An Intelligent Model to Predict Cardiovascular Diseases Using Machine Learning Techniques

Author

Zawaria Sadaf

318477

Supervisor

Dr. Sajid Gul Khawaja

DEPARTMENT OF COMPUTER SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

OCTOBER, 2019

An Intelligent Model to Predict Cardiovascular Diseases Using Machine Learning Techniques

Author

Zawaria Sadaf

318477

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor

Dr. Sajid Gul Khawaja

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

OCTOBER, 2019

Declaration

I certify that this research work titled "*An Intelligent Model to Predict Cardiovascular Diseases Using Machine Learning Techniques"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/ referred.

Signature of Student

Zawaria Sadaf

318477

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical, and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Zawaria Sadaf
318477

Signature of Supervisor
Dr. Sajid Gul Khawaja

Copyright Statement

Acknowledgments

I attribute all recognition and honor to the divine power of Allah (the most compassionate, the most gracious). Allah bestowed upon me the bravery, resilience, understanding, and capability to undertake and successfully accomplish this task. It is undeniable that Allah facilitated my journey, and without His blessings, I would not have achieved anything.

I would like to express my earnest appreciation to my advisor Dr. Sajid Gul Khawaja for boosting my morale and for his continual support, motivation, dedication, and invaluable guidance in my quest for knowledge. I am blessed to have such a co-operative and kind mentor for my research.

Along with my advisor, I would like to acknowledge my entire thesis committee: Dr. Muhammad Usman Akram, Dr. Arslan Shaukat and Dr. Farhan Hussain for their cooperation and prudent suggestions.

I express my deep gratitude to my family, who have been the primary source of my strength and support. In particular, I am immensely thankful to my dear parents, who cared for me from the earliest stages of my life, even when I could not walk, and continued to provide unwavering support in all aspects of my journey. My acknowledgment would be incomplete without recognizing their significant role in shaping my life.

Finally, I would like to express my gratitude to all my friends and the individuals who have encouraged and supported me through this entire period.

Abstract

Human life is the most important asset of human beings. Every year millions of people lose their lives to cardiovascular diseases. It is a group of diseases related to blood vessels and the heart. Chances of developing cardiovascular diseases in a person can be controlled by reducing some risk factors that cause them. If they are predicted timely in patients, the patients can take decisions and make changes to their lifestyles, and consequently reduce the risk of developing cardiovascular diseases. The traditional way to predict cardiovascular diseases is to consult a medical specialist. But this method can be inaccurate, consumes a lot of time and is expensive. In the past, work has been done regarding predicting cardiovascular diseases in patients using machine learning and deep learning techniques. But there was a gap in the literature that was filled by this study.

In the proposed research, a model consisting of five main modules; data collection, data preprocessing, data processing, performance layer, and validation has been used. The proposed model gave very promising results. It has proven to be very efficient in predicting cardiovascular disease in a person using Gradient Boosting Tree algorithm. The model had 78.78%, 76.78%, 81.10%, 82.43%, 18.90% accuracy, sensitivity, specificity, miss rate, and precision, respectively. Moreover, the fallout, LR+, LR-, and NPV were 18.90%, 4.06, 3.82, and 75.16% respectively. The classification time was a few milliseconds per record and the detection time was approximately 0.2137 seconds per record. The proposed model also outperformed various well-known machine learning algorithms and state-of-the-art models.

Key Words: *Cardiovascular Diseases, Prediction, Feature Extraction, Machine Learning, Principal Component Analysis, Artificial Neural Network*

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Cardiovascular diseases are diseases related to blood vessels and the heart. These include Coronary heart disease, Deep Vein Thrombosis, Congenital heart disease, Cerebrovascular disease, Peripheral Arterial disease, Rheumatic heart disease, and Pulmonary Embolism. Acute events of cardiovascular diseases like strokes and heart attacks are caused due to prevention of the flow of blood to the brain or heart because of some blockage in blood vessels. These blockages can result from a build-up of fatty deposits or blood clots [1]. Cardiovascular diseases are one of the top causes of death worldwide and are increasing day by day in our modern world. Figure 1.1 shows the top five causes of death globally in 2020. It can be seen that there were two cardiovascular diseases among these five causes. Around 17.3 million people lose their lives to cardiovascular diseases annually [2]. Moreover, it is estimated that by 2030, this figure will cross 23 million [3].

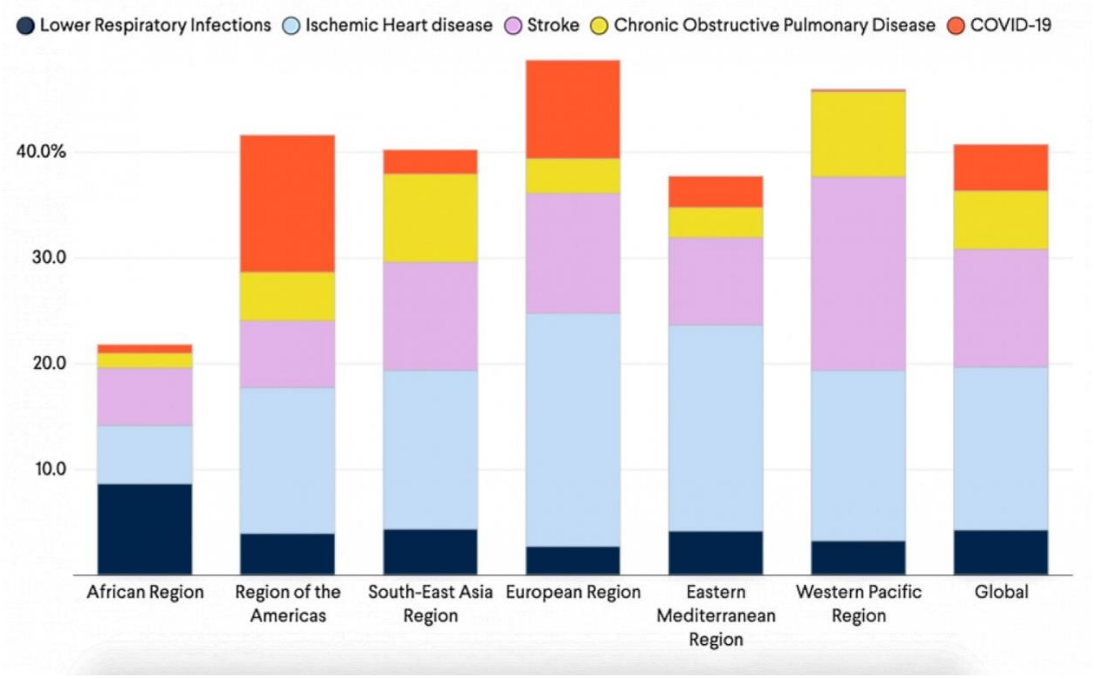

Figure 1.1 Top Five Causes of Death Globally in 2020 [4]

Behavioral factors and biochemical factors can cause cardiovascular diseases. Behavioral factors include diabetes, obesity, physical inactivity, food habits, smoking, lifestyle, and harmful use of alcohol. Whereas biochemical factors include glycemia, blood pressure, and raised blood lipids [5]. Any person with the conditions mentioned above can be at a higher risk of developing cardiovascular diseases. So, that person should take precautionary measures to reduce the risk.

Early methods of prognosing cardiovascular diseases help make decisions about high-risk patients' changes in their lifestyles that will reduce their risks [6], [7]. This brings to light the immense need for a system that can identify the risk of developing cardiovascular diseases in a person so that precautions can be taken timely.

A solution is to predict cardiovascular diseases in people somehow using behavioral factors and biochemical factors through machine learning. Machine learning is a branch of artificial intelligence that has been seen emerging in the past few years. Machine learning automates the building of analytical models. It is a data analysis technique. The main concept behind machine learning is that a system identifies patterns, learns from data, and take decisions with minimum human interference. Algorithms of machine learning improve automatically by using data or through experience. Machine learning is of three diverse types: reinforcement learning, supervised learning, and unsupervised learning. Various methods include using random forests [8], support vector machines (SVM) [9], convolutional neural network (CNN) [10], logistic linear regression [11]. To assess behavioral factors and biochemical factors, clinical data [8], coronary computerized tomography angiography (CCTA) [12], cardiac magnetic resonance imaging (MRI) [13], electrocardiogram (ECG) [14], urinary proteomic analysis [15], etc. have been used in the past. This research presents an improved intelligent model to predict cardiovascular diseases using clinical data. Distinct factors including age, weight, height, gender, glucose, smoking, alcohol intake, cholesterol, physical activity, diastolic blood pressure, and systolic blood pressure have been gathered. Machine learning algorithms have been used to detect the risk of patients developing cardiovascular diseases.

1.1 Motivation

The traditional way to predict cardiovascular diseases is to consult a medical specialist. They perform tests like electrocardiogram (ECG), coronary calcium scan (CCS), and cardiovascular magnetic resonance (CMR) imaging or take measures of cholesterol, blood sugar, and blood pressure [16]. However, this method consumes a lot of time and is expensive. Moreover, some medical specialists may mark a patient at substantial risk for cardiovascular disease. In contrast, another medical specialist may mark him at low or no risk for the same disease. Wrong prognosis can have very heavy consequences as cardiovascular diseases are fatal illnesses. For this purpose, the availability of a medical specialist who has years of professional experience is highly required.

These challenges can be reduced by using machine learning algorithms. This brings to light the immense need for a system that can automatically identify the risk of developing cardiovascular diseases in a person. The health care industry keeps records of a lot of medical data. So, machine learning techniques can be used to make effective decisions in predicting heart diseases. They can help medical experts improve the accuracy of the prognosis of diseases by giving a second opinion and decreasing the processing time. This is done by transferring man's knowledge to machine intelligence.

1.2 Problem Statement

Cardiovascular diseases are considered threatening diseases. They have an exceedingly high mortality rate. Specific behavioral and biochemical factors can increase a person's risk of developing them. High-risk people need to change their lifestyles to avoid or reduce the risk of developing them. This research aims to design a machine learning-based intelligent model for the prognosis of cardiovascular disease, which can prevent, mitigate, and reduce the risk of people developing them.

1.3 Aims and Objectives

The primary objectives of the research are mentioned below:

- To use preprocessing layer for processing raw data, normalizing data, and noise removal
- To analyze the dataset to extract useful features.
- To use machine learning techniques to develop a complete system for the prognosis of cardiovascular diseases.

1.4 Structure of Thesis

The structure of the remaining research work is as follows:

Chapter 2 covers the analysis of cardiovascular diseases.

Chapter 3 reviews the literature and the significant work done by researchers for predicting cardiovascular diseases using different machine learning techniques.

Chapter 4 consists of the proposed methodology in detail. It includes dataset explanation (raw data), data preprocessing (prepared data), data processing (analysis results), performance layer (finalized results), and finally, visualization.

Chapter 5 includes all the results of the experiments. All desired figures and tables are also included.

Chapter 6 concludes the thesis and reveals the future scope of this research.

Chapter 2 Analysis of Cardiovascular Diseases

Cardiovascular diseases are one of the top causes of deaths worldwide and are increasing daily. It is a group of diseases related to the heart and blood vessels, including rheumatic heart disease, coronary heart disease, cerebrovascular disease, etc. Heart attacks and strokes are the most common causes of death among cardiovascular diseases, causing more than four out of five deaths in cardiovascular patients. One out of three such patients dies prematurely under the age of seventy [1].

2.1 Cardiovascular System

The cardiovascular system has three major components: the heart, a closed network of blood vessels (veins, arteries, and capillaries), and blood. The responsibility of this system is to deliver blood to the entire body. The cells of the body get their nutrients and oxygen from the blood. The cells and tissues would function correctly and malfunction and die if they do not get sufficient blood supply.

The pulmonary circulatory system and the systemic circulatory system are two subsystems of the blood circulation system. Blood is delivered to the body's cells, tissues, and organs through the systemic circulatory system. The pulmonary circulatory system, however, is responsible for moving blood to and from the lungs. Here, oxygen and carbon dioxide are exchanged [17]. The pulmonary and systemic circulatory systems are depicted in Figure 2.1.

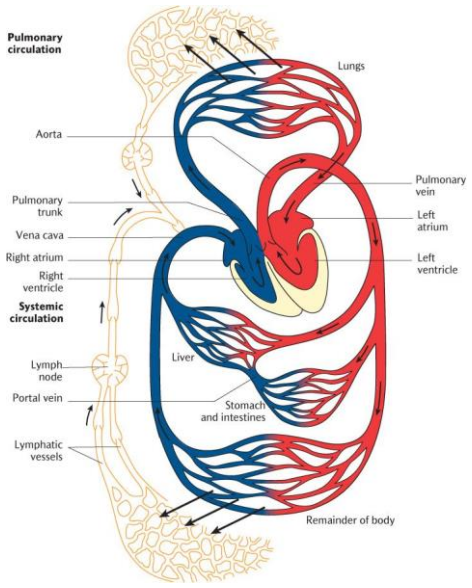

Figure 2.1 Systemic Circulatory System and Pulmonary Circulatory System [17]

2.1.1 Components of Cardiovascular System

The heart is the cardiovascular system's primary building block. Its main job is to provide oxygenated and deoxygenated blood to the body's organs and lungs. A human heart has two atriums and two ventricles, making up its four chambers [18]. The atriums are where blood enters the heart, and the ventricles are where it leaves. The right atrium sends deoxygenated blood to the right ventricle and lungs. The left atrium sends oxygenated blood to the left ventricle, which sends it to the rest of the body. The circulation of blood throughout the heart is seen in Figure 2.2.

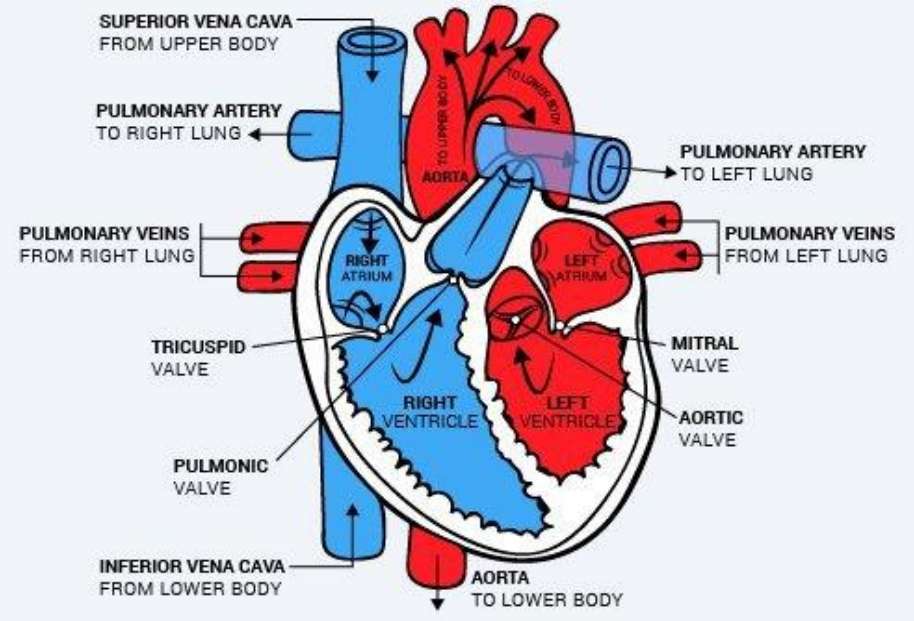

Figure 2.2 Circulation of Blood in the Heart [18]

The blood vessels are the second main part of the cardiovascular system. They come in three main varieties: capillaries, arteries, and veins. Deoxygenated blood travels from the body through veins to the heart. They have valves that only allow the blood to flow in one way and are thinner and weaker than arteries. The body receives oxygenated blood from the heart through arteries. Their muscular inner layer and hard outer layer assist in pumping blood throughout the body. The blood flows more easily because of the inner layer's smooth surface. Veins and arteries are connected by capillaries. They pull oxygenated blood from the arteries, transport it to the organs, where it is exchanged for deoxygenated blood, and then return it to the veins [19]. Figure 2.3 shows the flow and exchange of blood in the blood vessels.

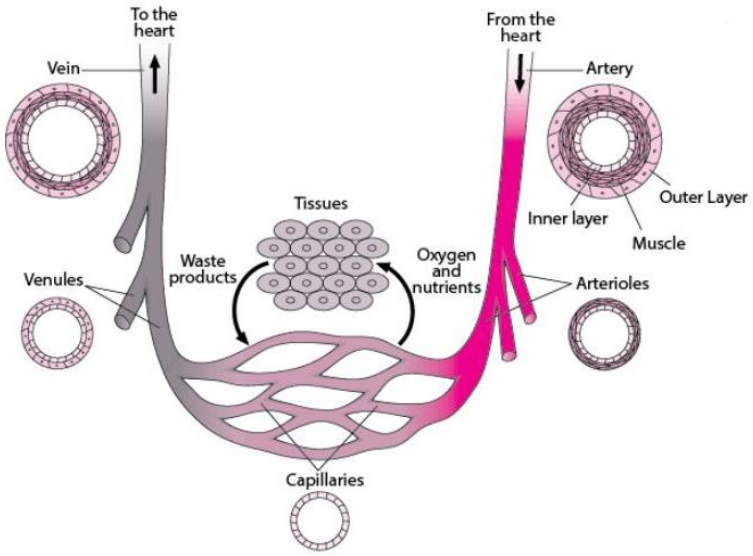

Figure 2.3 The Flow and Exchange of Blood in Blood Vessels [19]

Blood is the cardiovascular system's last major component. Blood is a bodily fluid made up of suspended red blood cells, white blood cells, and platelet-rich plasma. Aside from blood cells, blood also includes hormones, minerals ions, carbohydrates, proteins, and carbon dioxide. The body's organs and cells receive oxygen and nutrition from the blood, which also removes metabolic waste products from them. Blood that has been oxygenated is bright red, while blood that has been deoxygenated is dark red [20]. Figure 2.4 shows the composition of blood.

