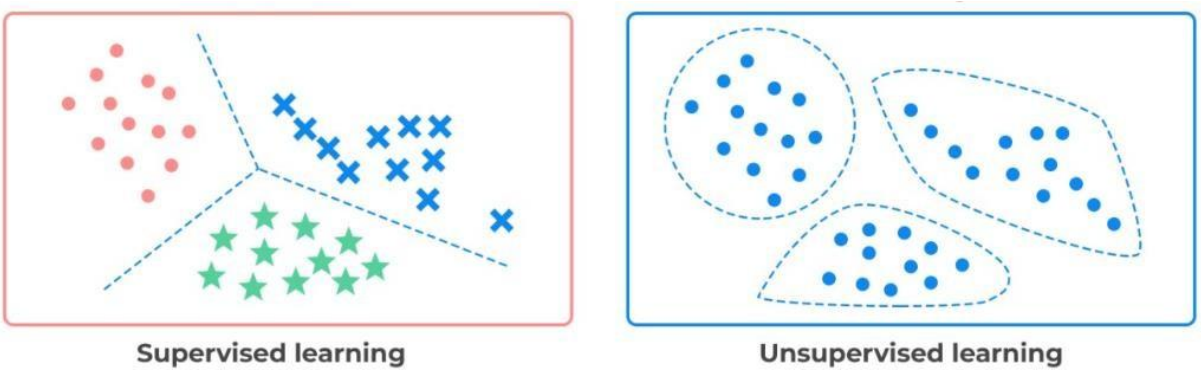

Figure 1.3: Supervised Learning and Unsupervised Learning [29]

1.8 Unsupervised Learning

Only inputs are used in unsupervised learning, and there are no outputs. The main objective of this form of learning is to model the structure of unlabeled data to get a better understanding of it. The data is left to the algorithm to analyze the relevant patterns in the data because there is no specific output. The relationships between the data are typically found here. The same collection of the dependent variable is used to obtain all of the outputs. However, in the case of supervised learning, the reason for an output set of data is the input set of data.

1.9 Problem Statement

Heart Disease (HD) is one of biggest reason in increasing death ratio. The most recent WHO study shows that heart disease is one of the major causes of this percentage of mortality [28]. There is a need to predicate accurately heart disease that would help to overcome the heart problem in a quick way and will save expensive life of people on time treatment. In various studies, machine learning techniques have already been used to predict heart diseases; however improvements in prediction result are still there by extracting significant features and using latest machine learning classifiers [29]. So, there is a need for a study that can predict heart disease better in its early stages. In existing studies various features are used to predict the heart disease but prediction of heart disease can be improved by studying more feature and apply new techniques [30].

1.10 Research Questions

The main goals of this study are to review the literature, analyze the efficiency of current heart disease prediction models and determine how various characteristics October impact heart disease performance.

RQ1: Which features can predict the heart disease at early stage?

RQ2: Which machine learning methods accurately predict heart disease?

1.11 Aims and Objectives

The research's key objectives are

OBJ 1: To identify the feature which can predict the heart disease at early-stage.

OBJ 2: To identify the machine learning methods which accurately predict heart disease.

1.12 Scope of the Study

In this study the primary focus is to predict the heart disease with higher accuracy using various machine earning approaches Machine Learning approaches have been performed but still to improve the prediction result and accuracy. Mutual information

classic/gain methods have been used for feature extraction. Random Forest, XG Boost, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Machine learning techniques and statistical methods have been used to search out the best feature extraction method for heart disease prediction at an early stage and to develop an even more accurate ensemble model with higher accuracy.

1.13 Overview of Thesis

Chapter 1 introduces Heart disease prediction and reviews past research on recognizing important features. It outlines the thesis's goals, scope, and reasons, setting the stage for the following chapters.

Chapter 2 covers essential background information; this includes different types of heart diseases, their risk factors, and the machine learning techniques used for predicting them.This chapter establishes the necessary knowledge base for the study.

Chapter 3 integrates methodologies into a framework for the research. Focuses on practical aspects, detailing tests on the system, insights into the dataset, and software programs used. This chapter offers a hands-on view of the study.

Chapter 4 analyzes the study's results, using various methods and highlighting significant findings. It provides a deep understanding of the research outcomes.

Chapter 5 A conclusions is the final part of thesis, such as an essay, presentation, or argument, where the main points are summarized, and a final judgment or decision is made. It serves to reinforce the key ideas, provide closure, and often suggests implications or future considerations. In a broader sense, a conclusion can refer to the outcome or resolution of any situation or inquiry.

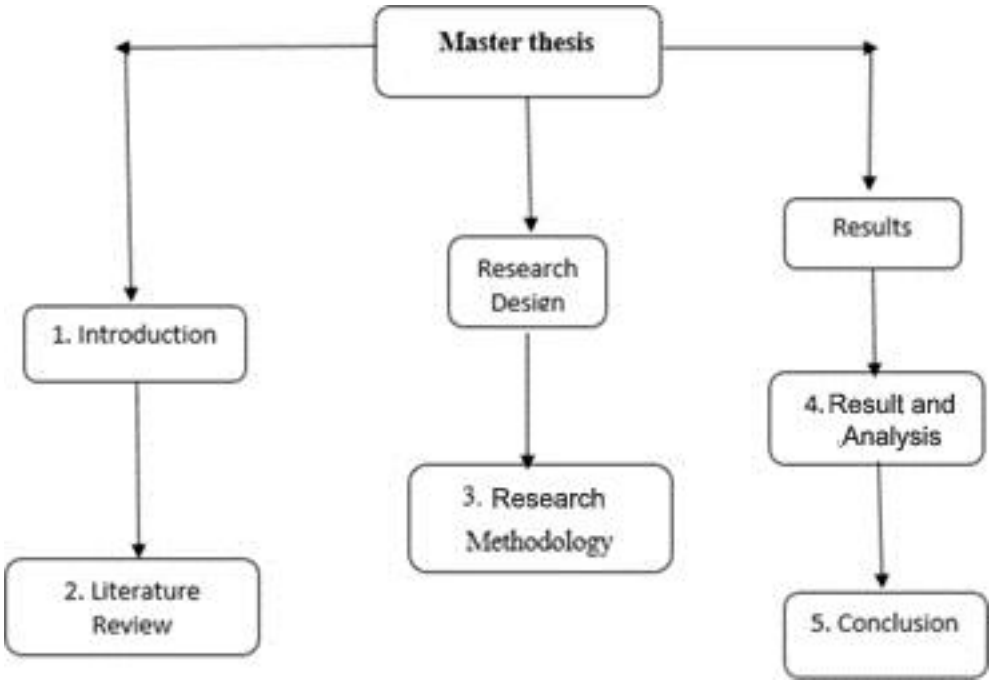

Figure 1.4: Thesis Structure

1.16 Summary

This chapter dives into predicting heart disease with machine learning. It emphasizes the importance of early detection for this leading cause of death. Machine learning offers a promising approach to analyzing medical data and predicting heart disease risk. The research aims to identify key features for early prediction and explore powerful machine learning methods to improve accuracy. It lays the foundation for the upcoming chapters, which will explore existing research, delve into the methodology, analyze the results, and discuss future directions in this field.

CHAPTER 2

LITERATURE REVIEW

In this chapter, literature review was discussed. The literature review is a critical component of any research because it offers a thorough understanding of the collection of literature already available on a particular topic. This chapter provides an overview of the related studies on heart disease, regarding the topic of this research.

2.1 Signs of Heart Disease

The most common sign of heart disease is angina or chest pain, which is one of the close indicators of heart issues. Angina is characterized by chest discomfort, a feeling of heaviness, pulsing, tightness, squeezing, or an uncomfortable sensation. Additionally, heart problems October affect the shoulders, arms, neck, throat, jaw, or back. Interestingly, compared to men, most women over 50 are more likely to experience heart disease. In contrast, males October suffer from this disease at an early young age [39]. The most common sign of Heart Disease like Sweating, Trouble breathing, Chest pain or discomfort, Feeling weak, dizzy, Faster heartbeat, Discomfort or pain in the shoulders or arms, Weakness or lightheadedness, Pain or pressure in the chest, Vomiting, feeling sick, drowsiness, or sweating, Severe anxiety and weakness, Irregular or fast heartbeats.

2.2 Causes of Heart Disease

Heart disease occurs when the heart or blood vessels are damaged. This damage can happen because of blocked arteries or the buildup of fatty material. When these deposits accumulate, the walls of the arteries become thicker and less flexible, restricting blood flow to the organs and tissues. Cholesterol plays a significant role in this issue, and it often worsens due to factors it can be controlled, such as smoking, being overweight, consuming unhealthy food, and not getting enough exercise [40]. Issues with blood vessels around heart and high sugar levels in the blood.

2.3 Risk Factors of Heart Disease

A risk factor for heart disease is any attribute or condition that increases the likelihood of developing cardiovascular issues. Common risk factors include high blood pressure, high

cholesterol, smoking, obesity, and a sedentary lifestyle.

Table 2.1: Different Types of Heart Disease Risk Factors

Factors	Description
Age	As age, there's an increased likelihood of damaging the arteries, becoming narrow, and heart muscle weakening or thickening.
Sex	Heart disease is normally more common in men, but in women, the risk increases after age.
Family history Smoking	The possibility of coronary artery disease is increased if you have a family history of heart disease, particularly if one of your parents had the condition while they were young (before the age of 55 for a male relative, like your brother or father, and 65 for a female relative, like your mother or sister).
Smoking	Smoking can make blood vessels narrower and harm their inner lining, which raises the risk of atherosclerosis. Heart attacks happen more often in smokers than in non-smokers.
Poor diet	Eating a lot of fatty, salty, sugary, and cholesterol-rich foods can lead to the development of heart disease.
High blood pressure Uncontrolled	When a person ignored high blood pressure, the arteries October thickensand narrow, causing blood vessels to shrink.
Diabetes	Heart disease is more likely for people with diabetes. Risk factors for both conditions are comparable, including high blood pressure and obesity
Obesity	Obesity is a health condition where a person has too much body fat, often leading to health problems.
Physical inactivity	In addition, a number of heart disease subcategories and other risk factors are linked to exercise.
Stress	Stress which is not controlled October increase other heart disease riskfactors and damage the arteries.
Poor hygiene	The chance of getting heart infection become increase if a person don’t wash hands on a regular basis and don't adopt other habits that can help stop bacterial or viral infections, particularly if a person already have a heart condition. Heart disease October also be increased by poor teeth health.

2.4 Different Types of Heart Diseases

Heart-type" typically refers to the various categories or classifications of heart conditions based on factors such as anatomy, function, and characteristics. Heart disease

comes in various forms [28]. Heart attack, Heart Failure, Idiopathic Heart Disease, Cardiomyopathy, Cardiac ischemia.

2.4.1 Coronary Artery Disease

This condition causes blockages in the coronary arteries. The heart receives the necessary oxygen and nutrients from these arteries, but it is not able to provide them properly because of blockage that causes cholesterol to pass out of the arteries.

2.4.2 Heart Failure

This type of heart failure, also known as congestive heart failure, happens when the heart can't pump enough blood to all parts of the body. This heart disease situation is quite severe.

2.4.3 Congenital Heart Disease

This heart problem is there from when a person is born. For example, there could be holes between the two sides of the heart, causing problems like breathing difficulties or blockages that partially or completely stop the blood flow through different parts of the heart.

2.4.4 Cardiomyopathy

This condition makes the heart muscles weak or changes their structure, leading to a decrease in the heart's ability to pump blood. This can eventually result in heart failure.

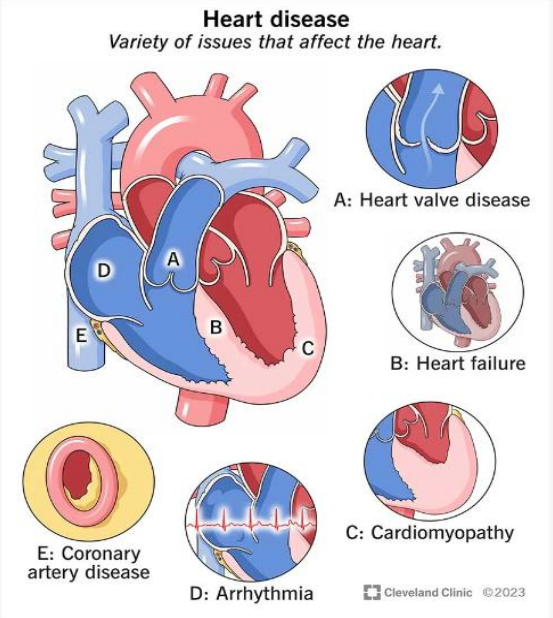

Figure 2.1: Type of Heart Disease. [22]

2.5 Work Related to Heart Disease

Heart disease encompasses various conditions affecting the heart, such as issues with its muscles, valves, rhythm, or blood vessels. These conditions include coronary artery disease, heart failure, arrhythmias, and heart valve problems. Symptoms October include chest pain, shortness of breath, fatigue, and palpitations. Heart disease is a leading cause of death globally and can greatly affect a person's well-being if not managed effectively [31].

Coronary artery disease (CAD) is a common heart condition worldwide, also known as ischemic heart disease or a heart attack. It's a significant health issue and can be life-threatening.

The diagnosis and treatment of heart disease can be complex, especially in developingcountries. This is due to limited access to effective diagnostic tools and a shortage of medical professionals and resources, which affects accurate predictions and patient care [32].

Coronary Heart Disease (CHD) affects the heart's ability to pump blood throughout the body. It can lead to narrowed arteries, reducing oxygen and nutrient supply. This narrowing is often caused by a buildup of calcium and fatty deposits called atherosclerosis [33]. According to the World Health Organization (WHO), cardiovascular disease is a leading cause of death globally, with heart disease and stroke being major contributors. Risk factors include gender, age, genetics, obesity, diabetes, stress, and dietary habits [34].

The heart is a vital organ that circulates blood throughout the body. Inadequate blood flow can affect organs like the brain, and complete heart failure leads to death. Proper heart function is essential for life. Heart disease refers to conditions related to the heart and blood vessels [35].

Most deaths are caused by coronary artery disease, also known as cardiovascular disease, which accounts for 20% of all deaths, mainly due to stroke or heart attack [36].

Machine learning is a technique used to extract important information from large amounts of data. It involves predictive models, which forecast specific outcomes, and descriptive models, which enhance data understanding without highlighting specific variables. Various machine learning methods, such as Multi-layer Perceptron (MLP),

Decision Tree (DT), K-nearest Neighbor (K-NN), Support Vector Machine (SVM), and Naive Bayes (NB), are valuable for analyzing large volumes of information [37].

The study aims to predict cardiac disease diagnosis using data mining techniques. By analyzing variables like gender, age, cholesterol levels, blood sugar, and blood pressure, the Naive Bayes algorithm is applied for classification. This helps doctors identify heart disease quickly, saving time and resources, and ensuring prompt treatment [38].

A literature review on heart disease involves examining and summarizing existing research and scholarly work related to heart health. It explores studies, articles, and publications on topics such as risk factors, preventive measures, treatment approaches, and recent advancements in understanding heart diseases. The review aims to offer a thorough overview of the current state of knowledge, identify gaps or inconsistencies in research, and lay the foundation for new studies or interventions related to heart health**.**

As per reports, there's a heart disease prediction system that uses machine learning. This system forecasts the risk of getting heart disease beforehand. It utilizes KNN and Decision Tree Algorithms to assess the heart disease risk level. The system depends on 13 medical factors like age, gender, fasting blood sugar, and chest pain. The system's outcome presents the likelihood of getting heart disease as a percentage and shows the accuracy of the two algorithms. The decision tree algorithm, with an 81% accuracy rate, detects the probability of heart disease in patients based on the dataset. Meanwhile, the nearest neighbor method, applied to the same dataset, predicts the likelihood of heart disease in patients with a 75% accuracy rate [41] [42].

This section discusses how machine learning classifiers are important for detecting and identifying heart diseases. Different tools like naive Bayes, logistic regression, support vectormachines, nearest neighbor, K-Means clustering, decision trees, and random forests are used to predict various illnesses. Compare these classifiers to see which one is best at finding diseases, Comparisons are made regarding these tools, considering how they are evaluated the results. Lastly, discussions about findings from previous studies in this field are presented. This article explains how machine learning classifiers and data analysis are used to predict heart disease. Many research have explored using these techniques to identify heart problems. This paper summarizes recent studies that show how well these machine learning techniques can predict heart disease, to understand more, many tools and methods are used

and how many effective these are new approaches [42].

Another research study developed a machine learning technique to predict heart disease. Heart disease is a significant concern due to its increasing rate. Which patients were more likely to be affected by heart disease by analyzing changes in health symptoms. Various techniques, including LR and KNN, were employed to forecast and identify heart disease patients. The proposed framework demonstrated good accuracy in identifying specific heart disease symptoms. This predictive method for cardiac disease could assist in analyzing large datasets simultaneously, improving disease diagnosis, and enhancing patient treatment [43].

Many studies have examined how machine learning can be used to identify heart disease. These studies have resulted in numerous publications focusing on the role of machine learning in detecting cardiac disease. This summary of recent research serves as an introductory guide to understanding the complexity of this field, including the tools and techniques utilized by researchers, as well as the performance levels achieved by various advanced methodologies [44].

The decision tree technique is used to analyze the type of coronary heart disease. Everyyear, three million people die from heart and blood vessel problems. Out of these, three million die from a specific heart issue, and 6.2 million die from strokes. Different things like age, gender, family history, smoking, not moving around much, being overweight, diabetes, high blood pressure, and what you eat can make you more likely to have heart problems. In this study, they made a set of information to help sort out and identify the type of heart disease by looking only at whether there is a problem with the blood vessels [45].

Based on a survey, a study is performed to forecast heart disease using machine learning. They evaluate the efficacy of different algorithms and methods in their research. The researchers built models using supervised learning algorithms such as K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Decision Trees, Random Forest, and others [46].

A recent study on Coronary Heart Disease Statistics from 1961 to 2011, released to celebrate the 50th anniversary of the British Heart Foundation (BHF), offers a thorough examination of how the disease has changed in the UK over the last 50 years. This document is an addition to the series "Coronary Heart Disease Statistics," regularly shared with the public with support from the BHF. It delves into CHD, long-term trends in other

cardiovascular diseases like heart failure and stroke, and risk factors for CHD such as obesity, poor diet, and smoking. "Trends in Coronary Heart Disease Statistics 1961-2011" primarily aims to identify these patterns rather than explain them. It is designed for health professionals, scientific researchers, journalists, students, and anyone interested in CHD, providing a broad overview of the biology of CHD in the UK over the past fifty years [47].

The heart is really important for keeping living things alive. But figuring out and predicting heart disease needs to be very accurate because even a small mistake can make people tired or, in serious cases, cause death. Unfortunately, more and more people are dying from heart problems every day Need a system that can predict when someone might get sick. Machine learning, which is part of Artificial Intelligence (AI), can help a lot by learning from natural events. This study checked how good machine learning is at predicting heart disease. They applied various techniques such as K-nearest neighbor, random forest, logistic regression, and support vector machine (SVM) algorithms. They tested these methods using data from the UCI repository. The researchers found that using a tool called Notebook for Python Anaconda (Jupyter) programming was the best way to do this because it has lots of different tools that make the job more accurate [48].

Inflammatory disease, also known as coronary heart disease (CHD), affects the heart, which pumps blood throughout the body. It occurs when the arteries that supply blood to the heart become narrow, reducing the flow of oxygen and nutrients needed for the heart to function properly. This narrowing is often due to a buildup of calcium and fatty deposits. According to the World Health Organization (WHO), cardiovascular disease is the leading cause of death worldwide, claiming approximately 17.3 million lives annually. Heart disease and stroke are the main causes, resulting in 7.3 million and 6.2 million deaths, respectively. These factors increased the risk factors for heart disease include age, gender, genetics, smoking, physical inactivity, stress, obesity, diabetes, and diet [49].

Another researcher R. Subramanian applied neural networks to introduce blood pressure, heart disease prediction and prognosis, and other features. Using the provided disease-related attributes, a very complex neural network with almost 120 hidden layers was built. This is a fundamental approach to ensure accuracy in predicting heart disease when applying the model to the test dataset. When a doctor used the model with new data, it used the information it had learned before to train and make predictions, helping to determine its accuracy. [50].

