Comprehensive Analysis and Comparison of Machine Learning Algorithms on Different Datasets to Predict the Risk of Cardiovascular Diseases

by

Tanvir Rahman
Md. Muiz Shahriar Hossein
Mostafa Mohiuddin Jalal

A THESIS SUBMITTED FOR THE DEGREE OF
BACHELOR OF SCIENCE

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY
FACULTY OF SCIENCE AND TECHNOLOGY
BANGLADESH UNIVERSITY OF PROFESSIONALS
FEBRUARY 2021

APPROVAL

The thesis titled "Exploring the Machine Learning Algorithms to Find the Best Features for Predicting the Cardiovascular Diseases" Submitted by Tanvir Rahman, ID: 17511047, Md. Muiz Shahriar Hossein, ID:17511078, Mostafa Mohiuddin Jalal, ID:17511083 Session 2016-17 has been approved as successful in fulfilment of the requirement for the degree of Bachelor of Science in Information and Communication Engineering.

Zarin Tasnim
Lecturer
Department of Information and Communication Technology
Bangladesh University of Professionals.

DECLARATION

We hereby declare that this thesis is authentic, and it has been drafted by us in its absoluteness. We have clearly mentioned and acknowledged each sources of information which have been used in the thesis work. This thesis has also not been submitted for any degree in any university previously. This work has not been submitted partly or fully to any University for award, degree, or any personal gain.

Tanvir Rahman
ID: 17511047
Department of Information & Communication Technology
Faculty of Science and Technology
Bangladesh University of Professionals
08 February 2020

Md. Muiz Shahriar Hossein
ID: 17511078
Department of Information & Communication Technology
Faculty of Science and Technology
Bangladesh University of Professionals
08 February 2020

Mostafa Mohiuddin Jalal
ID: 17511083
Department of Information & Communication Technology
Faculty of Science and Technology
Bangladesh University of Professionals
08 February 2020

ACKNOWLEDGEMENTS

Firstly, we would like to express our sincere gratefulness to Almighty Allah for giving us the strength, patience, and knowledge to complete this thesis in due time. Secondly, we would like to pay our deep sense of gratitude to our supervisor Lecturer Zarin Tasnim for taking us under his supervision. He played an active role in encouraging us and providing us the opportunity to conduct the research. We feel to acknowledge our indebtedness to Lecturer Zarin Tasnim for her valuable guidance and motivation throughout the whole process which shaped the outcome as a success. We would also like to thank our teachers and seniors for helping us in our time of need. We are greatly obliged towards our friends and classmates for their words of motivation, support, and care whenever we felt like giving up. We would also like to thank every BUP officials and staff members for letting us use university equipment and facilities whenever we asked for it. Last but not the least, we take this opportunity to convey our deep appreciativeness to our parents for understanding our labor for this work and for their patience throughout the journey.

ABSTRACT

Nowadays, cardiovascular diseases are considered as one of the fatal and main reasons for mortality all around the globe. The mortality or high-risk rate can be reduced if an early detection system for cardiovascular disease is introduced. A massive amount of data gets collected by healthcare organizations. A proper and careful study regarding the data can be carried out to extract some important and interesting insight that may help out the professionals. Keeping that in mind, in this paper, at first six distinct machine learning algorithms(Logistic Regression, SVM, KNN, Naïve Bayes, Random Forest, Gradient Boosting) were applied to four different datasets encompasses different set of features to show their performance over them. Secondly, the prediction accuracy of the ML algorithms was analyzed to find out the best set of features and the best algorithm to predict cardiovascular diseases. The results find out the best suited eleven feature and also showed that Random Forest performs well in terms of accuracy in predicting cardiovascular diseases.

Keywords: Prediction, Machine learning, Cardiovascular disease, Classification, Healthcare, Feature identification

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML Machine Learning

CVD Cardiovascular Disease

CAD Coronary Artery Disease

CHD Coronary Heart Disease

TIA Transient Ischaemic Attack

PAD Peripheral Arterial Disease

PVD: Peripheral Vascular Disease

AAA Abdominal Aortic Aneurysm

LDL Low-Density Lipoprotein

HDL High-Density Lipoprotein

CRP C-Reactive Protein

EKG/ECG Electrocardiogram

ECHO Echocardiogram

K-NN K-th Nearest Neighbour

SVM Support Vector Machine

AI Artificial Intelligence

TP True Positive

TN True Negative

FP False Positive

FN False Negative

LR Logistic Regression

GB Gradient Boosting

RF Random Forest

NB Naive Bayes

UN United Nations

ROC Region of Convergence

AUC Area under the ROC Curve

UCI University of California Irvine

NLP Natural language processing

RnD Research and Development

OCR Optical character recognition

API Application Programming Interface

MATLAB Matrix Laboratory

DT Decision Tree

FRS Functional Requirement Specification

GAM Generalized Additive Model

GBT Gradient Boosted Tree

RBF Radial Basis Function

UK United Kingdom

WHO World Health Organization

CPU Central Processing Unit

RAM Random-Access Memory

DDR Double Data Rate

GPU Graphics Processing Unit

UHD Ultra High Definition

LLC Limited Liability Company

BMI Body Mass Index

SMOTE Synthetic Minority Over-sampling Technique

CHAPTER 1

INTRODUCTION

1.1 Synopsis of this chapter

This chapter illustrates the motivation behind the research and the background concerning this study. It will provide a clear understanding of our objectives along with the statement of problems elaborately. In the end, our contribution towards this study will be covered.

1.2 Motivation behind the Research

Cardiovascular diseases are considered to be one of the main reasons of fatality around the world. The prevalence of this disease is ballooning. A large number of people are being affected by cardiovascular diseases without being concerned. In the last few years, an average of 19 million people [1] have been affected with this "secret epidemic". As per the experts, if the diseases are detected in the earlier stage, the remedy becomes simpler and easy to afford. Such detections are usually performed manually by one or more clinicians based on reports and test results which is both time consuming and overpriced. For the problem discussed above, this study can be utterly helpful. In this study we have introduced ML to detect the presence of cardiovascular diseases. This study includes Ensemble approach combining machine

learning algorithms for having the improved prediction of the presence of cardiovascular diseases based on parameter tuning.

1.3 Background

Cardiovascular disease (CVD) is the name for the group of disorders of heart and blood vessels, and include: hypertension (high blood pressure), coronary heart disease (heart attack) [2]. It's usually associated with a build-up of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots. It can also be associated with damage to arteries in organs such as the brain, heart, kidneys and eyes.

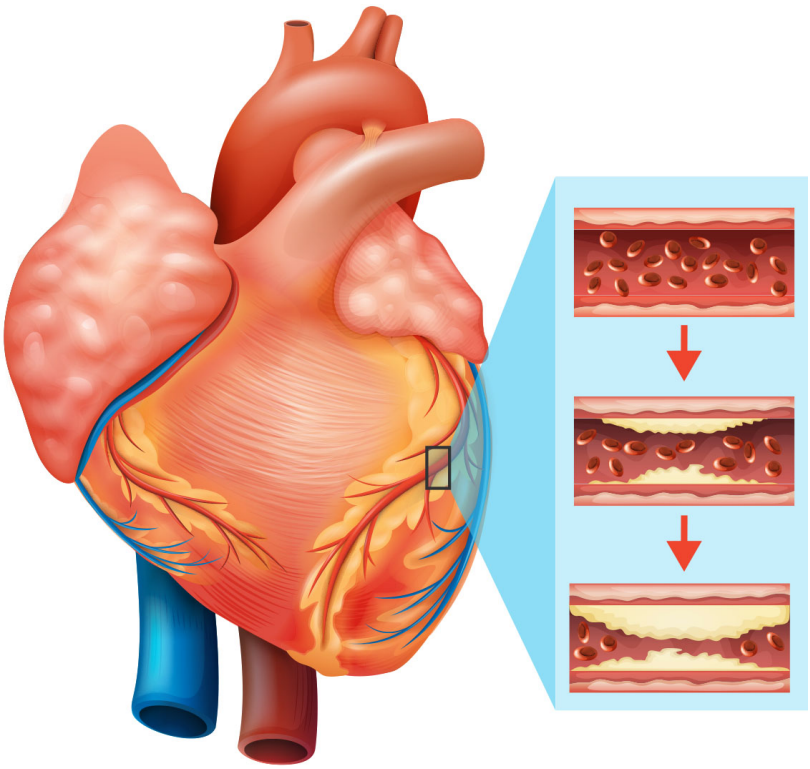

Fig. 1.1. Cardiovascular Diseases

CVD is one of the main causes of death and disability all over the world, but it can often largely be prevented by leading a healthy lifestyle or if the disease is identified in early stages [3]. The major risk factors for these disorders were recognized over

many years, and they include:

- High levels of low-density lipoprotein (LDL)
- Cholesterol
- Smoking
- Hypertension
- Diabetes
- Abdominal obesity
- Psychosocial factors
- Insufficient consumption of fruits and vegetables
- Excess consumption of alcohol
- Lack of regular physical activity

Though people of all age groups are vulnerable to these diseases, it mainly affects people that are adult and older (more than 65 years old).

1.3.1 Types and symptoms of Cardiovascular Disease

There are a number of types of cardiovascular diseases. These all can be categorized into four main types:

1. **Coronary heart disease**: CHD is the term that describes what happens when your heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries. Over time, the walls of your arteries can become furred up with fatty deposits. This process is known as ATHEROSCLEROSIS and the fatty deposits are called ATHEROMA. Atherosclerosis can be caused

by lifestyle factors, such as smoking and regularly drinking excessive amounts of alcohol. The main symptoms of CHD are:

(a) Chest pain (Angina)

(b) Shortness of breath

(c) Pain throughout the body

(d) Feeling faint

(e) Feeling sick (nausea)

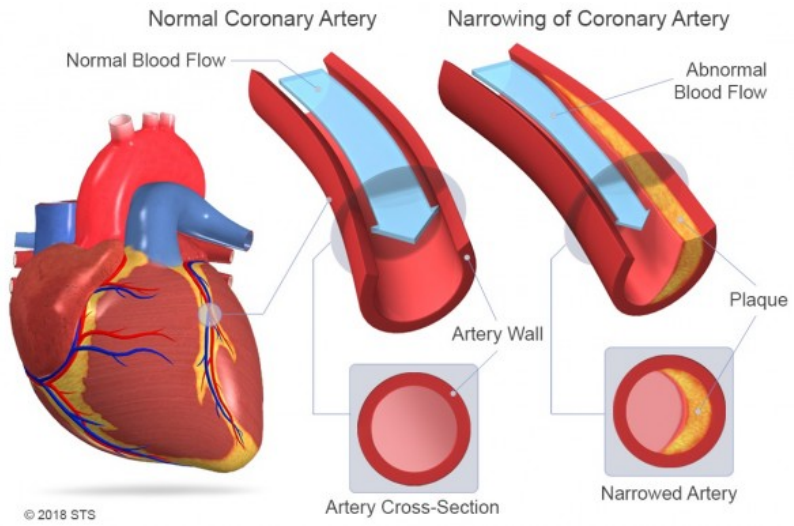

Fig. 1.2. Coronary heart disease

But not everyone has the same symptoms and some people may not have any before CHD is diagnosed. CHD puts increased strain on the heart, and can lead to:

(a) Angina: Chest pain caused by restricted blood flow to the heart muscle.

(b) Heart Attacks: Where the blood flow to the heart muscle is suddenly blocked.

(c) Heart Failure: Where the heart is unable to pump blood around the body properly.

2. **Stroke and TIAs**: A stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off. Strokes are a medical emergency and urgent treatment is essential. The sooner a person receives treatment for a stroke, the less damage is likely to happen.

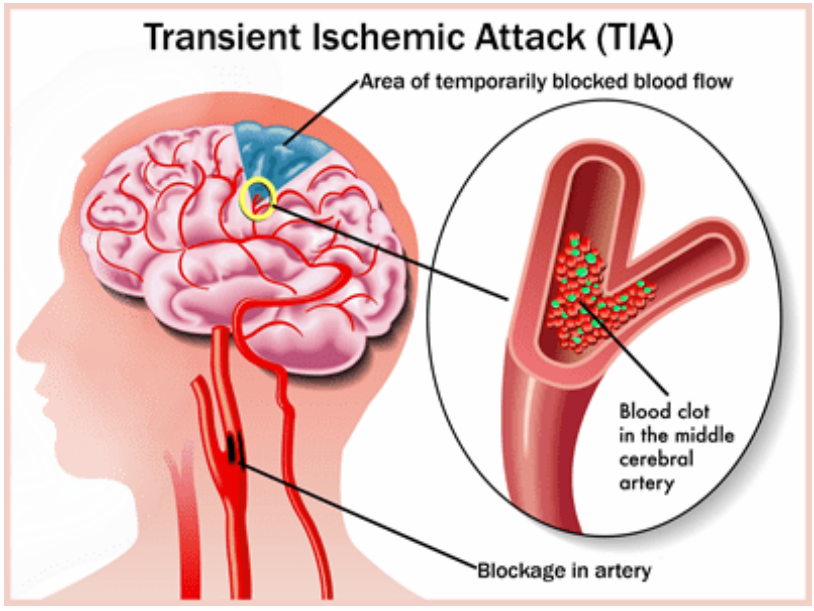

Fig. 1.3. Stroke and TIA

A transient ischaemic attack (TIA) or "mini stroke" is caused by a temporary disruption in the blood supply to part of the brain. The disruption in blood supply results in a lack of oxygen to the brain. This can cause sudden symptoms similar to a stroke, such as speech and visual disturbance, and numbness or weakness in the face, arms and legs. But a TIA does not last as long as a stroke. The effects last a few minutes to a few hours and fully resolve within 24 hours.