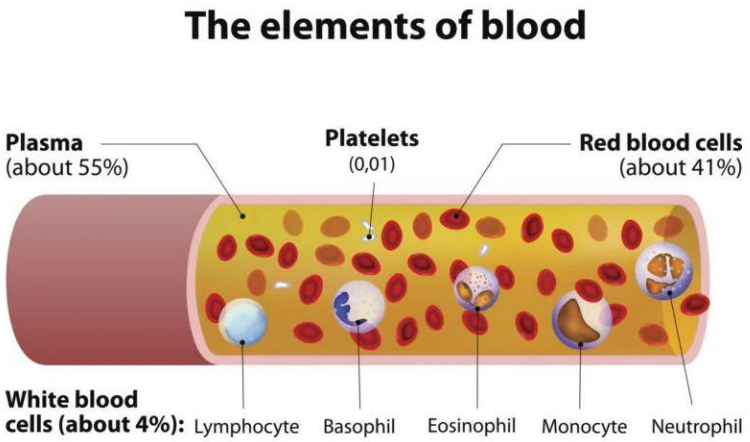

Figure 2.4 Composition of Blood [20]

2.2 Cardiovascular Diseases

Cardiovascular diseases are a group of diseases related to blood vessels and the heart. They are caused due to prevention of the flow of blood to the brain or heart because of some sort of blockage in blood vessels. These blockages can result from a build-up of fatty deposits or blood clots. Cardiovascular diseases can also be accompanied by damage to arteries in the eyes, kidneys, heart, brain, and other organs. Cardiovascular diseases are among the leading causes of death worldwide, but they can be prevented by adopting a healthy lifestyle [21].

2.2.1 Causes of Cardiovascular Diseases

The leading cause of many types of cardiovascular diseases is atherosclerosis. It occurs due to the build-up of fatty materials, including lipids, calcium, and cellular debris inside the blood vessels causing blockage, as shown in Figure 2.5 [22]. Health conditions like diabetes, virus, myocarditis, or structural problems present by birth can also damage the circulatory system. High blood pressure can also cause cardiovascular diseases without showing any symptoms, so people should get it checked regularly.

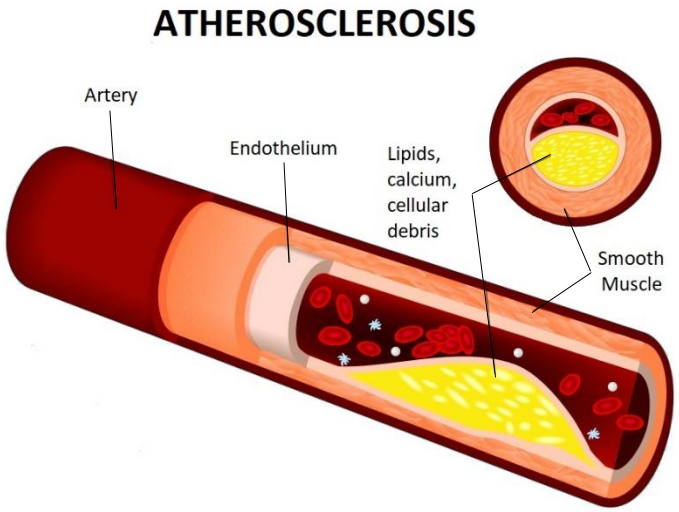

Figure 2.5 Build-up of Fatty Materials in Atherosclerosis [22]

2.2.2 Symptoms of Cardiovascular Diseases

Symptoms of cardiovascular diseases may vary from person to person. The same people having the same cardiovascular disease may have a difference in symptoms. The reason is that some

conditions like hypertension and type 2 diabetes may cause no symptoms to show in the initial stages. However, there are some symptoms that most patients may show. These symptoms include [23]:

- Cold sweats
- Dizziness or lightheadedness
- Fatigue and nausea
- Shortness of breath
- Pain or discomfort in the back, jaw, elbows, left shoulder, or arms.
- Pressure or pain in the chest

2.2.3 Types of Cardiovascular Diseases

Cardiovascular diseases can be divided into two categories. The first category is the diseases affecting the heart. They include Aortic Stenosis, Heart Attack, Stroke, Arrhythmia, Angina, Radiation Heart Disease, Rheumatic Heart Disease, Atrial Fibrillation, Dilated Cardiomyopathy, Pulmonary Stenosis, Hypertrophic Cardiomyopathy, Mitral Valve Prolapse, Coronary Artery Disease, Mitral Regurgitation, Heart Failure, and congenital heart disease.

The second category is the diseases affecting the blood vessels. They include Abdominal Aortic Aneurysm, Peripheral Artery Disease, Buerger's Disease, Blood Clotting Disorders, Venous Blood Clots, Peripheral Venous Disease, Raynaud's Disease, Renal Artery Disease, Atherosclerosis, and Aneurysm. A person can develop more than one cardiovascular disease simultaneously, one leading to the other. Some of these diseases are discussed below.

- **Aortic Stenosis**

The aortic valve of the heart becomes narrow, which causes Aortic Stenosis. The valve is not able to open fully. Figure 2.6 shows the opening and closing of valves of a normal person and a patient suffering from aortic stenosis. This prevents or blocks blood flow from the heart to the aorta, which is the body's main artery. This consequently blocks the flow of blood to the entire body. This causes the heart to pump harder for the blood to flow from the aorta, thickening and enlarging the left ventricle. Eventually, this weakens the heart muscles and can lead to other serious problems like heart failure [24].

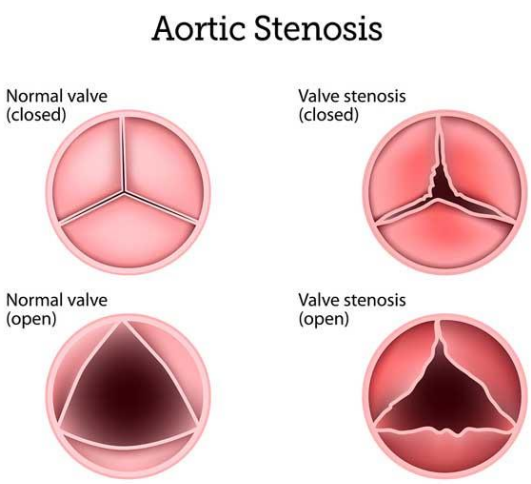

Figure 2.6 Opening and Closing of Valves of a Normal Person and a Patient Suffering from Aortic Stenosis [24]

- **Heart Attack**

Sometimes blood clots can cause blockage in blood flow to the heart muscle, leading to a heart attack. If the blood flow is cut off completely, the part of the heart muscle not receiving blood starts to die because it is not getting the supply of oxygen. Figure 2.7 shows dying heart muscles because of a blood clot. The first heart attack is not fatal, and people may continue to enjoy a productive life after their first heart attack. But changes need to be made in their lifestyle by them to avoid any future heart attack that could prove fatal [25].

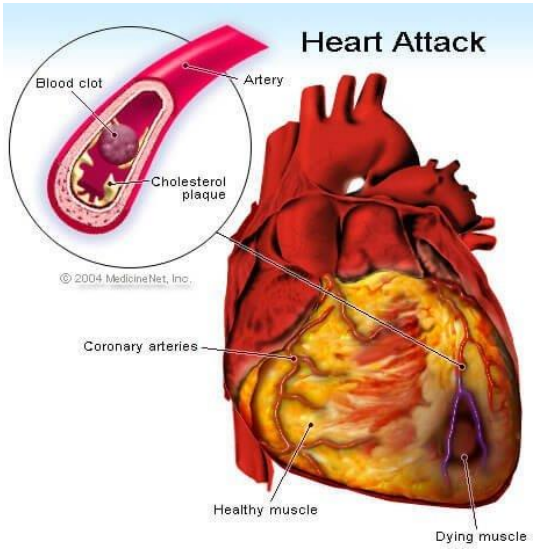

Figure 2.7 Dying Heart Muscles Because of a Blood Clot [25]

- **Stroke**

The reduction of blood flow to the brain can lead to the occurrence of a stroke, which can result in permanent brain damage or even death. Transient Ischemic Attack (TIA) refers to the temporary disruption of blood supply to the brain. The most prevalent form of stroke is known as Ischemic stroke, which occurs when a blood clot or any other obstruction blocks the blood vessel responsible for supplying blood to the brain. The insufficient blood supply to the brain leads to the death of certain brain cells, causing a loss of function controlled by the affected region of the brain. Another type of stroke is hemorrhagic stroke, which happens when a blood vessel ruptures within the brain, often due to hypertension. Figure 2.8 shows the two main types of strokes. The brain cells that die due to a stroke are never replaced, causing the effects of the stroke to be permanent [26].

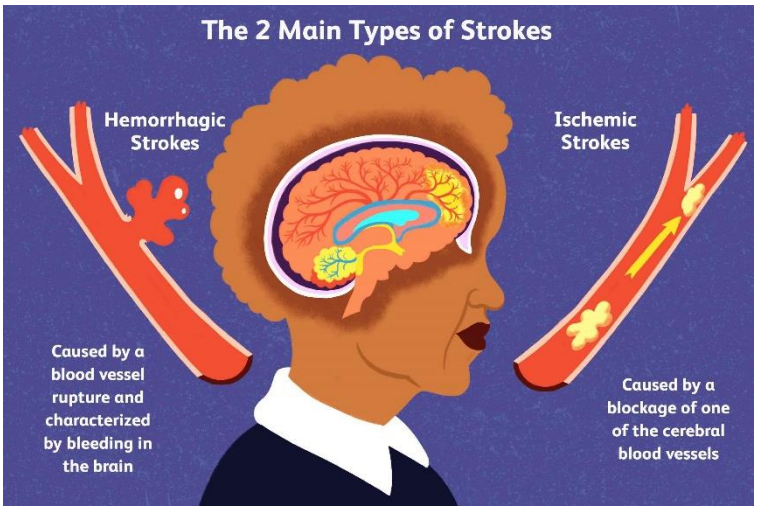

Figure 2.8 The Two Main Types of Strokes [26]

- **Arrhythmia**

If the heartbeat is abnormal, the heart cannot meet the body's requirements by pumping enough blood. The abnormal rhythm of the heartbeat is referred to as Arrhythmia. Arrhythmia can be of several types, including single Premature Ventricular Contractions (PVC), paired PVC, heteromorphic paired PVC, trigeminy, bigeminy, etc. Figure 2.9 shows the Electrocardiogram (ECG) of several types of Arrhythmias. These types can cause the heart to beat too, too slow, or irregularly. Types of Arrhythmias can be divided into two categories: Bradycardia and Tachycardia. The first causes the heart to operate too slowly, at less than 60 beats a minute. The latter causes the heart rate to accelerate in excess of 100 beats a minute [27].

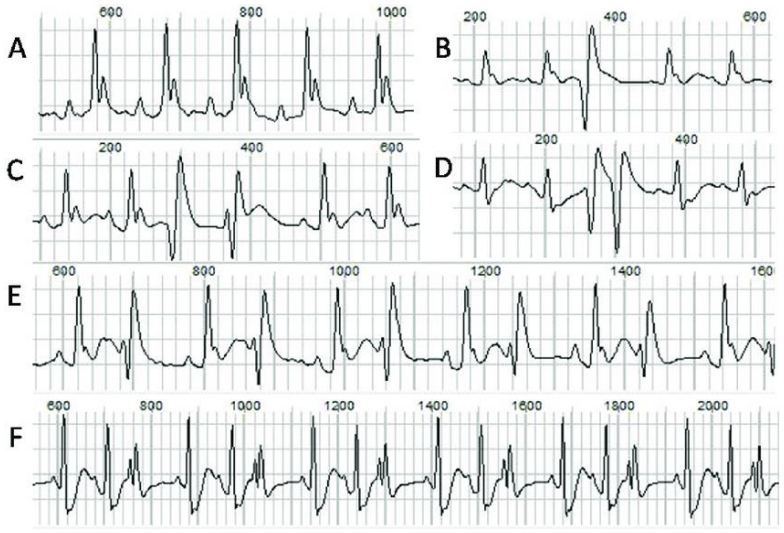

Figure 2.9 Examples of Electrocardiogram (ECG) Patterns. (A) Normal pattern of ECG signal, (B) Signal Variation due to Single Premature Ventricular Contraction (PVC), (C) Signal Variation due to Heteromorphic Paired PVC, (D) Signal Variation due to Paired PVC, (E) Signal Variation due to Bigeminy, (F) Signal Variation due to Trigeminy [27].

- **Abdominal Aortic Aneurysm**

When the aorta walls become progressively weak, the vessels start 'ballooning.' This causes Abdominal Aortic Aneurysm. Sometimes the ballooning of vessels is not that severe, therefore does not cause any symptoms and is not a problem. But other times, ballooning of vessels is so severe that it causes the vessels to expand and eventually rupture if proper diagnosis and treatment are not made on time. Aneurysm mostly occurs in the main artery of the abdomen and chest, called the abdominal aorta. This artery carries blood to all body parts, including legs and feet. So, if this artery ruptures, the supply of blood to the entire body is disrupted [28]. Figure 2.10 shows a healthy abdominal aorta and a ballooned abdominal aorta.

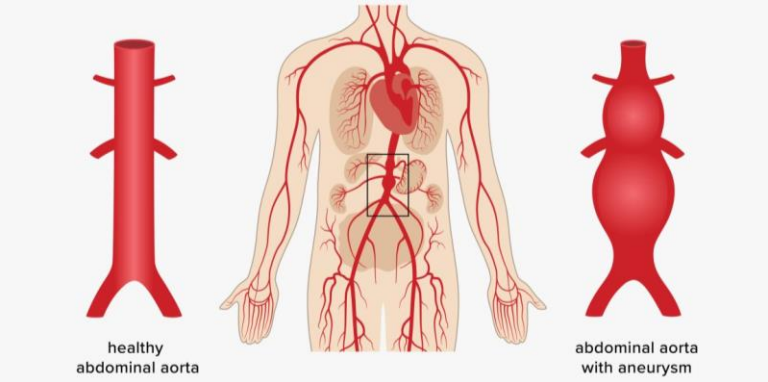

Figure 2.10 A Healthy Abdominal Aorta and a Ballooned Abdominal Aorta [28]

- **Peripheral Artery Disease**

The arteries to the legs and feet can be blocked due to the thickening of the inner walls because of a build-up of fatty deposits and plaque, causing Peripheral Artery Disease. It is also termed as "hardening of the arteries." Figure 2.11 shows a buildup of plaque in the arteries. This disease can also cause ulcers on the legs and feet, weakness or numbness in the feet, loss of hair from the legs and feet, and pain in the legs. If the surface of plaque build-up becomes ulcerated or irregular, it can also cause blood clots. When these blood clots travel through tiny vessels in the circulatory system, they can also reach and damage sensitive organs like the brain and cause a stroke [29].

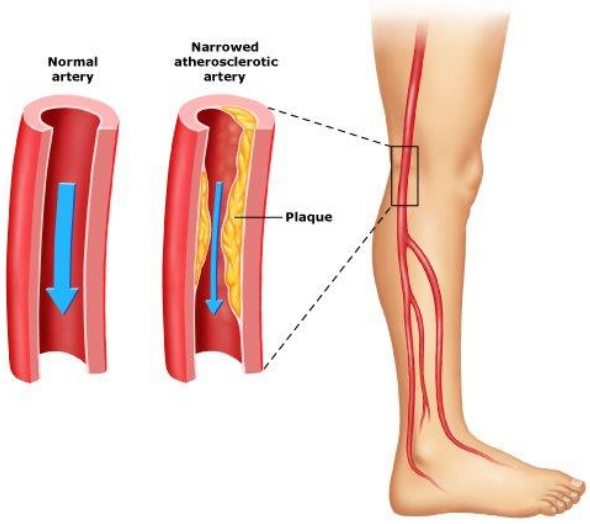

Figure 2.11 Buildup of Plaque in the Arteries [29]

2.3 Risk Factors that can Cause Cardiovascular Diseases

Some risk factors can increase the chances of developing cardiovascular diseases in a person. The more the person has risk factors, the greater his chances of developing cardiovascular diseases are. A person should get an NHS Health Check every five years after crossing the age of forty. This health check assesses the risk factors a person has. The person is then advised to take necessary preventive measures to reduce the risk of developing cardiovascular diseases. The main risk factors for cardiovascular diseases are discussed below [21].

2.3.1 High Blood Pressure

Hypertension or high blood pressure is among the most significant risk factors for developing cardiovascular diseases. Damage to blood vessels can be caused by high blood pressure leading to various cardiovascular diseases.

2.3.2 Smoking

Smoking is another main risk factor associated with developing cardiovascular diseases. Tobacco and other harmful substances may be deposited in the blood vessels, causing them to narrow. This can damage blood vessels, eventually leading to various cardiovascular diseases.

2.3.3 High Cholesterol

A fatty substance called cholesterol is found in the human blood. Prominent levels of cholesterol increase the risk of developing cardiovascular diseases. The blood vessels get narrow because of cholesterol deposits in them. Blood clots can also be formed because of this. This narrowing of blood vessels and forming blood clots can lead to various cardiovascular diseases.