All over the world, medical institutions collect datasets on many different kinds of health-related fields. From these data, a variety of machine-learning algorithms can be applied to select the valuable data. Nevertheless, a large amount of noisy data is generally gathered. These datasets size was too much for human analysis, then applied the preprocessing approaches can be quickly studied with a range of machine-learning methods. According to Liu et al, these algorithms have been recently shown to be highly helpful in properly forecasting the existence or absence of heart-related diseases [51].

Another research study with the help of doctors to establish an accurate and useful prediction system, and another investigation was carried out globally, the primary cause of death is heart disease. It is quite challenging for researcher to predict coronary artery disease because it is such an involved and very expensive process. Offering a cardiac disease prediction system to help physicians diagnose patients and make better decisions. Using medical parameters from medical records, the author of this work applies different machine learning classifiers, including KNN, SVM, NB, RF, and DT, to predict cardiac disorders. Utilizing many UCI datasets [52].

In this research author Used supervised machine learning techniques, this study provides an evaluation and comparison of heart disease prediction. In many nations, there is a lack of cardiovascular knowledge, resulting in high rates of misdiagnoses and preventable cases. The Random Forest (RF) approach achieved 100% sensitivity when applied to heart disease datasets from Kaggle website. The study suggests that even a basic monitoring machine-learning method can yield significant results [53].

For large dimensions and merged data, a more effective SVM classifier is produced, especially for high-dimensional data. The KNN method is used for classifying objects with a high pairing of continuous and categorical data, demonstrating superior performance for challenging surfaces of mixed data. Evaluation results indicate the proposed technique's better performance in terms of categorization.

In another research study, the author proposes a decision support system for categorizing cardiac diseases based on a Support Vector Machine (SVM) with an integer-coded genetic algorithm (GA). The Simple SVM method quickly finds support vectors iteratively. Utilizing data from the Cleveland Heart Disease Database, the classification problem shows increased overall accuracy up to 72.55%, suggesting its effectiveness as a

decision support system [54].

Machine learning is extensively used for identifying cardiac conditions, offering improved comprehension and prediction of medical datasets. One study applies the K-nearest neighbor method to predict coronary artery disease using Python Flask. The accuracy of the method is 65.93%, increasing to 82.41% after z normalization. The research demonstrates how the z-score can significantly enhance the accuracy of the KNN algorithm [55].

In a study involving 11 different machine learning models, researchers trained these models to check for heart disease using a large set of features from the Cleveland heart dataset. Testing with different feature sets chosen by three methods showed varying accuracy, with the SMO model achieving about 85.148% accuracy when using all features. Another method, Meta classifier bagging plus logistic regression, outperformed, showing a value of 0.91 for ROC area [55].

Heart disease is a major health problem around the world. In one study, researchers came up with a new model to predict heart disease. They combined two methods, Random Forest and Support Vector Machines. This combined model worked better than using either method alone. It was really accurate, reaching a success rate of 98.3%. Also, they found that adjusting certain things, like C and gamma, made the Support Vector Machine work even better [50].

Many research studies tried to find out the main things that make people more likely to get heart problems, especially coronary heart disease. They used different computer methods like Naive Bayes, Bernoulli Naive Bayes, and Random Forest to analyze a big set of data from the UCI Cleveland database. They found that Gaussian and Bernoulli Naive Bayes were better than Random Forest at getting accurate results and remembering important information [56].

To enhance the accuracy of clinical assessments for heart disease, researchers have suggested different techniques, like principle component analyses and chi-square. The smote-x-boost technique was introduced for predicting cardiac disease, demonstrating advantages over other algorithms in precision, accuracy, recall, F1-score, and AUC [57].

This study aimed to accurately determine overall risk and identify key symbols of heart disease risk using common machine learning techniques. It explores combining these

methods with hybrid machine learning methods and suggests a rule-based strategy to evaluate accuracy using rules from LR, DT, and SVM on the Cleveland Dataset [58]

Extracting a specific feature from an heart diseases that is appropriate for the study is known as feature extraction. It is essential components of recognition using heart disease only extract the relevant characteristic that corresponds to ensure the accuracy of the classification. A brief overview of the field of disease identification via heart disease was put out by author [5]. The reliable features reflect the relevant, practical data for the tasks' unique requirements. In general, having excellent features helps to identify disease states effectively and practically, as it significantly determines a system's unique capacity to identify objects via heart disease [111]

Feature selection is a general term for an automated process that selects only the most important and relevant features of an object. Redundant or irrelevant features are discarded or ignored before they have a chance to degrade the performance of the classification system. The use of feature selection thereby produces higher accuracy and reduces the overall computational cost of modeling [2].

Feature selection October be performed by a human expert, semi-automatic or explicitly through an automated process. The automatic feature or attribute selection process tags the most important and relevant attributes using many types of algorithms: Filter Methods, Wrapper Methods, Principle component method, Chi-square method and Embedded Methods. Filter methods use a scoring mechanism that calculates the statistical score of each feature, ranking the most probable highest.

Higher the score of a feature more probable the feature to be selected. Examples of Filter methods are correlation coefficients, Fisher scores and information gain. Wrapper-based feature selection looks for a combination of features, with each combination being labeled as a feature subset. Subsets can be used for prediction, predicting performance is calculated using some metric.

A subset with the highest performance metric is the one that is termed the required feature subset. Wrapper methods include: forward selection, best first search and random hill-climbing algorithms. Embedded Methods of feature selection select features through learning. When the model is first created, the features must be present, or the accuracy of future inspections will be defeated. Embedded Methods include Elastic Net and Lasso.

Feature Extraction differs from each of the previously discussed methodologies. Feature extraction does not look for individual features or feature subsets. Rather, feature extraction transforms the original feature set from higher dimensional space to lower one.

The features are not selected but the given feature space is projected to a new feature space. As an example, principal component analysis (PCA) is a feature extraction technique which is used in current research as well. PCA computes a covariance matrix for the input features, finds eigenvalues and eigenvectors for the matrix, selects the eigenvectors corresponding to high eigenvalues, creates a projection matrix from selected eigenvectors and applies the projection matrix over input feature set. This projects the original feature space into a lower dimension space. Linear discriminant analysis (LDA) is also used for feature extraction.

LDA maximizes inter-class distances and minimizes intra-class distances in order to create a new projection. Support vector machines are proved to be used in solving the problem of classification [3], [4], [5], [6]. Least square support vector machines (SVMs) were used for predicting and modeling of gases in liquids [7], [8].

On the other hand, to predict, classify, model water conditions, depth around bridge pilers and velocity prediction in sewer pipes, support vector machines are applied and proved to be efficient in the water science domain along with flow measurements [9], [10], [11], [12], [13] while sign language categorization problem was solved using these machines as well [14].

The performance of SVMs in cross-domain problems and handling the heterogeneous type of data makes it a good candidate for classification among the other ones.

Table 2.2: Relevant Research Studies

Year	Dataset	Features Selection	Classifiers Used	Accuracy	Limitations
2024 [59]	UCI dataset	Stack of Transformer Encoder layers.	Transformer model	Achieved the highest accuracy of 96.51%	Requires significant computational resources and memory complexity

2023 [60]	UCI dataset	Hybrid Feature Selection Framework	Hybrid multi-stage stacking classification framework	Accuracy achieved 95.3%	Framework can be complex to developand implement
2023 [61]	UCI dataset	Filter based feature selection	KNN, RF, SVM, NB and LR	RF and NBwas performed better as compared to other Classifiers	Ensemble methodsare complex and require more resources
2021 [62]	UCI dataset	Feature Importance technique	KNN and Decision Tree Algorithms	KNN 75% Decision Tree 81%	Accuracy relies heavily on havingsuper clean data
2016 [63]	UCI dataset	Forward , Backward Chi-square statistics test	KNN,SMO,J48 ,Naïve Bayes	J48 achieves higher accuracy compared tothree other Techniques	Accuracy relies heavily on having super clean data
2019 [64]	UCI dataset	PCA Minimum Redundancy Maximum Relevance	Modified K-means algorithm and Naive Bayes algorithm	NB achieves higher accuracy	Difficult to understand how algorithms arrive atpredictions, accuracy relies onhigh-quality data
2018 [65]	UCI dataset	RF used to extract features	Decision Tree, KNN and K-Means	Accuracy87.4%	Models can be complex to developand maintain due to combining two different techniques
2020 [66]	UCI dataset	Filter based PCA	KNNJ48 SMO Multilayer perception	KNN 78% J48 86% SMO 89% Multilayer perception86%	WEKA is not a complete solution for heart disease prediction, requiresexpertise in algorithm selection
2022 [67]	UCI dataset	PCA LASSO	J48 Logistic model tree algorithm Random forest algorithm	J48 is more accurate at 56.76% compared toLMT algorithm's 55.75%.	Models can fail to generalize well to unseen data, leadingto inaccurate predictions

2016 [68]	UCI dataset	Forward , Backward Chi-square	Neural network WETA Tool MATLAB	achievesaccuracy 84%	Interpretability challenges, data quality dependence, overfitting risk
2021/ [70]	Cleveland Heart Disease Dataset	PCA	DT,RF,SVM, LDA,LR	SVM achieved higher accuracy as compared toother algorithms	Limited Data Availability, Interpretability Challenge are facing
2020 [71]	UCI dataset	LASSO Relief,	KNNSVMNB	KNN 90%	Limited data availability, interpretabilit y Challenges
2021 [72]	Kaggle dataset	PCA LASSO Relief,	KNN, RF, ANN, Ada, GBA	ANN achieved higher accuracy as compared toother algorithms	Complex algorithmsOctober fit too closely to training data, leading to poor generalization
2019 [73]	Kaggle dataset	Relief, MRMR, and LASSO	Logistic Regression, KNN,SVM, NB,DT, and RF,	The Logistic Regression algorithm had the best performance and achievedhigh accuracy 89%	Reliance on ML models October introduce errors orbiases, lack of interpretability in some models
2020 [74]	Cleveland dataset	Correlation base	ANN SVMKNN,NB ,	ANN andSVM achieved highest result	Ensemble methodscan be computationally intensive, complex to implement
2018 [75]	Cleveland dataset	Principal Component Analysis (PCA) chi-square statistics test	Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models	Ensemblemodels archives higher accuracy	Over fitting occurs when models become overly tailored to training data, capturing noiseor irrelevant patterns, resulting inpoor real-world performance with new data.

2019 [76]	UCI dataset	Chi-square, Fisher's co-relation,	RF,SMO, AdaBoost,NB	Random Forest 70%.	ML for heart attack prediction needs lotsof clean data, can becomplex for doctors to understand
2021 [77]	Kaggle dataset	Maximum and Minimum	LR RF CNN	LR 94% RF 93% CNN 93%	Limited data access, interpretability challenges, generalization issues, ethical consideratio ns

2.7 Summary

The literature review in this chapter provides a comprehensive overview of existing research on heart disease, focusing on various aspects such as its definition, symptoms, causes, risk factors, and related studies. It highlights the significance of understanding heart disease due to its global prevalence and impact on health. The review discusses different techniques used in research, including machine learning algorithms like Naive Bayes, Decision Trees, Support Vector Machines, and K-nearest Neighbor, for predicting heart disease based on factors such as age, gender, cholesterol levels, and blood pressure.

CHAPTER 3
RESEARCH METHODOLOGY

In this chapter, a brief overview of the methodological approach adopted for predicting heart disease using machine learning is given. The approach involves several steps, each aimed at addressing different aspects of the problem.

3.1 Introduction

Firstly, a comprehensive dataset is collected/ download from internet. Which containing various attributes related to patient's health and medical history, including factors like age, gender, blood pressure, cholesterol levels, and presence of other medical conditions.

Next, exploratory data analysis is conducted to obtain awareness into the distribution and relationships between different features in the dataset. This helped to identify patterns, outliers, and potential challenges that needed to be addressed during preprocessing.

After cleaning and preprocessing, feature engineering was done to extract relevant information and new features were created that could improve the predictive performance of the models.

A range of machine learning methods, like logistic regression, decision trees, random forests, and support vector machines, were employed to build predictive models. These models' performance was measured using relevant techniques such as cross-validation and hyper parameters tuning.

Finally, results are interpreted; findings are discussed and identified potential avenues for future research.

The primary goal of the methodology chapter is to provide an understandable and comprehensive description of the study technique used.

3.2 Dataset

This research is based on quantitative methods, which involve working with a numerical dataset to make predictions. The UCI dataset, or University of California,

Irvine (UCI) Machine Learning Repository, is a freely available online resource.

The dataset contains numerical data about 1025 patients, with 76 attributes. These attributes help analyze various aspects of patient health. Predicting heart disease based on this data is crucial. Extracting insights from this dataset is challenging but necessary for better understanding the problem.

Table 3.1: Heart Disease Dataset's Features.

Sr. No.	Features / attribute	Overlapped	Information
1	Age	Yes	Patient's age in years
2	Sex	Yes	The person's sex
			0 = female
			1 = male
3	Cp	Yes	Chest pain experienced
			0 = typical angina
			1 = atypical angina
			2 = non-anginal pain
			3 = asymptomatic
4	Trestbps	Yes	Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	Yes	Serum cholesterol in mg/dl
6	FBS	Yes	(Fasting blood sugar >120 mg/dl)
			1 = true
			0 = false
7	Rest Ecg	Yes	Resting electrocardiographic measurement
			0 = normal
			1 = having ST-T wave abnormality,
			2 = showing probable or definite left ventricular hypertrophy by Estes' criteria.
8	Thalch	Yes	Maximum heart rate achieved
9	Exang	Yes	Exercise-induced angina
			0 = false
			1 = true

10	Oldpeak	Yes	ST depression induced by exercise relative to rest
11	Slope	Yes	the slope of the peak exercise ST segment
			0 = upsloping
			1 = flat
			2 = downsloping
12	Ca	No	The number of major vessels (0–3)
13	Thal	No	A blood disorder called thalassemia
			3 = normal
			6 = fixed defect;
			7 = reversable defect)
14	Target		Has heart disease or not
			0 = no
			1 = risk

3.3 Dataset Description

In a heart disease dataset, data description involves explaining each piece of information. All different attributes are analyzed, including details about age, gender, blood pressure, cholesterol levels, smoker, target value, slope, family history, and other health factors. It's important to note the types of data (like numbers or categories), the range and distribution of values, and whether any information is missing. Understanding how the data is organized and providing any background details helps researchers make sense of the information. If there's a specific thing the analysis aims to predict, like the presence of heart disease, that's the target variable. In this research, the dataset is downloaded from the UCI website.

This thesis focuses on accurately predicting heart disease. Furtherdetails about patient attribute.

3.3.1 Number of Instances

The dataset includes information from patients, with each record representing an individual case or instance. These instances include information about patients' age, gender, medical history, and diagnostic test results, among other factors. The dataset size is sufficient to perform robust analysis and develop predictive models for heart disease.

3.3.2 Number of Features

The dataset consists of multiple features or variables that describe different aspects of patients' health and lifestyle. These features include demographic information such as age and gender, physiological measurements such as blood pressure and cholesterol levels, and medical history indicators such as presence of diabetes or hypertension. The features capture a diverse range of factors that October influence the risk of heart disease, enabling comprehensive analysis and prediction.

3.3.3 Types of Features

The features in the dataset can be classified into different types based on their nature:\

3.3.4 Categorical Features

These attributes describe qualitative aspects, such as whether someone is male or female, and the different types of chest pain they might feel during a heart issue), and presence of hypertension (yes/no).

3.3.5 Numerical Features

These represent quantitative measurements such as age, blood pressure, cholesterol levels, and heart rate.

3.3.6 Binary Features

These are special categorical features with only two possible values, often representing binary outcomes such as presence of heart disease (yes/no).

3.3.7 Labels

Labels in machine learning are the outcomes or results. When working with labeled data, each data point in dataset is associated with target labels or specific category we want our model to learn to predict. In heart disease prediction, the labels might indicate whether a person has heart disease or not. These labels guide the training process, helping the model learn to make accurate predictions.

3.3.8 Target Labels

In the heart disease prediction task, the target label is defined as a binary variable indicating whether a patient has been diagnosed with heart disease. This binary label is assigned based on diagnostic criteria and medical assessments, with '1' indicating the presence of heart disease and '0' indicating its absence.

3.4 Result of Preprocessing

Before conducting analysis, it is essential to clean the data and handle any missing or inconsistent values. This involves several steps

- Check the data type
- Identifying missing values
- Imputation
- Removal of duplicates
- Outlier detection

The preprocessing step ensures that the data is ready for analysis and modeling without the need for imputation techniques. It can lead to more reliable results in machine learning models trained on the data.

3.5 Data Cleaning and Handling Missing Values

Before conducting analysis, it is essential to clean the data and handle any missing or inconsistent values. This involves several steps

- **Identifying missing values**
  Missing values in the dataset are find out, and their impact on the analysis isevaluated.

- **Imputation**

  To fill in missing values for numerical features, methods such as using the average, middle value, or most common value are used. For categorical features, the most common value is used for imputation.

- **Removal of duplicates**

  Duplicate records are checked and removed to avoid redundancy in the dataset.

- **Outlier detection**

Found and fixed unusual data points that could mess up the analysis or how well the model works.

3.6 Data Analysis

Exploratory Data Analysis (EDA) is the first and prepressing step in analyzing data. It's all about the dataset to really understand its features before applying any techniques. EDA uses various techniques to uncover the structure, patterns, and relationships within the data.

3.7 Testing and Training Dataset

In the context of using a heart disease dataset for machine learning, data testing and training with an 80/20 split refers to the process of dividing your data into two sets for training a model to predict the presence or absence of heart disease. The machine learning based framework heart disease prediction to automatically analyze the heart disease. The dataset consists of 1025 patients some of which have positive patients' dataset and the other have negative patient's dataset. This dataset can be used to train for prediction of heart disease. An experiment uses a ratio of 8:2 in which 80% of data is used for training and 20% of data is used for testing purpose.

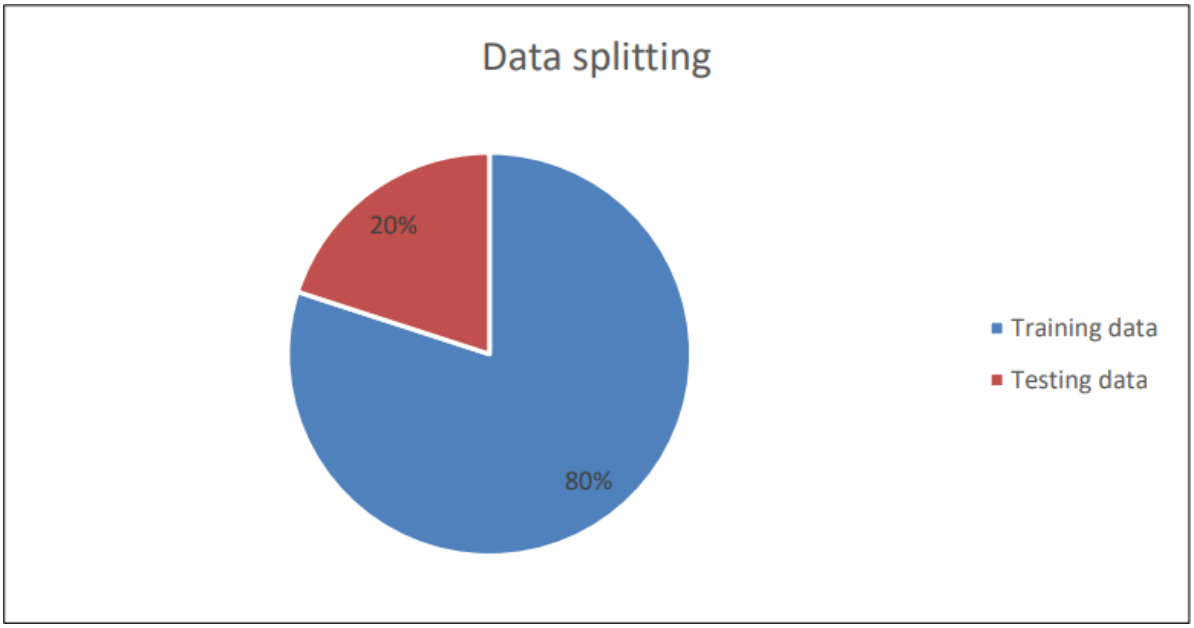

Figure 3.1: Data Splitting

Research methodology is the way to explain how the research was conducted by the researcher which methods were used to identify the process, select the information about the topic. Methodology refers to the systematic, theoretical analysis of the methods applied to a field of study. It encompasses the principles and rules that guide research, including the techniques and procedures used to collect and analyze data.