The main symptoms of stroke and TIA can be remembered with the word FAST:

(a) Face: The face may have dropped on 1 side, the person may not be able to smile, or their mouth or eye may have dropped.

(b) Arms: The person with suspected stroke may not be able to lift both arms and keep them there because of weakness or numbness in one arm.

(c) Speech: Their speech may be slurred or garbled, or the person may not be able to talk at all despite appearing to be awake; they may also have problems understanding what they are listening to.

(d) Time: The person should be taken to a hospital as soon as possible.

3. **Peripheral Arterial Disease**: PAD is a common condition where a build-up of fatty deposits in the arteries restricts blood supply to leg muscles. It's also known as peripheral vascular disease (PVD).

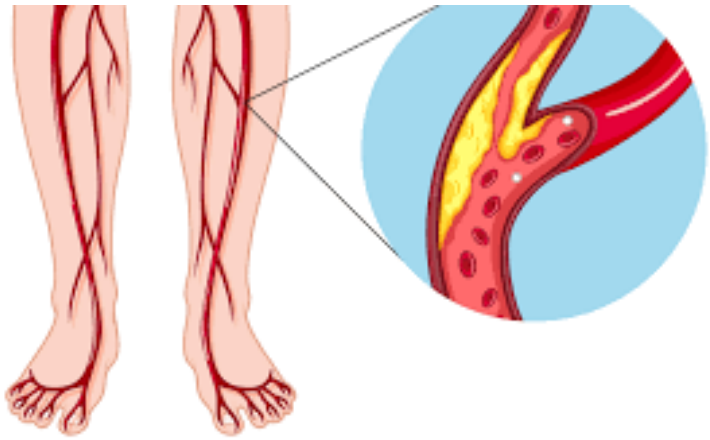

Fig. 1.4. Peripheral Arterial Disease

Many people with PAD have no symptoms. However,

(a) some develop a painful ache in their legs when they walk, which usually disappears after a few minutes' rest. The medical term for this is "intermittent claudication". The pain can range from mild to severe, and usually goes away after a few minutes when you rest your legs. Both legs are often affected at the same time, although the pain may be worse in one leg.

(b) Hair loss on your legs and feet

(c) Numbness or weakness in the legs

(d) Brittle, slow-growing toenails

(e) Ulcers (open sores) on your feet and legs, which do not heal

(f) Changing skin colour on your legs, such as turning pale or blue

(g) Shiny skin

(h) In men, erectile dysfunction

(i) The muscles in your legs shrinking (wasting)

The symptoms of PAD often develop slowly, over time. If your symptoms develop quickly, or get suddenly worse, it could be a sign of a serious problem requiring immediate treatment. This can cause:

(a) Dull or cramping leg pain, which is worse when walking and gets better with rest

(b) Hair loss on the legs and feet

(c) Numbness or weakness in the legs

(d) Persistent ulcers (open sores) on the feet and legs

4. **Aortic disease**: An abdominal aortic aneurysm (AAA) is a bulge or swelling in the aorta, the main blood vessel that runs from the heart down through the chest and tummy. An AAA can be dangerous if it is not spotted early on. It can get bigger over time and could burst (rupture), causing life-threatening bleeding. Men aged 65 and over are suggested to screen for AAA Women aged 70 or over, who have underlying risk factors such as high blood pressure, may also be advised to attend screening for AAA.

   AAAs do not usually cause any obvious symptoms, and are often only picked up during screening or tests carried out for another reason[4].

   Some people with an AAA have:

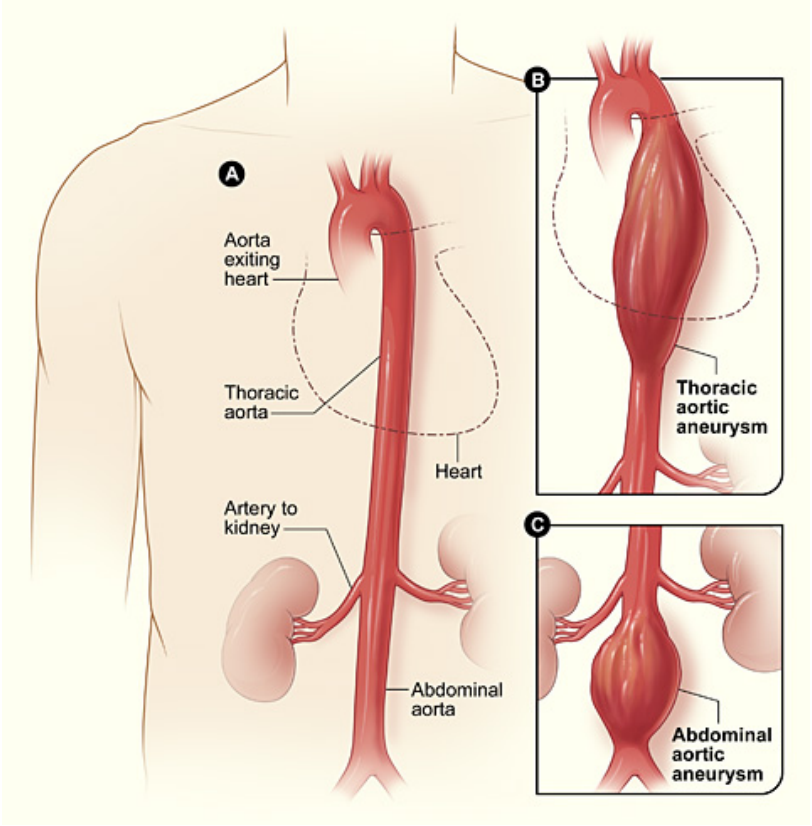

Fig. 1.5. Aortic Diseases

(a) A pulsing sensation in the tummy (like a heartbeat)

(b) Tummy pain that does not go away

(c) Lower back pain that does not go away

If an AAA bursts, it can cause:

(a) Sudden, severe pain in the tummy or lower back

(b) Dizziness

(c) Sweaty, pale and clammy skin

(d) A fast heartbeat

(e) Shortness of breath

(f) Fainting or passing out

1.3.2 Diagnosis of cardiovascular diseases

Cardiovascular diseases are diagnosed using an array of laboratory tests and imaging studies. The primary part of diagnosis is medical and family histories of the patient, risk factors, physical examination and coordination of these findings with the results from tests and procedures. Some of these tests are non-invasive, which means no instruments are inserted into the body. Other tests are invasive and require inserting instruments into the body.

- **Blood Tests for Heart Disease**

  1. **Lipid profile**

     The lipid profile includes:

     - Total cholesterol
     - LDL (low-density lipoprotein), also known as "bad" cholesterol
     - HDL (high-density lipoprotein), also known as "good" cholesterol
     - Triglycerides

  2. **Lipoprotein (a), or Lp (a)**

     Lipoprotein (a) is a special type of lipid-containing protein. The genes mainly play the main role in determining your level of Lp (a).

  3. **C-reactive protein (CRP)**

     Human liver produces C-reactive protein (CRP) as part of body's response to injury or infection. Inflammation plays a central role in the process of atherosclerosis, in which fatty deposits clog the arteries. CRP test results combined with other blood test results and risk factors for heart disease help create an overall picture of heart health.

  4. **Homocysteine**

     Human body uses homocysteine to make protein and to build and maintain

tissue. However, too much homocysteine may increase the risk of heart disease and stroke. Homocysteine is usually ordered for people who have a high risk for developing heart disease or have a known history of heart disease. It is also used for people with a family history of heart disease but no other known risk factors [5].

- **Non-Invasive Tests**

1. **Electrocardiogram**

   This is a simple and a painless test that records the heart's electrical activity. The patient is strapped to the instrument with several patches or leads placed over his or her chest, wrists and ankles. A small portable machine records the activities of the heart on a strip of graph paper.The test shows how fast the heart is beating and its rhythm. The strength and timing of the electrical signals as they pass through the heart are also seen. An EKG/ECG can help detect a heart attack, attacks of angina, arrhythmias etc.

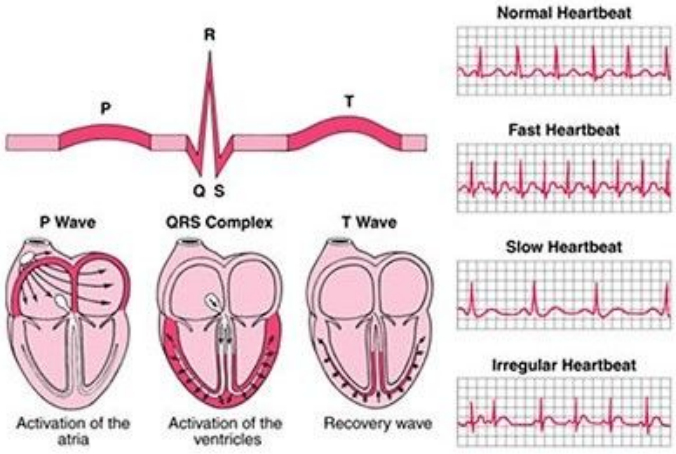

Fig. 1.6. Electrocardiogram

2. **Echocardiogram**

   An echocardiogram ("echo") is an ultrasound of the heart. A small probe

like a microphone, called a transducer, is placed on the chest in various places. The ultrasound waves sent by the transducer bounce off the various parts of the heart. A computer in the machine determines the time it takes for the sound wave to return to the transducer and generates a picture with the data. During the test, the patient lies on back or left side on a stretcher for about 45 minutes while the pictures are being recorded. The echocardiographer will review the pictures before sending a patient home to be sure that all the necessary information has been obtained.

3. **Stress EKG or Echocardiogram**

   Stress tests are performed to see how the heart performs under physical stress. The heart can be stressed with exercise on a treadmill or in a few instances, a bicycle. If a person cannot exercise on a treadmill or bicycle, medications can be used to cause the heart rate to increase, simulating normal reactions of the heart to exercise. During the stress test, the patient will wear EKG leads and wires while exercising so that the electrical signals of the heart can be recorded at the same time. The blood pressure is monitored throughout the test. The stress test can be performed together with the echocardiogram, described above.

4. **Nuclear Stress Test**

   Nuclear stress tests have two components to them: a treadmill (or chemical) stress test and scanning of the heart after injection of a radionuclide material. This material has been used safely for many years to determine the amount of blood the heart muscle is receiving during rest and stress. The scanning is done with a nuclear camera.

5. **Carotid Ultrasound**

   Carotid ultrasound is done to evaluate the risk of stroke. The sonographer presses the transducer gently against the sides of neck, which sends images

of the arteries to a computer screen for the technician to see. The technician monitors the blood flow through the carotid arteries on both sides of neck to check for stenosis. During the exam,a patient lies on back on an examination table and a small amount of warm gel is applied to the skin. The test usually takes about 15 to 30 minutes.

6. **Abdominal Ultrasound**

   The doctor may also want to have an abdominal ultrasound to screen for potential abdominal aortic aneurysm. The sonographer presses the transducer against the skin over the abdomen, moving from one area to another. The transducer sends images to a computer screen that the technician monitors. The technician monitors blood flow through the abdominal aorta to check for an aneurysm. During the exam, the patient lies on your back on an examination table and a small amount of warm gel is applied to the abdomen. The test usually takes between 20 minutes to an hour.

7. **Holter Monitor**

   A Holter monitor is a small, portable machine that a patient wears for 24 to 48 hours. It enables continuous recording of the EKG as he or she goes about daily activities. The patient will be asked to keep a diary log of the activities and symptoms. This monitor can detect arrhythmias that might not show up on a resting EKG that only records for a few seconds.

8. **Event Recorder**

   An event recorder (loop recorder) is a small, portable transtelephonic monitor that may be worn for several weeks. This type of recorder is good for patients whose symptoms are infrequent.

   The monitor 'loops' a two- to five-minute recording into its memory which is continually overwritten. When the patient experiences symptoms he or

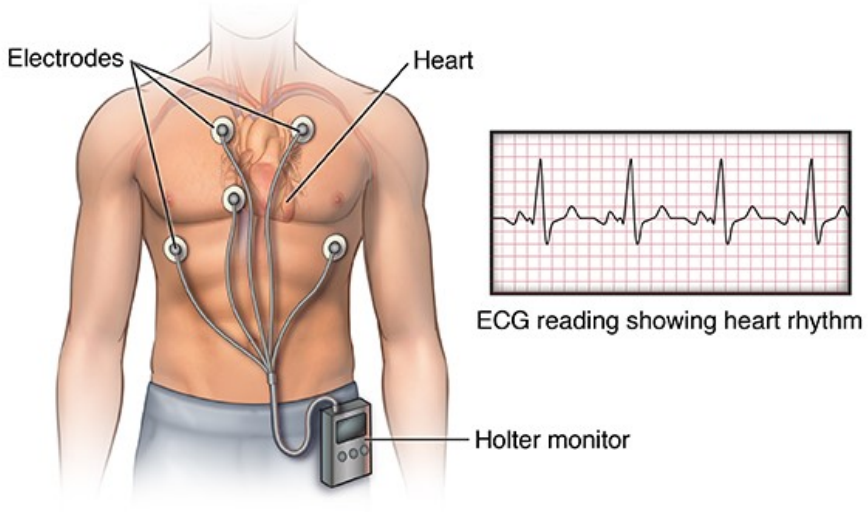

Fig. 1.7. Holter Monitor

she presses a 'record' button on the monitor, which stores a correlating strip of EKG. The recordings are telephoned through to a 24-hour monitoring station and faxed directly to the corresponding doctor.

- **Invasive Tests**

  1. **Cardiac Catheterization and Coronary Angiography**

     Cardiac catheterization is a common procedure that can help diagnose heart disease. In some cases, catheterization is also used to treat heart disease by opening blocked arteries with balloon angioplasty and stent placement.