2.3.4 Diabetes

Diabetes causes the level of sugar in the blood to rise exceedingly high. It is a lifelong condition and a crucial risk factor for developing cardiovascular diseases. High sugar levels in the blood can narrow the blood vessels and damage them, eventually leading to various cardiovascular diseases. Obesity is another risk factor for developing cardiovascular diseases associated with diabetes. Type two diabetes is caused more in people with obesity.

2.3.5 Being Overweight or Obese

Obesity can develop high blood pressure or diabetes, two of the most significant risk factors for developing cardiovascular diseases. If a person is obese and his body mass index (BMI) is more than twenty-five, he is at a substantial risk of developing cardiovascular diseases. Suppose a man's waist measurement is about thirty-seven inches or ninety-four centimeters, or a woman's waist measurement is about thirty-one and a half inches or eighty centimeters. In that case, they are also at a substantial risk of developing cardiovascular diseases.

2.3.6 Inactivity

Regular exercise and a healthy diet help keep the heart healthy and maintain a healthy weight. Suppose a person is inactive and does not exercise regularly. In that case, he can get obese or have high cholesterol or high blood pressure, leading to an increased risk of developing cardiovascular diseases.

2.3.7 Family History of Cardiovascular Diseases

Cardiovascular diseases can be inherent. If there is a history of cardiovascular diseases in a person's immediate family, he can also be more likely to develop them. But suppose the family member got diagnosed with cardiovascular disease after a particular age, fifty-five for men and sixty-five for women. In that case, this does not put the person at risk of developing them. The person is only at risk of developing them because of family history if the family members got diagnosed with it before that particular age.

2.3.8 Ethnic Background

Cardiovascular diseases are more common in areas like the United Kingdom than the Caribbean, Africa, or South Asia. The reason is that people from these areas of the world are more likely to have other risk factors associated with cardiovascular diseases like type 2 diabetes or high blood pressure. So ethnic background also plays a role in increasing or decreasing the risk of developing cardiovascular diseases.

2.3.9 Other Risk Factors

Some other risk factors associated with developing cardiovascular diseases are:

- Alcohol – Cholesterol, blood pressure and weight can increase due to the consumption of alcohol.
- Diet – High blood pressure or high cholesterol can be caused by an unhealthy diet.
- Gender – men are more likely to develop cardiovascular diseases than women at an earlier age.
- Age – people of age more than fifty years are more likely to develop cardiovascular diseases. This risk increases with the increase in age.

2.4 Prevention of Cardiovascular Diseases

The risks of developing cardiovascular diseases can be lowered by following a healthy lifestyle. Even if a person has cardiovascular disease, it can be stopped from worsening by following a healthy lifestyle. Ways to prevent cardiovascular diseases are discussed below [21].

2.4.1 Quit Smoking

Smoking is a significant risk factor that can cause all cardiovascular diseases. A person should try to give up smoking as soon as possible to reduce the risk of developing cardiovascular diseases. Quitting an addiction like smoking can be difficult, but trying to quit it in steps can drastically reduce the risk factor for developing cardiovascular diseases.

2.4.2 Have a Balanced Diet

A person should follow a balanced and healthy diet to stay fit and reduce the risk of developing cardiovascular diseases. A balanced diet should include:

- At least four to five portions of vegetables and fruits daily
- Plenty of whole-grain foods and fiber
- A low level of sugar and salt
- A low level of saturated fat
- Healthier sources of fat like olive oil, nuts, and seeds

2.4.3 Exercise Regularly

Exercise is particularly important for leading a healthy life. Adults should do at least two and a half hours of moderate-to-intense activity weekly, like brisk walking or cycling [30]. Suppose two

and a half hours seem difficult to a person. In that case, he should start at a lower level but try to reach this intensity and duration of exercise.

2.4.4 Maintain a Healthy Weight

A person should have a BMI lower than 25. If it is higher than that, he should try to lose weight by following a healthy diet and exercising regularly. According to the National Institute of Diabetes and Digestive and Kidney Disorders, developing cardiovascular diseases reduces if a person loses five to ten percent of their body weight [31].

2.4.5 Medication

Proper medication prescribed by a medical expert can help to reduce the risk of developing cardiovascular diseases if it is particularly high. Medications may include tablets for reducing blood pressure, a low-dose aspirin for preventing blood clots, and statins for lowering blood cholesterol levels.

2.5 Summary

Cardiovascular diseases are a prominent cause of mortality worldwide, with an escalating incidence. This group of diseases encompasses conditions affecting both the heart and blood vessels. Symptoms associated with cardiovascular diseases include cold sweats, dizziness or lightheadedness, fatigue and nausea, shortness of breath, as well as pain or discomfort in areas such as the back, jaw, elbows, left shoulder, or arms. Pressure or chest pain is also commonly observed. Various risk factors contribute to the development of cardiovascular diseases, such as high blood pressure, smoking, elevated cholesterol levels, and diabetes, among others. Implementing effective preventive measures is crucial, including smoking cessation, adoption of a balanced diet, regular exercise, maintenance of a healthy weight, and adherence to prescribed medications.

Chapter 3 - Literature Review

Recent years have seen an upsurge in the healthcare industry's medical data. That has led to a need for automated methods to predict the risk of developing cardiovascular diseases in a person. For this reason, researchers have applied machine-learning and deep-learning techniques for solving different problems. This chapter summarizes some of the valuable contributions that have been made to this field.

3.1 Traditional Machine Learning Techniques

Dinh *et al.* [32] used the National Health and Nutrition Examination Survey (NHANES) dataset to predict diabetes and cardiovascular diseases as these are the two leading causes of death. They performed an exhaustive search and selected the maximum number of features they could get for cardiovascular diseases in that dataset. A total of 131 features were selected from 2007-2014 because the maximum number of features were available in this period. The features were related to physical activities, as inactivity is a significant risk factor for causing cardiovascular diseases. Age, diastolic blood pressure, systolic blood pressure, the occurrence of chest pain, and self-reported weight were found to be the key contributors. A weighted ensemble model was developed by combining machine learning algorithms, including gradient boosting, random forests, SVM, and logistic regression.

In a different study, Alaa *et al.* [33] proposed an AutoPrognosis model based on a collection of modelling pipelines for machine learning. The pipelines consisted of four steps: feature processing, calibration, classification, and data imputation. Data imputation was done using eight algorithms, such as missForest and matrix completion. There were nine algorithms employed, such as Principal Component Analysis (PCA), fast Independent Component Analysis (ICA), and Linear SVM. Twenty algorithms, such as Linear SVM, AdaBoost, decision trees, logistic regression, and linear discriminant analysis (LDA), were employed. While the sigmoid function was applied for calibration.

Mohan, Thirumalai, and Srivastava [34] presented a novel approach for predicting heart disease, employing a hybrid model that combines machine learning techniques. It incorporated a set of thirteen risk factors as crucial variables in the prediction process including age, gender, chest pain, blood pressure, serum cholesterol level, fasting blood sugar level, ECG, highest achieved heart rate, exercise-induced angina, ST depression, exercise ST segment, the result of colored

Fluoroscopy, and status of the heart. Hybrid Random Forest with a Linear Model (HRFLM) was used for feature selection and prediction.

Pahwa *et al.* [35] used Support Vector Machine Recursive Feature Elimination (SVM-RFE) for feature selection among thirteen features, including age, gender, result of colored Fluoroscopy, exercise ST segment, ST depression, exercise-induced angina, highest achieved heart rate, results of resting electrographic, fasting blood sugar, serum cholesterol, blood pressure, chest pain type, and status of the heart. They then used a hybrid of two machine learning techniques, namely random forest, and Naïve Bayes, to predict heart disease and accurately perform this task.

Rajathi and Radhamani [36] used fifteen risk factors to assess the risk of developing Rheumatic Heart Disease using the Streptococcus Pyogenes dataset. The factors used by them were id, age, gender, chest pain location, pain provoked by exertion, pain relieved after rest, chest pain type, resting blood pressure, family history of coronary artery disease, serum cholesterol level, resting ECG result, diagnosis of heart disease, resting heart rate, duration of the exercise test, and (month of exercise ECG reading. The authors created a feature vector by combining these factors and utilized the k-Nearest Neighbors (kNN) algorithm for classifying them into risk and non-risk categories. Subsequently, they employed the Ant Colony Optimization (ACO) algorithm to initialize the population and conduct an optimized solution search.

Chethana [42] used a dataset of 7,639 patients to predict CVD risk using a kNN algorithm. kNN is a simple and intuitive machine learning algorithm that classifies new data points based on the class of the k-nearest neighbors in the training data. The author used feature selection techniques to select the most notable features for CVD risk prediction. The study concluded that the KNN algorithm could be an effective tool for predicting CVD risk.

Farooq *et al.* [37] selected eighteen clinical features from the Rapid Access Chest Pain Clinics (RACPC) dataset to predict whether chest pain was due to some developing cardiovascular problem or any other problem. Some of the clinical data was missing for various instances in the dataset. They used the Expectation Maximization (EM) technique and the Bernoulli mixture model for learning this missing data. They then used a decision tree-based classifier for feature section, pattern recognition, and classification of chest pain into cardiac pain and other pain.

Mahboob, Irfan, and Ghaffar [38] adopted a voting technique by using three machine learning approaches: the kNN, SVM, and Artificial Neural Network (ANN) to predict Coronary Heart Disease. They used the main detrimental features associated with cardiovascular diseases like smoking, alcohol, etc., and passed them to the three machine learning models. The class (risk or non-risk) predicted by the most classifiers out of the three was selected as the final class through the voting technique.

Raja *et al.* [39] proposed a framework for predicting Cardiovascular Syncope Disease. They used an ensemble of SVM, Gini Index, and Naïve Bayes classifiers. The data was obtained from the Armed Forces Institute of Cardiology & the National Institute of Heart Diseases (AFIC & NIHD). The classifiers used different combinations of features out of the thirty-one features extracted from the data. The features included age, gender, blurred vision, shortness of breath, fainting after a meal, volume depletion, traumatic injury, nausea, etc. The ensemble used the majority voting technique and the class predicted by the most classifiers out of the three was selected as the final class.

Wang [40] used a dataset of 70,000 patients to predict CVD risk using an SVM algorithm. He also considered twelve additional aspects of the patients, such as age and sensor-measured diastolic and systolic blood pressure. The dataset's dimensionality was reduced by the author using feature selection, and the most pertinent features for predicting CVD risk were chosen. He also used data augmentation techniques to increase the size of the dataset, which can improve the performance of the model.

Using a dataset of 303 patients, each with 14 clinical variables, Pal and Parija [41] employed a random forest algorithm to predict the risk of CVD. Random forest is an ensemble machine learning approach that mixes numerous decision trees to increase the accuracy and generalization of the model. The authors employed feature selection techniques to select the most notable features for CVD risk prediction. They evaluated the performance of the model using sensitivity, specificity, and accuracy measures. The study concluded that the random forest algorithm was effective in predicting CVD risk and could be used as a screening tool.

Singh and Jain [43] used a dataset of 13 attributes of patients to predict CVD risk using an ANN algorithm. The study employed genetic algorithm optimization techniques to improve the performance of the ANN model. Genetic algorithm is a metaheuristic optimization technique that

uses concepts from natural selection and genetics to find the best combination of model parameters. The study reported an accuracy of 90.3% in predicting CVD risk, suggesting that ANN could be an effective tool for predicting CVD risk.

Maji and Arora [44] used a dataset of 270 patients containing 13 attributes for each patient to predict CVD risk using a decision tree algorithm. The authors used feature selection techniques to select the most notable features for CVD risk prediction. The study reported an accuracy of 83.5% in predicting CVD risk. The study concluded that the DT algorithm was a useful tool for predicting CVD risk and could be used as a screening tool. Table 3.1 provides a literature review of some additional research that has used machine learning algorithms to predict the risk of developing cardiovascular diseases.

Table 3.1 Literature Review of Some Additional Research

Authors	Machine Learning Technique	Detection Accuracy (%)	Detection Miss Rate (%)
Dinesh *et al.* [45]	SVM	77	23
Frizzell *et al.* [46]	C Statistic	58.9	41.1
Rosendael *et al.* [47]	XG-Boost	70.1	29.9
Weng *et al.* [48]	Random Forest	68	32
Quesada *et al.* [49]	Naïve Bayes	69	31

3.2 Research Gaps

The literature found on the prediction of cardiovascular diseases is lacking in many ways. There are a lot of shortcomings in the research that has been conducted so far. These shortcomings include limited classifier variation and the absence of a data preprocessing-cleaning layer. Moreover, the results of the research have a lot of room for improvement. As most cardiovascular diseases are dangerous or fatal if not diagnosed at an early stage, there is a desperate need to decrease the rate of false negatives.

3.3 Summary

Work has been done regarding predicting cardiovascular diseases in patients using machine learning and deep learning techniques. The machine learning techniques used include gradient boosting, random forests, SVM, logistic regression, PCA, LDA, etc. Similarly, deep learning techniques including LSTM network, DBN, deep neural network, and RNN have also been used. But there is a gap in the literature that needs to be filled.

Chapter 4 Methodology

Cardiovascular diseases are considered threatening diseases. They have an exceedingly high mortality rate. Certain behavioral and biochemical factors can increase a person's risk of developing them. High-risk people need to change their lifestyles to avoid or reduce the risk of developing them. So, there is a great need for an intelligent system for the prognosis of cardiovascular disease, which can prevent, mitigate, and reduce the risk of people developing them. Traditional methods for predicting cardiovascular diseases are time-consuming, expensive, and prone to error. To overcome these challenges, machine learning algorithms are being used to predict cardiovascular diseases. The health care industry contains lots of medical data. Therefore, machine learning algorithms make effective decisions in predicting heart diseases. They help medical experts improve the accuracy of the prognosis of diseases by giving a second opinion and decreasing the processing time. This thesis proposes a novel intelligent model that finds appropriate features using machine learning-based techniques to find an accurate model to predict cardiovascular diseases. Different algorithms including Logistic Regression, kNN, Decision Tree Classifier, Ada Boost, SVM, SVM Polynomial Kernel, SVM Sigmoid Kernel, SVM Stochastic Gradient Descent, Random Forest Classifier, XG Boost, Naïve Bayes Classifier, Cat Boost, and ANN were implemented with different combinations of features.

The proposed intelligent model reduced the number of false positives and false negatives. This was done by transferring man’s knowledge to machine intelligence by changing the learning model accordingly. Moreover, labelled data was given to the model in the training process, so the model also incorporated changes over multiple iterations. The decision of human specialists was evaluated and streamlined using the proposed model. Consequently, predefined observations were used to predict the disease. Thus, the model used supervised feature learning for detecting human errors while giving labels and correcting the data. Moreover, new segments of data were also detected through the scoring method. The computing cost of the proposed model was also very efficient.

4.1 The Proposed Model for the Prediction of Cardiovascular Diseases

Figure 4.1 shows a visual representation of the proposed model for predicting cardiovascular diseases. The legend shows how the steps are represented in the visual representation. The details

of those steps are shown in the same-colored blocks below the legend. For example, the red block is for the data preprocessing module.

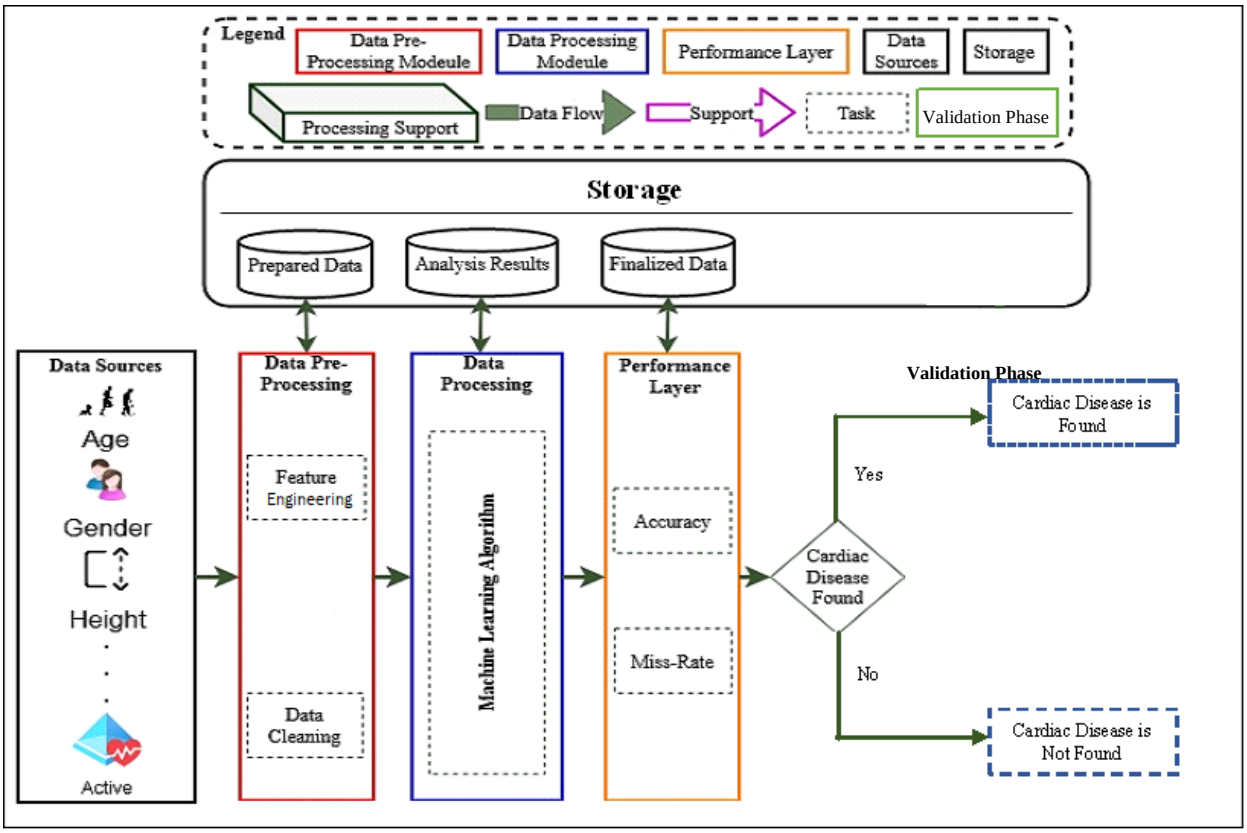

Figure 4.1 Visual Representation of the Proposed Model

"Data source" was connecting a system to external data sources. Data was collected and saved in the database using various methods from this module's given sources. It was then prepared according to the requirements for the next modules. This step was called "Data preprocessing." This step involved feature selection, feature extraction, and data cleaning. Data cleaning means deleting duplicate and faulty data, validating it, and standardizing it. Prepared data was again saved in the database.