This research is based on quantitative methods, which involve working with a numerical dataset to make predictions. The UCI dataset, or UCI Machine Learning Repository, is a freely available online resource. It provides open access to a vast collection of datasets covering various diseases and encompasses a wide range of databases and research conducted by scientists. I identified specific data points and selected a category for predictions. To analyze the data, we employed the Matplotlib and Seaborn framework, designed for quantitative machine learning implementations. Using various machine learning classification methods, Tests are conducted to evaluate the accuracy of different classifiers on quantitative datasets [79].

3.8 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial phase in analyzing data. It involves thoroughly examining and understanding the dataset's features before moving on to formal testing or modeling. EDA includes a range of techniques and methods designed to reveal the data's structure, patterns, and relationships.

During EDA, analysts typically visualize the data using charts, graphs, and statistical summaries to identify key features, trends, and anomalies. This exploration helps researchers understand the distribution of variables, detect outliers, assess the presence of missing values, and explore potential relationships between variables.

During EDA, researchers employ a variety of techniques to uncover key insights and patterns in the dataset.

3.8.1 Distribution of Classes

Class distribution analysis focuses on examining the distribution of the target label (the presence or absence of heart disease refers to whether a person has it or not.) within the

dataset. It helps us understand the prevalence of heart disease and assess potential class imbalances. Frequency or proportion of each class is calculated and visualize the distribution using histograms or pie charts. Understanding the class distribution is crucial for selecting appropriate modeling techniques and evaluating model performance.

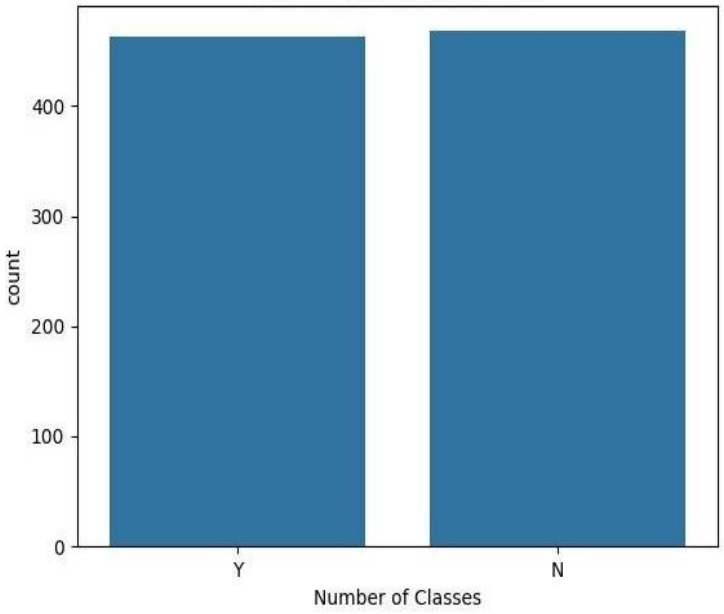

Figure 3.2: Class Distribution Analysis

3.8.2 Distribution of Ages

The figure below visualizes the frequency of different age groups within a dataset. The horizontal axis, or x-axis, shows age intervals, the vertical axis, or y-axis, indicates how frequently occurs. Or count of individuals in each age group. Each bar in the histogram corresponds to an age interval, with the height of the bar reflecting the number of individuals in that particular age range.

Examining this visualization provides several insights into the dataset's age distribution. Firstly, it allows us to observe which age groups are more prevalent or dominant within the dataset. Peaks in the histogram indicate age intervals with higher frequencies, suggesting that these age groups are more represented in the dataset compared to others.

Additionally, the visualization helps in identifying the central tendency of the age distribution, indicating the most common or average age within the dataset. This is often determined by locating the age interval with the highest frequency, representing the age group with the largest number of individuals.

Moreover, the spread of the histogram bars provides insights into the variability or dispersion of ages within the dataset. A wider spread indicates a broader range of age's present, while a narrower spread suggests a more concentrated distribution around specific age groups.

By examining any outliers or unusual patterns in the histogram, potential anomalies within the dataset can be identified. Outliers October represent unique characteristics or exceptional cases within the dataset, warranting further investigation.

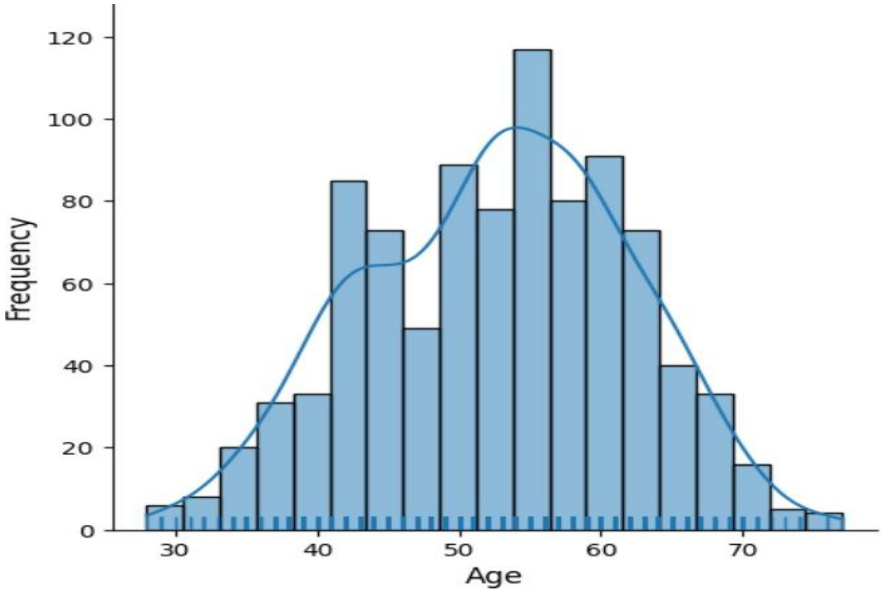

Figure3.3: Ages Distribution Analysis

3.8.3 Distribution of Fasting Blood Sugar

The "Distribution of Fasting Blood Sugar" visualization, represented in the form of a pie chart, displays the proportion of individuals within a dataset categorized into two groups based on their fasting blood sugar levels "Yes" and "No".

In this visualization, the pie chart is divided into two segments, each representing one of the two possible options "Yes" and "No". The size of each segment corresponds to the proportion of individuals within the dataset belonging to that particular category.

For instance, if the "Yes" segment of the pie chart is larger, it indicates that a greater percentage of individuals in the dataset have fasting blood sugar levels categorized as "Yes". Conversely, if the "No" segment is larger, it signifies that a higher proportion of individuals have fasting blood sugar levels categorized as "No".

This visualization provides a clear and intuitive way to understand the distribution of fasting blood sugar levels within the dataset, highlighting the relative prevalence of individuals with and without elevated fasting blood sugar levels.

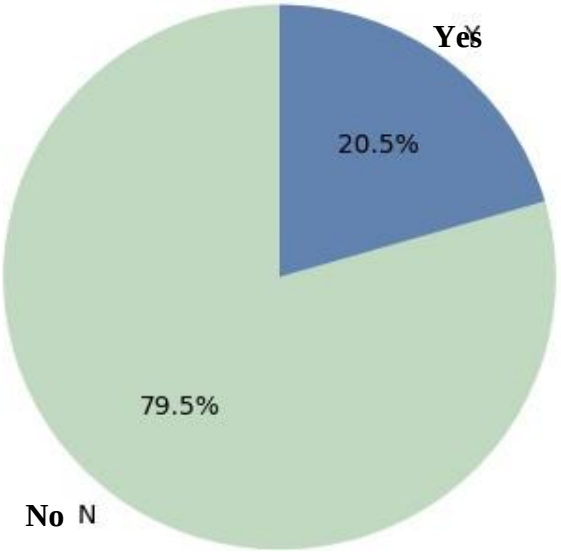

Figure 3.4: Fasting Blood Sugar Distribution Analysis

3.8.4 Distribution of Resting Blood Pressure

The distribution of resting blood pressure in a heart disease dataset describes how the measurements of blood pressure are spread out across individuals in the dataset. In simpler, it shows the range and frequency of different blood pressure values when people are at rest. Analyzing the distribution might involve creating a graph, like a histogram, to visualize how common different resting blood pressure values are in the dataset. This can provide insights into whether certain ranges of blood pressure are more prevalent among individuals with or without heart disease.

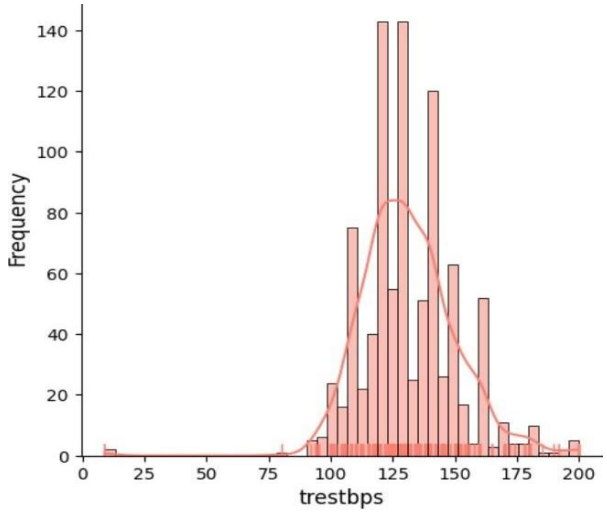

Figure 3.5: Resting Blood Pressure Distribution Analysis

3.8.5 Distribution of Serum Cholesterol

The distribution of serum cholesterol refers to how cholesterol levels are spread across a group of individuals. It helps identify common and varying levels in the population. Analyzing this distribution provides insights into potential patterns and associations, particularly in conditions like heart disease. Visual representations, such as histograms, display the frequency and range of cholesterol values.

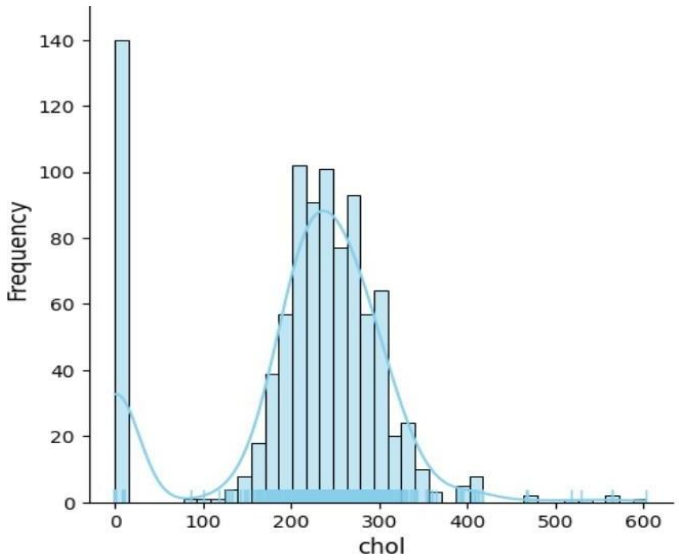

Figure 3.6: Serum Cholesterol Distribution Analysis

3.8.6 Distribution of Maximum Heart Rate Achieved

The distribution of maximum heart rate achieved in a heart disease dataset shows how the highest heart rate in graphic. It helps us understand the range and how often different maximum heart rates occur in the dataset, giving insights into cardiovascular health and fitness levels.

3.8.7 Distribution of Gender

In Figure below, it is displayed that the data of approximately 27.7% females and 72.3% males are used in the dataset. The quantity of males is greater than the number of females.

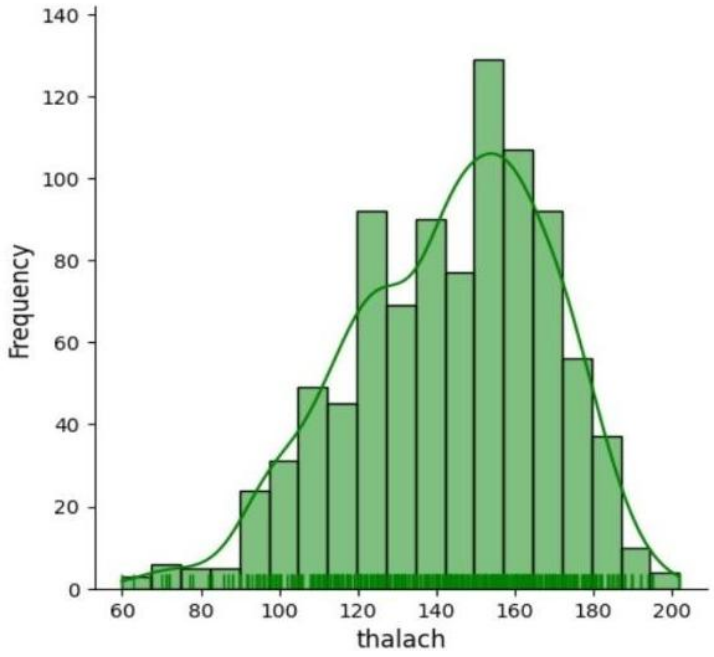

Figure 3.7: Maximum Heart Rate Distribution Analysis

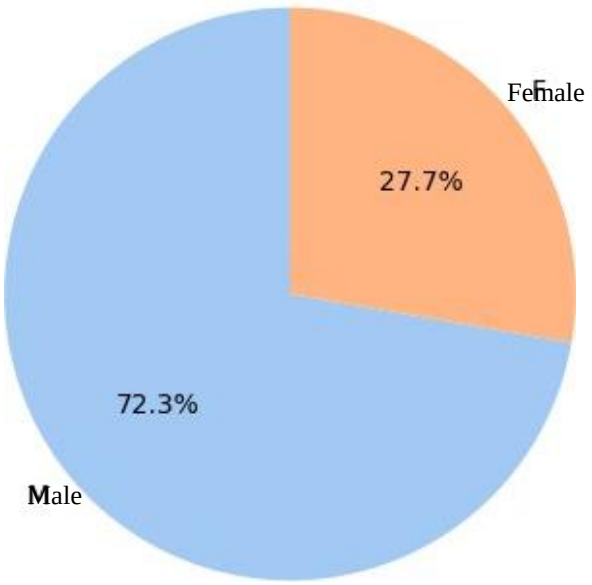

Figure 3.8: Gender Distribution Analyses

3.8.8 Distribution of Target

The distribution of the "target" variable, where 50.3% individuals have the disease (yes) and 49.7% do not have the disease (no), describes how the presence of the heart disease or absence of the disease is distributed in a dataset.

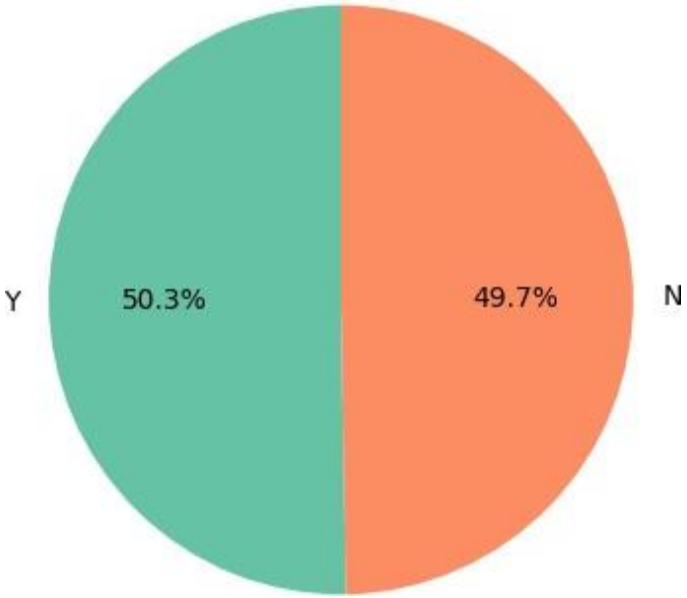

Figure 3.9: Target Distribution Analysis

3.8.9 Distribution of Chest Pain

Chest pain is a discomfort in the chest area that October signal various health issues, including heart-related concerns or digestive problems. Seeking medical help is important to identify the cause and receive proper care.

Value 1. Typical Angina: Classic Symptoms of Heart-Related Chest Pain

Value 2. Atypical Angina: Unusual Characteristics in Chest Discomfort

Value 3. Non-Anginal Pain: Chest Discomfort Unrelated to Heart Issues

Value 4. Asymptomatic: Absence of Noticeable Heart-Related Symptoms occur

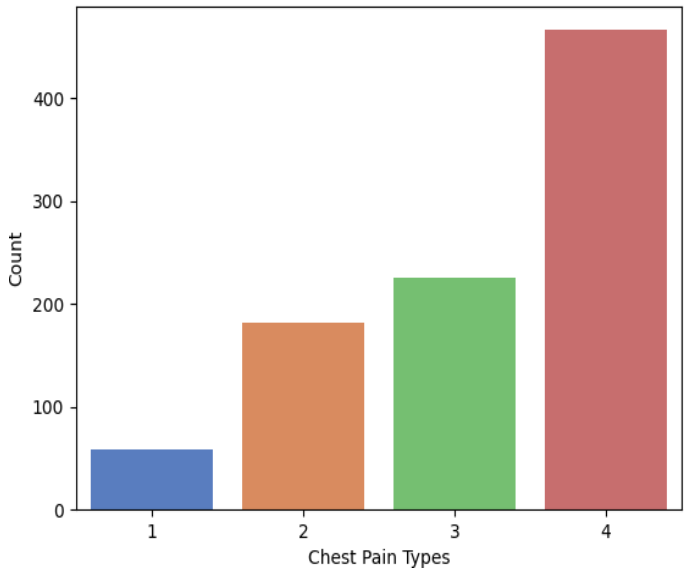

Figure 3.10: Chest Pain Distribution Analysis

3.8.10 Distribution of Resting Electrocardiographic

The distribution of resting electrocardiographic (ECG) results in a heart disease dataset shows how frequently various heart-related findings occur among individuals. ECG is a test that records the heart's electrical activity at rest, and the results are categorized numerically (e.g., 0, 1, 2, 9), each representing different patterns or abnormalities.

Category 0. Normal ECG.

Category 1. Abnormality not related to a heart condition.

Category 2. Possible or definite heart-related abnormality.

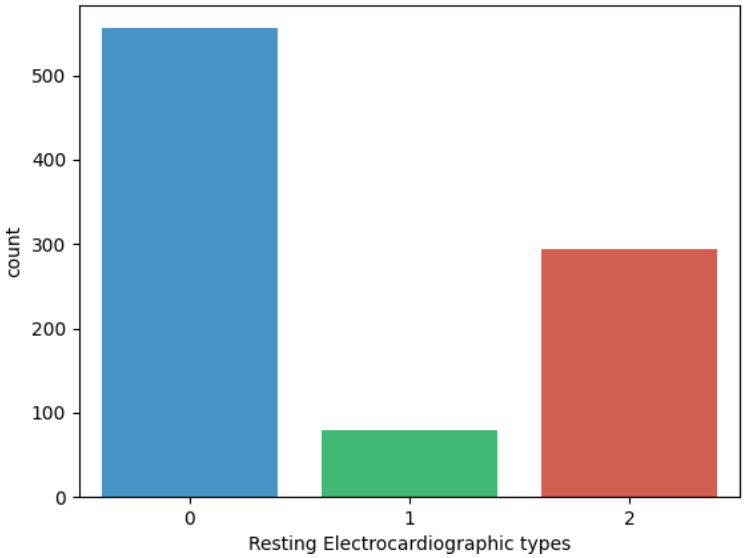

Figure 3.11: Resting Electrocardiographic Distribution Analysis

3.8.11 Distribution of Slope

The slop distribution consists of bars representing each unique slope type, with the height of each bar indicating how many occurrences there are in the dataset. Identify which slope types are more or less common.