     Cardiac catheterization can show:

     - If the blood vessels in your heart have narrowed.
     - If your heart is pumping normally and blood is flowing correctly.
     - If the valves in your heart are functioning normally.
     - If you were born with any heart abnormalities.
     - If the pressures in the heart and lung are normal. If not, catheteriza-

tion can help assess the problem.

During catheterization, the cardiologist inserts a long, flexible tube called a catheter into a blood vessel, either through the wrist artery or the groin artery, and gently guides it towards the heart under X-ray guidance. Once the catheter is in place, X-rays and other tests are done to help the doctor evaluate whether the coronary arteries are blocked and how well the heart is working. At times, it might also be necessary to insert a small catheter into a vein to allow measurement of specific pressures in the heart and lung. This procedure can be done either through a neck vein, arm vein or groin vein.

2. **Electrophysiology Study**

   An electrophysiology study (EP) is a recording of the electrical activity of the heart. This test helps the doctor determine the cause of rhythm disturbance (arrhythmia) and the best treatment. During the test, the doctor may safely reproduce the arrhythmia, and then give certain medications to see which one controls it best.

   An EP study is performed in the Electrophysiology Laboratory, where the patient will lie on an X-ray table. As with a cardiac catheterization, the doctor inserts a long, flexible tube — an electrode catheter — into a blood vessel, usually in the groin.

   There are stages in an EP study:

   - Recording the heart's electrical signals to assess electrical function.
   - Pacing the heart to bring on certain abnormal rhythms for observation under controlled conditions.
   - If the valves in your heart are functioning normally.

1.4 Problem Statement

The problem due to cardiovascular diseases is ever rising in the developing countries. It can cause untold suffering if it is not timely recognized, reported or understood. The impact of this problem on the social and economic sector have been manifold because of the lack of awareness of people. People are being deprived of proper treatment because of ignorance. The problems due to cardiovascular diseases can be reduced and for this, detection of cardiovascular disease, especially at the initial stages, is compulsory. This sort of diagnosis is normally performed by medical experts in a non-automated way. The outcome of the tests are analyzed depending on the test result, medical and family history of the patients. At times, this process is not feasible because of being time consuming and putting a huge effect on the patient's money. Over the last few years, data mining has been performing an incredible job in detecting these diseases. Many works regarding detection of this disease using machine learning algorithms have already been performed by many experimenters but most of them failed to provide an optimized accuracy and the best set of attributes in the prediction process. So, this is an issue that is yet to be effectively solved with optimized results.

1.5 Aim of this Research

Mainly, the aim of this research is to detect cardiovascular diseases using ML algorithms. Besides, our research work has multiple objectives, namely-

a) To offer an alternative method of detection that may provide an automated process and valuable insights into diagnosis and classification.

b) explore the use of demographic, family and medical data to sort people into vulnerable and non-vulnerable classes through the implementation of six different

ML models (Logistic Regression, K-NN, Naive Bayes, Gradient Boosting, SVM and Random Forest).

c) Further, the performance of the ML algorithms and the dataset was computed to get the best outcome from this study.

1.6 Contribution of this Research

Our research work has some significant contributions towards predicting a disease which is normally unknown to most of the people. The major contributions of our research work can be summarized as follows:

1. **Finding the best parameter by using hyper parameter tuning for SVM**
   We did parameter tuning of SVM using parameters such as C, gamma and kernel to find out the hyperplane which will effectively classify the dataset into demented and non-demented class.

2. **Proposing more generalized and normalized dataset**
   We took SMOTE technique to remove the class imbalance situation so that the outcome cannot be biased on a particular class.

3. **Proposing a more generalized model for optimum accuracy**
   We took outcome performance from six ML algorithms on four different dataset and compared them to find out the best set of features and algorithms.

1.7 Organization of the Thesis Paper

This thesis is composed of five chapters. **Chapter 1** addresses the introduction and background study of cardiovascular diseases. The chapter concludes with a concise summary of this study's problem statement and aim, accompanied by a research contribution. **Chapter 2** describes the definition of Machine learning, Machine

Learning methods, Contribution of Machine Learning in medical field accompanied by related works. **Chapter 3** addresses the required experimental setup. the ML libraries to be used. The information related to the used datasets along with the data preprocessing, heatmap analysis and the ML algorithms that have been used to predict cardiovascular diseases. **Chapter 4** addresses the theoretical result of the entire ML process. Lastly, the chapter discusses the performance of the ML algorithms. **Chapter 5** discusses the findings of our work along with its limitations and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Synopsis of this chapter

This chapter illustrates the basic idea of machine learning. Besides, it also discusses the ML methods along with use of machine learning in the medical field. Finally, this chapter covers the related research work in this domain.

2.2 Definition of Machine Learning

Machine learning is a sub-field of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people [6].

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

2.3 Machine Learning methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed (citeAPAGE). Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

2.3.1 Supervised Learning

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly [7].

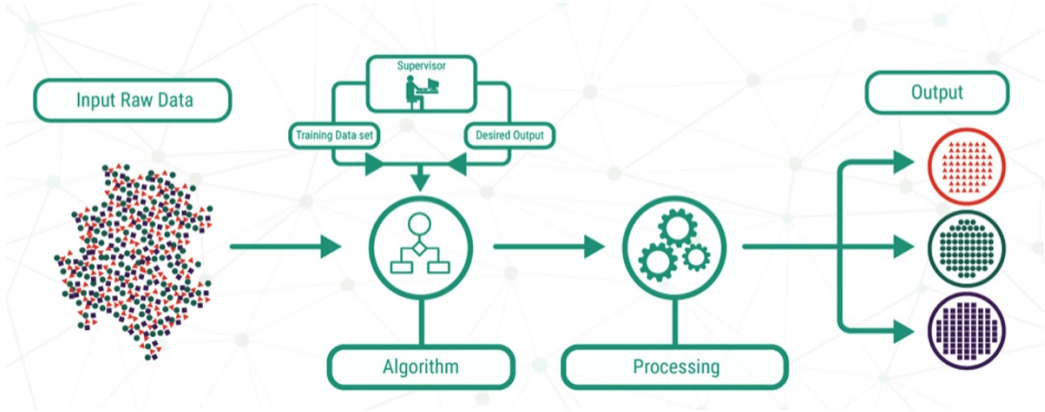

Fig. 2.1. Supervised Learning

2.3.2 Unsupervised Learning

unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data [8].

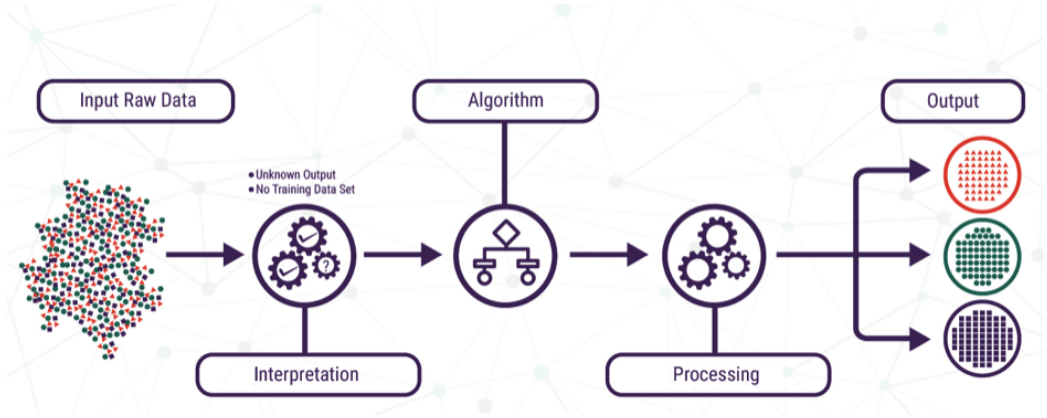

Fig. 2.2. Unsupervised Learning

2.3.3 Semi-supervised machine learning

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources [9].

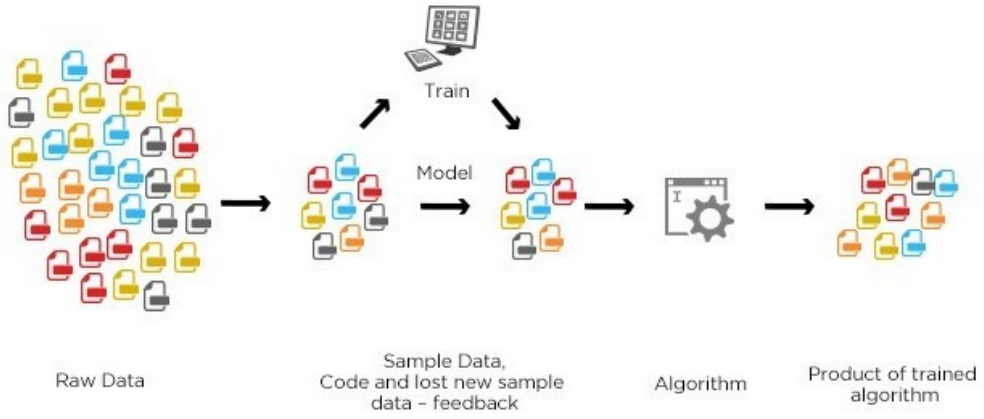

Fig. 2.3. Supervised Learning

2.3.4 Reinforcement machine learning

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance [10]. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Fig. 2.4. Reinforcement Learning

2.4 Machine Learning in medical science

Machine learning, a type of artificial intelligence when computers are programmed to learn information without human intervention. In machine learning, the development of the underlying algorithms rely on computational statistics. Computers are provided data and then the computers “learn” from that data. The data actually “teaches" the computer by revealing its complex patterns and underlying algorithms. The larger the sample of data the “machine" is provided, the more precise the machine’s output becomes. Machine learning in healthcare is becoming more widely used and is helping patients and clinicians in many different ways. The most common healthcare use cases for machine learning are automating medical billing, clinical decision support and the development of clinical care guidelines. There are many notable examples of machine learning and healthcare concepts being applied in medicine. The performance of machine learning-based automatic detection and diagnosis systems has shown to be equivalent to that of an experienced physician. Healthcare use cases for machine learning are many. For example, the same NLP technology that is used to determine creditworthiness for a consumer or sentiment analysis of someone’s social media post can now be used to read a patient’s chart to extract important data elements like the patient’s medications, treatment plans and medical conditions. The increasingly growing number of applications of machine learning in healthcare allows us to glimpse at a future where data, analysis, and innovation work hand-in-hand to help countless patients without them ever realizing it. Soon, it will be quite common to find ML-based applications embedded with real-time patient data available from different healthcare systems in multiple countries, thereby increasing the efficacy of new treatment options which were unavailable before. In the healthcare industry there are a number of tasks that have been easier than ever, once after ML technologies got attached with it, like:

- **Identifying Diseases and Diagnosis**

  One of the main ML applications in healthcare is the identification and diagnosis of diseases and ailments which are otherwise considered hard-to-diagnose. This can include anything from cancers to any other diseases which are tough to catch during the initial stages like dementia or cardiovascular diseases, to other genetic diseases.

- **Drug Discovery and Manufacturing**

  One of the primary clinical applications of machine learning lies in the early-stage drug discovery process. This also includes R&D technologies such as next-generation sequencing and precision medicine which can help in finding alternative paths for therapy of multi-factorial diseases. Currently, the machine learning techniques involve unsupervised learning which can identify patterns in data without providing any predictions

- **Medical Imaging Diagnosis**

  Machine learning and deep learning are both responsible for the breakthrough technology called Computer Vision. This has found acceptance in the InnerEye initiative developed by Microsoft which works on image diagnostic tools for image analysis. As machine learning becomes more accessible and as they grow in their explanatory capacity, expect to see more data sources from varied medical imagery become a part of this AI-driven diagnostic process.

- **Personalized Medicine**

  Personalized treatments can not only be more effective by pairing individual health with predictive analytics but are also ripe for further research and better disease assessment. Currently, physicians are limited to choosing from a specific set of diagnoses or estimate the risk to the patient based on his symptomatic history and available genetic information. In the coming years, we will see more

devices and biosensors with sophisticated health measurement capabilities hit the market, allowing more data to become readily available for such cutting-edge ML-based healthcare technologies.

- **Machine Learning-based Behavioral Modification**

  Behavioral modification is an important part of preventive medicine, and ever since the proliferation of machine learning in healthcare, countless startups are cropping up in the fields of cancer prevention and identification, patient treatment, etc.

- **Smart Health Records**

  Maintaining up-to-date health records is an exhaustive process, and while technology has played its part in easing the data entry process, the truth is that even now, a majority of the processes take a lot of time to complete. The main role of machine learning in healthcare is to ease processes to save time, effort, and money. Document classification methods using vector machines and ML-based OCR recognition techniques are slowly gathering steam, such as Google's Cloud Vision API and MATLAB's machine learning-based handwriting recognition technology.

- **Clinical Trial and Research**

  Machine learning has several potential applications in the field of clinical trials and research. Clinical trials cost a lot of time and money and can take years to complete in many cases. Applying ML-based predictive analytics to identify potential clinical trial candidates can help researchers draw a pool from a wide variety of data points, such as previous doctor visits, social media, etc. Machine learning has also found usage in ensuring real-time monitoring and data access of the trial participants, finding the best sample size to be tested, and leveraging the power of electronic records to reduce data-based errors.

- **Crowdsourced Data Collection**

  Crowdsourcing is all the rage in the medical field nowadays, allowing researchers and practitioners to access a vast amount of information uploaded by people based on their own consent. This live health data has great ramifications in the way medicine will be perceived down the line. Now, users can access interactive apps which apply ML-based facial recognition to try and treat Asperger's and Parkinson's disease. With the advancements being made in IoT, the healthcare industry is still discovering new ways in which to use this data and tackle tough-to-diagnose cases and help in the overall improvement of diagnosis and medication.