The next module was "Data Processing" which involved extracting useful information from the data using machine learning algorithms and techniques. These algorithms and techniques were stored in the database as analysis results. The following module was the "Performance Layer" that improved the findings of the previous module using big data techniques. Miss rate and accuracy

were calculated in this step. The database then stored the finalized data that told whether there was a prediction of cardiovascular diseases in the patient or not.

4.2 Data Source and Data Set

The data set for this study had been taken from the Kaggle Repository [55]. It is publicly available and divided into two classes: positive and negative. The positive class meant a prediction of cardiovascular disease in the patient. In contrast, the negative class meant that there was no prediction of cardiovascular disease in the patient. The dataset consisted of eleven input features. The proposed model gave one feature as an output: "yes" or "no." But the height and weight given in the dataset were used to calculate Body Mass Index (BMI) and fed as a single feature. Table 1.1 shows the detailed structure of the dataset.

Table 4.1 Detailed Structure of the Dataset

Original Dataset			**Preprocessed**		
Sr. No.	**Features**	**Datatype**	**Formula**	**Features**	**Datatype**
1	Age	Quantitative	Days-years	Age	Quantitative
2	Gender	Qualitative		Gender	Qualitative
3	Height	Quantitative	(weight/height) /100) *2)	BMI	Quantitative
4	Weight	Quantitative			
5	ap_hi	Quantitative		ap_hi	Quantitative
6	ap_lo	Quantitative		ap_lo	Quantitative
7	cholesterol	Qualitative		cholesterol	Qualitative
8	Gluc	Qualitative		Gluc	Qualitative
9	Smoke	Qualitative		Smoke	Qualitative
10	Alco	Qualitative		Alco	Qualitative
11	Active	Qualitative		Active	Qualitative
12	Cardio	Qualitative		Cardio	Qualitative

4.2.1 Data Visualization

The dataset used in this study was pre-labelled. It had eleven features, out of which ten were independent and given input to the model. In contrast, one was dependent and was taken as an output from the model, which was then compared with the given label in the dataset. There was a total of 70,000 cases taken for this study.

4.2.1.1 Histogram of Classes

Figure 4.2 shows the histogram of the classes. The blue bar represents the negative class, which means that there was no prediction of cardiovascular disease in the patient, whereas the yellow bar represents the positive class which means that there was a prediction of cardiovascular disease in the patient. It can be seen in the figure that there is a 1:1 ratio of both classes. This shows that the dataset was balanced.

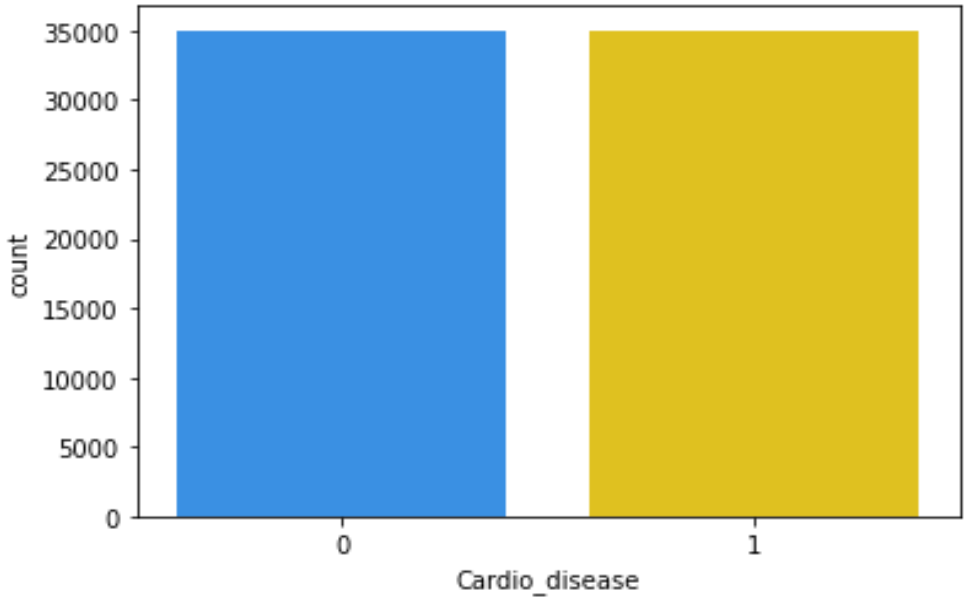

Figure 4.2 Histogram of the Classes

4.2.1.2 Relationship Between Age and the Risk of Developing Cardiovascular Diseases

Figure 4.3 shows the relationship between age and the risk of developing cardiovascular diseases. It can be seen that people of age between fifty and sixty-five are at a higher risk of developing

cardiovascular diseases as compared to any other age group. People of age less than forty-five have a lower risk. This is because middle-aged people are very concentrated on their work and providing their family with a good living standard. They often ignore the need for annual medical tests. Hence, the heavy work pressure only allows them to visit the hospital when they are sick. Thus, they have a high prevalence of risk of developing cardiovascular diseases.

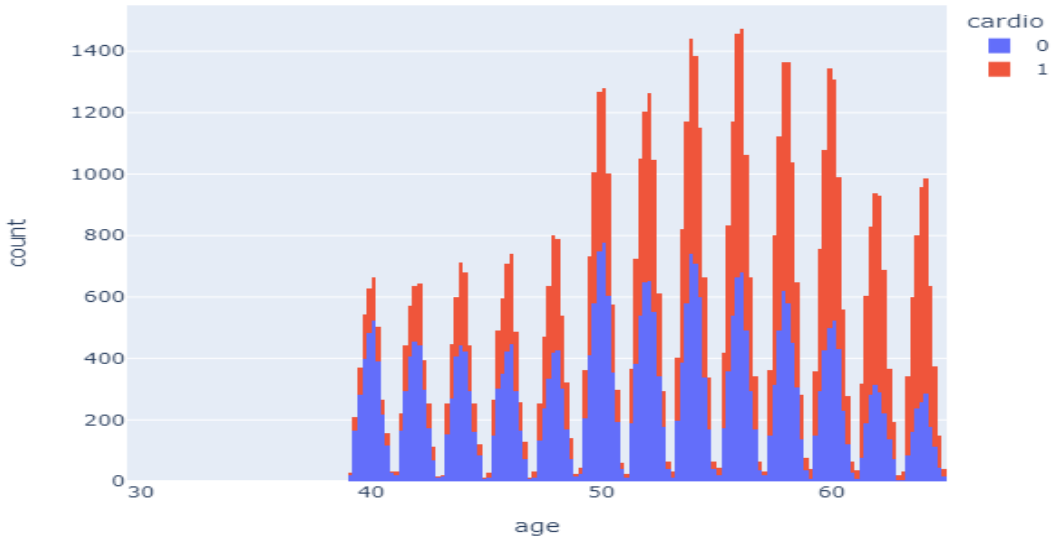

Figure 4.3 Relationship Between Age and the Risk of Developing Cardiovascular Diseases

4.2.1.3 Distribution of Independent Features

Figure 4.4 shows the distribution of seven independent features in the dataset with three outcomes at maximum. The top left column shows the distribution of gender. There are 45,530 male patients and 24,470 female patients. Next to it is the distribution of cholesterol. 52,385 patients have normal cholesterol, 9,549 have above normal cholesterol, whereas 8,066 patients have well above normal cholesterol.

Moreover, 59,479 have normal glucose level, 5,190 have above normal glucose level, whereas 5,331 patients have well above normal glucose level. Moving on, 63,831 patients are non-smokers, and 6,169 patients are smokers. On the other hand, 66,236 patients are non-alcoholic, and the rest 3,764 patients are alcoholic. 13,739 patients are not active whereas 56,261 patients are active.

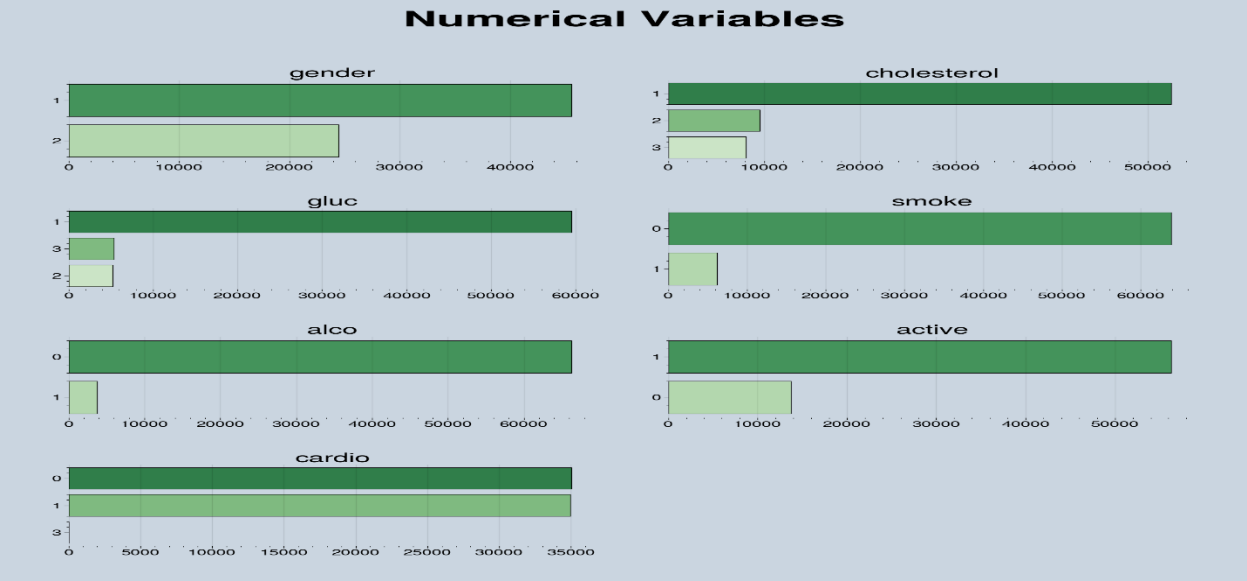

Figure 4.4 Distribution of Independent Features in the Dataset

4.2.1.4 Parallel Projection of the Dataset

Figure 4.5 shows the graph of standard deviation of all the features. It is the measure of the variability or dispersion of the data points. It indicates how spread out the data is around the mean. From the figure it can be seen that ap_ lo and cholesterol had exceedingly high standard deviations. This indicated that there might be some outliers in the data. Outliers were dealt with further in the study.

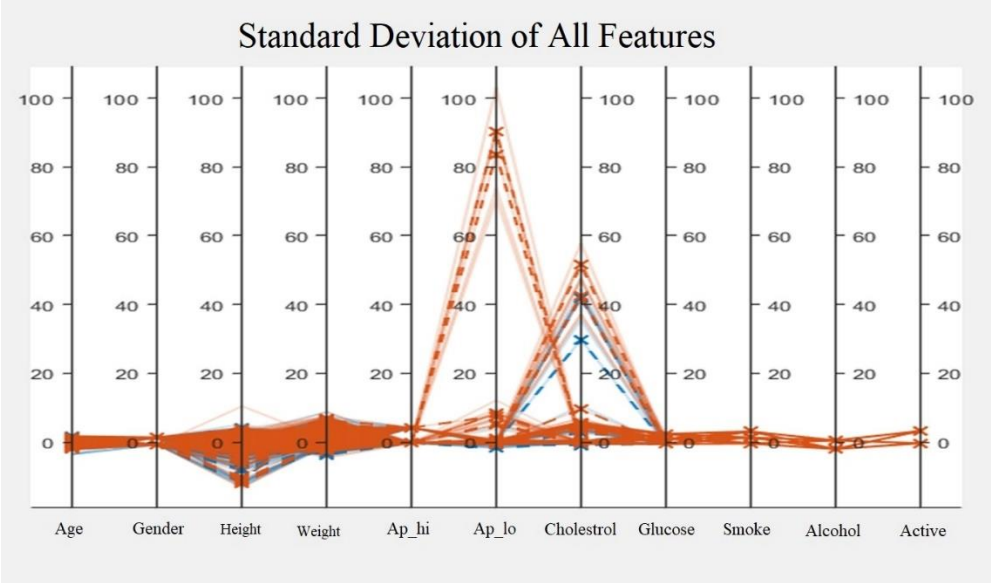

Figure 4.5 Standard Deviation of All Features

Figure 4.6 shows the parallel projection of the dataset. It can be seen that a higher ratio of males is inactive as compared to females. Moreover, from the patients with well above normal cholesterol, a higher ratio belongs to the positive class compared to the negative class. The same is the case with patients that have above normal cholesterol. This means that cholesterol plays a particularly important role in increasing the risk of developing cardiovascular diseases.

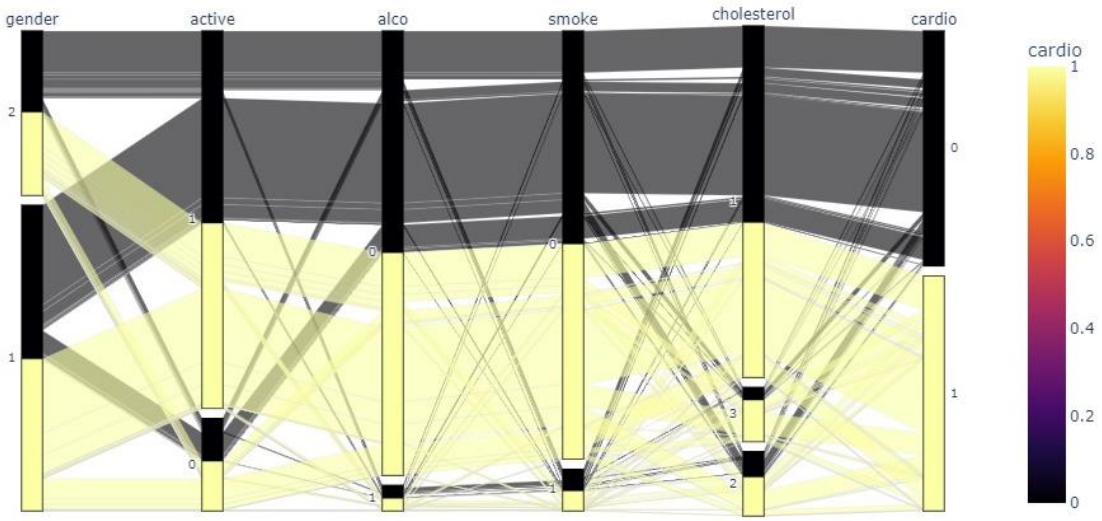

Figure 4.6 Parallel Projection of the Dataset

4.2.1.5 Correlation of Variables

Figure 4.7 shows the correlation matrix of the independent variables. We can see that height and gender have slightly strong correlations. What is more, the correlation between gender and age is -0.025. This indicates that men are more likely to develop cardiovascular diseases as compared to women at an earlier age. Moreover, ap_hi and ap_lo have a connection between them as both of them are concerned with the heart. Their correlation also suggests the same as they have a correlation of 0.7. Slightly strong correlations can also be seen between gender and smoke; height and weight; cholesterol and glucose; and smoke and alcohol. Finally, this simple visualization does not clearly show the relationship of some variables. Their connections will be evaluated using various other techniques in the later part.