Flat or No Slope	0 might represent a flat surface or no slope.
Down Slope	1 might indicate a gentle or slight slope.
Upslope Slope	2 might represent a moderate incline.
High Slope	3 might be used to represent a high or severe slope

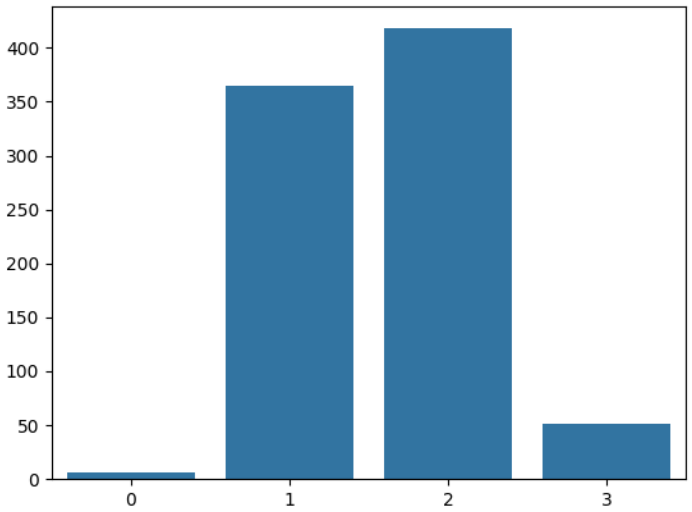

Figure 3.12: Slope Distribution Analyses

3.8.12 Distribution of Numbers of Major Vessels

The distribution of major vessels in a heart disease dataset shows how many people have different numbers of these important blood vessels. It helps us see how common or uncommon each vessel count is and understand its connection to heart-related issues.

Indicates that none of the major blood vessels are affected or showabnormalities.

Suggests that one major blood vessel is affected or exhibits some form ofpathology.

Implies that two major blood vessels are involved or show signs of disease.

Indicates that three major blood vessels are affected in sever condition.

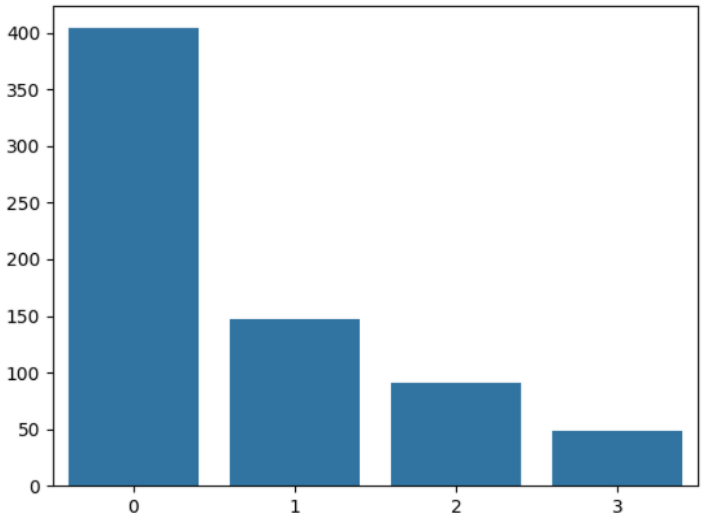

Figure 3.13: Number of Major Vessels Distribution Analysis

3.8.13 Distribution of Thalassemia

The distribution of Thalassemia in a heart disease dataset represents how frequently different types of thalassemia are observed among individuals. Thalassemia is a genetic blood disorder affecting hemoglobin production. Analyzing this distribution helps understand the prevalence of various thalassemia types in the dataset and their potential associations with heart disease.

Thal 3 (Normal) Indicates normal blood flow to the heart muscle.

Thal 6 (Fixed Defect) suggests a permanent lack of blood flow, possibly due to scartissue.

Thal 9 (Reversible Defect) indicates a temporary reduction in blood flow during stress,which returns to normal at rest, suggesting potential ischemia.

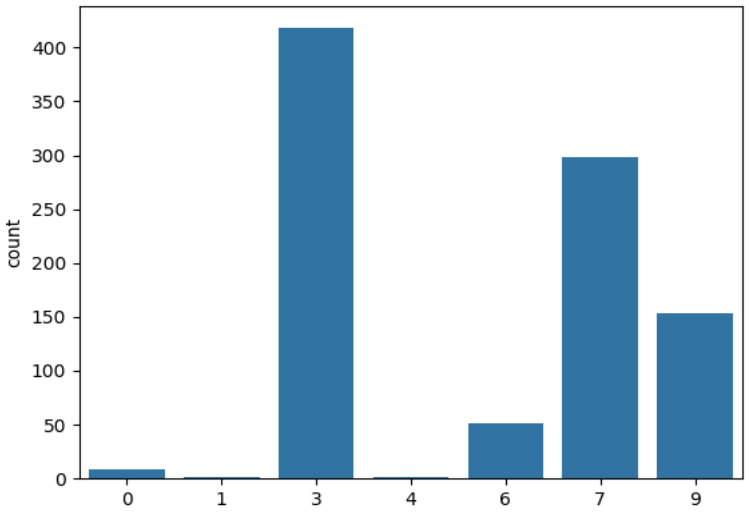

Figure 3.14: Thalassemia Distribution Analyses

3.8.14 Distribution of Exercise Induced Angina

The distribution of Exercise Induced Angina in a heart disease dataset, with 33.4% having "no" and 66.6% having "yes," describes how frequently the occurrence of angina during exercise is observed among individuals. This distribution provides insights into the prevalence of exercise-induced angina in the dataset and its potential association with heart disease.

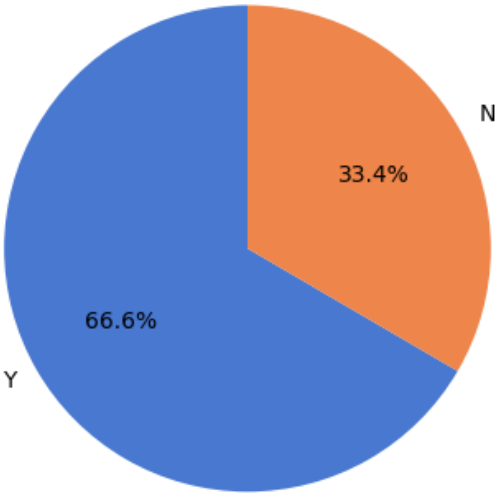

Figure 3.15: Exercise Induced Angina Distribution Analysis

3.8.15 Feature Analysis with Respect to Target Variable

Analyzing characteristics in relation to the target variable compared to gender in a dataset, where 0 stands for female and 1 stands for male involves examining how different features relate to the target variable (outcome of interest) considering the gender of individuals. According to the below figure, it is seen that the rate of heart disease in females is less than the rate of heart disease in males. Approximately 70 females have heart disease out of a total of 300 females. The total number of males in the dataset was 700, in the 700 males list, 300 males are healthy and 400 have heart disease Further, the dark brown line bar represents healthy persons and the lite brown bar represents heart disease persons.

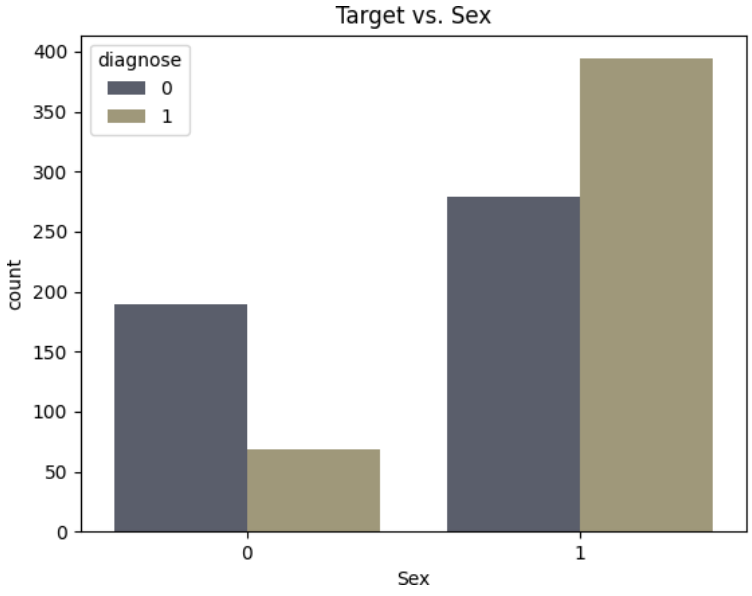

Figure 3.16: Sex Feature Analyses with Respect to Target Variable

The following figure shows the visualization of Chest Pain relation with Heart Disease, thevalues for chest pain represents.

- Typical Angina (Value 1) Chest pain that is expected and characteristic of heart-relatedissues.
- Atypical Angina (Value 2) Chest pain that deviates from the usual or expected patternsrelated to heart issues.
- Non-Anginal Pain (Value 3) Chest discomfort that is not associated with typical anginapatterns.
- Asymptomatic (Value 4) Chest pain or discomfort reported of heart related issue occur.

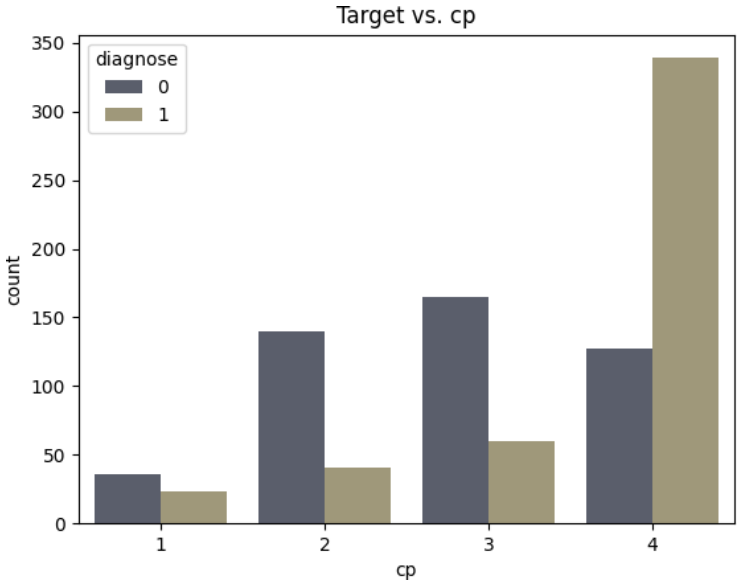

Figure 3.17: Chest Pain Feature Analysis with respect to Target Variable

In Figure below, heart disease frequency based on Fasting Blood Sugar (FBS) is presented. Dark brown line means no heart disease and the lite brown bar line means to have heart disease. 0 represents a false (no FBS) issue and 1 represents a true (have FBS issue) in the below figure. The people who have no FBS issue have more heart disease. So, it is not necessary that people who have high blood sugar also have heart disease.

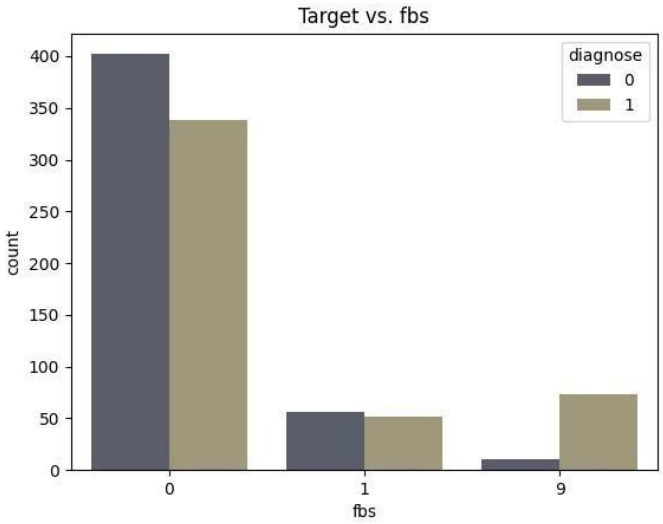

Figure 3.18: Fasting Blood Sugar Feature Analysis with respect to Target Variable

In the dataset, the data of different persons of different ages is collected. In Figure below, heart disease frequency analysis w.r.t ages is performed. People with older ages have more heart diseases as compared to smaller ages.

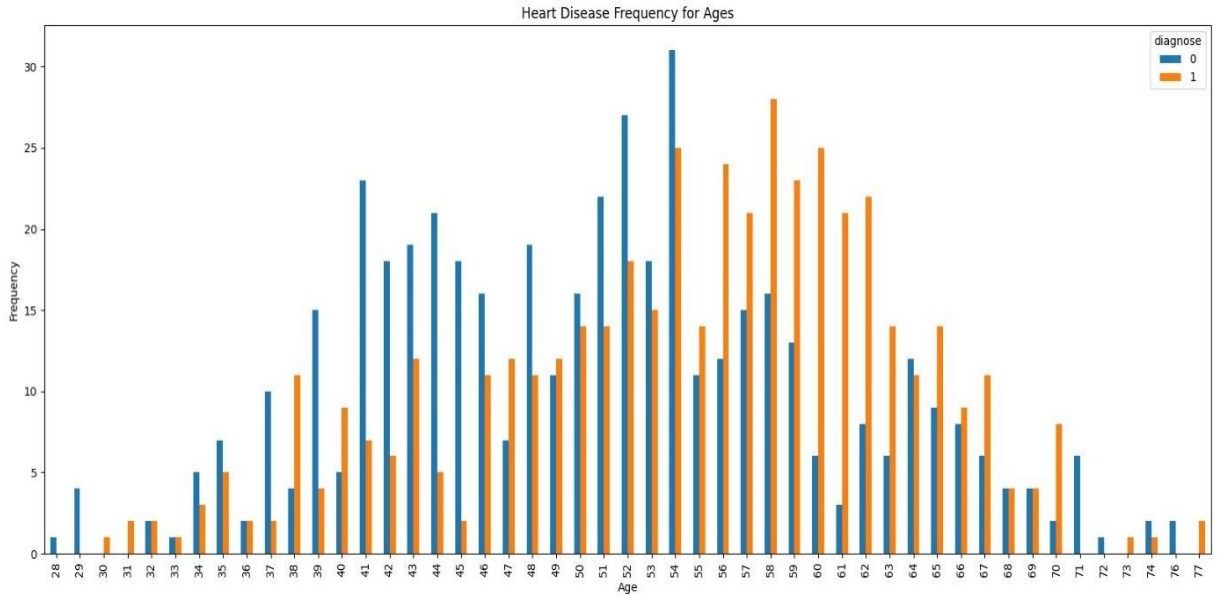

Figure 3.19: Ages Feature Analysis with respect to Target Variable

3.8.16 Heatmaps

Heatmaps visualize the correlation matrix between features, with brighter colors indicating stronger correlations. Heatmaps help identify clusters of highly correlated features and assess their relationship with the target variable.

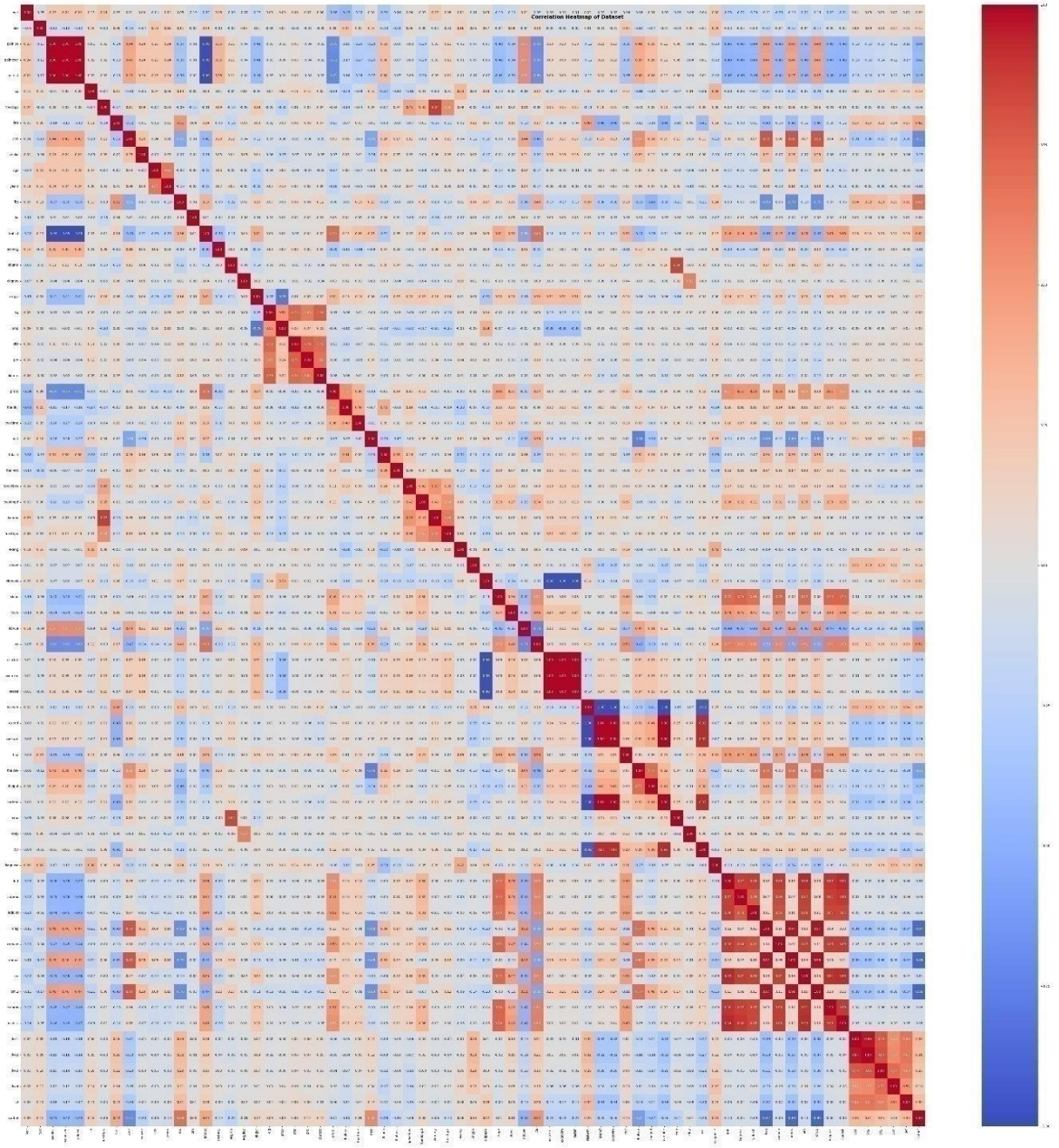

Figure 3.20: Heart Maps Visualize the Correlation Matrix Analysis

3.9 Feature Analysis and Selection

Feature analysis and important feature section is crucial steps in the machine learning pipeline that involve understanding the characteristics of features and identifying the most relevant ones for predictive modeling.

Feature description involves providing detailed information about each feature in the dataset, including its name, data type, and potential relevance to the target variable. In a heart disease prediction task, the features consist of age, gender, blood pressure, and cholesterol

levels. And presence of other medical conditions.

3.10 Feature Importance

Feature importance analysis aims to determine which features have the most important impact on predicting the target variable.

3.11 Feature Extraction

As the study for predicting of heart diseases is getting an increased attention so there is also an increase in the combination of algorithms and features explored. This section lay down the common feature extraction techniques are used in this system. Also call them as word embedding techniques.

The predictive models for heart disease, the quality and selection of features (variables like age, cholesterol level, blood pressure, etc.) are crucial for accurate predictions. These techniques commonly used for feature extraction are Mutual information (MI) Principal Component Analysis (PCA) and the Chi-Square test.

In this study, utilized the PCA, Mutual information and Chi-square to extract features from the dataset. The feature extraction process involves converting the raw text into numerical representations, often referred to as word embedding or features, which can be processed by the respective models.

3.12 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method that transforms a large set of numerical variables into a smaller set that still contains most of the information from the original data. This is especially useful when we have many features (such as patient data with multiple medical measurements) that are highly correlated or redundant [4].

In heart disease prediction, datasets often contain many numerical features that might be correlated. For instance, features like systolic and diastolic blood pressure, cholesterol levels, and age could be interrelated. Using all these variables as-is might make the predictive model complex and prone to over fitting. PCA helps to reduce dimensionality by finding a smaller set of principal components that capture the essential patterns of the data [2] [3].

Standardizing the Data

Before applying PCA, the numerical features (like cholesterol, blood pressure, age) need to be standardized, meaning that all the variables are scaled to have a mean of 0 and a standard deviation of 1. This ensures that features with different units (e.g., cholesterol in mg/dL and age in years) are treated equally in PCA.

Finding Principal Components

PCA identifies principal components—new features that are linear combinations of the original variables. These components are chosen based on the amount of variance they explain in the data.

- The first principal component explains the largest amount of variance in the dataset, essentially summarizing the most important patterns.
- The second principal component explains the next largest amount of variance but is orthogonal (uncorrelated) to the first one, meaning it captures new information.