- **Better Radiotherapy**

  One of the most sought-after applications of machine learning in healthcare is in the field of Radiology. Medical image analysis has many discrete variables which can arise at any particular moment of time. There are many lesions, cancer foci, etc. which cannot be simply modeled using complex equations. Since ML-based algorithms learn from the multitude of different samples available on-hand, it becomes easier to diagnose and find the variables. One of the most popular uses of machine learning in medical image analysis is the classification of objects such as lesions into categories such as normal or abnormal, lesion or non-lesion, etc. Now the developed algorithms can detect the difference between healthy and cancerous tissue and improve radiation treatment for the same.

- **Outbreak Prediction**

  AI-based technologies and machine learning are today also being put to use in monitoring and predicting epidemics around the world. Today, scientists have access to a large amount of data collected from satellites, real-time social media updates, website information, etc. Artificial neural networks help to collate this

information and predict everything from malaria outbreaks to severe chronic infectious diseases. Predicting these outbreaks is especially helpful in third-world countries as they lack in crucial medical infrastructure and educational systems.

2.5 Related works

Several studies have been conducted using the UCI machine learning repository heart disease dataset to classify the presence of cardiovascular disease in the human body. This multivariate dataset involves 303 instances along with 75 features. For example, Rajyalakshmi et al [11] explored different machine learning algorithms including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and a hybrid model named HRFLM that combines Random Forest and Linear method for predicting cardiovascular diseases. The hybrid model showed the best accuracy among the implemented algorithms. Similarly, Sen et al [12] implemented the Naïve Bayes, SVM, Decision Tree, and KNN algorithms and found that SVM shows the best accuracy (83%). Again, Dinesh et al [13] predicted the possibility of cardiovascular diseases using six different algorithms, that includes Logistic Regression, Random Forest, SVM, Gradient Boosting, and an ensemble model, while Logistic regression showed the best accuracy (91%). In another study, Maji et al [14] proposed the use of hybridization Author techniques to reduce the test cases to predict the outcome using ANN, C4.5, and hybrid Decision Tree algorithms. The result also showed that the hybrid Decision tree performed better in terms of accuracy.

Prediction of cardiovascular diseases has also been conducted using different datasets to bring out the best accuracy like UCI Statlog Dataset having 270 instances along with 13 features. Dwivedi et al [16] proposed an automatic medicinal system using advanced data mining techniques like the multilayer perceptron model and applied several classification techniques like Naïve Bayes, SVM, logistic Regression, and KNN.

Table 2.1: Summary of related studies

Ref	No. of Features	Objectives	ML Techniques	Results			
				Accuracy	Specificity	Sensitivity	Precision
[11]	13	Predicting whether a person has heart disease or not by applying several ML algorithms and provides diagnosis.	Logistic Regression	87.00%			
			Random Forest	81.00%			
			Naive Bayes	84.00%			
			Gradinet Boosting	84.00%			
			SVM	78.00%			
[15]	13	Presenting a survey of various models based on algorithm and their performance.	Naive Bayes	84.16%			
			SVM	85.77%			
			KNN	83.16%			
			Decision Tree	77.55%			
			Random Forest	91.60%			
[16]	13	Evaluating six potential ML algorithms based on eight performance indices and finding the best algorithm for prediction.	ANN	84%	79%	87%	84%
			SVM	82%	89%	77%	90%
			Logistic Regression	85%	81%	89%	85%
			KNN	80%	76%	84%	81%
			Classification Tree	77%	73%	79%	79%
			Naive Bayes	83%	80%	85%	84%
[17]	13	Comparing different algorithms of decision tree classification in heart disease diagnosis.	J48 with Reduced Errorpruning Algorithm	57%			
			Logistic Mode Tree	56%			
			Random Forest				
[18]	14	Applying traditional ML algorithms and ensemble models to find out the best classifier for disease prediction.	Decision Tree	78%		83%	77%
			Naive Bayes	83%		87%	84%
			Knn, k=1	76%		78%	78%
			Knn, k=3	81%		84%	82%
			Knn, k=9	83%		84%	85%
			Knn, k=15	83%		84%	85%
			MLP	83%		82%	82%
			Radial Basic Function	84%		86%	85%
			Single Conjunctive Rule Learner	70%		70%	73%
			SVM	84%		90%	83%
[19]	14	Identifying significant features and mining techniques to improve CVD prediction.	SVM	85%			
			Vote	86%			
			Naive Bayes	86%			
			Logistic Regression	86%			
			NN	85%			
			KNN	83%			
			Decision Tree	83%			
[13]	13	Finding significant features and introducing several combinations with ML techniques.	KNN	59%			
			SVM	72%			
			Logistic Regression	77%			
			Naive Bayes	70%			
			Random Forest	74%			
[12]	14	Comparing performance of various ML algorithms and predicting CVD.	Naive Bayes	83%			
			SVM	84%			
			Decision Tree	76%			
			KNN	76%			
[14]	13	Proposing hybridization technique and validating using several performance measures to predict CVD.	ANN	77%			
			C4.5	77%			
			Hybrid-DT	78%			

Another study conducted by Georga et al [20] explained AI methods(Random Forest, Logistic Regression, FRS, GAM, GBT algorithms) to find out the most effective predictors to detect Coronary Heart Disease (CAD). In this study, a more coherent and clustered dataset along with hybridization were thought to be incorporated to find out the best result. UCI Cleveland data has also been used in classification to predict the presence of cardiovascular disease in the human body in many studies. For example, Pouriyeh et al [18] investigated and compared the performance of the accuracy of different classification algorithms including Decision Tree, Naïve Bayes, SVM, MLP, KNN, Single Conjugative Rule Learner, and Radial Basis Function. Here, the hybrid model of SVM and MLP produced the best result. Latha [21] suggested a comparative analytical approach to determine the performance accuracy of ensemble techniques and found that the ensemble was a good strategy to improve accuracy. Again, Amin et al [19] identified significant features and improve the prediction accuracy by using different algorithms including, KNN, Decision Tree, Naïve Bayes, Logistic Regression, SVM, Neural Network, and an ensemble model that combines Na¨ıve Bayes and Logistic regression. It used the most impactful 9 features to calculate the prediction and the Na¨ıve Bayes gave the best accuracy. Alaa et al [22] used UK Biobank data to develop an algorithmic tool using Auto-Prognosis that automatically selects features and tunes ensemble models. Six ML algorithms were used in this study, namely, SVM, Random Forest, Neural Network, AdaBoost, and Gradient Boosting. The Auto-Prognosis showed a higher AUC-ROC compared to all other standard ML models. In sum, the literature review showed that most of the studies focusing on cardiovascular diseases and ML were striving to explore the algorithm that showed the best performance in predicting the possibility of cardiovascular diseases. The summary of the literature review is shown in Table 2.1. Again, though different studies are conducted using different sets of datasets having a different number of features, no study has been conducted to explore how the performance

accuracy of different algorithms are varying due to different set of features used in different datasets. Thus, this study focuses on this issue.

CHAPTER 3

Methodology

3.1 Synopsis of this Chapter

This chapter represents the methodology of this research work. It will introduce the experimental set up configuration, dataset along with data preprocessing and will give a brief description machine learning algorithms used here for predicting. Figure 3.1 addresses the overall methodology of this research.

3.2 Experimental Setup

All experiments in this study were conducted on a laptop computer with Intel Core (TM) i7-8565U CPU @ 4.60GHz, 16GB of DDR4 RAM, and Intel UHD Graphics 620 GPU. The proposed architecture requires more complex computation for performing prediction which requires very high CPU and GPU configuration. As the laptop used in this study did not match the required hardware configuration to execute the proposed architecture, Google Colaboratory was used to facilitate the executions and the computations also. Google Colaboratory is a cloud-based service which allows to perform any kinds of Machine Learning research and can be used for any kind of education purposes. It is designed basically to replace a desktop-based Python Jupyter notebook into a cloud domain environment which requires no setup to use.

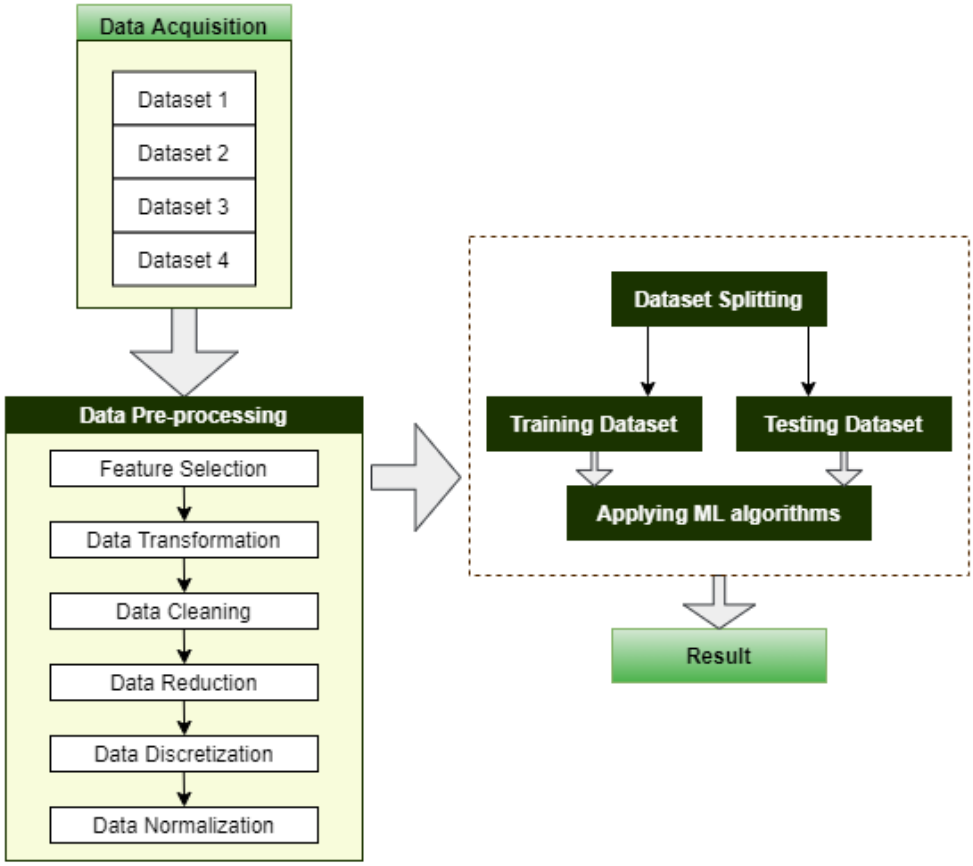

Fig. 3.1. The overview of research methodology

It concedes an opportunity to develop Machine Learning applications by using any python libraries. To perform smoothly and al quickly, it provides free Graphics Processing Unit (GPU) memory. It also holds a configuration that provides- Tesla k80 GPU- 12 GB, Xenon Processors 2.3Ghz., 12 GB of RAM and 33 GB of Disk for free usage.

3.3 Machine Learning Library

3.3.1 Pandas

In this study, Pandas has been used for data analysis and modeling. It allows fast, flexible, and also expressive data structures which are designed to work with the

"relational" or "labeled" data. It helps to manipulate the data by performing various tasks such as - handling of missing data, by providing automatic and explicit data alignment, providing mutability and so on.

3.3.2 Numpy

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of arrays. In this study, this library has been used to manipulate and process the array of data. Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities.

3.3.3 Matplotlib

Matplotlib is a package that graphs the data on Figures (i.e., windows, Jupyter widgets, etc.), each of which can contain one or more Axes (i.e., an area where points can be specified in terms of x-y coordinates (or theta-r in a polar plot, or x-y-z in a 3D plot, etc.). The most simple way of creating a figure with an axis is using pyplot.subplots. We can then use Axes.plot to draw some data on the axes. In this study, we have used this package to visualize the data into a graph.

3.3.4 Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps to explore and understand the data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets to focus on what the different elements of the plots mean, rather

than on the details of how to draw them.

3.3.5 Sklearn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps to explore and understand the data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets to focus on what the different elements of the plots mean, rather than on the details of how to draw them.

3.4 Dataset

In this study, there are four datasets that have been used in detecting and predicting CVD in this research work. Datasets have been acquired from both UCI machine learning repository and Kaggle [23]. The following table contains the source of the dataset along with the number of features and instances. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. It is used by students, educators, and researchers all over the world as a primary source of machine learning data sets. As an indication of the impact of the archive, it has been cited over 1000 times. Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

The datasets used here for supervised learning contain different sets of attributes

Table 3.1: Summary of datasets

Dataset	Source	Features	Instances
Dataset 1	Kaggle Cardiovascular Disease Dataset.	13	70000
Dataset 2	Kaggle Cleveland dataset	14	303
Dataset 3	Kaggle Integrated Heart Disease dataset	11	899
Dataset 4	UCI Framingham Dataset	14	4240

and they possess different amounts of instances.At first, we acquired data from various sources namely UCI machine learning repository [23] and Kaggle. Four different datasets were used in this research having different numbers of instances and the features of these datasets are not similar. As Dataset 1, we have used the "Kaggle Cardiovascular Disease Dataset" while having 13 features and 70000 instances. The " Cleveland dataset" from Kaggle having 14 attribues and 303 instances as Dataset 2. Likewise, we used "Kaggle Integrated Heart Disease dataset" as Dataset 3 that contains 76 features and 899 instances. As Dataset 4, we used "UCI Framingham Dataset" having 14 features and 4240 instances. However, we have considered 11 and 14 features for computational purposes from dataset 3 and dataset 4 respectively. There are 8 common features in each of the datasets.