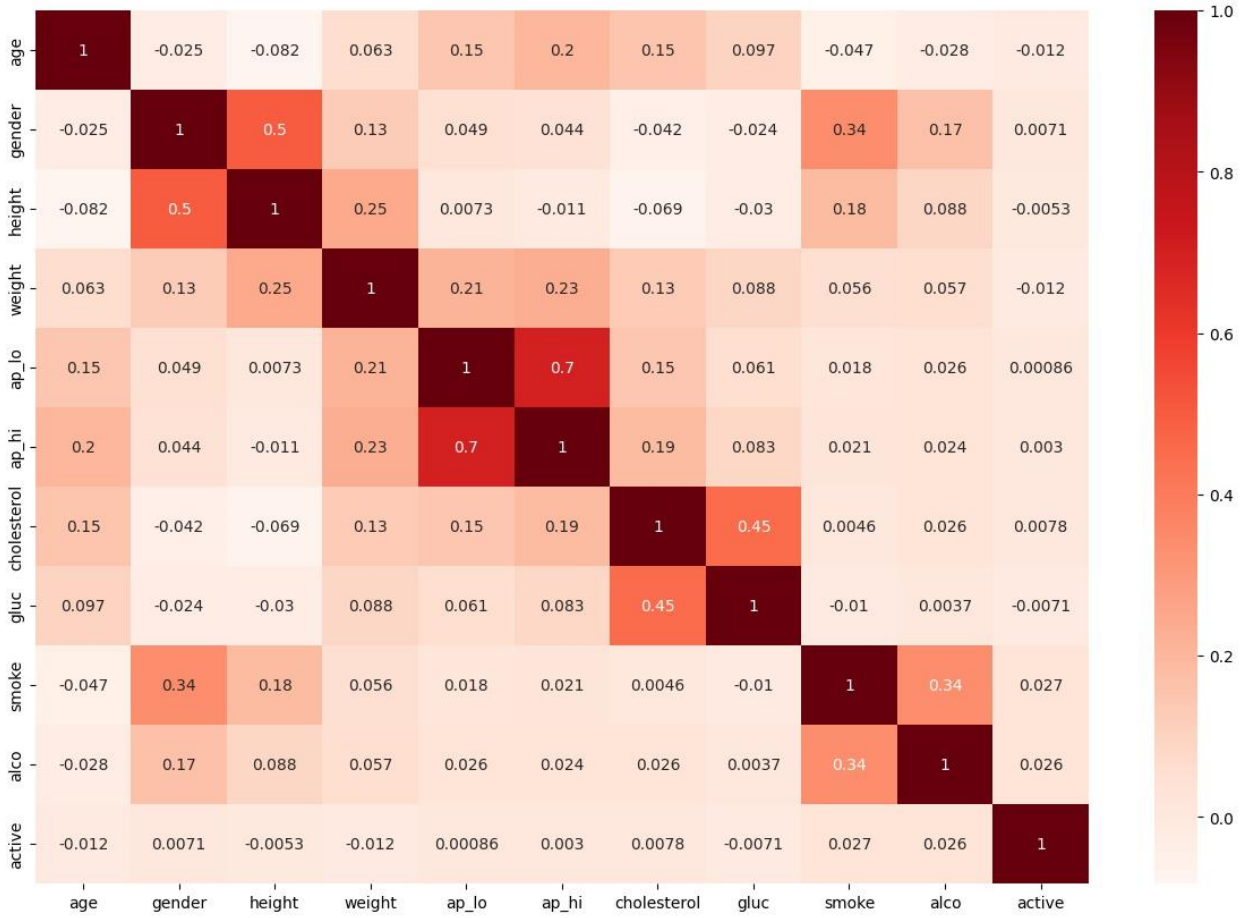

Figure 4.7 Correlation Matrix of Independent Variables

4.2.1.6 Box Plot of the Relative Risks

The box plot of the relative risks in Figure 4.8 represents the data outliers. The vertical lines inside the boxes show medians of the distribution of relative risk. Their lower and upper quartiles are represented by hinges that are at the left and right of the boxes. Extreme data points are connected through horizontal lines to their respective hinges. After collection of data, it was sent to the next module, the data preprocessing module.

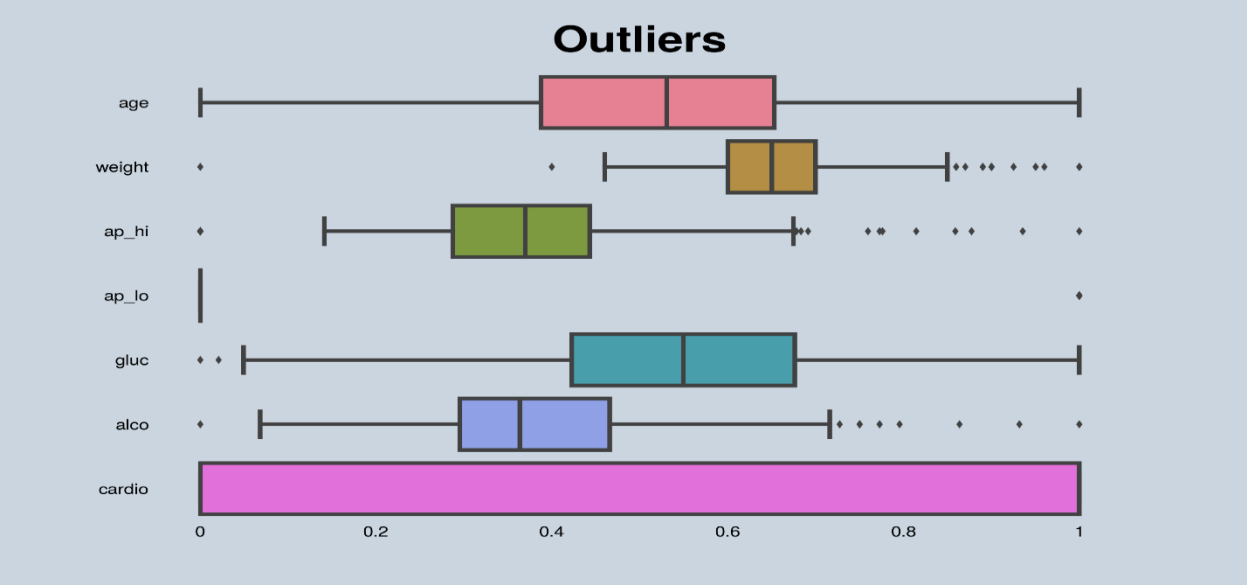

Figure 4.8 Outliers of Features

4.2.1.7 Data Visualization in Two Dimensions

The dataset was visualized in two dimensions using PCA. It is one of the best techniques that is being used for data visualization along with dimensionality reduction. Transforming the data into a lower-dimensional space makes it easier to interpret and visualize. The inherent patterns and structures within the data are more visible and clearer using PCA.

PCA identifies the orthogonal directions capturing the maximum variance in the data. These orthogonal directions are called principal components. The singular value decomposition or a covariance matrix of the dataset is used to compute the orthogonal directions. The principal components are arranged in a manner that the one for the highest amount of variation comes at the start, whereas the one for the lowest amount of variation comes at the end.

These components are then used to project the dataset onto the subspace. In this way, only the most significant patterns and features are retained, omitting the others. It then becomes easier to analyze and explore the trends, clusters, and relationships in the transformed data.

Using PCA, the dataset for this study was visualized in a two-dimensional representation, as compared to the twelve-dimensional representation. Figure 4.9 shows this representation.

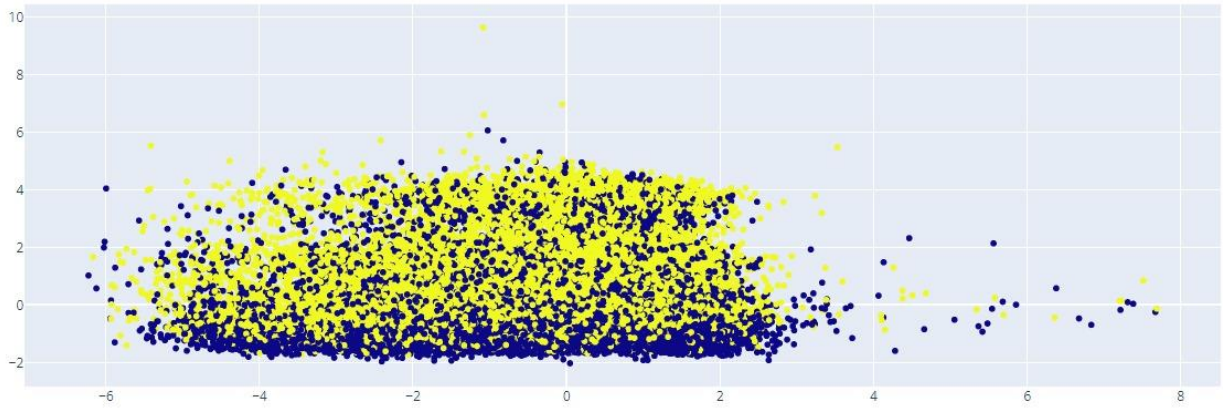

Figure 4.9 Two-Dimensional Representation of Dataset using PCA.

It can be seen in Figure 4.9 that there were some clusters in the dataset where the data points were grouped together. However, despite the presence of clusters, a considerable amount of variance and disturbance was also present in the dataset. Clustering was not sufficient to explain them. There could be many reasons behind this disturbance like overlapping data points, noise, or outliers. It caused heterogeneity and complexity in the overall dataset.

4.2.2 Outlier Detection

Scatter plots can be useful for outlier detection because they allow us to visually examine the distribution of data points and identify any observations that deviate significantly from the overall pattern. Outliers, as the name suggests, are data points that lie far away from the majority of the other points. Figure 4.9 shows the scatter plots of some independent variables; age, ap_hi, ap_lo, height, and weight. Using these scatter plots, outliers were easily detected. The points further away from the cluster or pattern were the outliers. They have been marked in Figure 4.10.

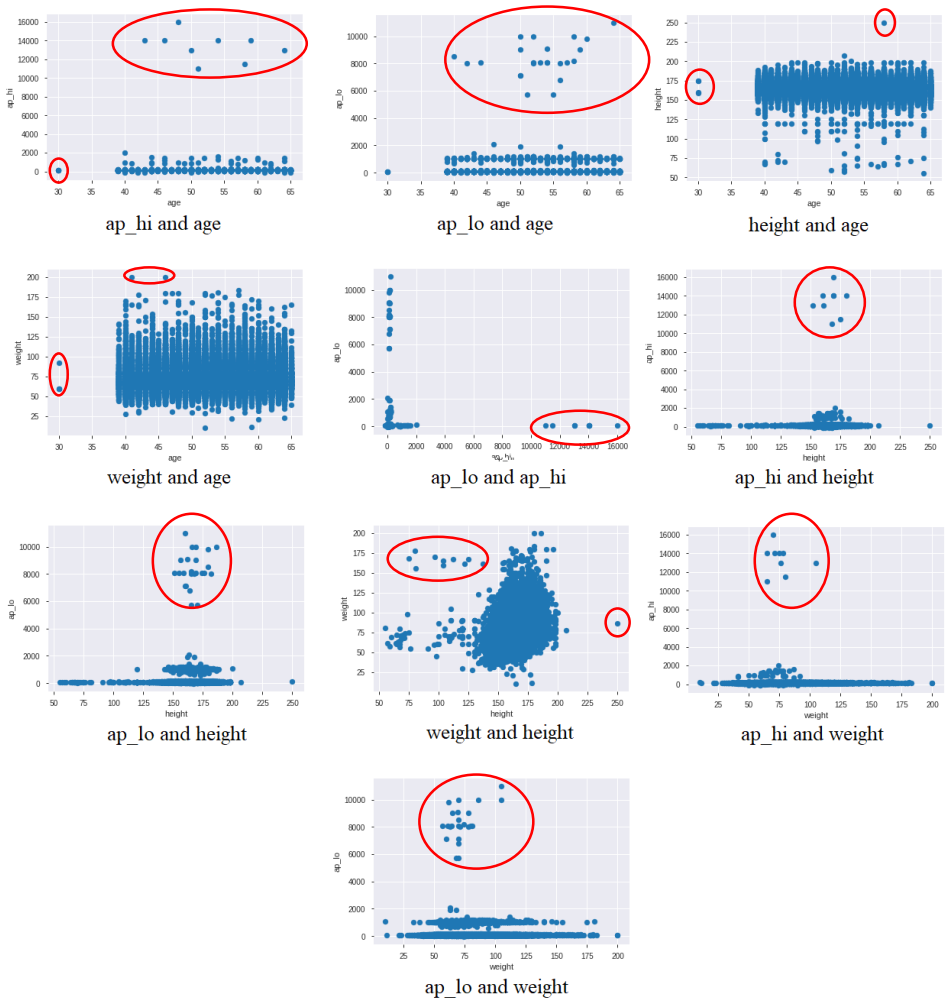

Figure 4.10 Scatter Plots of Independent Variables

The following algorithm was used to detect outliers.

Step 1: Calculate the interquartile range for the outlier threshold by subtracting the first quartile from third quartile.

Step 2: Calculate the upper limit for outliers by adding 1.5 times the interquartile range to the third quartile.

Step 3: Calculate the lower limit for outliers by subtracting 1.5 times the interquartile range from the first quartile.

Step 4: Identify the outliers; the data points that were below the lower limit, or above the upper limit.

4.2.3 Outlier Removal

The data points that were below the lower limit were replaced with the lower limit calculated in the previous step using interquartile range and the first quartile. Similarly, the values that were above the upper limit were replaced with the upper limit calculated in the previous step using interquartile range and the third quartile. Once the outliers had been removed, box plots were created to analyze the distribution of the remaining data. Figure 4.11 shows box plots of some independent variables; age, ap_hi, ap_lo, height, and weight after outlier removal. The vertical lines inside the boxes show medians of the distribution of relative risk. Their lower and upper quartiles are represented by hinges that are at the left and right of the boxes. Extreme data points are connected through horizontal lines to their respective hinges.

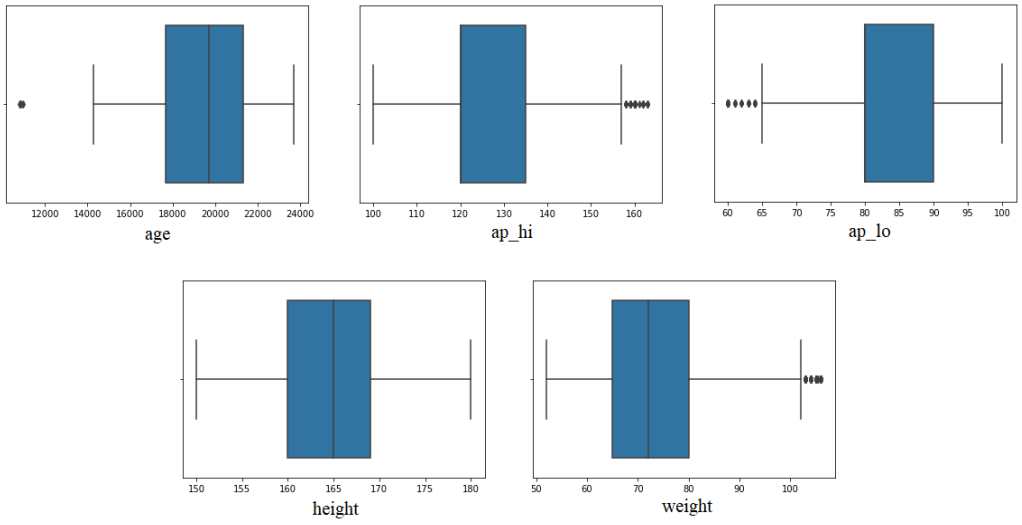

Figure 4.11 Box Plots of Independent Variables after Outlier Removal

4.2.4 **Standardization**

Z-score normalization is performed for the scaling of Quantitative features to have a mean of 0 and standard deviation 1. This transformation makes features directly comparable. Also, Z-score normalization is highly affected by the outliers in the dataset as mean is gets effected by the extreme values therefore outliers need to be removed before hand.

```
Z-Score DataFrame:
             age    height    weight     ap_hi     ap_lo
0      -0.494871  0.504405 -0.983822 -1.145970 -0.126993
1       0.246220 -1.252519  0.996426  1.033989  1.086733
2      -0.198434  0.065174 -0.811626  0.307336 -1.340720
3      -0.791307  0.650816  0.738133  1.760641  2.300460
4      -0.791307 -1.252519 -1.500408 -1.872622 -2.554446
...          ...       ...       ...       ...       ...
69993  0.098002  1.090047 -0.295040  0.307336  1.086733
69994  0.690875  0.065174  0.565937  1.760641 -0.126993
69995 -0.050216  0.504405  0.221546 -0.419317 -0.126993
69998  1.135530 -0.227647 -0.122845  0.670662 -0.126993
69999  0.394439  0.797226 -0.122845 -0.419317 -0.126993

[60142 rows x 5 columns]
```

Figure 4.1212 Z-Score normalization of the qualitative features of dataset

4.3 The Proposed Framework for the Flow of Data

Figure 4.12 shows the framework for flow of data in the proposed methodology. The dataset, after cleaning of outliers was passed to the preprocessing layer where features were selected, extracted, and cleaned. The dataset was then split into training and testing data. The training data was sent to the training layer where different machine learning algorithms were applied, and the performance layer checked whether the learning criteria was met or not. Both these steps on the data were iterative until the learning criteria was met. The testing data that had been stored in the storage was then retrieved and passed to the validation phase where the final trained model was also loaded. Here the model predicted whether there was a risk of developing cardiovascular disease or not in the testing data.