Dimensionality Reduction

After finding the principal components, we can reduce the dataset's dimensions by keeping only the components that explain most of the variance. For instance, instead of working with 10 original features, we might only keep 3 or 4 principal components that summarize most of the information about heart disease risk.

Using Principal Components for Prediction

Once PCA has been applied, the transformed data (principal components) is used as input for a machine learning model (such as logistic regression, decision trees, or neural networks) to predict heart disease. This process reduces redundancy and multicollinearity in the data, making the model simpler and often more accurate.

3.13 Chi-Square Test

Chi-Square Test The Chi-Square (χ^2) test is a statistical method used to determine whether there is a significant relationship between a categorical feature and the target variable (in this case, the presence or absence of heart disease). This test is commonly used in feature selection when dealing with categorical data [41].

In heart disease prediction, some features are categorical, such as gender, chest pain type, or exercise-induced angina. The Chi-Square test helps identify which of these features are most relevant to predicting whether a patient has heart disease.

Setting Up the Hypotheses

- **Null Hypothesis (H_0)**

  The feature (e.g., gender) and the target (heart disease or not) are independent, meaning there is no association between them.

- **Alternative Hypothesis (H_1)**

  There is an association between the feature and the target, meaning the feature might be important for predicting heart disease.

Observed vs. Expected Frequencies

The test looks at the observed frequency of different categories of the feature (e.g., the number of males and females who have heart disease) and compares them with the expected frequencies if the feature and the target were independent.

For example, if there's no association between **gender** and heart disease, the number of males and females with heart disease should be proportional to the overall distribution of males and females in the dataset.

Calculating the Chi-Square Statistic:

The Chi-Square statistic measures the difference between the observed and expected frequencies. A large difference indicates that the feature and the target are likely related.

Mathematically, the formula for the Chi-Square statistic is:

$$\chi 2 = \sum \frac{(O-E)^2}{E}$$

Where O is the observed frequency, and E is the expected frequency.

If the calculated statistic is larger than a certain threshold, it suggests that the feature is significantly related to heart disease.

Selecting Features Based on Chi-Square Values

Once the Chi-Square statistics are calculated for each categorical feature (e.g., gender, chest pain type, smoking status), those with the highest Chi-Square values are considered the most relevant features for predicting heart disease.

These selected features are then used in building the predictive model, helping to eliminate irrelevant or less important variables.

Using **PCA and Chi-Square Together in Heart Disease Prediction**

In practice, PCA and the Chi-Square test are often used together to handle both numerical and categorical features:

PCA: Applied to numerical features like age, cholesterol, and blood pressure to reduce dimensionality and avoid multicollinearity.

Chi-Square: Applied to categorical features like gender, chest pain type, and smoking status to select the most important features. By combining these techniques, you can build a more efficient and accurate predictive model for heart disease. PCA reduces complexity and captures essential numerical patterns, while Chi-Square helps identify the most relevant categorical variables, making the model both interpretable and powerful in predicting heart disease outcomes.

3.14 Feature Selection through Mutual Information

Mutual information is a statistical tool that helps measure the amount of information gained about one variable by observing another variable. It's used to understand how much knowing the value of one variable can reduce uncertainty about the other variable. In simpler terms, mutual information tells us how much one thing can tell us about another.

Feature selection through mutual information in the context of heart disease involves using the statistical measure of mutual information to identify which features or risk factors are most informative for predicting the presence or risk of heart disease. Researchers can analyze various patient characteristics, such as age, gender, blood pressure, cholesterol levels, and other relevant factors, to determine their mutual information with the occurrence of heart disease. Features with higher mutual information values are considered more informative and October be selected for inclusion in predictive models or further analysis to better understand

andmanage heart disease.

3.15 Calculation of Mutual Information

Mutual information between two variables X and Y is calculated using the following formula

MI(x, y) = H(x) +H(y) –H(x)(y)

Where

p(x, y) is the probability distribution of variables X and Y.

p(x) and p(y) probability distributions of variables X and Y,

respectively.

The mutual information score goes from 0 to positive infinity. A higher mutual information score specify a stronger relationship between the variables, with 1 indicating no relationship and higher values indicating stronger dependencies [64].

3.16 Interpretation of Mutual Information

Mutual information measures both linear and nonlinear relationships between variables. It is particularly useful for identifying nonlinear dependencies that October not be captured by correlation measures. A high mutual information score indicates that the feature provides valuable information about the target variable and is therefore likely to be relevant for predictive modeling.

3.17 Method of Mutual Information in Feature Selection

The method of Mutual Information in Feature Selection involves evaluate the statistical dependence between each feature and the target variable. This is done by calculating the mutual information score, which measures the amount of information that one variable (feature) contains about another (target). Features with high mutual information scores are considered relevant to the target variable and are selected for further analysis or model training. This method helps in identifying the most informative features for predictive modeling while reducing the dimensionality of the dataset.

Below are the scores of features calculated through the mutual information.

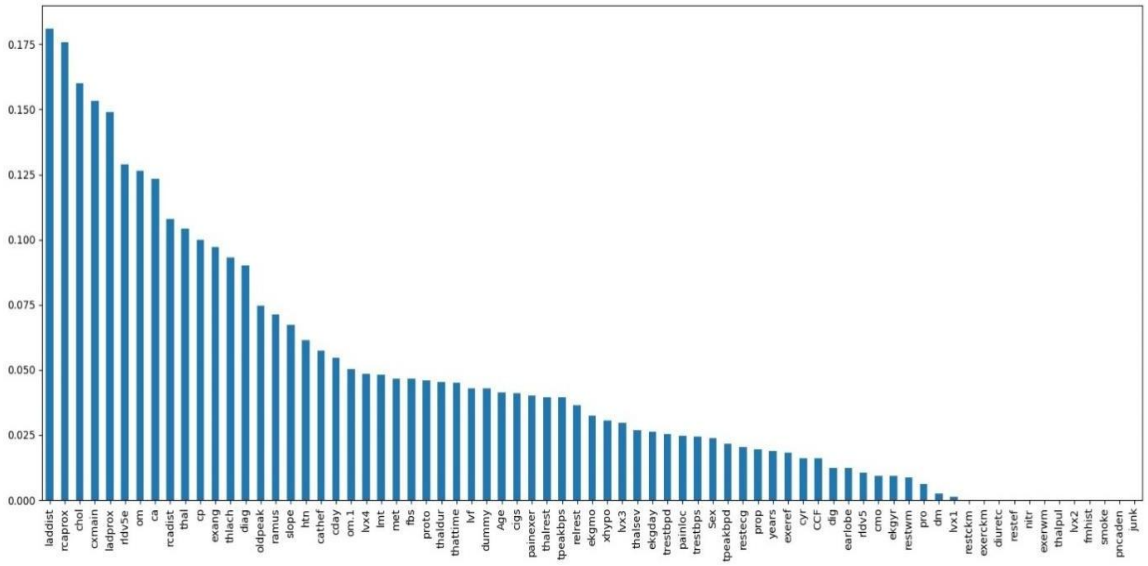

Figure 3.21: Feature Selections Method

3.18 Advantages and Limitations of Mutual

Advantages

- Captures both linear and nonlinear dependencies between variables.
- Not easily affected by strange or incorrect data in the information.
- Does not assume any specific distribution of the data.

Limitations

- Takes a lot of time and resources to process when dealing with big datasets with lots ofdifferent things to look at.
- October overestimate the importance of redundant or irrelevant features.
- Requires careful tuning of parameters such as bin size for discretization.

3.19 Proposed Stacking Model in Machine learning

3.19.1 Introduction

Stacking is one of the most used and best-performing ensemble techniques used in the field of machine learning. Stacking is one of the most used and best-performing ensemble techniques used in the field of machine learning. It is very similar to the stacking ensembles but also assigns the weights to the machine learning algorithms, where two layers of models

are present ground models and Meta models. Due to this, Stacking tends to perform best of all the other ensemble techniques used in machine learning [80].

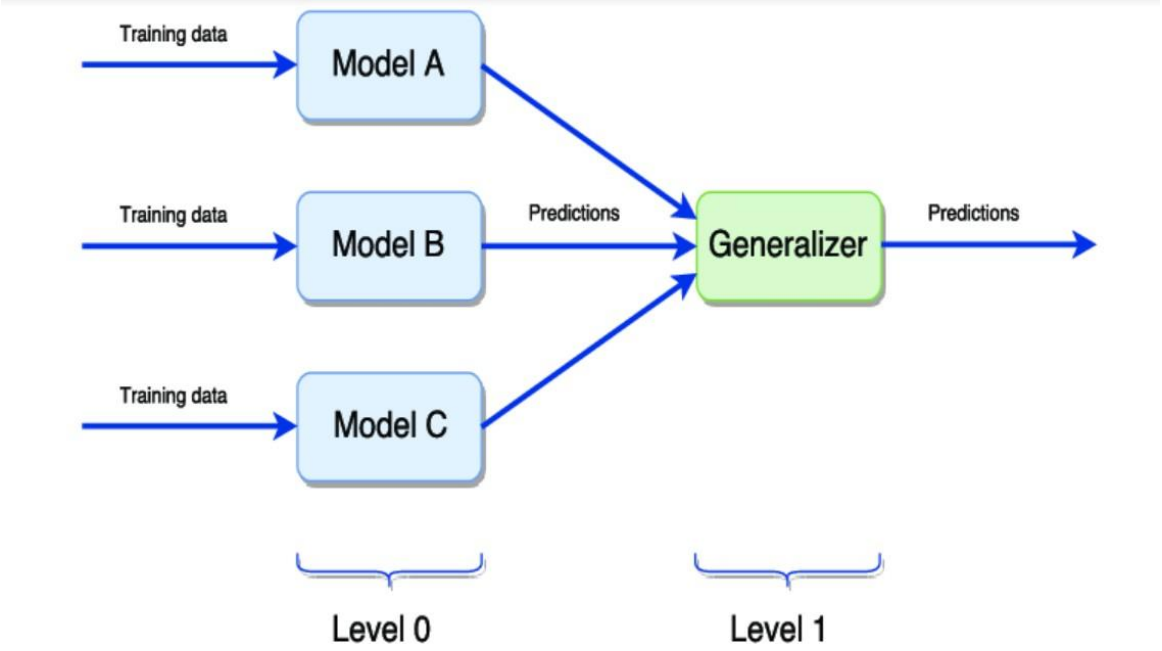

Figure 3.22: Stacking Model Methodology [49]

In Stacking, multiple machine learning algorithms are used as the ground models, but here there is also a further layer of the model called meta models. This model assigns differentweights to the ground models, and then the prediction task is performed in stacking.

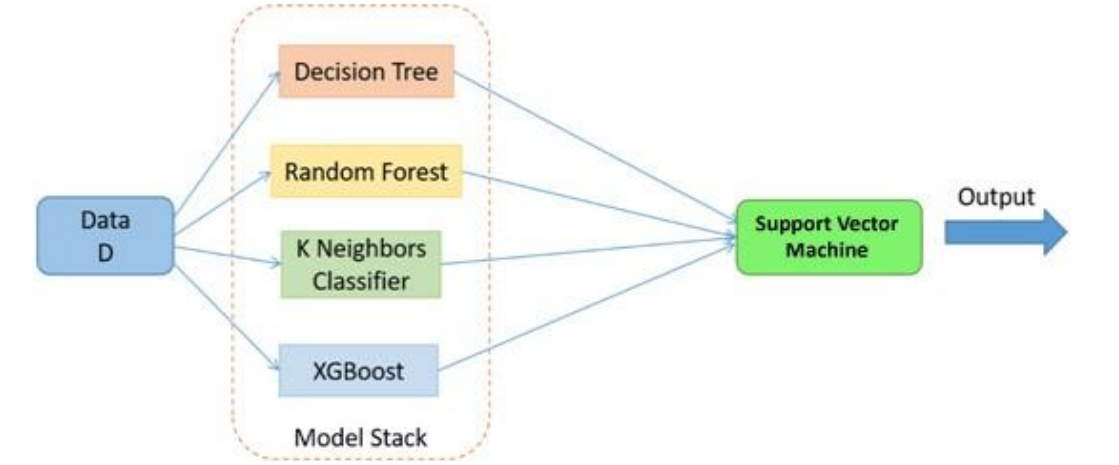

Figure 3.23: The Propose of Stacking Model

Suppose we have dataset D, and we have three machine learning ground models Decision Tree, Random Forest, KNN, and the XGBoost, and the meta-model in the second layer is the Logistic Regression. Now we are feeding the dataset D to every respective ground model. We train the ground models on the same dataset, and one trained model is able to predict for our test dataset. Once we introduce the ground models, then we take the prediction data of every ground model and use that data to train the meta-model Decision tree. So here, the training data for the decision tree is different. Once we introduce the meta-models, it assigns the weights to the ground models, and the output from the meta-models has been considered the final output of the stacking algorithm. [81].

The meta-model, typically a simpler machine learning algorithm such as logistic regression learns to combine the predictions of the base models to generate the final prediction for heart disease risk. Through a process of cross-validation or validation on a separate dataset, the Stacking Model is optimized to maximize predictive accuracy while minimizing over fitting.

In the methodology chapter, we discussed the rationale behind selecting and training the base models, emphasizing their complementary strengths in capturing different aspects of the data. Additionally, we outlined the methodology for creating the stacking dataset, training the meta-model, and evaluating the performance of the Stacking Model [82].

3.19.2 Training Process

Stacking model technique is employed for training the model. Stacking, or stacked generalization, is a method in machine learning where we combine the predictions of many different models to make better predictions overall. It utilizes the variety of individual base models to construct a stronger and more accurate predictive model. Stacking is particularly effective when individual base models have different power and defect, as it can use the complementary nature of their predictions.

The basic idea behind stacking is to train a meta-model, also known as a blender or a combiner, on the predictions generated by a set of diverse base models. Instead of relying on a single model's predictions, stacking aggregates the predictions from multiple models to make final predictions. This approach often leads to better generalization and higher predictive accuracy compared to any single base model.

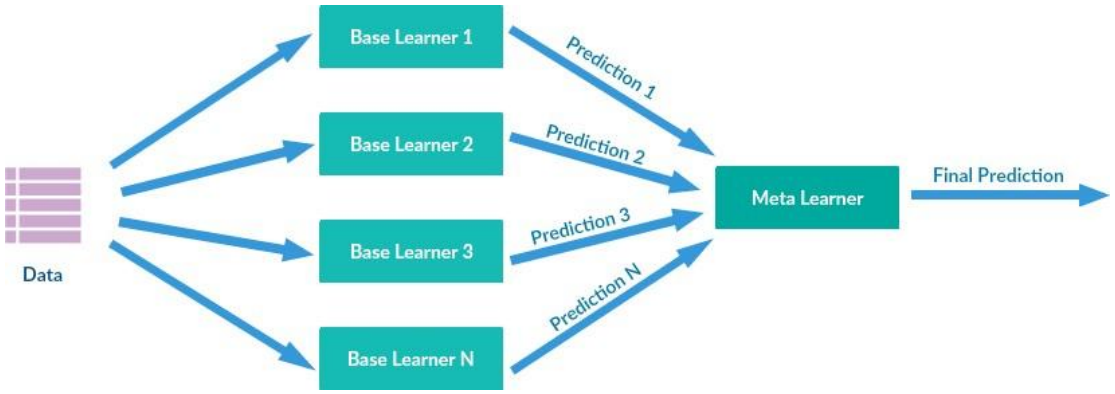

Figure 3.24: The Process of Stacking Model [25]

3.19.3 Combinations of Base Estimators

In the stacking approach, multiple combinations predictions of base estimators are tried to find a stronger predictive model. To do this, different combinations of base estimators are created and the data was trained on it. Each combination consists of three or more base estimators, and each combination is trained separately to generate predictions.

The process of creating combinations of base estimators involves systematically selecting a subset of base estimators from the pool of available models. Different combinations are explored to leverage the diversity of the base estimators and improve the overall predictive performance of the stacking model.

3.19.4 Meta-Model Selection

In stacking, a meta-model plays a crucial role in combining predictions from multiple base estimators to make final predictions. Let's break down the concept of a meta-model and understand its role in the stacking ensemble.

3.19.5 Exploring the Concept of a Meta-Model

A meta-model acts as a team leader, gathering insights from different experts (base models) and blending their perspectives to make the final decision. Its primary role is to understand and merge the varied views of these base models to enhance prediction accuracy. By learning from the strengths and weaknesses of each base model, the meta-model generates

a more informed prediction. During training, it adjusts its strategy based on the collective wisdom of the base models to enhance predictions. Essentially, the meta-model harnesses the combined knowledge of the base models to deliver more dependable predictions.

A meta-model combines predictions from base models to make a final prediction. It assesses each base model's performance, assigns weights to their predictions, and then combines them to produce the final output. This approach leverages the strengths of each base model and enhances predictive accuracy.

3.20 Procedural Flow

The process was started by identifying the pool of base estimators available for the stacking model. In this case, five different machine learning models are under consideration these are Random Forest, XG Boost, Decision Tree, Support Vector Machine (SVM), and KK Neighbors Classifier.

Next, combinations of three or more base estimators from the pool are generated. All possible combinations are considered, ensuring that each combination includes at least three base estimators. For example, a combination consisting of Random Forest, XG Boost, and Decision Tree, or another combination consisting of XG Boost, Decision Tree, and SVM.

Once all the possible combinations are created, each combination is trained separately using the training data to find out the best combination of models.

After training, the performance of each combination-based stacking model using validation data is evaluated. Metrics such as accuracy, precision, recall, or F1-score are used to see which combination of rules works best for making predictions.

Finally, combination-based stacking model is selected with the highest performance as final model for making predictions on new, unseen data.

By exploring different combinations of base estimators and leveraging their diversity, a more robust and accurate stacking model can be built. Each combination contributes unique insights and perspectives that allow capturing a broader range of patterns and relationships in the data. This approach helps to improve the overall predictive performance of stacking ensemble and create a more reliable model for making predictions.

3.20.1 Selection of Base Models

Base models are individual machine learning models that serve as the building blocks of the stacking ensemble. These models can vary in complexity and diversity, and they are trained on the same dataset.

It is common to select a diverse set of base models to capture different aspects of the data and improve overall predictive performance.

In the stacking approach, different machine learning models were choose to act as the basic building blocks of the predictive model. These models are called base estimators, and they form the foundation upon which the final prediction is built. These models are.

- Random Forest
- XG Boost
- Decision Tree
- Support Vector Machine
- K-Nearest Neighbor

3.20.2 Random Forest (RF)

Random Forest is a type of machine learning tool that creates many decision trees during training. It works by randomly choosing parts of the training data and features to build each tree. This randomness helps prevent the model from fitting too closely to the training data, making it better at making predictions about new data [81]. Each decision tree in the Random Forest is trained separately. Once they're all trained, their predictions are combined to make the final prediction. In regression tasks, this final prediction is usually the average of all the tree predictions, while in classification tasks, it's decided by what most trees predict [83].

One key aspect of Random Forest is its robustness to noise and outliers in the data. This robustness makes it suitable for handling real-world datasets that October contain irregularities. Additionally, Random Forest provides a measure of feature importance, which helps in understanding which features contribute the most to the model's predictions. This feature importance analysis can be valuable for feature selection and understanding the

underlying patterns in the data. Overall, Random Forest is favored for its accuracy, flexibility, and ability to perform well on various types of datasets. Its ensemble approach leverages the diversity of individual decision trees to create a more robust and accurate predictive model compared to single decision trees. [84].

3.20.3 Extreme Gradient Boosting (XGB)

XGBoost, also called Extreme Gradient Boosting, is a popular and effective method for both classification and regression. It's an advanced version of gradient boosting that uses a tree-based model and modern techniques to improve efficiency and reduce over fitting [85].

The main concept of XGBoost is to repeatedly train a series of simple decision trees on the mistakes of the previous trees. Each tree focuses on the data points that were classified incorrectly by the preceding trees. XGBoost introduces regularization terms to the loss function to avoid over fitting and employs a unique method called "gradient boosting with randomized features" to enhance generalization [86].

XGBoost includes additional advanced features like tree pruning, early stopping, and automatic handling of missing values, making it a strong and efficient algorithm. It also has an efficient implementation that can benefit from parallel processing and distributed computing.

XGBoost has many benefits. It is highly accurate, capable of managing large datasets and high-dimensional features, and flexible with various types of data. However, it needs precise adjustment of hyper parameters and can be influenced by noisy data [87].