In the dataset 1, there are 13 attributes and 70000 instances. It contains 13 columns as features attribute which are- ID (Identification number of subject), Age, Gender, Height (Height of the subject in inch), Weight (Weight of the subject in KG), Systolic blood pressure (Pressure exerted when blood is ejected into arteries), Diastolic blood pressure (Pressure within arteries between heartbeats), Cholesterol (Cholesterol level in blood in mg/dl), Glucose (Blood sugar level in blood in mg/dl), Smoke (if the subject is involved in smoking or not), Alcohol intake (if the subject consumes alcohol or not), Physical Activity (if the subject is involved in any form of physical activity or not) and at last, the class attribute, Presence or absence of cardiovascular disease. At a glance, the attributes of the dataset are show in table 3.2.

Table 3.2: Attributes of Dataset 1

Kaggle Cardiovascular Disease Dataset	
No.	**Attribute**
1	ID
2	Age
3	Gender
4	Height
5	Weight
6	Systolic blood pressure
7	Diastolic blood pressure
8	Cholesterol
9	Glucose
10	Smoke
11	Alcohol intake
12	Physical Activity
13	Presence or absence of cardiovascular disease

In the dataset 2, there are 14 attributes and 303 instances. This dataset is obtained from Kaggle. It contains 14 columns as features attribute which are- Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholestrol, Fasting Blood Sugar, Resting Electrocardiographic Results, Maximum Heart Rate, Exercise Induced Angina, St Depression Induced By Exercise Relative To Rest, The Slope of the Peak Exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, thal(Thalamus: gray matter located at the top of the brainstem. The thalamus derives its blood supply from a number of arteries),diagnosis of disease (Presence of the disease). The overall dataset attributes are shown in table 3.3.

In the dataset 3, there are 11 attributes and 978 instances. This dataset is obtained from Kaggle. It contains 11 columns as features attribute which are- Age, Sex, Chest Pain Type, Resting Blood Pressure(Blood pressure of the subject while resting), Smoking Year(If actively smoking, the number of years the subject has been active), Fasting Blood Sugar, Diabetes History(If the subject has any history

Table 3.3: Attributes of Dataset 2

Kaggle Cleveland dataset	
No.	**Attribute**
1	Age
2	Sex
3	Chest Pain Type
4	Resting Blood Pressure
5	Serum Cholesterol
6	Fasting Blood Sugar
7	Resting Electrocardiographic Results
8	Maximum Heart Rate
9	Exercise Induced Angina
10	St Depression Induced by Exercise ST segment
11	The Slope of the Peak Exercise ST segment
12	number of major vessels (0-3) colored by fluoroscopy
13	thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14	diagnosis of disease

of diabetes in the past. I.e- Diabetes during pregnancy), Family History Coronary, ECG(Electrocardiogram output of the subject), Pulse Rate(The number of heartbeats per minute), Presence of Disease. At a glance, the attributes of the dataset are show in table 3.4.

In the dataset 4, there are 14 attributes and 4240 instances. This dataset is obtained from Kaggle. It contains 11 columns as features attribute which are- Sex, Age, Currently Smoking,Cigarettes per day,Blood Pressure Medication (If the subject is on any sort of medication because of blood pressure), Prevalent Stroke (If the subject has previous experience of stroke or not), Prevalent Hyp (If the subject has previous experience of hypertension or not), Diabetes, Total Cholesterol Level, Systolic Pressure, Diastolic Pressure, BMI, Heartrate, risk of coronary heart disease. The overall dataset attributes are shown in table 3.5.

Table 3.4: Attributes of Dataset 3

Kaggle Integrated Heart Disease dataset	
No.	**Attribute**
1	Age
2	Sex
3	Chest Pain Type
4	Resting Blood Pressure
5	Smoking Years
6	Fasting Blood Sugar
7	Diabetes History
8	Family History Coronary
9	ECG
10	Pulse Rate
11	Presence of the Disease

Table 3.5: Attributes of Dataset 4

UCI Framingham Dataset	
No.	**Attribute**
1	Age
2	Sex
3	Currently Smoking
4	Cigarettes Per Day
5	Blood Pressure Medication
6	Prevalent Stroke
7	Prevalent Hyp
8	Diabetes
9	Total Cholesterol Level
10	Systolic Pressure
11	Diastolic Pressure
12	BMI
13	Heart Rate
14	risk of coronary heart disease

3.5 Data Preprocess

While using real world data, it is often found that the collected data are often noisy, inconsistent and have a lot of missing value or ambiguous value. With this sort of record, outcome is often poor in the data mining process. To have a desired outcome from a dataset, the dataset must be clean. The noisy data, ambiguous data should be eliminated. Ambiguous data are often generated by human or machine made mistakes like, typing error or unavailability during data collection or process. Having null or missing values in the dataset leads to inconsistent performance or inaccurate predictivity. Performance of ML algorithms hugely relies on data preprocessing [24, 25]. Thus, the noise, ambiguity and redundancy of data need to be reduced for better classification accuracy. This study consists of four datasets. Datasets carry features and attributes that hold their own significance. But this raw dataset was not used in this research work. The raw dataset was being transformed into an understandable format which was formed by fulfilling the following criteria-

- Duplicate rows in each of the datasets are removed.
- The rows having ambiguous data are removed.
- Missing numerical values in each of the datasets are replaced with the mean value of the particular column.
- The columns having text or string values are converted to numeric data to apply ML algorithms.
- Data having numeric values are normalized to obtain values between one and zero, since ML algorithms show better outcomes if values of numeric columns are changed in the dataset to a common scale, without distorting differences in the ranges of values.

- Outcomes of ML algorithms can be biased if an equal number of class instances doesn't exist in the training set of data [26]. To eradicate class instances imbalance problems, Synthetic Minority Over-sampling Technique (SMOTE) [27] was used.

- training and testing are performed with a random train-test split of 70-30.

3.6 Heatmap analysis

We've generated heatmaps of our data. Heatmap is a useful visual representation for a large data set. Because color shade consumes less space than numbers to represent data. The diagonal reflects the feature that relates to itself. It is obvious that it would be 1. By definition, heatmap visualization or heatmap data visualization is a type of graphical representation of numeric data where colors are used to signify the meaning of each data point. The most widely used color scheme used for heatmap visualization is a warm-to-cold color scheme of warm colors representing high-value data points and cool colors representing low-value data points. In data science, A heatmap is a graphical representation where individual values of a matrix are represented as colors. A heatmap is very useful in visualizing the concentration of values between two dimensions of a matrix. This helps in finding patterns and gives a perspective of depth. Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are that is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1/off-white because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest of the number and color, the bar chart on the right

defines the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

3.6.1 Degree of correlation

- **Perfect:** If the value is near ± 1, then it is said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

3.6.2 Heatmap analysis of the datasets

Heatmap analysis of the datasets are shown in Figure 3.2, 3.3, 3.4, 3.5.

3.7 Analysing ML algorithms

The objectives of the analysis are to find the optimum set of features and an algorithm for the prediction of cardiovascular diseases. For this, different machine learning algorithms which used mostly for predicting cardiovascular disease were chosen that includes, Logistic Regression, Support Vector Machine, K-th Nearest Neighbor, Naïve Bayes, Random Forest and Gradient Boosting. These algorithms were applied on the selected four datasets by splitting the dataset into 70% as training set and 30% as testing set.

3.7.1 Logistic Regression (LR)

Logistic Regression is one of the widely used ML algorithms. It is used in modeling or predicting because of its less computational complexity [28]. LR is considered as

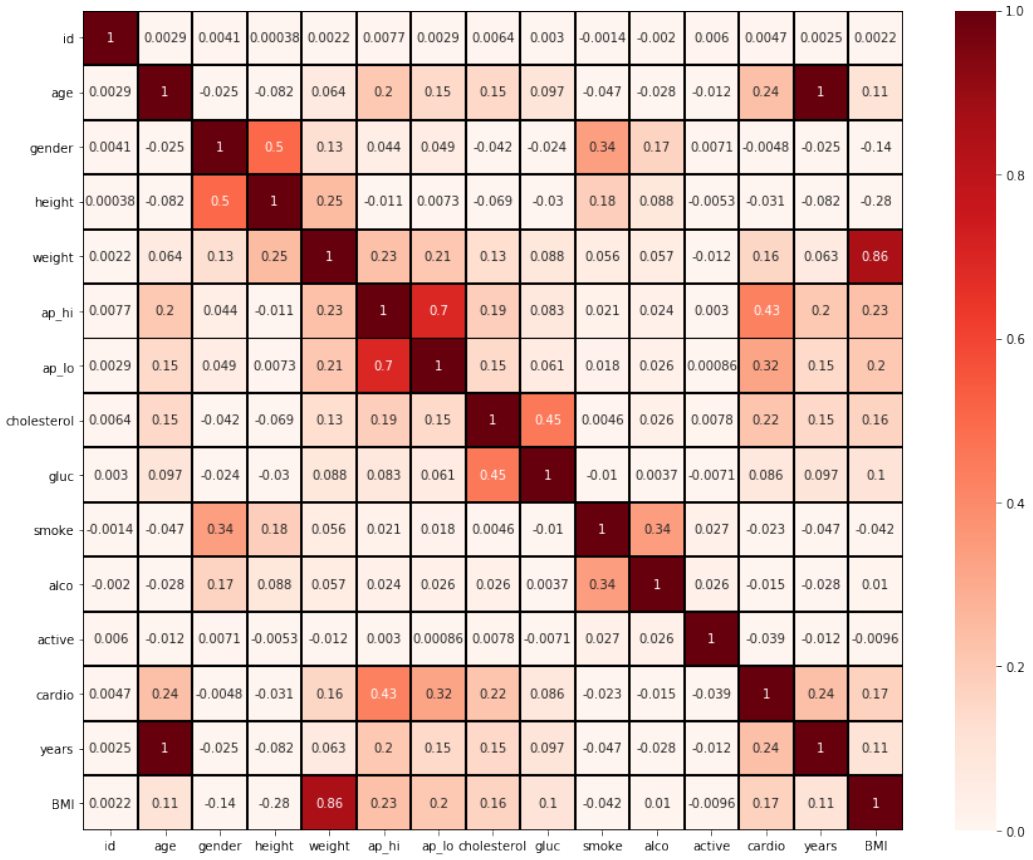

Fig. 3.2. Heatmap Analysis of Dataset 1

the standard statistical approach to modeling binary data. It is a better alternative for a linear regression which assigns a linear model to each of the class and predicts unseen instances basing on majority vote of the models [29].Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The LR process is addressed in figure 3.6

Some interesting findings can be observed by using the Logistic Regression method as the prediction model considering predefined classes of features and results are shown in Figure 3.7. Here, datasets are shown in X axis and performance measures are shown in Y axis. Different datasets come with different features and outcomes. The highest accuracy 93.8% was attained by Dataset 3 with 93.7% precision, 100%

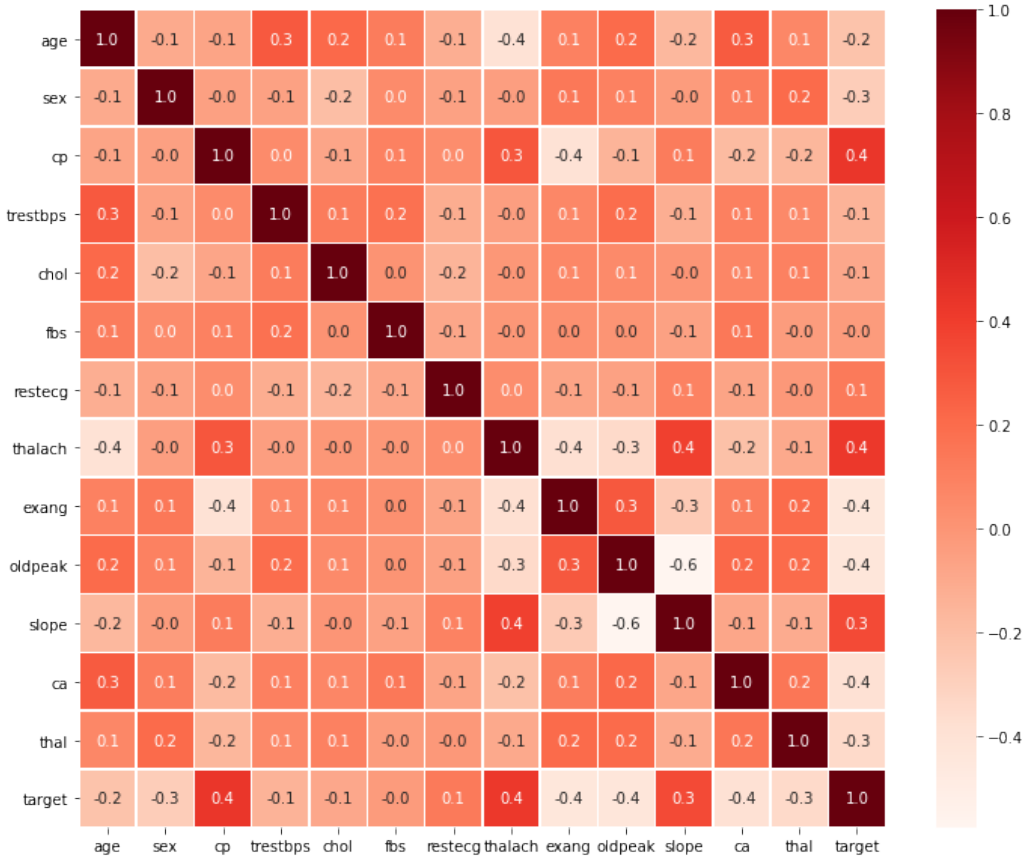

Fig. 3.3. Heatmap Analysis of Dataset 2

sensitivity and 0% specicity.