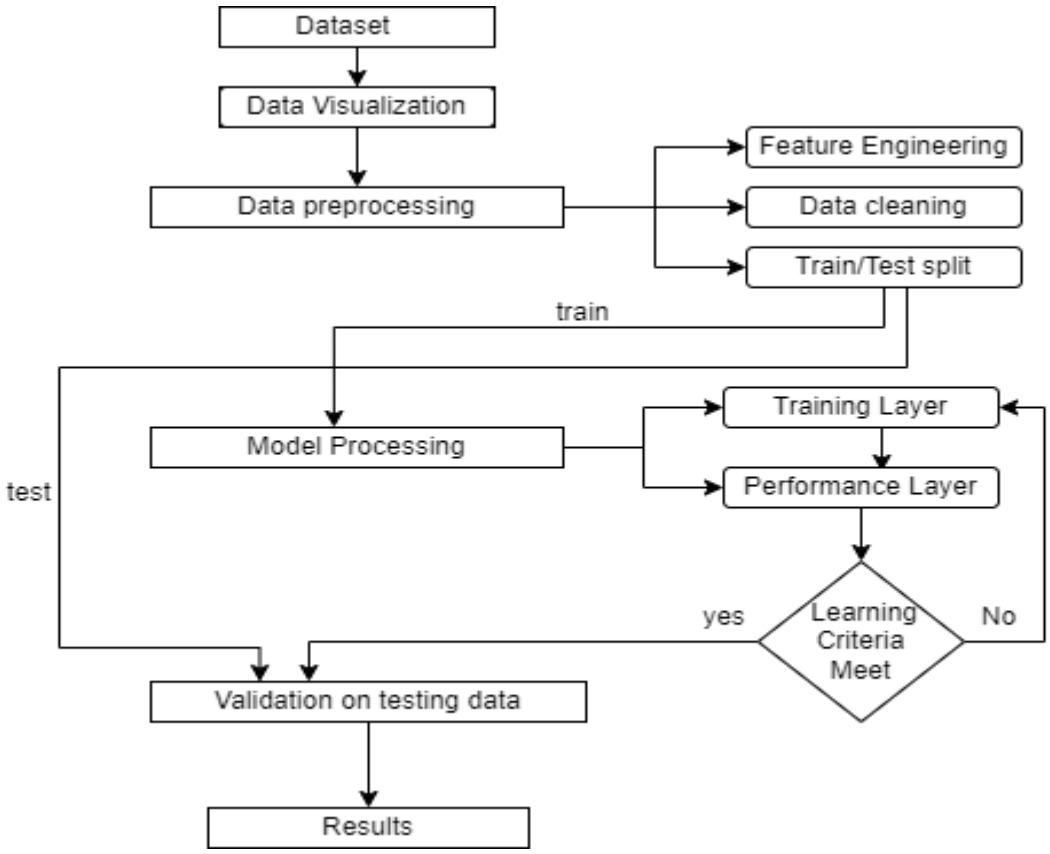

Figure 4.13 Framework for the Flow of Data in the Proposed Methodology

4.4 Data Preprocessing

The preprocessing module divided the dataset into training and testing. The preprocessed data was split into 80% training data set and 20% testing data set from each class. After splitting, the prepared testing data was stored in the storage and the training data was sent to the next module, the data processing module. The dataset comprised 70,000 cases.

4.5 Data Processing and Performance Layer

The data processing module predicted cardiovascular disease using different machine learning algorithms. They are discussed in detail below. Whereas the performance layer checked whether the learning criteria had been met or not. These two modules worked iteratively until the learning criteria had been met using accuracy and miss rate. There were two phases in these modules: the training and validation phases.

The machine learning algorithms used in this study can be divided into two categories: non-tree-based algorithms and tree-based algorithms. Non-tree-based algorithms included Fine kNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, Weighted KNN, Subspace KNN, Subspace Discriminant, Gaussian Naïve Bayes, Kernal Naïve Bayes, and Extreme Learning Machine (ELM). Whereas tree-based algorithms included Decision Tree, Random Forest, Fine Tree, Medium Tree, Coarse Tree, Boosted Tree, Bagged Tree, RUSBoosted Tree, Optimizable Tree, and Gradient Boosting Tree. The results of all these classifiers were compared and the Gradient Boosting Tree algorithm was selected as the final classifier for this study because it produced the best results. The results and comparison are discussed in the next chapter.

4.5.1 Non-Tree Based Algorithms

- **Fine kNN**

kNN algorithm groups the data by using the proximity between the data points. It is assumed that the data points that are similar to each other have less distance between them. The training data is only stored in the kNN algorithm while it is in the training stage. The computations are done at the time of prediction. Fine kNN algorithm uses a single neighbor to classify a data point in a class while predicting. This helps to distinguish the classes finely. Fine kNNs are difficult to interpret.

- **Medium KNN**

Medium kNN algorithm is a variation of the kNN algorithm just like the fine kNN algorithm. The difference between the two is that the medium kNN algorithm uses ten neighbors to classify a data point in a class while predicting. This helps to distinguish the classes moderately. Medium kNNs are difficult to interpret.

- **Coarse KNN**

In the coarse kNN variation, a hundred neighbors are used to classify a data point in a class while predicting. This helps to distinguish the classes coarsely. Coarse kNNs are also difficult to interpret.

- **Cosine KNN**

A Cosine kNN algorithm is similar to a medium kNN algorithm. It also uses ten neighbors to classify a data point in a class while predicting. This helps to distinguish the classes moderately. The difference is that it uses cosine similarity for finding the distance between two neighboring datapoints. Cosine similarity is calculated by finding the cosine of the angle that is made in between two vectors. This distance decides whether the two data points are to be considered neighbors or not. This is done on the basis of cosine principles. These principles state that two data points are considered as less similar to each other as the distance between them increases. Cosine kNNs are difficult to interpret. Equation 4.1 shows the formula to calculate the cosine distance between two vectors u and v.

$$1 - \frac{u \,.v}{|u| \,.|v|} \qquad 4.1$$

where |u| and |v| are the norms of u and v vectors, respectively.

- **Cubic KNN**

A Cubic kNN algorithm is also similar to a medium kNN algorithm. It also uses ten neighbors to classify a data point in a class while predicting. This helps to distinguish the classes moderately. The difference is that it uses cubic distance for finding the distance between two neighboring datapoints. If there are two n-dimensional vectors; u and v, then the cubic distance between them can be calculated by Equation 4.2. It is difficult to interpret cubic kNN.

$$\sqrt[3]{\sum_{i=1}^{n} |u_i - v_i|^3} \qquad 4.2$$

- **Weighted KNN**

Like the previous two algorithms, Cubic kNN algorithm is also similar to a medium kNN algorithm. It also uses ten neighbors to classify a data point in a class while predicting. This helps to distinguish the classes moderately. The difference is that it uses weighted Euclidean distance for finding the distance between two neighboring datapoints. Euclidean distance is the actual length of the distance between two points. Weights are assigned to them in weighted Euclidean distance. If there are two n-dimensional vectors; u and v, then the weighted Euclidean distance between them can be calculated by Equation 4.3. Weighted kNNs are difficult to interpret.

$$\sqrt[2]{\sum_{i=1}^{n} w_i (x_i - y_i)^2} \qquad 4.3$$

where $0 < w_i < 1$ and $\sum_{i=1}^{n} w_i = 1$.

- **Subspace KNN**

Subspace kNN algorithm is an ensemble of kNNs. It uses the stochastic process for selecting a random number of components from the feature vector. Only the features that have been selected play a role in calculating the distance while comparing a test sample to a prototype. This means that a new set of neighbors will be computed every time because every time a new subspace is selected. The test sample is assigned to a subspace through a majority vote on all the sets of neighbors. The algorithm has medium flexibility and is difficult to interpret.

- **Subspace Discriminant**

Subspace Discriminant algorithm is an ensemble of multiple discriminant learners. The algorithm has medium flexibility and is difficult to interpret.

- **Gaussian Naïve Bayes**

Gaussian Naïve Bayes algorithm uses Gaussian distribution and probabilistic approach. An assumption is made while using this algorithm that all the features play an independent role in predicting the class of the testing data. Predictions using each feature are made separately and then all the predictions are combined to give a probability for each class that how much chance there is that the testing data belongs to that class. The testing data is finally assigned to the class with the highest probability. As there is no option to change any parameters, the flexibility of the algorithm is low, and it is easy to interpret.

- **Kernal Naïve Bayes**

The Kernal Naïve Bayes algorithm uses a weighing function to estimate the density functions or conditional expectations of random variables. The weighing function itself does not have a fixed structure so it uses all the data points to reach a prediction. The flexibility of the algorithm is medium, and it is easy to interpret.

- **Extreme Learning Machine**

Extreme Learning Machine (ELM) is a feedforward neural network. It can be used for performing the tasks of feature learning, compression, sparse approximation, clustering, regression, and classification. There are different variations of ELM. The number of layers of hidden nodes can

range from one-to-many layers. The parameters of these hidden nodes can be tuned. They can also be used as is by randomly initiating them or inheriting them from their ancestors and then never tuning them. Different activation functions are used to decide whether a neuron should be activated or not.

4.5.2 Tree Based Algorithms

- **Decision Tree Algorithm**

A Decision Tree is a hierarchical model that has leaf nodes, internal nodes, branches, and root nodes. The structure of a tree is like a flowchart. All the tests done on the data are denoted by the internal nodes and all its outcomes are denoted by the branches respectively. The class labels are denoted by the leaf nodes. The construction of a decision tree involves splitting the training dataset into parts on the basis of the attributes unless some stopping criterion is fulfilled. The stopping criterion could include that the maximum depth of the tree is reached, or the minimum limit of samples required for splitting a node is met, etc. Different metrics like Gini impurity or entropy are used to select the best attribute for splitting dataset. These metrics measure the level of randomness or impurity in the dataset. The purpose is to reduce the impurity or increase the gain of information after every split.

- **Random Forest Algorithm**

The Random Forest algorithm is a combination of various decision trees. The combined output of all the decision trees is used to find a single result. In other words, we can say that this algorithm is an extended form of the bagging method as an uncorrelated forest consisting of multiple decision trees is created by this algorithm using feature randomness and bagging. The set of features are divided randomly which reduces correlation among decision trees. Random forests select features from a random subset of them, whereas a decision tree splits features into all subsets and then selects them. This is the main difference between them.

- **Fine Tree Algorithm**

Fine Tree is a decision tree in which precise categorization is done by creating numerous divisions. The maximum number of splits that can be done in a fine Tree are hundred. It can have many leaves and its flexibility is high.as this tree has numerous leaves, so it gives a good accuracy on

training dataset, but not so good on the testing dataset. The reason is that a leafy tree may overfit, resulting in the testing accuracy dropping.

- **Medium Tree Algorithm**

Medium Tree is a decision tree which uses a moderate number of branches to finely differentiate classes. It has a maximum limit of twenty splits. The flexibility of a medium tree is moderate.

- **Coarse Tree Algorithm**

Coarse Tree is a decision tree in which there are only a few leaves. The maximum number of splits that can be done in a Coarse Tree are four. The flexibility of a coarse tree is low because of the few leaves. This makes it easy to interpret. The dwindling number of leaves causes the coarse tree to not achieve a high accuracy. This makes the algorithm robust so that the testing accuracy is almost equal to the training accuracy.

- **Boosted Trees Algorithm**

Boosted Trees algorithm is an ensemble of multiple dependent decision trees; typically, AdaBoost with Decision Tree learners They are connected in such a way that each subsequent tree is dependent on the previous trees. The trees work in a chain, and each tree fits the residual of the previous trees, thus aiding it to learn. There is a minor risk of less coverage in boosted trees, but this type of tree has the capability to improve training and testing accuracy. One drawback of boosted tree is that it requires more learners and parameter tuning. The algorithm has medium to high flexibility and is difficult to interpret.

- **Bagged Trees Algorithm**

Boosted Trees algorithm is also an ensemble of multiple random forest trees and decision trees. The dataset is divided into random subsets and the base classifiers are each fit on one of these subsets. Every classifier gives an individual prediction, and all these predictions are aggregated to give a final prediction either by averaging or voting. The variance of an individual decision tree is reduced by using bagged tree. The algorithm has high flexibility and is difficult to interpret.

- **RUSBoosted Trees Algorithm**

RUSBoosted Trees Algorithm is an efficient and simple algorithm that is a combination of data boosting and sampling algorithms. It works best for imbalanced data, providing high training and testing accuracy. This algorithm is remarkably effective for skewed data. It has medium flexibility and is difficult to interpret.

- **Optimizable Tree Algorithm**

Optimizable Tree algorithm is the one in which hyperparameters are tuned in order to search the space for the values that will provide optimized results. The most commonly optimized hyperparameters are the maximum number of splits and the split criterion. The maximum number of splits can range from 1 to max (2, n–1) where n is the total number of observations. For split criterion, there are a few options from Maximum Deviance Reduction (MDR), Twoing rule, and Gini's diversity index. Some additional hyperparameters of optimizable tree algorithm include maximum surrogates per node and surrogate decision splits.

- **Gradient Boosting Tree Algorithm**

The Gradient Boosting Tree algorithm is used to optimize the prediction value of some model. This is done through a learning process that consists of multiple successive steps. Every iteration of this learning process actually adjusts the values of biases, weights, or coefficients that are being used for the input values of the variables for predicting the target values. The end-goal is to minimize the value of the loss function. The loss function is used to calculate the difference between the targeted and predicted values. The word gradient is used in the name of this algorithm because incremental adjustments are made to improve the result of the algorithm in every step of the learning process. Moreover, the word boosting is used in the name of this algorithm because the process of improving the training and testing accuracies until an optimum threshold value is reached, is accelerated using the boosting method. Figure 4.13 shows that working of a Gradient Boosting Tree algorithm.

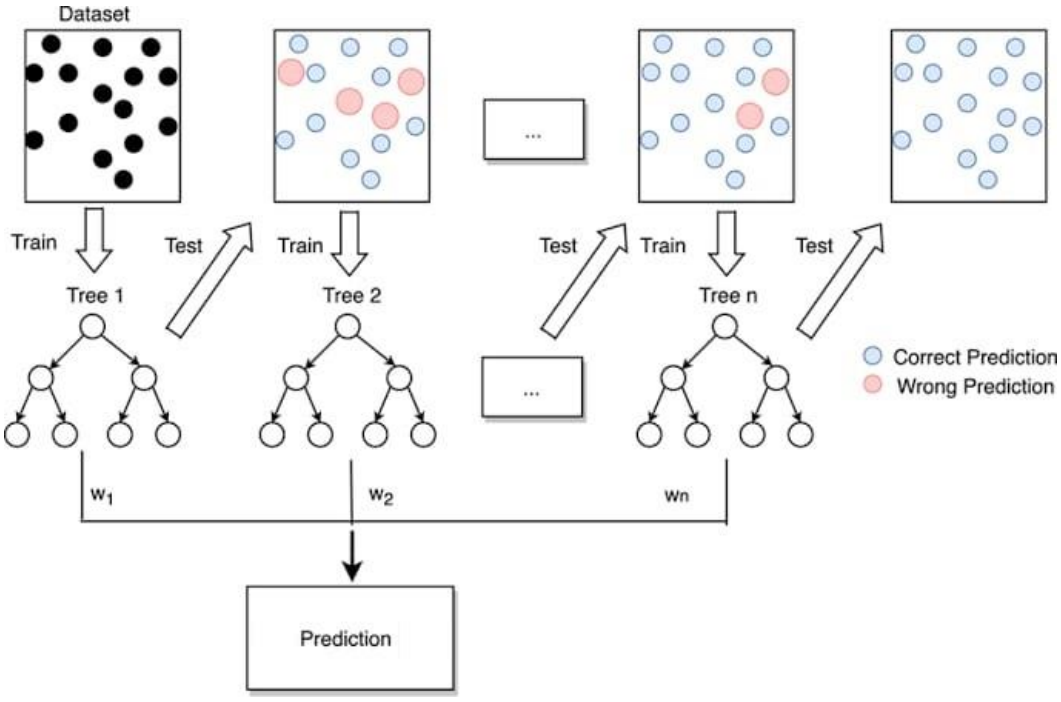

Figure 4.14 Working of a Gradient Boosting Tree Algorithm

4.6 Summary

The model proposed in this study consisted of five main modules. The first module was data collection. The dataset had been taken from the Kaggle Repository. There was a total of 70,000 cases. Ten features were given as input and binary classification was done i.e., absence or presence of the risk of developing cardiovascular diseases. The second module was data preprocessing. In this module noisy data was mitigated using various techniques including data cleaning, normalization, and moving average that use mean imputation method. The next two modules were data processing and performance layer. These two modules worked iteratively until the learning criteria had been met. The data processing module predicted cardiovascular disease using multiple machine learning algorithms. Whereas the performance layers checked whether the learning criteria had been met or not using accuracy and miss rate. The last module was validation module. This module predicted whether there was a risk of developing cardiovascular disease or not in the testing data.

Chapter 5 Experimental Results

The experimentation and results chapter focuses on the data set used for this study and the process of evaluating the proposed model's performance. The data set was obtained from the Kaggle Repository and is widely recognized as a reliable source for researching cardiovascular diseases [55]. To ensure a robust evaluation, the preprocessed data was divided into an 80% training data set and a 20% testing data set. Data preprocessing was performed beforehand to eliminate errors and ensure consistency in the data.

During the training process, the model was provided with labeled data, allowing it to adapt and learn from the input to improve its performance over multiple iterations. In contrast, the testing data set was unlabeled and contained additional types of attacks do not present in the training data set. This setup enhanced the proposed model's credibility and accuracy in identifying various scenarios. To assess the system's efficiency, simulation results played a crucial role. These results provided valuable insights into the model's performance and helped in accurately predicting the system's effectiveness. Throughout this chapter, we will delve into the specifics of the experimentation process, present the results obtained, and analyze the performance of the proposed model based on various evaluation metrics.

5.1 Experimental Setup

The performance of the proposed model was evaluated using Python programming language [57]. The training and testing experiments were performed on Jupyter Notebook tool [58] on CPU Intel(R) Core i7-5th Gen with 16GB RAM. The classification time was a few milliseconds per record.

5.2 Performance Metrics

The performance evaluation of the proposed model involved ten key metrics. These metrics provided insights into various aspects of the model's performance, including the confusion matrix, sensitivity, specificity, accuracy, miss rate, fallout, Likelihood Positive Ratio (LR+), Likelihood Negative Ratio (LR-), Positive Predictive Value (PPV), and Negative Predictive Value (NPV).