In conclusion, XGBoost is a well-liked and potent algorithm used across domains like web search, finance, and healthcare. It has excelled in numerous data science competitions, thanks to its outstanding performance, efficient implementation, and user-friendly nature.

3.20.4 Decision Tree (DT)

A popular tool for classification tasks is the decision tree classifier (DTC). It's a type of model that recursively divides the feature space into smaller areas based on simple decision rules. The decision tree identifies features that best distinguish between classes,

creating homogenous groups. It continues this process iteratively, selecting features that provide the best impurity reduction or the most information gain. This continues until a stopping criterion, like maximum tree depth or minimum samples per leaf, is met [88].

After construction, the decision tree predicts the class label of new data points by traversing the tree from root to leaf. Each leaf node represents a class label, and internal nodes represent decision rules based on features. Decision tree classifiers offer interpretability, handle numerical and categorical data, and capture non-linear correlations. However, they can be sensitive to slight data changes and prone to over fitting [89].

Despite these limitations, DTC finds applications in healthcare, finance, and marketing, offering insights into decision-making processes. They can be enhanced by combining with other models like RD and XGB to improve better presentation and reduce over fitting.

Decision Tree is a simple yet effective model that is easy to interpret.

It can capture nonlinear relationships and interactions between features.

Decision trees are computationally efficient and can handle both numerical and categorical data.

Despite its simplicity, decision trees can perform well in many machine learning tasks.

3.20.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a versatile machine learning tool used for classification and regression tasks. It works by finding the best line or boundary between different groups of data points to accurately classify new data.

One of the key features of SVM is its ability to find the optimal hyper plane that separates different classes with the maximum margin. This margin represents the gap between the nearest data points from each class and the hyperplane [90].

SVM is advantageous because it can handle high-dimensional data, is robust to noise and outliers, and has good generalizability to new data. However, it October require

significant computational resources and hyper parameter tuning to achieve optimal performance.

3.20.6 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple machine learning technique used for grouping items or making predictions based on similar items nearby. It works by categorizing a new data point based on its proximity to labeled training data points. Essentially, it looks at the K nearest data points and assigns the new data point the same label as the majority of its closest neighbors in the training dataset [90].

In KNN, the user gets to choose the value of K, which determines how many nearby neighbors to consider. Once K is set, the algorithm finds the K closest neighbors based on their distance from the new data point. For classification tasks, the label assigned to the new data point is the one that occurs most frequently among its nearest neighbors. In regression tasks, the new data point is assigned the average value of the target variable across its closest neighbors [85].

KNN has advantages like being easy to create, dealing with non-linear data, and functioning well with data from multiple classes. However, it has downsides, such as sensitivity to outliers, reliance on the choice of K and the distance measure, and inefficiency in handling high-dimensional data. Overall, KNN is a widely used and practical algorithm suitable for various scenarios, including recommendation systems, image and text classification [91].

3.21 Software and Libraries

Python is a highly popular general-purpose programming language used for both small and large-scale systems. It is known for its simplicity, ease of interpretation, and support for various programming methods, including functional, object-oriented, and procedural programming. Python's efficiency is enhanced by eliminating unnecessary complexities.

Anaconda is free software that simplifies coding in languages like Python or R. It functions on operating systems such as Windows, Linux, and macOS. Globally, people use Anaconda for coding, testing, and machine learning training. It is well-known in the industry and aids in tasks like system building, running tests, and AI training. Anaconda supports

various crucial tasks.

- It manages all imported libraries, their dependencies, and development environments within Anaconda.
- Anaconda supports the creation of machine learning techniques using tools like Tensor Flow and scikit-learn.
- Analyzing and manipulating datasets through tools like NumPy and pandas.
- Anaconda enables users to visualize data results using tools like Matplotlib and Datashader.
- Additionally, Anaconda includes Jupyter notebooks with pre-installed libraries, reducing coding stress and enhancing efficiency.

Python, a highly popular general-purpose programming language used for both small-and large-scale systems, is known for its simplicity, ease of interpretation, and support for various programming methods, including functional, object-oriented, and procedural programming. Its effectiveness is enhanced by eliminating unnecessary complexities.

Numpy is a Python programming library designed to simplify the handling of large datasets, matrices, and multidimensional arrays. It provides users with a range of mathematical operations, making calculations simpler. It's freely available software accessible to everyone.

Pandas, another library written in Python, is dedicated to data analysis. It empowers users to effectively handle and manipulate large volumes of data using a variety of tools and functions.

Scikit-learn is a Python library that focuses on data analysis. It enables users to efficiently manage and manipulate large volumes of data using various tools and functions.

Uploading and accessing a dataset in Anaconda involves a few steps. Anaconda is a software distribution that includes tools for managing environments, libraries, and packages.

In Anaconda Navigator, you can launch Jupyter Notebook or JupyterLab, or you can use Anaconda Prompt to directly launch Python.

3.22 Challenges and Limitations

The process of developing predictive models for heart disease prediction using machine learning techniques is not without its challenges and limitations.

Data Quality

One of the main hurdles in predicting heart disease is verifying the quality of the input data. Healthcare datasets might have missing or wrong information, which can affect how well our predictions work. To make accurate predictions, it's important to clean and fix the data before using it.

Feature Selection

Identifying the most relevant features or risk factors associated with heart disease is crucial for developing accurate predictive models. However, feature selection can be challenging due to the complex and multifaceted nature of heart disease etiology. Selecting unrelated or unnecessary attribute can show to model over fitting and reduced generalization ability

Model Interpretability

While machine learning models offer powerful predictive capabilities, their inherent complexity often compromises interpretability. Interpretable models are essential in healthcare settings to facilitate clinical decision-making and patient communication. Balancing model complexity with interpretability is a significant challenge in developing predictive models for heart disease prediction.

Validation and Evaluation

Ensuring the effectiveness of predictive models requires thorough validation and evaluation to measure their performance accurately, reliability, and clinical utility. Challenges arise in selecting appropriate evaluation metrics, designing robust validation strategies, and establishing benchmark performance thresholds. It's important to have thorough checks to make sure predictive models are safe and work well in real healthcare situations.

Addressing the challenges and limitations in developing predictive models for heart disease prediction requires a comprehensive approach that encompasses data quality assurance, feature selection, model interpretability, generalizability to diverse populations, ethical considerations, and rigorous validation and evaluation protocols. By overcoming these challenges, researchers and healthcare professionals can harness the potential of machine learning to advance the early detection and prevention of heart disease, ultimately improving patient outcomes and population health.

The limitations and challenges of the methodology are discussed and suggest avenues for future research to further enhance the accuracy and applicability of heart disease prediction models.

The methodology provides a comprehensive overview of the steps undertaken, key findings obtained, and implications for heart disease prediction. By summarizing approach and awareness.

It was aimed to contribute to the advancement of predictive modeling in healthcare and the improvement of heart disease risk assessment and management strategies.

3.23 Performance Evaluation Metrics Analysis

In this part, a closer focus is given at the numbers and measurements that help understand how well the models are doing. These evaluation metrics function as tools to measure the performance of machine learning models in predicting heart disease.

The Significance of Evaluation Metrics

3.23.1 Accuracy

Accuracy tells us how often our model's predictions are correct. It's like a scorecard that shows us the percentage of times our model got it right. Accuracy is a simple measure to check how well classification models perform. It measure the percentage of correct predictions out of all the predictions made. It measures the proportion of correctly classified instances among all instances.

$$\textbf{Accuracy} = \frac{\textbf{True Postives+False Positive}}{\textbf{True Postives+False Postives}} \quad \textbf{i}$$

3.23.2 Precision

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It indicates the model's ability to correctly identify positive instances while minimizing false positives.

$$\textbf{Precision} = \frac{\textbf{True Positive}}{\textbf{True Positive + False Positive}} \quad \textbf{ii}$$

3.23.3 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances. It indicates the model's ability to capture all positive instances.

$$\textbf{Recall} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}} \quad \textbf{iii}$$

3.23.4 F1-Score

The F1-Score is a combination of precision and recall, aiming to affect a balance between the two measures. It is helpful when there's an uneven distribution of classes or when both false positives and false negatives are equally significant.

It gives a single number that considers both precision and recall. So, if we want a balanced view of how well the model is doing, the mathematical representation of F1-Score is shown below.

$$\textbf{F1-Score} = 2 \times \frac{\textbf{Precision x Recall}}{\textbf{Precision + Recall}} \quad \textbf{iv}$$

3.24 Confusion Matrix

The confusion matrix is like a detailed report card. It shows exactly where the model got it right and where it got it wrong. It can be seen how many true positives, true negatives, false positives, and false negatives the model had, giving a clear picture of its performance.

A confusion matrix shows how many correct and incorrect and type of heart disease predictions a classificationmodel makes, arranged by the actual and predicted classes. It helps understand how well the model performs.

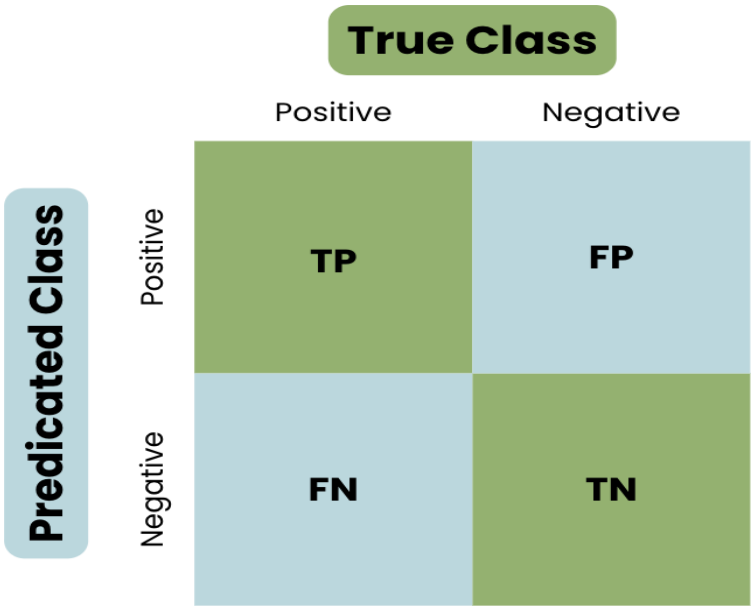

Figure 3.25: The Confusion Matrix Consists of a Grid

The confusion matrix consists of a grid that categorizes predictions made by the model into four categories based on their correspondence to the actual class labels

True Positive (TP) This refers to situation where the model accurately predicts positive cases. In the context of heart disease prediction, a true positive would occur when the model correctly identifies a patient as having heart disease, and the patient indeed has heart disease.

False Positive (FP) This occurs when the model incorrectly predicts positive cases. In simpler terms, it means the model see a patient has heart disease, but in truth, they don't actually have it. False positives are also known as Type I errors.

True Negative (TN) This describe situation where the model accurately predicts negative cases. In the context of heart disease prediction, a true negative would occur when the model correctly identifies a patient as not having heart disease, and the patient indeed does not have heart disease.

False Negative (FN) This occurs when the model incorrectly predicts negative cases. In other

words, the model predicts that a patient does not have heart disease, but in reality, the patient does have heart disease. False negatives are also known as Type II errors. By categorizing predictions into these four groups, the confusion matrix gives a clear and detailed view of the model's performance.

Model Adjustment for Multi-class Classification

The modifying or configuring the machine learning model so that it can predict more than two classes, especially when working with datasets involving multiple types of heart diseases. This is crucial when two task involves not just predicting the presence or absence of disease, but also distinguishing between different types of cardiovascular diseases, such as Coronary Artery Disease (CAD), Heart Failure (HF), Arrhythmia, etc.

1. Dataset Preparation for Multi-class Classification

To train the model for multi-class classification, a dataset where the target variable (labels) contains multiple classes. Where different types of heart diseases, the labels should represent those types, such as:

Class 0: Coronary Artery Disease (CAD)

Class 1: Heart Failure (HF)

Class 2: Arrhythmia

2. Description of the Confusion Matrix Graph

Figure 4.3 presents the confusion matrix for the classification of different types of heart diseases: Coronary Artery Disease (CAD), Heart Failure (HF), and Arrhythmia. The matrix is structured such that rows correspond to the actual labels and columns correspond to the predicted labels.

- True Positives are represented along the diagonal of the matrix. For instance, the model correctly classified instances of Coronary Artery Disease (CAD),
- False Positives are represented by the off-diagonal values in each row. There were some cases where the model incorrectly predicted Heart Failure when the actual condition was Coronary Artery Disease.

- False Negatives are represented by off-diagonal values in the columns. There were cases incorrectly predicted as Coronary Artery Disease when the actual diagnosis was Arrhythmia.

3.25 Summary

In this section, comprehensive summary of the methodology employed in heart disease prediction study is provided, highlighting key steps, findings, insights, and implications for heart disease prediction. Methodological steps followed in this study are summarized, starting with data acquisition and preprocessing, including data collection, cleaning, and feature engineering. The EDA helps understand the dataset by exploring its different aspects. Followed by feature selection, model training using machine learning algorithms such as XGBoost and Random Forest, and evaluation of model performance was explored. Additionally, the challenges encountered and strategies employed to address them throughoutthe methodology are discussed.

CHAPTER 4

RESULTS AND ANALYSIS

This chapter describes the implementation procedure and evaluation of outcomes. Traditional machine learning (ML) algorithms and statistical methods were employed. Specifically, we utilized machine learning algorithms such as Naive Bayes, Support Vector Machine, and Logistic Regression, alongside the statistical method. Our data analysis approach involved splitting the dataset into two parts, with 80% allocated for training our models and 20% for testing their performance. This allowed us to estimate how effectively our models generalized to unseen data. Overall, the chapter provides a detailed of our implementation strategy, evaluation process, and the various methods employed to analyze the data.

Initially, we discussed the datasets of the heart disease and how we imported them into our study. Following that, we detailed the steps involved in preprocessing and analyzing the data. This included converting text into numerical representations and extracting important features for further analysis. We then proceeded to select and implement a suitable model for our study. Finally, we concluded by examining and discussing the results obtained from our analysis.

4.1 Feature Extraction Result

4.1.1 Mutual Information

Feature extraction methods in machine learning change raw data into more helpful features for training. Mutual Information (MI) is a technique that measures how much one feature can tell us about another. It's useful for choosing the best features to make our models work better and simpler.

Table 4.1: Data Attributes

Sr.No.	Attribute	Description
1.	ID	A unique number is given to each patient
2.	AGE	Age of the person
3.	Sex	Male or Female
4.	Origin	Place
5.	Cp	Chest pain 0: Normal 1: anginal pain 2: anginal pain (atypical) 3: non-anginal pain 4: asymptomatic
6.	Trestbps	Rate of blood pressure
7.	FBS	Sugar in blood (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
8.	Chol	Serum cholesterol (mg)
9.	Restecg	0= normal 1= having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2= showing probable or definite left ventricular hypertrophy by Estes' criteria
10.	Ca	number of major vessels (0-3) colored by fluoroscopy
11.	Exang	exercise-induced angina (1 = yes; 0 = no)
12.	Oldpeak	ST depression induced by exercise relative to rest
13.	Slope	slope: the slope of the ST segment's maximum workout 3: down-slopping 2: flast, 1: up-slopping
14.	Thalach	Heart rate (maximum)
15.	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
16.	Target	Targeted value 0= healthy, 1= heart disease patient

	CCF	Age	Sex	painloc	painexer	relrest	pncaden	cp	trestbps	htn	...	om.1	rcaprox	rcadist	lvx1	lvx2	lvx3	lvx4	lvf	cathef	junk
0	0	65	1	1	1	1	9	4	115	0	...	1	1	1	1	1	1	1	1	75	9
1	0	32	1	0	0	0	9	1	95	1	...	1	1	1	1	1	1	5	1	63	9
2	0	61	1	1	1	1	9	4	105	0	...	1	2	1	1	1	1	1	1	67	9
3	0	50	1	1	1	1	9	4	145	0	...	1	1	1	1	1	1	5	4	36	9
4	0	57	1	1	1	1	9	4	110	0	...	1	2	1	1	1	1	1	1	60	9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
926	0	55	1	1	0	0	9	2	160	1	...	9	1	1	1	1	1	1	1	9	9
927	0	43	0	1	1	0	9	2	120	0	...	9	1	1	1	1	1	1	1	9	9
928	0	48	1	1	1	1	9	4	160	1	...	9	1	1	1	1	1	1	2	9	9
929	0	54	1	0	0	0	9	1	120	0	...	9	1	1	1	1	1	1	1	9	9
930	0	54	1	1	0	1	9	3	120	0	...	9	1	1	1	1	1	1	1	9	9

931 rows × 74 columns

Figure 4.1: Feature Selections through Mutual information.

4.2 Model Performance

Now that have a good understanding of what these evaluation metrics mean, After this the evaluation of machine learning model's performance in predicting heart disease.

4.2.1 Accuracy

This tells us how often the models are getting their predictions right. A higher accuracy means the models are doing a good job overall. In figure below, the accuracy of the final selected framework is calculated.

```
print("Accuracy:",acc,"%")

Accuracy: 98.4 %
```

Figure 4.2 the Accuracy Score Result

4.2.2 Precision

Precision isabout how precise the models are when they predict someone has heart disease. If precision is high, it means the models are doing a good job of not misclassifying healthy people as having heart disease. The following figure shows the Precision Score of the model.

```
print("Precision Score:",round(precision_score(y_test, y_pred)*100,2),"%")
```

```
Precision Score: 96.63 %
```

Figure 4.3: The Precision Score Result

4.2.3 Recall

Recall tells us how good the models are at finding all the people who actually have heart disease. A high recall means the models are doing a good job of not missing anyone who has heart disease. The figure below, represents the model's Recall Score.

```
print("Recall Score:",round(recall_score(y_test, y_pred)*100,2),"%")
```

```
Recall Score: 98.85 %
```

Figure 4.4: The Recall Score Result

4.2.4 F1-Score

F1-Score acts as a balance between precision and recall. It gives a number that considers both precision and recall. So, if a balanced view of how well the models are doing is required, the F1-Score can be considered. The figure below, shows the F1-Score achieved by the model.

```
print("F1-Score:",round(f1_score(y_test, y_pred)*100,2),"%")
```

```
F1-Score: 97.73 %
```

Figure 4.5: The F1-Score Result

4.3 Confusion Matrix

The confusion matrix gives us a detailed breakdown of where the models are getting it right and where they're getting it wrong. It shows how many true positives, true negatives, false positives, and false negatives the models have.

The figure below, shows how well the model performed and we can see there are only3 misclassified samples by the model out of which 2 were misclassified as ‘Patient have Heart Disease’ and only 1 misclassified as ‘Patient don’t have Heart Disease’.

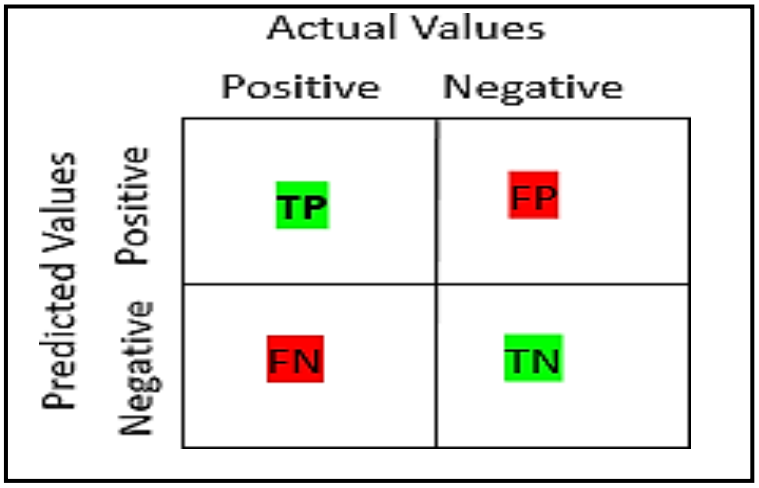

Figure 4.6: Confusion matrix

It offers a clear and simple method of understanding how effectively a model is working to categorize data into different groups or classes. When dealing with binary classification problems—where the objective e is to identify data into one of two groups, such as "positive" or "negative," or "yes" or "no". Confusion matrix is very helpful. A confusion matrix typically consists of four key values:

True Positives (TP): These are situations in which the model accurately predicted the positive category. In other words, the instances that are genuinely positive were correctly identified as positive by the model.

True Negatives (TN): These instances represent situations where the model made correct predictions for the negative class. In this scenario, instances that are truly negative were accurately classified as negative by the model.