3.7.2 Support Vector Machine (SVM)

SVM is a supervised pattern classification model which is used as a training algorithm for learning classification and regression rules from gathered data. The purpose of this method is to separate data until a hyperplane with high minimum distance is found. SVM is used to classify two or more data types [30]. SVM is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs. The SVM mechanism is

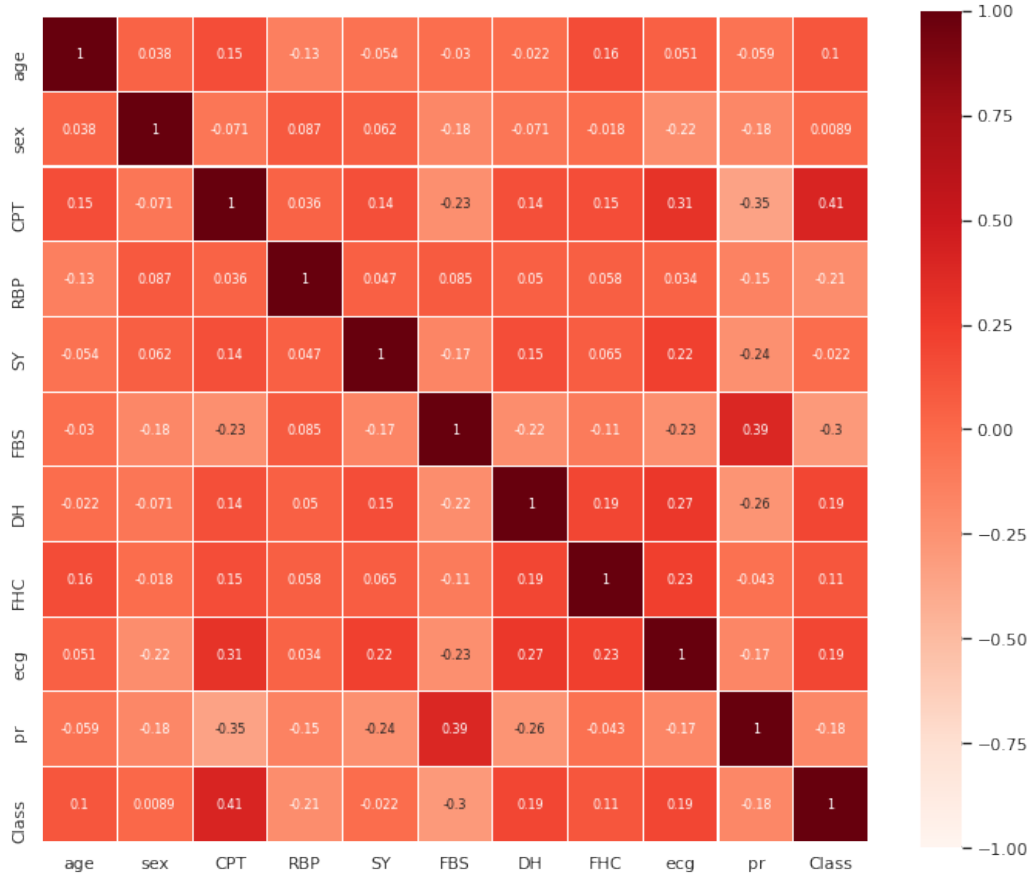

Fig. 3.4. Heatmap Analysis of Dataset 3

shown in figure 3.8

Results of applying SVM on predefined classes of features are shown in Figure 3.9. Here, the best performance measure was observed for dataset 3, having accuracy, precision, sensitivity, specificity of 96.39%, 93.79%, 100%, 0% respectively.

3.7.3 K-th Nearest Neighbor

KNN is a supervised learning method which is used for diagnosing and classifying cancer. In this method, the computer is trained in a specific field and new data is given to it. Additionally, similar data is used by the machine for detecting (K); hence, the machine starts nding KNN for the unknown data [30]. KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms.

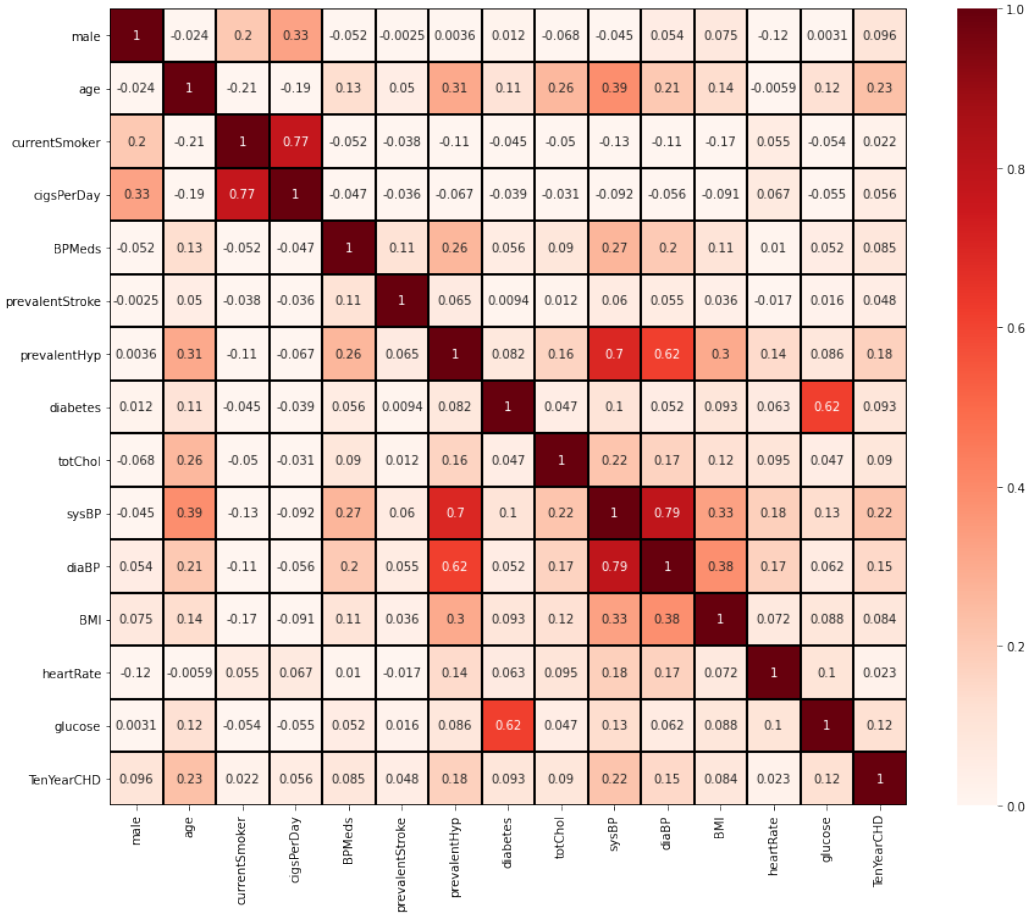

Fig. 3.5. Heatmap Analysis of Dataset 4

The KNN is a supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. The process of KNN is shown in figure 3.10

This algorithm has been performed based on different sets of features of different datasets. The performance measure of this algorithm is found to be less effective than previous algorithms. Here better performance was observed using dataset 3 (see Figure 3.11). The performance measures namely, accuracy, precision, sensitivity and specificity were 88.37%, 92.57%, 99.67%, 0% respectively.

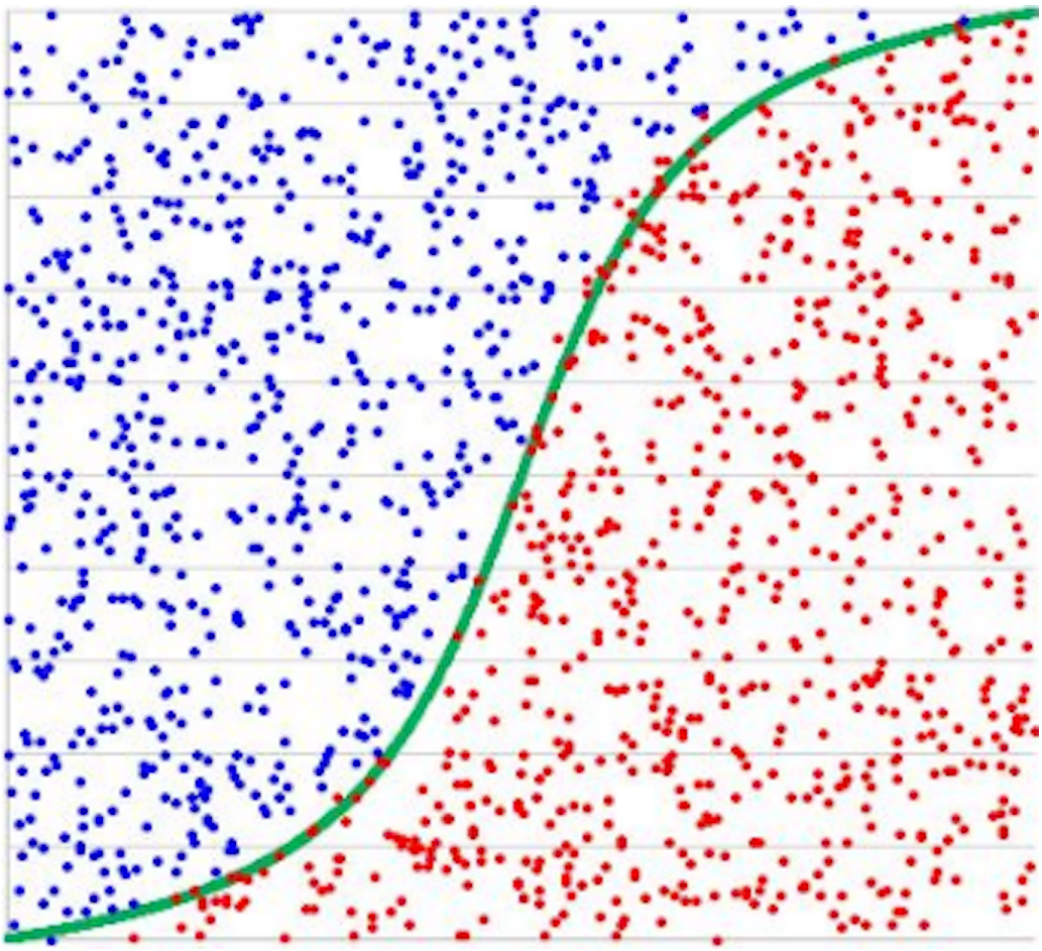

Fig. 3.6. Logistic Regression

3.7.4 Naïve Bayes

Naïve Bayes refers to a probabilistic classifier that applies Bayes' theorem with robust independence assumptions. In this model, all properties are considered separately to detect any existing relationship between them. It assumes that predictive features are conditionally independent given a class. Moreover, the values of the numeric features are distributed within each class. NB is fast and performs well even with a small dataset [30]. Naive Bayes (NB) is 'naive' because it makes the assumption that features of a measurement are independent of each other. This is naive because it is (almost) never true. Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly

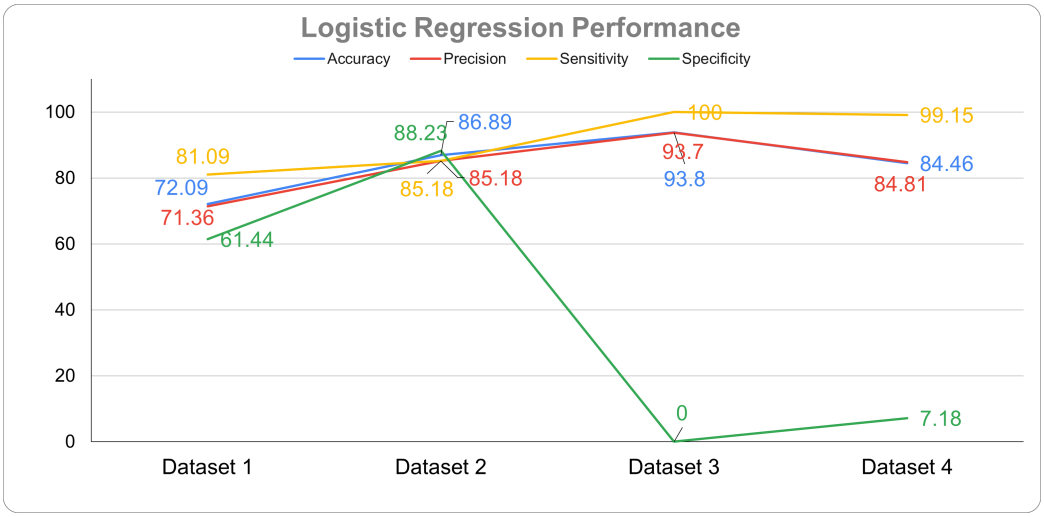

Fig. 3.7. Performance of Logistic Regression algorithm on different datasets

used in text classification and with problems having multiple classes. The NB process is addressed in figure 3.12

Findings from the performance of this algorithm from different datasets vary because of the selection of features. Results of these are shown in Figure 3.13 . The best accuracy can be obtained from Dataset 3 having accuracy, precision, sensitivity, specificity of 92.3, 93.9, 100, 0 percent respectively.