The confusion matrix is a useful tool for summarizing the number of correct and incorrect predictions made by the model for each class. It compares the model's predicted results with the actual input labels. The confusion matrix consists of four parameters: True Positive (TP), True

Negative (TN), False Positive (FP), and False Negative (FN). TP represents the number of records belonging to the positive class that are correctly predicted as positive, while TN represents the number of records belonging to the negative class that are correctly predicted as negative. FP indicates the number of records belonging to the negative class that are incorrectly predicted as positive, and FN represents the number of records belonging to the positive class that are incorrectly predicted as negative. Figure 5.1 provides a visual representation of the confusion matrix.

By analyzing these performance metrics, a comprehensive evaluation of the proposed model's effectiveness can be obtained, enabling a deeper understanding of its predictive capabilities and potential areas for improvement.

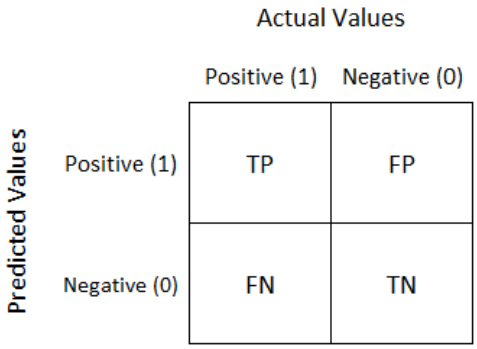

Figure 5.1 Confusion Matrix

Accuracy is the measure of how often the model has given a correct prediction. It measures the overall correctness of the model. It can be calculated using the formula in Equation 5.1.

$$Accuracy = \frac{(TN+TP)}{(TN+FN+TP+FP)} \quad 5.1$$

Sensitivity is also known as the True Positive Rate (TPR) or Recall. It is the measure of how often the model has correctly classified positive instances. It can be calculated using the formula in Equation 5.2.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad 5.2$$

Specificity is also known as the True Negative Rate (TNR). It is the measure of how often the model has correctly classified negative instances. It can be calculated using the formula in Equation 5.3.

$$Specificity = \frac{TN}{(TN+FP)} \quad 5.3$$

Miss rate is also known as the False Negative Rate (FNR). It is the measure of how often the model has given a wrong prediction. It measures the overall error of the model. It can be calculated using the formula in Equation 5.4.

$$Miss\ Rate = \frac{(FN+FP)}{(TN+FN+TP+FP)} \quad 5.4$$

The above equation can also be written as Equation 5.5.

$$Miss\ Rate = 1 - Accuracy \quad 5.5$$

Fallout is also known as the False Positive Rate (FPR). It is the measure of how often the model has incorrectly classified negative instances as positive. It can be calculated using the formula in Equation 5.6.

$$Fallout = \frac{FP}{(TN+FP)} \quad 5.6$$

Likelihood Positive Ratio (LR+) is the measure of how much chances of a positive prediction are to be a true positive case as compared to a false positive case. It can be calculated using the formula in Equation 5.7.

$$Likelihood\ Positive\ Ratio = \frac{True\ Positive\ Rate}{False\ Positive\ Rate} \quad 5.7$$

Likelihood Negative Ratio (LR+) is the measure of how much chances of a negative prediction are to be a true negative case as compared to a false negative case. It can be calculated using the formula in Equation 5.8.

$$Likelihood\ Negative\ Ratio = \frac{True\ Negative\ Rate}{False\ Negative\ Rate} \quad 5.8$$

Positive Predictive Value (PPV) is also known as Precision. It is the measure of how many times a positive prediction was actually a positive instance. It can be calculated using the formula in Equation 5.9.

$$Positive\ Predictive\ Value = \frac{TP}{(TP+FP)} \quad 5.9$$

Negative Predictive Value (NPV) is the measure of how many times a negative prediction was actually not a negative instance. It can be calculated using the formula in Equation 5.10.

$$Negative\ Predictive\ Value = \frac{TN}{(TN+FN)} \qquad 5.10$$

5.3 Results and Discussion

The performance layer in the proposed model checked whether the learning criteria had been met or not. The learning criteria were checked in terms of detection accuracy and miss rate. The dataset consisted of 70000 records that was further divided into 80% training data set (56000 records) and 20% testing data set (14000 records). Figure 5.2 and Figure 5.3 show the confusion matrices for the training and testing data set, respectively. Figure 5.2 shows that out of the total 56000 records selected for training phase, 29901 records belonged to the positive class and the remaining 26099 records belonged to the negative class. 24867 patients were correctly predicted to have cardiovascular disease. 4460 patients did not have a risk of developing cardiovascular disease, but the model incorrectly predicted them to have the risk. 21639 patients were correctly predicted to not have a risk for developing cardiovascular disease. And 5034 patients had the risk of developing cardiovascular disease, but the model incorrectly predicted them as safe. Overall, the model only predicted 16.95% records incorrectly.

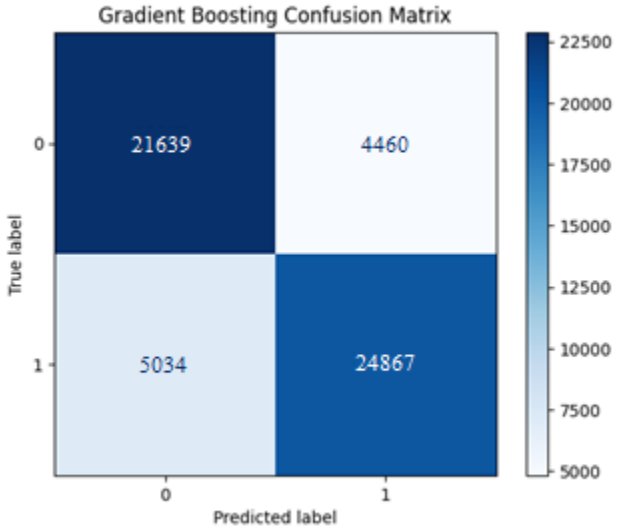

Figure 5.2 Confusion Matrix for Training Data Set

Figure 5.3 shows that out of the total 14000 records selected for testing phase, 7502 records belonged to the positive class and the remaining 6498 records belonged to the negative class. 5760 patients were correctly predicted to have cardiovascular disease. 1228 patients did not have a risk of developing cardiovascular disease, but the model incorrectly predicted them to have the risk.

5270 patients were correctly predicted to not have a risk for developing cardiovascular disease. And 1742 patients had the risk of developing cardiovascular disease, but the model incorrectly predicted them as safe. Overall, the model only predicted 21.2% records incorrectly.

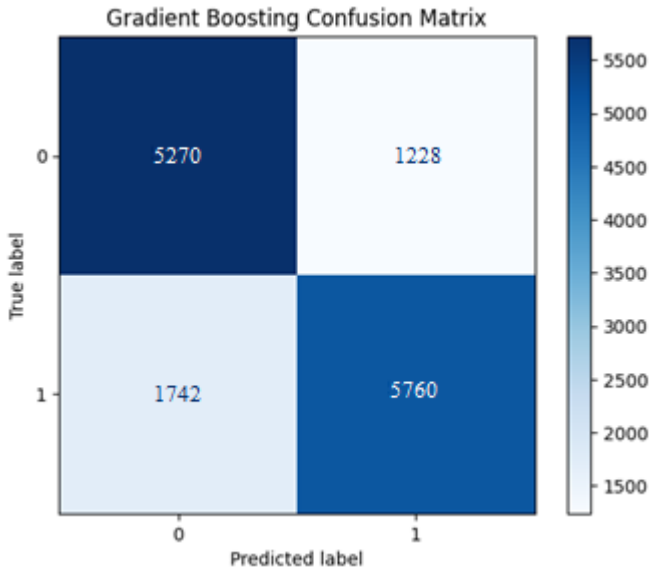

Figure 5.3 Confusion Matrix for Testing Data Set

Table 5.1 provides an overview of the proposed model's performance in both the training and testing phases, highlighting various performance metrics. In the training phase, the model exhibited accuracy, sensitivity, specificity, miss rate, and precision values of 83.05%, 83.16%, 82.91%, 16.95%, and 84.79% respectively. Similarly, in the testing phase, it achieved accuracy, sensitivity, specificity, miss rate, and precision values of 78.78%, 76.78%, 81.10%, 82.43%, and 18.90% respectively. Furthermore, the training phase results yielded a fallout, LR+ (Positive Likelihood Ratio), LR- (Negative Likelihood Ratio), and NPV (Negative Predictive Value) of 17.09%, 4.87, 4.89, and 81.13% respectively, while the testing phase resulted in a fallout, LR+, LR-, and NPV of 18.90%, 4.06, 3.82, and 75.16% respectively. These findings demonstrate the model's robust performance across different evaluation metrics in both the training and testing phases.

Table 5.1 Performance of the Proposed Model in the Training and Testing Phase (%)

	Training	**Testing**
Accuracy	83.05	78.78

Sensitivity (TPR)	83.16	76.78
Specificity (TNR)	82.91	81.10
Miss Rate (FNR)	16.95	21.22
Fallout (FPR)	17.09	18.90
Likelihood Positive Ratio (LR+)	4.87	4.06
Likelihood Negative Ratio (LR-)	4.89	3.82
Positive Predictive Value (Precision)	84.79	82.43
Negative Predictive Value (NPV)	81.13	75.16

5.4 Proposed Model vs. Other Machine Learning Algorithms and State-of-the-Art Models

The proposed model introduces an innovative approach that utilizes machine learning-based techniques to identify relevant features for accurate prediction of cardiovascular diseases. Several algorithms, including Fine Tree, Medium Tree, Coarse Tree, Fine kNN, Medium kNN, Coarse kNN, Cosine kNN, Cubic kNN, Weighted kNN, Boosted Trees, Bagged Trees, Subspace Discriminant, Subspace kNN, RUSBoosted Trees, Gaussian Naïve Bayes, Kernel Naïve Bayes, Optimizable Tree, Decision Tree, and Random Forest, were implemented with different feature combinations. Their performances were compared to the proposed Gradient Boosting Tree model.

Figure 5.4 illustrates the training and testing accuracies of all the models. The proposed model achieved training and testing accuracies of 83.05% and 78.78%, respectively. Furthermore, the proposed model exhibited training and testing miss rates of 16.95% and 21.22%, respectively. Notably, the testing accuracy of the proposed model outperformed other well-known machine learning techniques by 5.39% to 28.79%, as demonstrated in Figure 5.4.

Table 5.2 presents additional performance metrics of the implemented algorithms. The algorithm that demonstrated the best performance for each metric is highlighted in bold within the respective columns. It is evident that the Gradient Boosting Tree algorithm outperformed other techniques in terms of most performance metrics, except for LR-, PPV, and NPV.

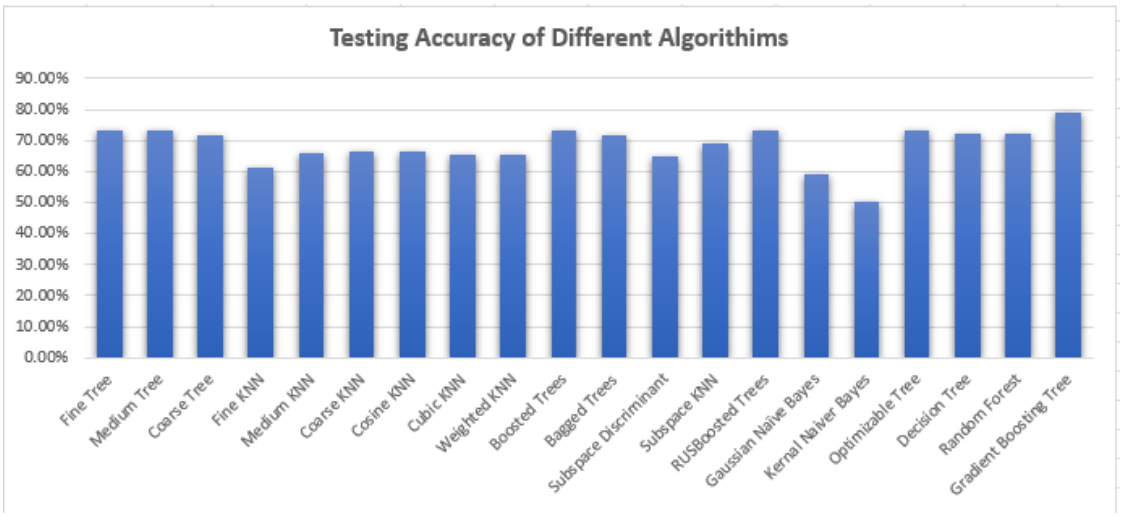

Figure 5.4 Comparison of Testing Accuracies of Different Well-Known Algorithms with the Proposed Model

Table 5.2 Comparison of Testing Performance Metrics of Different Well-Known Algorithms with the Proposed Model

	Sensitivity	Specificity	Miss Rate	Fallout	LR+	LR-	PPV	NPV
Fine Tree	70.93%	75.93%	26.84%	24.07%	2.95	2.83	78.54%	67.77%
Medium Tree	70.55%	76.38%	26.90%	23.62%	2.99	2.84	79.35%	66.84%
Coarse Tree	68.96%	75.39%	28.30%	24.61%	2.80	2.66	79.01%	64.38%
Fine KNN	61.00%	61.58%	38.72%	38.42%	1.59	1.59	62.70%	59.86%
Medium KNN	66.50%	65.49%	34.02%	34.51%	1.93	1.93	64.48%	67.48%
Coarse KNN	64.97%	67.71%	33.78%	32.29%	2.01	2.00	70.49%	61.95%
Cosine KNN	66.99%	66.01%	33.51%	33.99%	1.97	1.97	65.08%	67.90%
Cubic KNN	66.05%	64.86%	34.57%	35.14%	1.88	1.88	63.58%	67.28%
Weighted KNN	64.70%	65.77%	34.79%	34.23%	1.89	1.89	67.06%	63.36%
Boosted Trees	70.98%	76.58%	26.55%	23.42%	3.03	2.88	79.38%	67.50%
Bagged Trees	70.89%	72.73%	28.23%	27.27%	2.60	2.58	73.94%	69.60%
Subspace Discriminant	63.64%	65.98%	35.28%	34.02%	1.87	1.87	68.76%	60.67%
Subspace KNN	67.55%	70.49%	31.09%	29.51%	2.29	2.27	72.83%	64.98%
RUSBoosted Trees	70.55%	76.38%	26.90%	23.62%	2.99	2.84	79.35%	66.84%
Gaussian Naïve Bayes	55.70%	72.18%	40.95%	27.82%	2.00	1.76	**88.69%**	29.38%
Kernal Naiver Bayes	47.80%	49.96%	50.04%	50.04%	0.96	**1.00**	0.28%	**99.69%**
Optimizable Tree	70.69%	76.44%	26.79%	23.56%	3.00	2.85	79.35%	67.05%
Gradient Boosting Tree	**76.78%**	**81.10%**	**21.21%**	**18.90%**	**4.06**	3.82	82.43%	75.16%
Decision Tree	71.02%	75.21%	27.18%	24.79%	2.86	2.77	79.12%	66.23%
Random Forest	70.71%	74.69%	27.62%	25.31%	2.79	2.70	79.52%	64.73%

Table 5.2 shows the training and testing accuracies of the ELM classifier on the dataset for different combinations of activation functions, and number of hidden layers. The number of hidden layers selected for this experimentation ranged from twenty-five to two hundred. The four activation functions selected were hardlim, Radial Bases, Sigmoid, and Sin. It can be seen that the final classifier i.e., Gradient Boosting Tree algorithm outperformed all the combinations of the ELM classifiers in terms of accuracies.

Table 5.3 Training and Testing Accuracies of the ELM Classifier

Training Phase									
		Hidden Layers							
		25	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Activation Functions	**hardlim**	0.64	0.65	0.66	0.67	0.68	0.67	0.68	0.68
	Radial Bases	0.60	0.61	0.63	0.65	0.65	0.65	0.66	0.66
	Sigmoid	0.67	0.69	0.70	0.71	0.71	0.71	0.71	0.72
	Sin	0.62	0.64	0.64	0.66	0.67	0.67	0.69	0.69
Testing Phase									
		Hidden Layers							
		25	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Activation Functions	**hardlim**	0.64	0.65	0.66	0.67	0.67	0.68	0.68	0.68
	Radial Bases	0.60	0.62	0.63	0.65	0.65	0.65	0.67	0.66
	Sigmoid	0.67	0.69	0.70	0.70	0.71	0.71	0.71	0.71
	Sin	0.62	0.63	0.64	0.66	0.68	0.67	0.70	0.70

Table 5.3 shows a comparison of performance in terms of accuracy and miss rate in training and testing phases of the proposed model with some state-of-the-art models for prediction of cardiovascular diseases. The highest accuracies for both phases and lowest miss rates for both phases are in bold. It can be seen that in training phase, Dinesh *et al.*'s model [45] outperformed the other models but in testing phase, the proposed model performed the best. A clear improvement in the testing accuracy and miss rate can be seen. The proposed model performed around 1.78%

to 19.88% better in terms of both testing accuracy and miss rate as compared to other state-of-the-art models.