False Positives (FP): These are also referred to as Type I errors. They represent situations

where the model made incorrect predictions for the positive class when it should have been negative. In other words, instances that are truly negative were wrongly classified as positive by the model.

False Negatives (FN): These are also known as Type II errors. They occur when the model incorrectly predicts the negative class when the instance should have been positive. In other words, instances that are genuinely positive were wrongly classified as negative by the model.

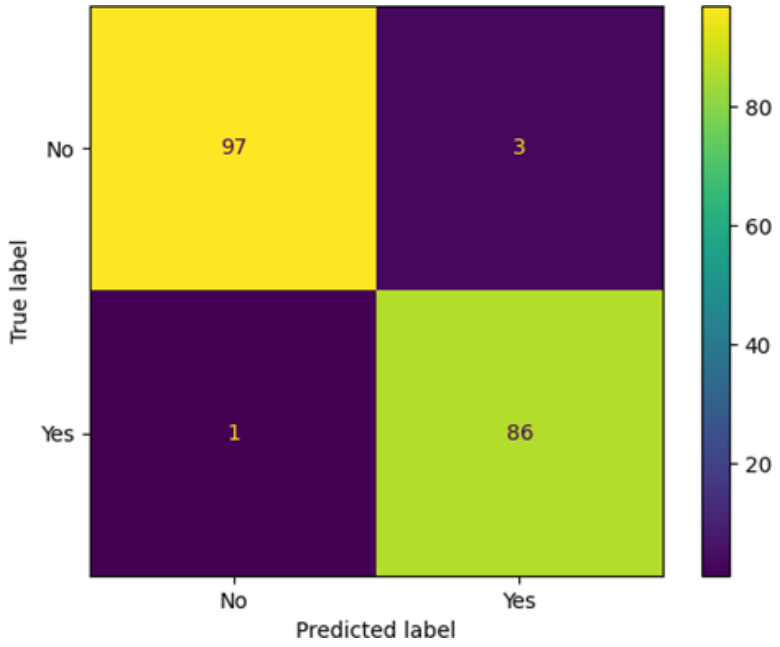

Figure 4.7: The Confusion Matrix Result

4.4 Machine Learning Results and Analysis

In this part, we used both traditional machine learning and statistical methods to set up experiments by adjusting features. We employed machine learning methods like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), XGB, Decision Tree (DT) and Random Forest (RF), along with statistical techniques such as Stacking model we divided our data into 80% for training and 20% for testing.

4.4.1 K-Nearest Neighbor (KKN)

The K-Nearest Neighbor Classification Performance the table below presents the effectiveness of a KNN in categorizing items into four groups. It highlights different evaluation metrics like (accuracy, F1 score, Recall and precision) contributing valuable

insights to the study.

Table 4.2: Result of KNN Classifier

Classifiers	Accuracy	Precisions	Recall	F1-score
KKN	90.56	85.67	88.67	84.11

4.4.2 Support Vector Machine (SVM)

The Support Vector Machine (SVM) Classification Performance the table below presents the effectiveness of a SVM in categorizing items into four groups. It highlights different evaluation metrics like (accuracy, f1 score, recall and precision) contributing valuable insights to the study.

Table 4.3: Result of SVM Classifier

Classifiers	Accuracy	Precisions	Recall	F1-score
SVM	85.09	89.19	78.66	83.59

4.4.3 Decision Tree (DT)

In this table below show the classification performance of and result Decision Tree. The different four columns show different evaluation metrics like (accuracy, f1 score, recall and precision) contributing valuable insights to the study.

Table 4.4: Result of DT Classifier

Classifiers	Accuracy	Precisions	Recall	F1-score
DT	90.65	90.57	90.71	90.64

4.4.4 Random Forest (RF)

The Random Forest (RF) classification performance the table below presents the effectiveness of a RF in categorizing items into four groups. It highlights different evaluation metrics like (accuracy, F1 score, Recall and precision) contributing valuable insights to the study.

Table 4.5: Result of RF Classifier

Classifiers	Accuracy	Precisions	Recall	F1-score
RF	95.35	97.08	90.19	92.80

4.4.5 XG Bosst (XGB)

In this table below show the classification of performance and result XGBosst. The different four columns show different evaluation metrics like (accuracy, f1 score, recall and precision) contributing valuable insights to the study.

Table 4.6: Result of XGB Classifier

Classifiers	Accuracy	Precisions	Recall	F1-score
XGBosst	96.11	94.56	95.66	95.89

4.4.6 Proposed Stacking Model

In this below given table represent the performance of stacking model on heart disease dataset. The model was trained with 80% of the dataset, with the remaining 20% used for testing. The performance evaluation metric (accuracy, f1 score, recall and precision).

Table 4.7: Result of Proposed Stacked Model

Classifiers	Accuracy	Precisions	Recall	F1-score
Proposed Stacked Model	98.4	94.56	95.66	95.89

4.5 Performance Comparison with Existing Studies

Table: 4.8: Comparisons with Existing Studies.

Classifiers	Accuracy	Precisions	Recall	F1-score
SVM	85.09	89.19	78.66	83.59
Decision Tree	90.65	90.57	90.71	90.64
K-Nearest Neighbor	90.56	85.67	88.67	84.11
Random Forest	95.35	97.08	90.19	92.80
XGBoost	96.11	94.56	95.66	95.89
Proposed Stacking Model	98.4	96.63	98.95	97.73

4.6 Visualizations of Model's Learning Progress

To better understand the model's learning progress during training, graphical representations of the training loss and accuracy curves was generated. These curves were plotted over the course of multiple epochs. The training loss curve demonstrated a steady decrease over epochs, signifying that the model was gradually improving its predictive capabilities and minimizing errors during the training process. This reduction in training loss indicates that the model was effectively adjusting its parameters to better fit the

training data. Concurrently, the accuracy curve displayed a consistent upward trend during the initial epochs, indicating an improvement in the model's ability to make correct predictions.

However, it's worth noting that after the 6th epoch, the accuracy curve might have started to flatten or even decrease slightly. This observation aligns with the convergence analysis mentioned earlier, indicating that the model was starting to over fit to the training data beyond the 6th epoch. The visualization of the learning progress through these curves is valuable for understanding how the model was evolving during training, providing insights into potential overfitting issues and the model's overall performance trends.

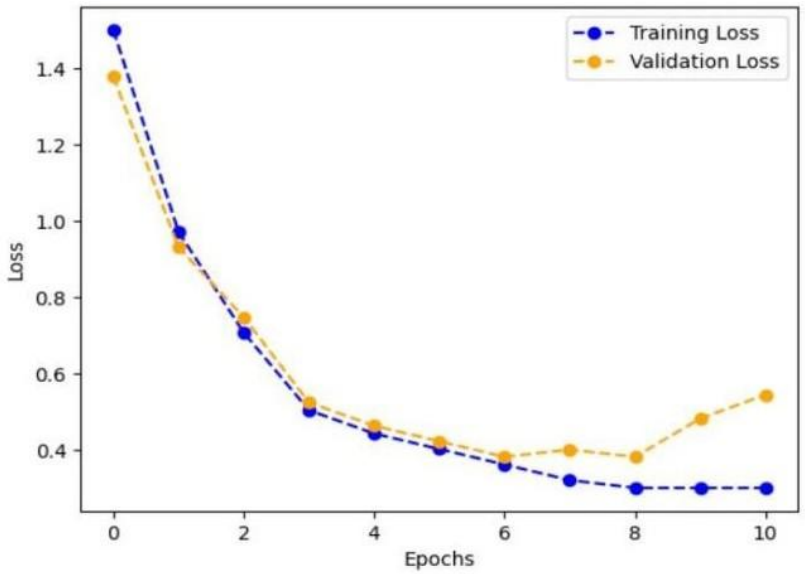

Figure 4.8: Training and Validation Loss

4.7 Discussion on Model Over fitting or Under fitting

Upon analyzing the results to the training performance, signs of over fitting in the model were observed. The slight decline in metrics on the set compared to the training set after the 6th epoch indicates that the model started to over fit to the training data beyond that point. During the training process monitored the training and validation loss trends closely. The observed divergence between the training and validation loss, with the training loss continuing to decrease while the validation loss started to increase slightly after the 6th epoch, is indeed indicative of over fitting.

While it's always beneficial to have explicit bias and variance data to quantify the degree of overfitting, the trends in the loss curves you described provide practical insight into the model's behavior. This divergence suggests that the model is starting to memorize the

training data rather than learning generalizable patterns, which aligns with the definition of overfitting.

To address this issue and prevent further over fitting, an early stopping mechanism is employed during training. This decision is based on convergence analysis, which revealed theover fitting trend. By implementing early stopping, model became capable to halt training when the performance ceased to improve, thereby preventing it from becoming overly specialized to the training data and helping it generalize better to new, unseen data.

Overall, the model's performance on the set provided valuable insights into its generalization capabilities and the need for measures to combat over fitting. The results on the confirmed the effectiveness of the model and its potential for practical applications in sentiment analysis tasks.

4.8 Performance Comparison with Machine Learning Model

Performance comparison with machine learning models involves evaluating and contrasting different algorithms based on metrics such as accuracy, precision, recall, and F1 score. This process helps determine which model best fits a specific dataset or task. Additionally, it include assessing training time, computational efficiency, and robustness to overfitting. Ultimately, the goal is to select the most effective model for optimal predictive performance.

Table 4.9: Performance Comparison with existing models studies and dataset

Model	Year	Data set	Accuracy	Precisions	Recall	F1-score
Hybrid model	2023	UCI data set	90.7%	84.0%	88.1%	85.0%
Transformer model	2024	Kaggledataset	96.51	88.4%	85.2%	88.4%
Proposed Stacking Model	2024	UCI data set	98.4	94.56	95.66	95.89

4.9 Multi-class Classification Result

The multi- class classification result compares various classifiers based on their performance in predicting different types of heart diseases (Coronary Artery Disease, Heart Failure, and Arrhythmia). The metrics included are **Accuracy**, **Precision**, **Recall**, and **F1-**

score.

Table 4.10: Multi-class classification result with Heart diseases

Classifier	Accuracy (%)	CAD Precision (%)	CAD Recall (%)	CAD F1-score (%)	HF Precision (%)	HF Recall (%)	HF F1-score (%)	Arrhythmia Precision (%)	Arrhythmia Recall (%)	Arrhythmia F1-score (%)
SVM	85.09	89.19	78.66	83.59	85.50	80.25	82.75	78.45	75.35	76.85
Decision Tree	90.65	90.57	90.71	90.64	89.40	88.10	88.75	85.25	85.10	85.15
K-Nearest Neighbor	90.56	85.67	88.67	84.11	82.35	81.45	81.90	88.10	87.50	87.80
Random Forest	95.35	97.08	90.19	92.80	95.25	92.50	93.75	92.35	90.85	91.35
XGBoost	96.11	94.56	95.66	95.89	94.25	93.90	94.07	93.80	92.60	93.10
Proposed Stacking Model	98.40	96.63	98.95	97.73	96.85	96.35	96.50	98.10	98.00	98.05

4.10 Overall Performance

The **Proposed Stacking Model** achieves the highest accuracy (98.40%) and maintains high precision, recall, and F1-scores across all heart disease types, indicating that this model is very effective in predicting heart disease.CAD, the Random Forest model has the highest precision (97.08%), which means it is very accurate when it predicts CAD. The Proposed Stacking Model also demonstrates high precision across all categories, suggesting that it makes fewer false positive predictions.

The Proposed Stacking Model also achieves the highest recall for CAD (98.95%) and Arrhythmia (98.00%), indicating that it successfully identifies most cases of these diseases. This is crucial in medical applications where missing a positive case could have severe consequences.

The F1-score provides a balance between precision and recall. The Proposed Stacking Model again excels with high F1-scores across all categories; making it a well-rounded choice for this classification problem. The SVM classifier has the lowest performance across all metrics, suggesting that it October not be suitable for this particular dataset or task. The Decision Tree and K-Nearest Neighbor classifiers also show competitive performance,

particularly in terms of accuracy and F1-scores, though they do not reach the levels achieved by the ensemble models.

4.11 Analysis

4.11.1 Performance of Machine Learning Classifiers

Various machine learning classifiers (Random Forest, XG Boost, Decision Tree, Support Vector Machine, and K-Nearest Neighbor) were employed to predict heart disease.

Metrics such as accuracy, precision, recall, and F1 score were utilized to evaluate the performance of each classifier.

The analysis aimed to identify the most efficient algorithm for heart disease prediction among the tested classifiers.

Stacking models were employed to combine the predictions of multiple base classifiers, potentially improving predictive performance.

4.11.2 Effect of Feature Extraction Methods

Feature extraction methods, particularly Mutual Information (MI), were explored to enhance heart disease prediction accuracy.

The impact of MI on the performance of machine learning models was assessed through comparative analysis.

The effectiveness of MI in extracting relevant features for heart disease prediction was evaluated.

4.11.3 Influence of Different Parameters on Algorithm Performance

Parameters such as the number of features, size of the training dataset, and size of the testing dataset were investigated for their influence on machine learning algorithm performance.

The analysis aimed to identify the optimal combination of parameters for accurate heart disease prediction.

Iterative refinement and improvement of machine learning models were facilitated by

adjusting these parameters based on performance metrics.

4.11.4 Utilization of Stacking Models

Stacking models were employed to combine the predictions of multiple base classifiers, leveraging their individual strengths.

The stacked ensemble model aimed to achieve higher predictive accuracy compared to individual classifiers.

The effectiveness of the stacking approach in improving heart disease prediction performance was evaluated through experimentation and comparative analysis.

Overall, the focused on evaluating the performance of machine learning classifiers, exploring feature extraction methods, assessing the impact of various parameters on algorithm performance, and utilizing stacking models to enhance heart disease prediction accuracy**.**

CHAPTER 5

CONCLUSION

5.1 Introduction

Heart disease remains a leading cause of mortality worldwide, necessitating the development of effective predictive models to enable early diagnosis and intervention. In this thesis, we present a comprehensive framework for heart disease prediction, leveraging the Stacking Model technique to enhance predictive accuracy.

5.2 Utilizing Stacking Model

The Stacking Model is a powerful ensemble learning technique that combines the predictions of multiple base models to improve overall performance. In our framework, we harness the Stacking Model's ability to leverage diverse machine learning algorithms and integrate their predictions effectively.

5.3 Framework Architecture

Our framework comprises several key components, including data preprocessing, feature selection, model training, and evaluation. During the preprocessing phase, we clean and prepare the dataset obtained from the UCI repository, ensuring its suitability for predictive modeling. Feature selection is performed using mutual information analysis to identify the most informative variables for heart disease prediction.

5.4 Model Selection and Integration

For the Stacking Model, we select five diverse machine learning algorithms as base models: Random Forest, XG Boost, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors. Each base model is trained independently on the dataset to capture different aspects of the data and generate individual predictions. These predictions are then combined using

the Stacking Model, with a meta-model (Support Vector Machine in our case) trained on the base models' outputs to make the final prediction.

5.5 Performance Evaluation

The performance of our framework is evaluated using standard metrics such as precision, recall, and F1-score on a separate test set. We compare the performance of the Stacking Model with individual base models and traditional ensemble techniques to assess itseffectiveness in heart disease prediction.

5.6 Key Findings and Achievements

Our results demonstrate the superior performance of the Stacking Model in predicting heart disease risk. The Stacking Model consistently outperforms individual base models and traditional ensemble techniques, achieving high precision, recall, and F1-score metrics. This highlights the efficacy of integrating diverse machine learning algorithms using the Stacking Model approach.

5.7 Contributions to Heart Disease Prediction

The primary contribution of our research lies in the development and validation of a comprehensive framework for heart disease prediction using the Stacking Model technique. By leveraging the Stacking Model's capabilities, we have created a robust predictive model that can accurately identify individuals at risk of heart disease.

5.8 Conclusion

In this heart disease study, researchers uncovered the severity of this condition and its impact on numerous lives. To help diagnose heart diseases early, machine learning models were used. But initially, right information about heart diseases was needed. Dataset about heart diseases was obtained from UCI. It contains info about different things like age, blood pressure, and cholesterol levels.

After preprocessing, it is intended to figure out which pieces of information are most important for predicting heart diseases. To achieve this, a special method called mutual information is used. This method turnout to be the most reliable in extracting features. Stacking Model was used for the prediction of heart diseases. This technique combines the strengths of different machine learning models to make more accurate predictions. Five different machine learning models were selected to be the building blocks of the Stacking Model i.e. Random Forest, XG Boost, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors. Each of these models has its own strengths and weaknesses, so by combining them, it was hoped to create a more reliable predictor for heart disease.

Different combinations of these five models were tried to see which one worked best. Each combination included three or more models, and all of these are trained separately to see which combination gave the most accurate predictions. The Stacking Model, built using a combination of XG Boost and Decision Tree as Estimators with Support Vector Machine as Meta Model, showed the highest accuracy in predicting heart disease.

It had a precision, recall, and F1-score that were all very high i.e. 98%, meaning it was great at both finding people with heart disease and not misclassifying healthy people.

This study demonstrates the effectiveness of using the Stacking Model technique, combined with a diverse set of machine learning models, to predict heart disease accurately. By utilizing the strengths of different models and combining their predictions, they was able to create a powerful tool for identifying individuals at risk of heart disease.

In conclusion, our thesis presents a novel framework for heart disease prediction utilizing the Stacking Model technique. Through comprehensive experimentation and evaluation, we have demonstrated the effectiveness of the Stacking Model in accurately predicting heart disease risk. This research contributes to the ongoing efforts to develop proactive and personalized healthcare interventions for mitigating the burden of heart disease on public health.

5.9 Future work

Looking ahead, future research directions October explore enhancements to our framework, such as incorporating additional features or exploring alternative machine learning algorithms. Additionally, the integration of advanced techniques such as deep learning or

ensemble methods could further improve predictive performance.

Further Training in Feature Engineering and Selection: In future endeavors, more training could be provided in advanced feature engineering and selection techniques to better choose the most relevant features. This could involve the use of advanced methods such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and feature importance techniques.

To enhance the performance of the stacking model, future work could focus on model optimization through hyper parameter tuning and model evaluation techniques. This might involve employing methods like grid search, random search, and cross-validation to improvethe model's accuracy, precision, and recall.

In addition to the stacking model, future work could explore the use of other ensemble techniques such as bagging and boosting to further improve the overall performance and robustness of the model.

Future work might involve the development of a real-time monitoring system that continuously monitors patients' health data and alerts them to potential heart disease risks in a timely manner. This could involve the use of advanced analytics and data streaming techniques.

Future work could involve the use of external datasets to improve the generalization of the model. Additionally, collecting data from diverse populations and regions could make the model more robust.

Validation of the proposed model in clinical settings could be conducted in future work to provide healthcare professionals with reliable and accurate predictions of heart disease risks. Subsequently, the model could be deployed in real-world scenarios to ensure patients receive timely and effective treatment

References

1. World Health Organization. Cardiovascular Diseases (CVDs). Available online: (2023). https://www.afro.who.int/health-topics/cardiovascular-diseases, (accessed on 5 May).

2. Alom, Z. et al. Early Stage Detection of Heart Failure Using Machine Learning Techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning, Cox's Bazar, Bangladesh, 23–25 September 2021.

3. Ouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.

4. Zhang, X., Zhang, Y., Du, X. & Li, B. Application of XGBoost algorithm in clinical prediction of coronary heart disease. Chin. J. Med. Instrum. 43 (1), 12–15 2019.

5. D. Liu, M. Görges, and S. A. Jenkins, "University of Queensland vital signs dataset: development of an accessible repository of anesthesia patient monitoring data for research," Anesthesia & Analgesia, vol. 114, no. 3, pp. 584–589, 2012.

6. B. Akkaya, E. Sener, and C. Gursu, "A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques," in 2022 International Congress on Human- Computer Interaction, Optimization and Robotic Applications (HORA), IEEE, 2022, pp.1–8.

7. J. Digumarthi, V. M. Gayathri, and R. Pitchai, "Early Prediction of Cardiac Arrhythmia using Novel Bio-inspired Algorithms," in 2022 8th International Conference on Smart Structures and Systems (ICS), IEEE, 2022, pp. 01–04.

8. L. Yahaya, N. David Oye, and E. Joshua Garba, "A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques," AJAI, vol. 4, no. 1, p. 20, 2020, doi: 10.11648/j.ajai.20200401.12.