3.7.5 Random Forest (RF)

RF algorithm is used at the regularization point where the model quality is highest. RF builds numerous numbers of Decision Trees using random samples with a replacement to overcome the problem of Decision Trees. RF is used in the unsupervised mode for assessing proximities among data points [30]. Random Forest is ensemble model made of many decision trees using bootstrapping, random subsets of features, and average voting to make predictions. This is an example of a bagging ensemble. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. The RF mechanism is shown

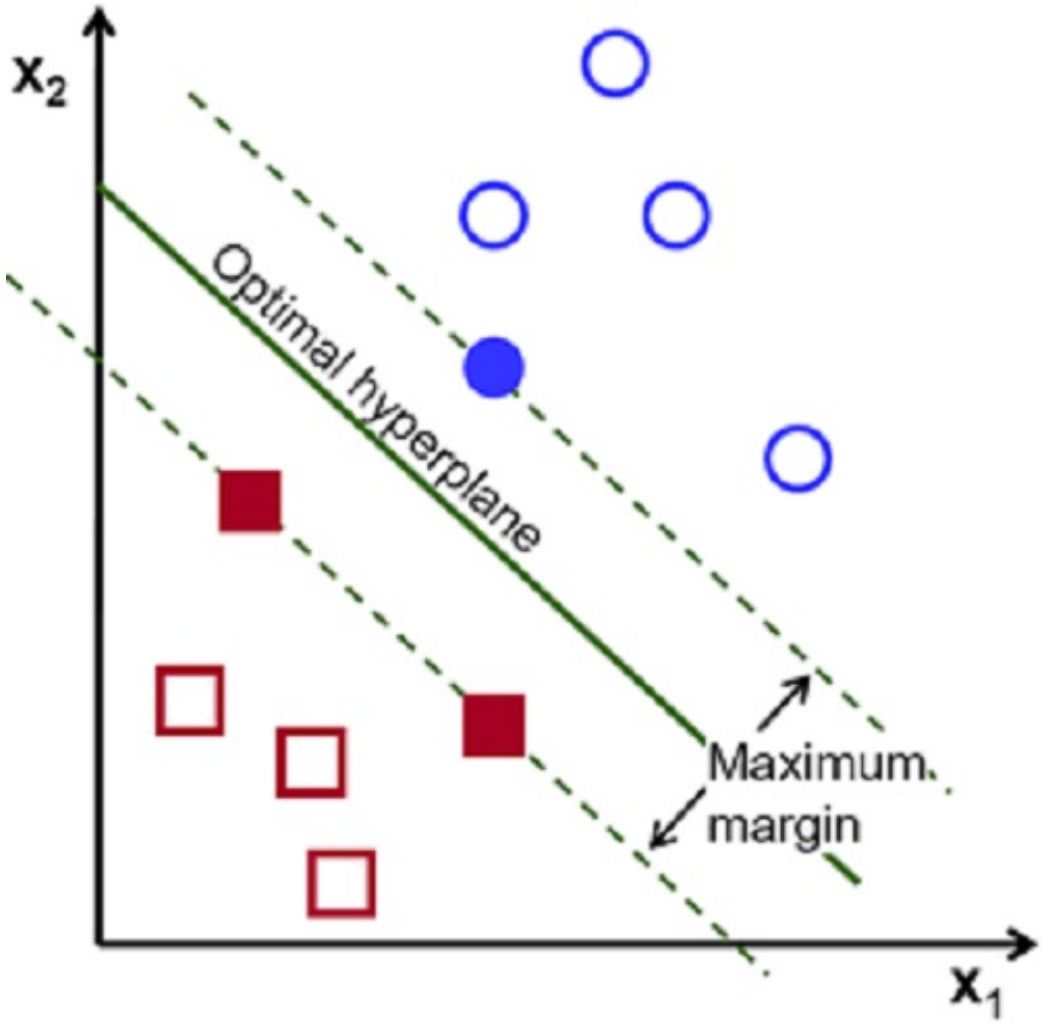

Fig. 3.8. SVM

in figure 3.14

By performing Python scripts on 4 different datasets on predefined features. Results of these are shown in Figure 3.15. The best accuracy of this algorithm can be observed for dataset 3. The accuracy, precision, sensitivity and specificity of this algorithm for this particular dataset are 93.85, 94.67, 100, 37.5 percent respectively.

3.7.6 Gradient Boosting

The gradient boosting is a machine learning technique for regression and classification [13]. Gradient boosting is a type of machine learning boosting. It relies on the

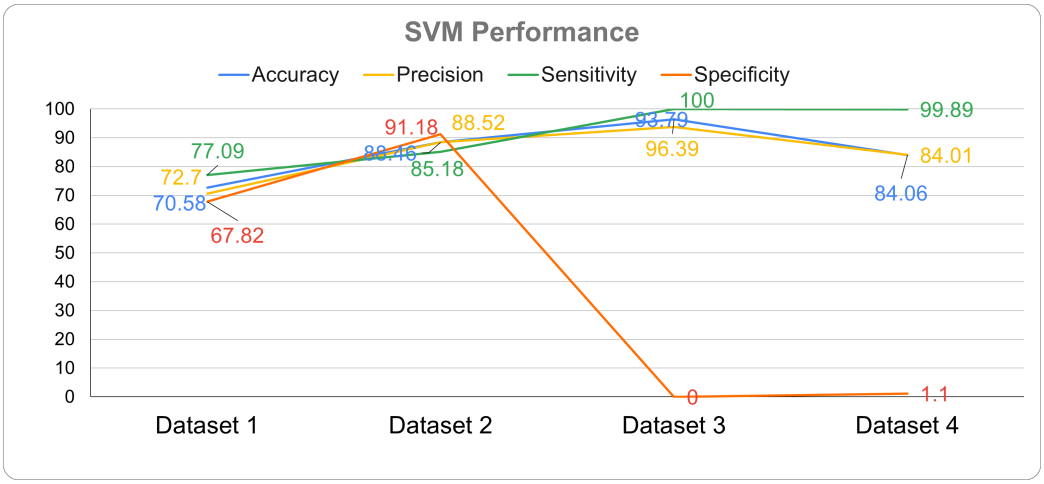

Fig. 3.9. Performance of SVM algorithm on different datasets

intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. Gradient boosting is a greedy algorithm and can overfit a training dataset quickly. It can benefit from regularization methods that penalize various parts of the algorithm and generally improve the performance of the algorithm by reducing overfitting.The process of GB is shown in figure 3.16

Different datasets with different sets of features were considered to compute the performances for this algorithm and the results are shown in Figure 3.17. The best outcome in terms of accuracy comes from dataset 3. Overall performance measures namely, accuracy, precision, sensitivity and specificity were 89.8, 91.65, 96.36, 13.7 percent respectively.