Table 5.4 Comparison of Accuracy and Miss Rate in Training and Testing Phases of the Proposed Model with Some State-Of-The-Art Models

Authors	**ML Technique**		**Accuracy (%)**	**Miss Rate (%)**
Dinesh *et al.* [45]	SVM	Training (60%)	88	12
		Testing (40%)	77	23
Frizzell *et al.* [46]	C statistic	Training (65%)	59.2	40.8
		Testing (35%)	58.9	41.1
Rosendael *et al.* [47]	XG-Boost	Training (80%)	77.1	22.9
		Testing (20%)	70.1	29.9
Weng *et al.* [48]	Random Forest	Training (75%)	70	30
		Testing (25%)	68	32
Quesada *et al.* [49]	Naïve Bayes	Training (80%)	73	27
		Testing (20%)	69	31
Proposed Model	Gradient Boosting Tree	Training (80%)	83.05	16.95
		Testing (20%)	**78.78**	**21.22**

Sometimes, the selection of dataset for the study can also affect the performance of a model. So, the results of the proposed methodology were also compared with the results of some other studies that can be found in literature using the same dataset as used in this study. This comparison can be seen in Table 5.5. The highest accuracy and lowest miss rate are in bold. It can be seen that the

proposed model performed the best. A clear improvement in accuracy and miss rate can be seen. The proposed model performed around 3.88% to 7.18% better in terms of both accuracy and miss rate as compared to other state-of-the-art models.

Table 5.5 Comparison of Accuracy and Miss Rate of the Proposed Model with Git-Repository Models

Studies	Classifier	Accuracy (%)	Miss Rate (%)
Niteesh [59]	Random Forest	73.04	26.96
Dritsas *et al.* [60]	Logistic Regression	72.06	27.94
Adeboye *et al.* [61]	Gradient Boosting	74.9	25.1
Yue *et al.* [62]	Random Forest	71.6	28.4
Proposed Model	**Gradient Boosting Tree**	**78.78**	**13.35**

5.5 Summary

The proposed model has demonstrated exceptional performance in predicting cardiovascular disease using machine learning algorithms. This success can be attributed to significant improvements made in the preprocessing layer and the careful selection of hyperparameters. The model achieved remarkable metrics, with accuracy, sensitivity, specificity, miss rate, and precision values of 78.78%, 76.78%, 81.10%, 82.43%, and 18.90% respectively. Additionally, the fallout, LR+ (Positive Likelihood Ratio), LR- (Negative Likelihood Ratio), and NPV (Negative Predictive Value) were recorded at 18.90%, 4.06, 3.82, and 75.16% respectively. The model's classification time per record was mere milliseconds. These findings highlight the model's exceptional efficiency and accuracy in predicting cardiovascular disease.

Chapter 6 Conclusion

The prevalence of cardiovascular diseases is rapidly increasing, making it one of the leading causes of death globally. This encompasses a range of conditions that affect the heart and blood vessels. Various risk factors contribute to the development of cardiovascular diseases, including diabetes, high cholesterol, smoking, and high blood pressure, among others. To prevent such diseases, individuals are advised to quit smoking, maintain a balanced diet, engage in regular exercise, manage their weight, and adhere to prescribed medications. Previous research has explored the use of machine learning and deep learning techniques to predict cardiovascular diseases in patients. However, this study aims to address a gap in existing research by providing further insights into the subject matter.

The model proposed in this study consisted of five main modules: data collection, data preprocessing, data processing, performance layer and prediction. The proposed model gave very promising results. It has proven to be very efficient in predicting cardiovascular disease in a person using Gradient Boosting algorithm. The model had 78.78%, 76.78%, 81.10%, 82.43%, 18.90% accuracy, sensitivity, specificity, miss rate, and precision, respectively. Furthermore, the study yielded promising results with regard to fallout, LR+ (Positive Likelihood Ratio), LR- (Negative Likelihood Ratio), and NPV (Negative Predictive Value), which demonstrated values of 18.90%, 4.06, 3.82, and 75.16% respectively. The classification process exhibited remarkable efficiency, with an average processing time of just a few milliseconds per record. Additionally, the detection time per record was approximately 0.2137 seconds. Notably, the proposed model highlighted superior performance compared to several established machine learning algorithms and state-of-the-art models, further reinforcing its efficacy and relevance in the field.

In future work, our focus will be on enhancing the performance of the model through the implementation of additional preprocessing techniques. Furthermore, the applicability of this model can be explored for predicting other diseases as well. Additionally, an alternative approach could involve replacing the machine learning algorithm in our proposed model with a deep learning algorithm. Transfer learning, a powerful technique in deep learning, has demonstrated remarkable results in various classification tasks. This approach involves leveraging knowledge acquired from a pre-trained model in one task and applying it to another related task. Two common methods for transfer learning are fine-tuning and feature extraction. In our future work, we aim to

investigate the impact of employing feature extraction or fine-tuning techniques on the proposed model by utilizing other pre-trained models. This analysis will provide valuable insights into the performance of the proposed model and its potential for improvement.

References

[1] World Health Organzation, "Cardiovascular Diseases (CVDs)," 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Jan. 07, 2022).

[2] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics - At-a-Glance: A report from the American Heart Association," *Circulation*, vol. 137, no. 12, pp. E67–E492, 2018, doi: 10.1161/CIR.0000000000000558.

[3] I. De Backer, "Cardiovascular Disease 2020-2030: Trends, Technologies & Outlook," 2020.

[4] A. Micah *et al.*, "Financing Global Health 2020," 2021.

[5] S. C. Smith, "Multiple Risk Factors for Cardiovascular Disease and Diabetes Mellitus," *Am. J. Med.*, vol. 120, no. 3 SUPPL. 1, pp. 3–11, 2007, doi: 10.1016/j.amjmed.2007.01.002.

[6] P. Joseph *et al.*, "Prognostic validation of a non-laboratory and a laboratory based cardiovascular disease risk score in multiple regions of the world," *Heart*, vol. 104, no. 7, pp. 581–587, 2018, doi: 10.1136/heartjnl-2017-311609.

[7] M. Kivimäki and A. Steptoe, "Effects of stress on the development and progression of cardiovascular disease," *Nat. Rev. Cardiol.*, vol. 15, no. 4, pp. 215–229, 2018, doi: 10.1038/nrcardio.2017.189.

[8] N. Hill *et al.*, "Machine Learning to Detect and Diagnose Atrial Fibrillation and Atrial Flutter (AF/F) Using Routine Clinical Data," *Value Heal.*, vol. 21, p. S213, 2018, doi: 10.1016/j.jval.2018.04.1448.

[9] H. Hae *et al.*, "Machine learning assessment of myocardial ischemia using angiography: Development and retrospective validation," *PLoS Med.*, vol. 15, no. 11, pp. 1–19, 2018, doi: 10.1371/journal.pmed.1002693.

[10] S. B. Golas *et al.*, "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, pp. 1–17, 2018, doi: 10.1186/s12911-018-

0620-z.

[11] S. Blecker *et al.*, "Early Identification of Patients With Acute Decompensated Heart Failure," *J. Card. Fail.*, vol. 24, no. 6, pp. 357–362, 2018, doi: 10.1016/j.cardfail.2017.08.458.

[12] E. Alderwish, P. Noack, J. Moore, M. A. Alamir, and M. Poon, "Remote Interpretation of Coronary Computed Tomographic Angiography (CCTA) Can Safely Expand Access To Advanced Cardiovascular Imaging To Evaluate Acute Chest Pain in Community Hospital Setting," *J. Am. Coll. Cardiol.*, vol. 71, no. 11, p. A1625, 2018, doi: 10.1016/s0735-1097(18)32166-1.

[13] D. Lu *et al.*, "Identification of Blood Circular RNAs as Potential Biomarkers for Acute Ischemic Stroke," *Front. Neurosci.*, vol. 14, no. February, pp. 1–15, 2020, doi: 10.3389/fnins.2020.00081.

[14] M. Z. Poh *et al.*, "Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms," *Heart*, pp. 1–8, 2018, doi: 10.1136/heartjnl-2018-313147.

[15] K. Rossing *et al.*, "Urinary proteomics pilot study for biomarker discovery and diagnosis in heart failure with reduced ejection fraction," *PLoS One*, vol. 11, no. 6, pp. 1–15, 2016, doi: 10.1371/journal.pone.0157167.

[16] V. Kouranos *et al.*, "Complementary Role of CMR to Conventional Screening in the Diagnosis and Prognosis of Cardiac Sarcoidosis," *JACC Cardiovasc. Imaging*, vol. 10, no. 12, pp. 1437–1447, 2017, doi: 10.1016/j.jcmg.2016.11.019.

[17] Pocket Dentistry, "The circulatory system," *Pocket Dentistry*, 2015. https://pocketdentistry.com/4-the-circulatory-system/ (accessed Jan. 11, 2022).

[18] C. P. Davis and J. R. Balentine, "How the Heart Works: Sides, Chambers, and Function," *MedicineNet*, 2020. https://www.medicinenet.com/heart_how_the_heart_works/article.htm (accessed Jan. 11, 2022).

[19] J. I. Gupta and M. J. Shea, "Biology of the Blood Vessels," *MSD Manual*, 2019. https://www.msdmanuals.com/home/heart-and-blood-vessel-disorders/biology-of-the-heart-and-blood-vessels/biology-of-the-blood-vessels (accessed Jan. 12, 2022).

[20] Broad Learnings, "Human Circulatory System," *Broad Learnings*, 2020. https://www.broadlearnings.com/courses/human-circulatory-system/ (accessed Jan. 12, 2022).

[21] NHS, "Cardiovascular disease," 2018. https://www.nhs.uk/conditions/cardiovascular-disease/ (accessed Jan. 13, 2022).

[22] S. Basu, "Atherosclerosis - Causes, Symptoms And Treatment," *NetMeds.com*, 2019. https://www.netmeds.com/health-library/post/atherosclerosis-causes-symptoms-and-treatment (accessed Jan. 13, 2022).

[23] A. Felman and P. Kohli, "What to know about cardiovascular disease," *Medical News Today*, 2019. https://www.medicalnewstoday.com/articles/257484#symptoms (accessed Jan. 17, 2022).

[24] A. Pick, "Aortic Stenosis: Symptoms, Diagnosis & Treatment," *HeartValveSurgery.com*, 2020. https://www.heart-valve-surgery.com/aortic-stenosis-valve-heart-narrowing.php (accessed Jan. 14, 2022).

[25] D. L. Kulick, J. W. Marks, and C. P. Davis, "Heart Attack (Myocardial Infarction)," *MedicineNet*, 2020. https://www.medicinenet.com/heart_attack/article.htm (accessed Jan. 14, 2022).

[26] H. Moawad and H. Sheikh, "Causes and Risk Factors of Stroke," *verywell health*, 2021. https://www.verywellhealth.com/stroke-causes-4014093 (accessed Jan. 15, 2022).

[27] Y. Feng, J. Cheng, B. Wei, and Y. Wang, "CaMKII inhibition reduces isoproterenol-induced ischemia and arrhythmias in hypertrophic mice," *Oncotarget*, vol. 8, no. 11, pp. 17504–17509, 2017, doi: 10.18632/oncotarget.15099.

[28] A. Kandola and K. Martinez, "Everything you need to know about abdominal aortic aneurysms," *Medical News Today*, 2020. https://www.medicalnewstoday.com/articles/abdominal-aortic-aneurysm (accessed Jan. 17, 2022).

[29] Comprehensive Integrated Care, "Peripheral Artery Disease Vs. Peripheral Vascular Disease: What's the Difference?," *Comprehensive Integrated Care*, 2019.

https://www.ciccenters.com/peripheral-artery-disease-vs-peripheral-vascular-disease-whats-the-difference/ (accessed Jan. 18, 2022).

[30] American Heart Association, “Recommendations for Physical Activity in Adults and Kids,” 2018.

[31] National Institute of Diabetes and Digestive and Kidney Diseases, “Health Risks of Overweight & Obesity,” 2018.

[32] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” *BMC Med. Inform. Decis. Mak.*, vol. 5, pp. 1–15, 2019, doi: 10.1186/s12911-019-0918-5.

[33] A. M. Alaa, T. Bolton, E. Di Angelantonio, H. James, F. Rudd, and M. Van Der Schaar, “Cardiovascular disease risk prediction using automated machine learning : A prospective study of 423,604 UK Biobank participants,” *PLoS One*, vol. 14, no. 5, pp. 1–17, 2019, doi: 10.1371/journal. pone.0213653.

[34] 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.

[35] K. Pahwa, A. C. S. Dangare, S. S. Apte, and I. Study, “Prediction of Heart Disease Using Hybrid Technique For Selecting Features,” in *4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University*, 2017, pp. 500–504, doi: 10.1109/upcon.2017.8251100.

[36] S. Rajathi and G. Radhamani, “Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO,” 2016, doi: 10.1109/sapience.2016.7684132.

[37] K. Farooq *et al.*, “A Novel Cardiovascular Decision Support Framework for Effective Clinical Risk Assessment,” in *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, 2014, pp. 117–124, doi: 10.1109/CICARE.2014.7007843.

[38] T. Mahboob, R. Irfan, and B. Ghaffar, “Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics,” *Internet Technol. Appl.*, pp. 110–115,

2017, doi: 10.1109/ITECHA.2017.8101920.

[39] A. A. Raja, M. Guftar, Irfan-Ul-Haq, T. A. Khan, and D. Greibl, "Intelligent Syncope Disease prediction framework using DM-ensemble techniques," *FTC 2016 - Proc. Futur. Technol. Conf.*, no. December, pp. 269–273, 2017, doi: 10.1109/FTC.2016.7821621.

[40] Y. Wang, "Identification of Cardiovascular Diseases Based on Machine Learning," *ACM Int. Conf. Proceeding Ser.*, vol. 1, no. 1, pp. 531–536, 2022, doi: 10.1145/3570773.3570855.

[41] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, 2021, doi: 10.1088/1742-6596/1817/1/012009.

[42] C. Chethana, "Prediction of heart disease using different KNN classifier," in *5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2021, pp. 1186–1194.

[43] A. Singh and A. Jain, "Prediction of Heart Disease using Dense Neural Network," *2022 IEEE Glob. Conf. Comput. Power Commun. Technol. GlobConPT 2022*, vol. 6, no. 1, pp. 51–61, 2022, doi: 10.1109/GlobConPT57482.2022.9938354.

[44] S. Maji and S. Arora, "Decision Tree Algorithms for Prediction of Heart Disease," in *Proceedings of Third International Conference on ICTCS 2017*, 2017, vol. 40, pp. 447–454, doi: 10.1007/978-981-13-0586-3.

[45] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of Cardiovascular Disease using Machine Learning," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 2021, no. 1, pp. 1–7, doi: 10.1109/ICCTCT.2018.8550857.

[46] J. D. Frizzell *et al.*, "Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches," *JAMA Cardiol.*, vol. 2, no. 2, pp. 204–209, 2017, doi: 10.1001/jamacardio.2016.3956.

[47] A. R. van Rosendael *et al.*, "Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry," *J. Cardiovasc. Comput. Tomogr.*, vol. 12, no. 3, pp. 204–

209, 2018, doi: 10.1016/j.jcct.2018.04.011.

[48] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can Machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, pp. 1–14, 2017, doi: 10.1371/journal.pone.0174944.

[49] J. A. Quesada *et al.*, "Machine learning to predict cardiovascular risk," *Int. J. Clin. Pract.*, vol. 73, no. 10, pp. 1–6, 2019, doi: 10.1111/ijcp.13389.

[50] K. Junwei, H. Yang, L. Junjiang, and Y. Zhijun, "Dynamic prediction of cardiovascular disease using improved LSTM," *Int. J. Crowd Sci.*, vol. 3, no. 1, pp. 14–25, 2019, doi: 10.1108/ijcs-01-2019-0002.

[51] P. Lu *et al.*, "Research on Improved Depth Belief Network-Based Prediction of Cardiovascular Diseases," *J. Healthc. Eng.*, vol. 2018, 2018, doi: 10.1155/2018/8954878.

[52] R. Poplin *et al.*, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nat. Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, 2018, doi: 10.1038/s41551-018-0195-0.

[53] F. Ali *et al.*, "A Smart Healthcare Monitoring System for Heart Disease Prediction Based On Ensemble Deep Learning and Feature Fusion," *Inf. Fusion*, 2020, doi: 10.1016/j.inffus.2020.06.008.

[54] J. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest," *J. Am. Heart Assoc.*, vol. 7, pp. 1–12, 2018, doi: 10.1161/JAHA.118.008678.

[55] "Kaggle." https://www.kaggle.com/ (accessed Apr. 11, 2022).

[56] M. F. Khan *et al.*, "An iomt-enabled smart healthcare model to monitor elderly people using machine learning technique," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/2487759.

[57] Python, "Python Programming Language." [Online]. Available: https://www.python.org/.

[58] Jupyter, "Jupyter Notebook." [Online]. Available: https://jupyter.org/.

[59] Niteesh, https://github/Niteesh95/cardiovascular_disease_prediction/blob/master/src/CVD

_model.ipynb

[60] Dritsas, Elias & Alexiou, Sotiris & Moustakas, Konstantinos. (2022). Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. 315-321. 10.5220/0011088300003188.

[61] Adeboye, Nureni Olawale and Abimbola, Olawale Victor. 'An Overview of Cardiovascular Disease Infection: A Comparative Analysis of Boosting Algorithms and Some Single Based Classifiers'. 1 Jan. 2020 : 1189 – 1198.

[62] W. Yue, L. I. Voronova and V. I. Voronov, "Design and Implementation of a Remote Monitoring Human Health System," 2020 Systems of Signals Generating and Processing in the Field of on Board Communications, Moscow, Russia, 2020, pp. 1-7, doi: 10.1109/IEEECONF48371.2020.9078574.