9. A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning," IEEE Access, vol. 11, pp. 61646–61659, 2023, doi: 10.1109/ACCESS.2023.3285247.

10. D. A. Anggoro and N. C. Aziz, "Implementation of K-Nearest Neighbors Algorithm for Predicting Heart Disease Using Python Flask," eijs, pp. 3196–3219, Sep. 2021, doi: 10.24996/ijs.2021.62.9.33.

11. M. B. Shah, M. Kaistha, and Y. Gupta, "Student Performance Assessment and PredictionSystem using Machine Learning," in 2019 4th International Conference on Information Systems and

Computer Networks (ISCON), Mathura, India: IEEE, Nov. 2019, pp. 386– 390. doi: 10.1109/ISCON47742.2019.9036250.

12. V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICANN), Greater Noida, India: IEEE, Dec. 2020, pp. 177–181. doi: 10.1109/ICACCCN51052.2020.9362842.

13. S. Jayaprakash, S. Krishnan, and V. Jaiganesh, "Predicting Students Academic Performance using an Improved Random Forest Classifier," in 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India: IEEE, Mar. 2020, pp. 238–243. doi: 10.1109/ESCI48226.2020.9167547.

14. X. Yuan, J. Chen, K. Zhang, Y. Wu, and T. Yang, "A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT," IEEE Trans. Ind. Inf., vol. 18, no. 3, pp. 2032–2040, Mar. 2022, doi: 10.1109/TII.2021.3098306.

15. V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques : a survey," IJET, vol. 7, no. 2.8, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

16. M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," in 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India: IEEE, Feb. 2015, pp. 520–525. doi: 10.1109/ABLAZE.2015.7154917.

17. K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," Applied Sciences, vol. 11, no. 18, p. 8352, Sep. 2021, doi: 10.3390/app11188352.

18. T. Suresh, T. A. Assegie, S. Rajkumar, and N. Komal Kumar, "A hybrid approach tomedical decision-making: diagnosis of heart disease with machine-learning model,"IJECE, vol. 12, no. 2, p. 1831, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838inkedin.com/pulse/supervised-vs-unsupervised-learning-whats-difference-smriti- saini/

19. C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012. [2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," pp. 108–115, 2008.

20. L. Yahaya, n. D. Oye, and a. Adamu, "performance analysis of some selected machine learning algorithms on heart disease prediction using the noble uci datasets," international journal of engineering, no. 1, pp. 36–46, 2020.

21. Y. E. Shao, C.-D. Hou, and C.-C. Chiu, “Hybrid intelligent modeling schemes for heart disease classification,” Applied Soft Computing, vol. 14, pp. 47–52, 2014.

22. Yang, M., Wang, X., Li, F. & Wu, J. A machine learning approach to identify risk factors for coronary heart disease: a big data analysis. *Comput. Methods Programs Biomed.* 127, 262–270 2016.

23. T. Silwattananusarn, “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,” IJDKP, vol. 2, no. 5, pp. 13–24, Sep. 2012, doi: 10.5121/ijdkp.2012.2502.

24. B. Leventhal, “An introduction to data mining and other techniques for advanced analytics,” Journal of Direct, Data and Digital Marketing Practice, vol. 12, no. 2, pp. 137–153, 2010.

25. B. Leventhal, “An introduction to data mining and other techniques for advanced analytics,” Journal of Direct, Data and Digital Marketing Practice, vol. 12, no. 2, pp. 137–153, 2010.

26. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” presented at the 2017 IEEE Symposium on computers and communications (ISCC), IEEE, 2017, pp. 204–207.

27. C. M. Weiner, “Patient and Professional Constructions of Familial Hypercholesterolaemia and Heart Disease: Testing the Limits of the Geneticisation Thesis”doi.org/10.1016/ j.matpr.2020.12.901.

28. Ngufor, C., Hossain, A., Ali, S. & Alqudah, A. Machine learning algorithms for heart disease prediction: a survey. *Int. J. Comput. Sci. Inform. Secur.* 14 (2), 7–29 (2016). Ngufor, C., Hossain, A., Ali, S. & Alqudah, A. Machine learning algorithms for heart disease prediction: a survey. Int. J. Comput. Sci. Inform. Secur. 14 (2), 7–29 2016.

29. Mariettou S, Koutsojannis C, Triantafillou V. Predicting Coronary Heart Disease through Machine Learning Algorithms. Easy Chair; 2024 Feb 13.

30. Saikumar, K., & Rajesh, V. A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. International Journal of System Assurance Engineering and Management, 15(1), 135-151. 2024.

31. Ibrahim, I., &Abdulazeez, A. The Role of Machine Learning Algorithms for Diagnosing Diseases. Journal of Applied Science and Technology Trends, 2(01), 10–19. https://doi.org/10.38094/jastt20179 2021.

32. Gu, Y., Zhang, S., Qiu, L., Wang, Z., & Zhang, L. A layered KNN-SVM approach to predict missing values of functional requirements in product customization. Applied Sciences

(Switzerland), 11(5). https://doi.org/10.3390/ app11052420 2021.

33. Menshawi, A., Hassan, M.M., Allheeib, N. and Fortino, G. A Hybrid Generic Framework for Heart Problem diagnosis based on a machine learning paradigm. Sensors, 23(3), p.1392 2023.

34. Ishak, A., Ginting, A., Siregar, K., &Junika, C. of Heart Disease using Decision Tree AlgorithmClasiffication. IOP Conference Series: Materials Science and Engineering, 1003(1). https://doi.org/10.1088/1757-899X/1003/1/ 012119 2020.

35. Adebiyi, A. A., &Olugbara, O. Optimized hybrid investigative-based dimensionality reduction methods for malaria vector using KNN classifier. Journal of Big Data. https://doi.org/10.1186/s40537-021-00415-z 2021.

36. V. Chaurasia and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques," 2013 Published under Caribbean Journal of Scince and Technology ISSN 0799-37572013 Published under.

37. Ukwuoma, C. C., Cai, D., Heyat, M. B. B., Bamisile, O., Adun, H., Al-Huda, Z., & Al- Antari, M. A. Deep learning framework for rapid and accurate respiratory COVID-19 prediction using chest X-ray images. Journal of King Saud University- Computer and Information Sciences, 35(7), 101596 2023.

38. Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi- functional machine learning platform development for better healthcare and precision medicine," Database, vol. 2020, p. baaa010, Jan. 2020, doi: 10.1093/database/baaa010.

39. X. Liu et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method," Computational and Mathematical Methods in Medicine, vol. 2017, pp. 1–11, 2017, doi: 10.1155/2017/8272091.

40. A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," IEEE Access, vol. 9, pp. 39707– 39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

41. Khandaker Mohammad Mohi Uddin a,*kaiya Ripa a , Nilufar Yeasmin a , Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset, Bangladesh b School of Science and Technology, Bangladesh Open University,Gazipur,1705, 2020.3064084.

42. Shoukat, A., Arshad, S., Ali, N. & Murtaza, G. Prediction of Cardiovascular diseases using machine learning: a systematic review. *J. Med. Syst.* 44 (8), 162. https://doi.org/10.1007/s10916-020-01563-1 2020.

43. S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

44. G. G. N. Geweid and M. A. Abdallah, “A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique,” IEEE Access, vol. 7, pp. 149595–149611, 2019, doi: 10.1109/ACCESS.2019.2945527.

45. J. Vijayashree and H. P. Sultana, “A Machine Learning Framework for Feature Selectionin Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier,” Program Comput Soft, vol. 44, no. 6, pp. 388–397, Nov. 2018, doi: 10.1134/S0361768818060129.

46. S. Sharanyaa, S. Lavanya, M. R. Chandhini, R. Bharathi, and K. Madhulekha, “Hybrid Machine Learning Techniques for Heart Disease Prediction,” IJAERS, vol. 7, no. 3, pp. 44–48, 2020, doi: 10.22161/ijaers.73.7.

47. SarangamKodati, and Dr. R Vivekanandam, “A Comparative Study on Open Source Data Mining Tool for Heart Disease”, International Journal of Innovations & Advancement in Computer Science, Vol. 7, Issue 3, March 2018

48. Asha Rajkumar, and Mrs. G. SophiaReena, “Diagnosis of Heart Disease Using Data Mining Algorithm”, Global Journal of Computer Science and Technology, Vol. 10, pp. 38-43, Sept. 2010.

49. K.Gomath, Dr. Shanmugapriyaa, “Heart Disease Prediction Using Data Mining Classification”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol.4, Issue 2, February 2016.

50. Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R. I., Tamam, N., Sulieman, A., &Pathan, R. K. The Efficacy of Machine-Learning- Supported Smart System for Heart Disease Prediction. Healthcare (Switzerland), 10(6), 1–19. //doi.org/10.3390/healthcare 10061137, 2022.

51. Adebiyi, A. A., &Olugbara, O. Optimized hybrid investigative-based dimensionality reduction methods for malaria vector using KNN classifier. Journal of Big Data. //doi.org/10.1186/s40537-021-00415-z 2021.

52. Alobed, M., Altrad, A. M. M., &Bakar, Z. B. A. A Comparative Analysis of Euclidean, Jaccard, and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring. Proceedings - CAMP 2021: 2021 5th International Conference on Information Retrieval and Knowledge Management: Digital Technology for IR 4.0 and Beyond, 70–74. //doi.org/10.1109/CAMP51653.2021.9498119 2021.

53. Amin, S., Shishavan, S., Kutlu, F., &Farrokhizadeh, E. Engineering Applications of Artificial Intelligence Novel similarity measures in spherical applications. Engineering Applications of Artificial Intelligence, 94(April), 103837. https://doi.org/10.1016/ j.engappai. 2020.103837, 2020.

54. Shankar, G. R., Chandrasekaran, K. & Babu, K. S. An analysis of the potential use of Machine Learning in Cardiovascular Disease Prediction. *J. Med. Syst.* 43 (12), 345. https://doi.org/10.1007/s10916-019-1524-8 2019.

55. Arvind, S., Sahay Prasad, P., Saraswathi, R. V., Vijayalata, Y., Tasneem, Z., AshlinDeepa, R. N., &Ramadevi, Y. Deep Learning Regression-Based Retinal Layer Segmentation ThingSpeak. Mobile Information Systems, 2022.

56. Ayon, S. I., Islam, M., & Hossain, R. Coronary Artery Heart Disease Prediction : A Comparative Study of Computational Intelligence Techniques Coronary Artery Heart Disease Prediction : A Comparative Study of. IETE Journal of Research, 0(0), 1–20. //doi.org/10.1080/03772063.2020. 1713916. 2020.

57. Bemando, C., Miranda, E., &Aryuni, M). Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms. Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and InformationManagement. 2021.

58. Besta, M., Kanakagiri, R., Mustafa, H., Karasikov, M., Ratsch, G., Hoefler, T., &Solomonik, E. Communication-Efficient Jaccard Similarity for High- Performance Distributed Genome Comparisons. Proceedings - 2020 IEEE 34th International Parallel and Distributed Processing Symposium, IPDPS 2020, 1122–1132.//doi.org/10.1109/IPDPS47924.2020.00118. 2020.

59. Budholiya, K., Shrivastava, S. K., & Sharma, V. An optimized XGBoost-based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences, 34(7), 4514–4523.//doi.org/10.1016/j.jksuci.2020.10.013. 2022.

60. R. Jenke, A. Peer, and M. J. I. T. o. A. c. Buss, "Feature extraction and selection for emotion recognition from EEG," vol. 5, no. 3, pp. 327-339, 2014.

61. Moon, S., Lee, W. & Hwang, J. Applying machine learning to Predict Cardiovascular diseases. Healthc. Inf. Res. 25 (2), 79–86. https://doi.org/10.4258/hir.2019.25.2.79 2019.

62. Lakshmi, M. & Ayeshamariyam, A. A layered KNN-SVM approach to A Stacking Ensemble Model to Predict Daily Numberof Hospital Admissions for Cardiovascular Diseasesof functional requirements in product customization. Applied Sciences (Switzerland), 11(5). https://doi.org/10.3390/ app11052420.

63. Md, R. et al. Early detection of cardiovascular autonomic neuropathy: A multi-class

classification model based on feature selection and deep learning feature fusion. Information Fusion, vol. 77, P 70–80, January 2022.

64. Hamdaoui, H. El, Boujraf, S., Chaoui, N. E. H., &Maaroufi, M. A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques. 2020 International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020. //doi.org/10.1109/ATSIP 49331.2020.9231760. 2020.

65. Ibrahim, I., &Abdulazeez, A. The Role of Machine Learning Algorithms for Diagnosing Diseases. Journal of Applied Science and Technology Trends, 2(01), 10–19. 2021.

66. Jawthari, M., &Stoffová, V. Predicting students' academic performance using a modified kNN algorithm. Pollack Periodica, 16(3), 20–26. https://doi.org/ 10.1556/606.2021.00374. 2021.

67. Farag, A., Farag, A. & Sallam, A. Improving Heart Disease prediction using boosting and bagging techniques. *Proc. Int. Conf. Innovative Trends Comput. Eng. (ITCE)*. 90-96 https://doi.org/10.1109/ITCE.2016.7473338 2016.

68. Mirza, Q. Z., Siddiqui, F. A. & Naqvi, S. R. The risk prediction of cardiac events using a decision Tree Algorithm. *Pakistan J. Med. Sci.* 36 (2), 85–89. https://doi.org/10.12669/pjms.36.2.1511 2020.

69. Jhajhria, S. & Kumar, R. Predicting the risk of Cardiovascular diseases using ensemble learning approaches. *Soft. Comput.* 24 (7), 4691–4705. https://doi.org/10.1007/s00500-019-04268-8 2020.

70. JA. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

71. Samadiani, N., Moghadam, E., Motamed, C. & A. M., & SVM-based classification of Cardiovascular diseases using feature selection: a high-dimensional dataset perspective. *J. Med. Syst.* 40 (11), 244. https://doi.org/10.1007/s10916-016-0573-7 2016.

72. Mujawar, S., &Devale, P. Heart Disease Prediction Using Modified K Means And Using Naive Baiyes. International Journal of Computer Sciences and Engineering, 3.Mathematical Problems in Engineering, vol. 2021, 2019, doi: 10.1155/2021/5594899.

73. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707. Diseases; Heart;Data mining; Support vector machines; Feature extraction; Machine learning.

74. Wongkoblap, A., Vadillo, M. A. & Curcin, V. Machine learning classifiers for early detection of

Cardiovascular Disease. J. Biomed. Inform. 88, 44–51. https://doi.org/10.1016/j.jbi.2018.09.003 2018.

75. Azmi, Javed, Muhammad Arif, MdTabrezNafis, M. AfsharAlam, SafdarTanweer, and Guojun Wang. "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data." Medical engineering & physics 105 (2022): 103825.

76. Shetty, Ashwini, and Chandra Naik. "Different data mining approaches for predicting heart disease." International Journal of Innovative in Science Engineering and Technology 5 2016: 277-281.

77. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest Mohamed G. El-Shafiey1 & Ahmed Hagag2 & El-Sayed A. El-Dahshan1,3 &Manal A. Ismail4 Received: 10 February 2021 / Revised: 10 June 2021 / Accepted: 25 January 2022 / # The Author(s) 2022.

78. Al Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J. (2019) 40:1975–86. doi: 10.1093/eurheartj/ehy404

79. A. S. Hassan, I. Malaserene, and A. A. Leema, "Diabetes Mellitus Prediction using Classification Techniques," no. 5, pp. 2080–2084, 2020, doi: 10.35940/ijitee.E2692.039520.

80. Ahn I, Gwon H, Kang H, Kim Y, Seo H, Choi H, et al. Machine learning–based hospital discharge prediction for patients with cardiovascular diseases: development and usability study. JMIR Med Inform. (2021) 9:e32662. doi: 10.2196/32662

81. Shah, D., Patel, S. & Bharti, S. K. Heart disease prediction using machine learning techniques. SN Comput. Sci. 1, 1–6 (2020).16. Nouman, A. &Muneer, S. A systematic literature review on heart disease prediction using blockchain and machine learning techniques. Int. J. Comput. Innov. Sci. 1(4), 1–6 2022.

82. Rao, GorapalliSrinivasa, and G. Muneeswari. "A Review: Machine Learning and Data Mining Approaches for Cardiovascular Disease Diagnosis and Prediction." EAI Endorsed Transactions on Pervasive Health and Technology 10 2024.

83. Y. X. Zhao, H. Yuan, and Y. Wu, "Prediction of Adverse Drug Reaction using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset," ACM International Conference Proceeding Series, pp. 17–21, 2021, doi: 10.1145/3472813.3472817.

84. A. Di Tang, S. Q. Tang, T. Han, H. Zhou, and L. Xie, "A Modified Slime Mould Algorithm for Global Optimization," Computational Intelligence and Neuroscience, vol. 2021, 2021, doi: 10.1155/2021/2298215.

85. M. Zahid et al., "Electricity Price and Load Forecasting using Enhanced Convolutional Neural Network and Enhanced Support Vector Regression in Smart Grids," Electronics, vol. 8, no. 2, p. 122, 2019, doi: 10.3390/electronics8020122.

86. Faieq, A., & Mijwil, M. M. Prediction of heart diseases utilizing support vector machine and artificial neural network. Indonesian Journal of Electrical Engineering and Computer Science, 26(1), 374–380. https://doi.org/ 10.11591/ijeecs.v26.i1.pp374-380. 2022.

87. N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R. I., Tamam, N., Sulieman, A., & Pathan, R. K. The Efficacy of Machine-Learning- Supported Smart System for Heart Disease Prediction. Healthcare (Switzerland), 10(6), 1–19. https://doi.org/10.3390/healthcare 10061137. 2022.

88. M., Altrad, A. M. M., & Bakar, Z. B. A. A Comparative Analysis of Euclidean, Jaccard, and Cosine Similarity Measure and Arabic Wordnet for Automated Arabic Essay Scoring. Proceedings - CAMP 2021: 2021 5th International Conference on Information Retrieval and Knowledge Management: Digital Technology for IR 4.0 and Beyond, 70–74. https://doi.org/10.1109/CAMP51653.2021.9498119. 2021.

89. Delavar, M. R., Motwani, M. & Sarrafzadeh, M. A. Comparative study on feature selection and classification methods for Cardiovascular Disease diagnosis. *J. Med. Syst.* 39 (9), 98. https://doi.org/10.1007/s10916-015-0333-5 2015.

90. Delavar, M. R., Motwani, M. & Sarrafzadeh, M. A. Comparative study on feature selection and classification methods for Cardiovascular Disease diagnosis. J. Med. Syst. 39 (9), 98. https://doi.org/10.1007/s10916-015-0333-5 2015.

91. Aflori, M. Craus, "Grid implementation of the Apriori algorithm Advances in Engineering Software", Volume 38, Issue 5, October 2007, pp. 295-300. A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Mining frequent patterns in image databases with 9DSPA representation", Journal of Systems and Software, Volume 82, Issue 4, April 2009, pp.603-618.

92. Delavar, M. R., Motwani, M. & Sarrafzadeh, M. A. Comparative study on feature selection and classification methods for Cardiovascular Disease diagnosis. J. Med. Syst. 39 (9), 98. https://doi.org/10.1007/s10916-015-0333-5 2015.

93. Rani P, Kumar R, Sid NMO, Anurag A. A decision support system for heart disease prediction based upon machine learning. J ReliabIntell Environ 7(3):263–275. https://doi.org/10.1007/ s40860-021-00133-6. 2021.

94. Krishnaiah V, Srinivas M, Narsimha G, Chandra NS Diagnosis of heart disease patients using fuzzy classification technique. IEEE Int Conf Comput Commun Technol. https://doi. org/10.1109/ICCCT2.2014.7066746. 2014.

95. Jousilahti P, Vartiainen E, Tuomilehto J, Puska P Sex, age, cardiovascular risk factors, and coronary heart disease. Circulation 99(9):1165–1172. https://doi.org/10.1161/01.cir.99.9. 1165. 1999.

96. Moallem P, Razmjooy N, Ashourian M Computer vision based potato defect detection using neural networks and support vector machine. Int J Robot Autom 28(2):137–145. https://doi.org/10.2316/Journal.206.2013.2.206-3746. 2013.

97. Prakash B, Debnath D, Midhun B A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. Distrib Parallel Databases. https://doi.org/10.1007/s10619-021-07329-y. 2021.