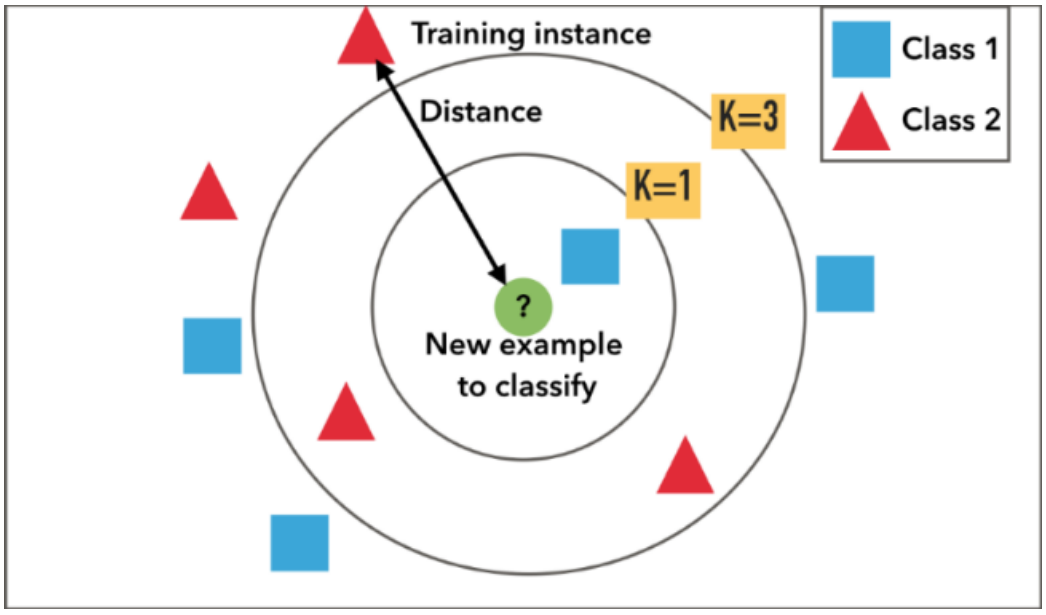

Fig. 3.10. KNN

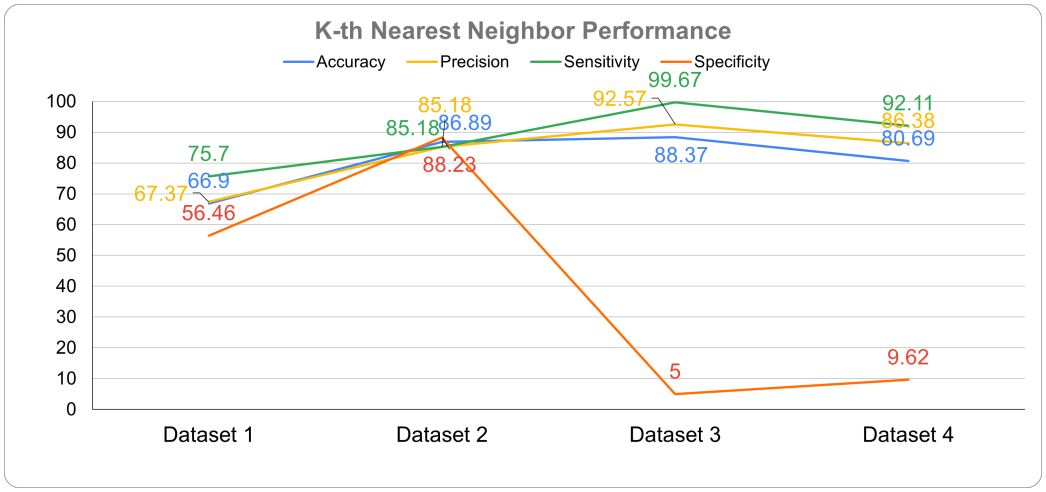

Fig. 3.11. Performance of KNN algorithm on different datasets

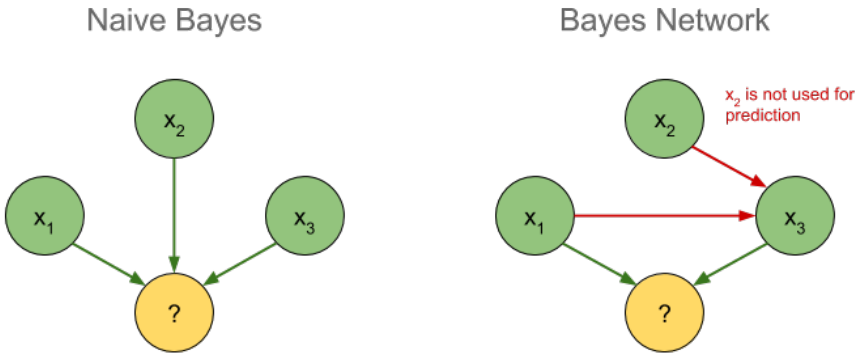

Fig. 3.12. Naive Bayes

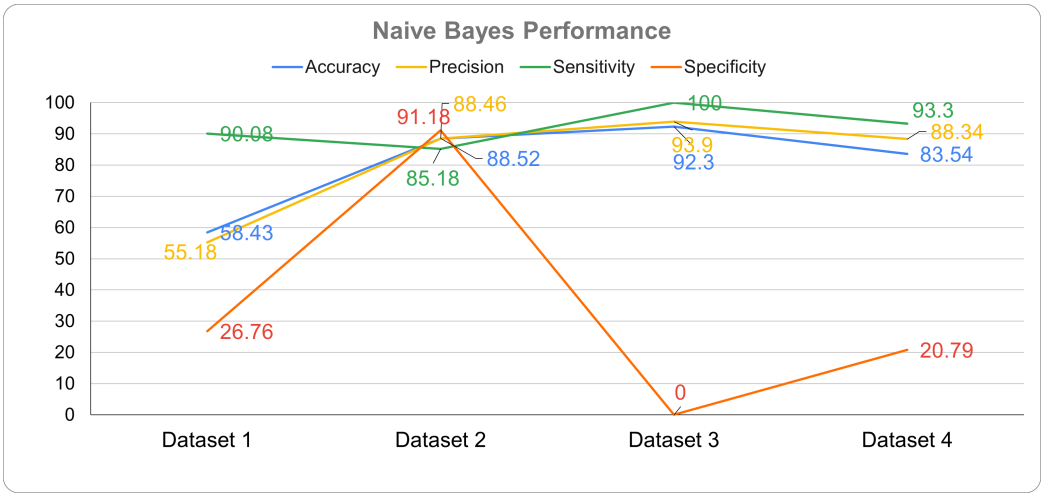

Fig. 3.13. Performance of Naive Bayes algorithm on different datasets

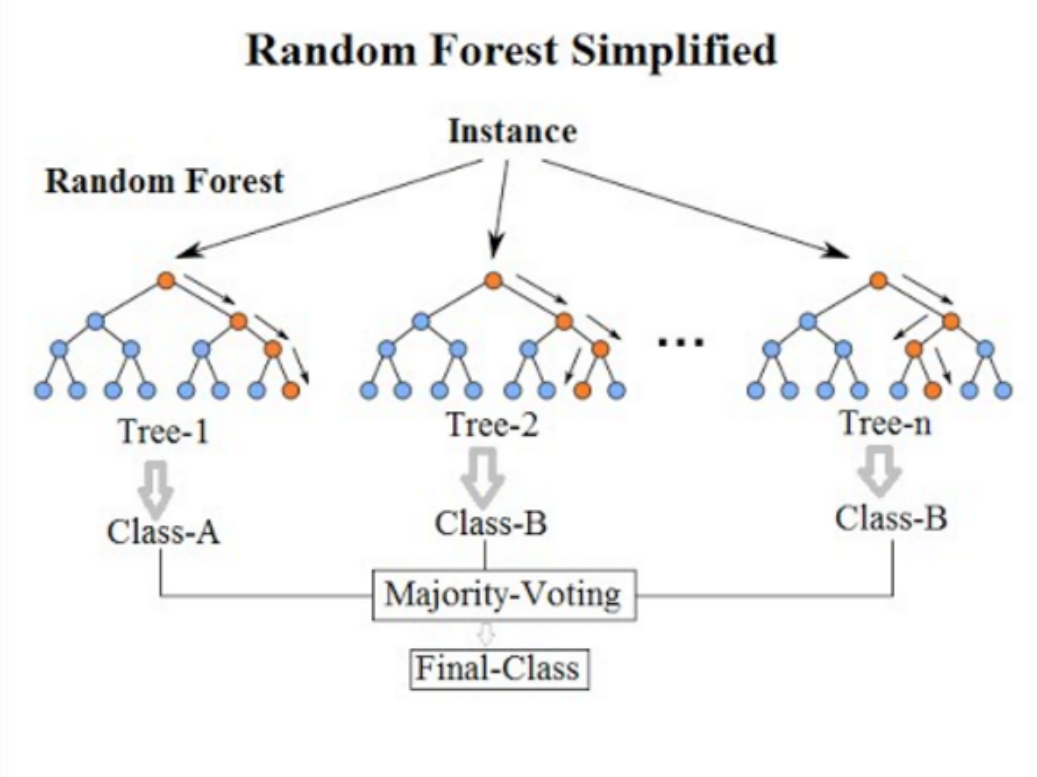

Fig. 3.14. Random Forest

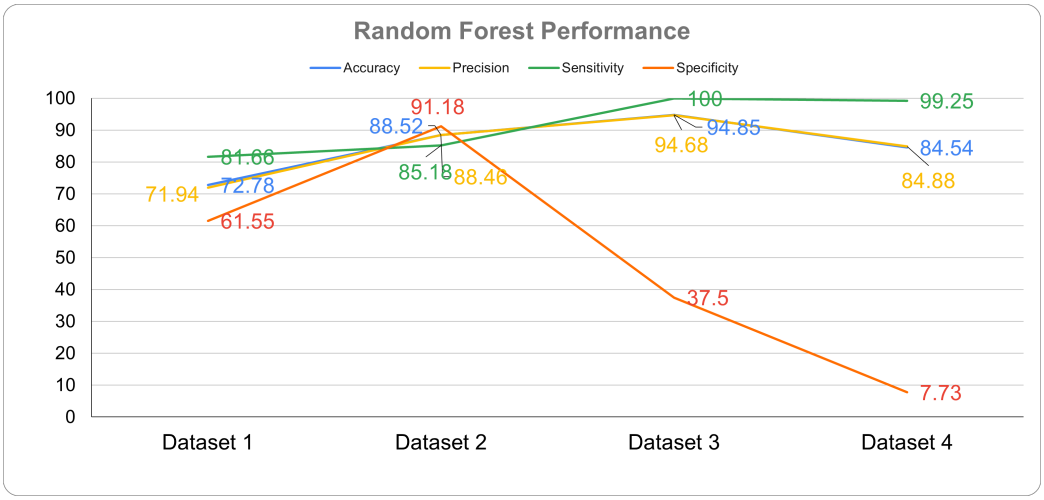

Fig. 3.15. Performance of Random Forest algorithm on different datasets

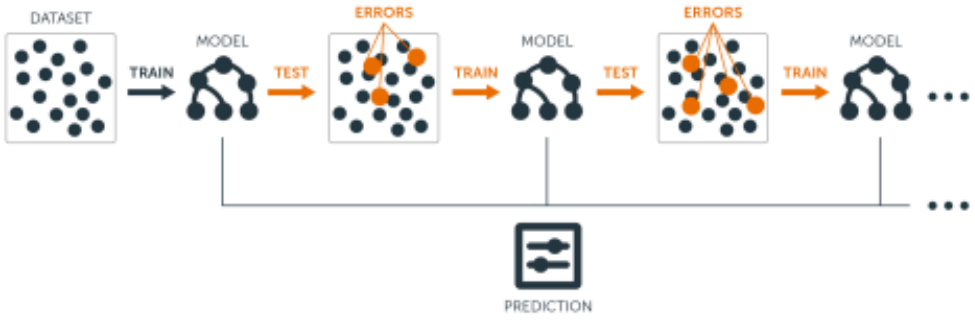

Fig. 3.16. Gradient Boosting

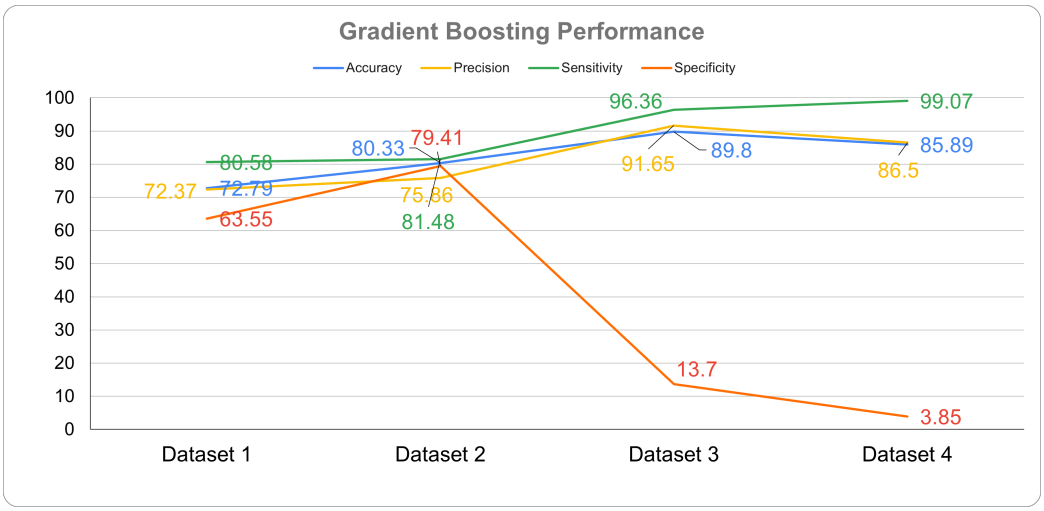

Fig. 3.17. Performance of Gradient Boosting algorithm on different datasets

CHAPTER 4

Results and Discussion

4.1 Synopsis of this Chapter

The study result highlights that with different dataset along with different features show diversity in the result. Datasets used here for computation purposes consist of different numbers of features. As they have different features, this could be one of the reasons of varying the prediction accuracy.

4.2 Result

In this research, we have implemented six machine machine learning algorithms in four different datasets. As the datasets come from different source and geographic location along with different set of attribute, it is normal to react differently on the ML algorithms that have been applied. In this section, we will be addressing the outcomes of the ML algorithms on the different datasets of cardiovascular diseases.

4.2.1 Performance of Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). In this study, the datasets contain binary class. Logistic regression is used to describe data and to explain the relationship

Table 4.1: Performance of the Logistic Regression for different datasets

Logistic Regression	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	72.09%	86.89%	93.8%	84.46%
Precision	71.36%	85.18%	93.7%	84.81%
Sensitivity	81.09%	85.18%	100%	99.15%
Specificity	61.44%	88.23%	0%	7.18%

between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The performance of the algorithms rely on the accuracy. In the logistic regression, the four datasets show diiferent acccuracy level. That are, for dataset 1, the accuracy is 72.09%. For dataset 2, it is 86.89%. While considering dataset 3 the accuracy is 93.8%. For the dataset 4, the obtained accuracy is 84.46%. The performance of this algorithm also comes in with other performance parameter like the precision, sensitivity and specificity. These information are shown in table 4.1.

4.2.2 Performance of SVM

SVM can be used for classification (distinguishing between several groups or classes) and regression (obtaining a mathematical model to predict something). They can be applied to both linear and non linear problems. In this study, we have used linear SVM technique. The performance of this algorithm has also been considered based on the accuracy. Because of different background and set of attribute, the results are different. 72.7%, 88.52%, 96.39%, 84.01% are the accuracy of the dataset 1, 2, 3 and 4 accordingly. There are other parameters that have also been measured during the implementation. Other performance parameters namely the precision, sensitivity and specificity are shown in table 4.2.

Table 4.2: Performance of the Support Vector Machine for different datasets

Support Vector Machine	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	72.7%	88.52%	96.39%	84.01%
Precision	70.58%	88.46%	93.79%	84.06%
Sensitivity	77.09%	85.18%	100%	99.89%
Specificity	67.82%	91.18%	0%	1.1%

Table 4.3: Performance of the K-th Nearest Neighbour for different datasets

KNN	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	66.9%	86.89%	88.37%	80.69%
Precision	67.37%	85.18%	92.57%	86.38%
Sensitivity	75.7%	85.18%	99.67%	92.11%
Specificity	56.46%	88.23%	5%	9.62%

4.2.3 Performance of KNN

KNN is a supervised ML algorithm which is preferred because of its quick calculation time, simplicity in terms of interpretation, versatility – usefulness for regression and classification and high accuracy – that do not need to compare with better-supervised learning models mostly. There are four different accuracy that have been obtained as a result of the implementation of this very algorithm. For the dataset 1, the accuracy is 66.9%, The dataset 2 shows 86.89% accuracy. For the dataset 3, the reading of accuracy parameter is 88.37% and for the dataset 4, this reading is 80.69%. The table 4.3 shows the value of other parameters namely the precision, sensitivity and specificity.

4.2.4 Performance of Naive Bayes

Naive bayes does quite well when the training data doesn't contain all possibilities so it can be very good with low amounts of data. The performance of the algorithms rely on the accuracy. In the logistic regression, the four datasets show diiferent acccuracy level. That are, for dataset 1, the accuracy is 58.43%. For dataset 2, it

Table 4.4: Performance of the Naive Bayes for different datasets

Naive Bayes	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	58.43%	88.52%	92.3%	83.54%
Precision	55.18%	88.46%	93.9%	88.34%
Sensitivity	90.08%	85.18%	100%	93.3%
Specificity	26.76%	91.18%	0%	20.79%

Table 4.5: Performance of the Random Forest for different datasets

Random Forest	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	72.78%	88.52%	94.85%	84.54%
Precision	71.94%	88.46%	94.68%	84.88%
Sensitivity	81.66%	85.18%	100%	99.25%
Specificity	61.55%	91.18%	37.5%	7.73%

is 88.52%. While considering dataset 3 the accuracy is 92.3%. For the dataset 4, the obtained accuracy is 83.54%. The performance of this algorithm also comes in with other performance parameter like the precision, sensitivity and specificity. These information are shown in table 4.4.

4.2.5 Performance of Random Forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). The performance of this algorithm has also been considered based on the accuracy. Because of different background and set of attribute, the results are different. 72.78%, 88.52%, 94.85%, 84.54% are the accuracy of the dataset 1, 2, 3 and 4 accordingly. There are other parameters that have also been measured during the implementation. Other performance parameters namely the precision, sensitivity and specificity are shown in table 4.5.

Table 4.6: Performance of the Gradient Boosting for different datasets

Gradient Boosting	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Accuracy	72.79%	80.33%	89.8%	85.89%
Precision	72.37%	75.86%	91.65%	86.5%
Sensitivity	80.58%	81.48%	96.36%	99.07%
Specificity	63.55%	79.41%	13.7%	3.85%

4.2.6 Performance of Gradient Boosting

Gradient Boosting is a special type of Ensemble Learning technique that works by combining several weak learners(predictors with poor accuracy) into a strong learner(a model with strong accuracy). This works by each model paying attention to its predecessor's mistakes. There are four different accuracy that have been obtained as a result of the implementation of this very algorithm. For the dataset 1, the accuracy is 72.79%, The dataset 2 shows 80.33% accuracy. For the dataset 3, the reading of accuracy parameter is 89.8% and for the dataset 4, this reading is 85.89%. The table 4.6 shows the value of other parameters namely the precision, sensitivity and specificity.

4.3 Discussion

The study result highlights that with different dataset along with different features show diversity in the result. Datasets used here for computation purposes consist of different numbers of features. As they have different features, this could be one of the reasons of varying the prediction accuracy.

Again, different datasets show difference in accuracy as shown in Table 4.7 and Figure 4.1 . For dataset 1, 2, 3 and 4 comparatively better accuracy was obtained using Gradient Boosting; SVM, Naïve Bayes and Random Forest; Random Forest; Random Forest and Gradient Boosting respectively. The results thus indicate that Random Forest shows the best accuracy in most of the datasets.

Table 4.7: Performance of the selected ML techniques for different datasets

Dataset	Performance Measures	Logistic Regression	SVM	K-NN	Naïve Bayes	Random Forest	Gradient Boosting
Dataset 1	Accuracy	71.36%	72.7%	66.9%	58.43%	72.78%	72.79%
	Precision	81.09%	70.58%	67.37%	55.18%	71.94%	72.37%
	Sensitivity	61.44%	77.09%	75.7%	90.08%	81.66%	80.58%
	Specificity	71.36%	67.82%	56.46%	26.76%	61.55%	63.55%
Dataset 2	Accuracy	86.89%	88.52%	86.89%	88.52%	88.52%	80.33%
	Precision	85.18%	88.46%	85.18%	88.46%	88.46%	75.86%
	Sensitivity	85.18%	85.18%	85.18%	85.18%	85.18%	81.48%
	Specificity	88.23%	91.18%	88.23%	91.18%	91.18%	79.41%
Dataset 3	Accuracy	93.8%	96.39%	88.37%	92.3%	94.85%	89.8%
	Precision	93.7%	93.79%	92.57%	93.9%	94.68%	91.65%
	Sensitivity	100%	100%	99.67%	100%	100%	96.36%
	Specificity	0%	0%	5%	0%	37.5%	13.7%
Dataset 4	Accuracy	84.46%	84.01%	80.69%	83.54%	84.54%	85.89%
	Precision	84.81%	84.06%	86.38%	88.34%	84.88%	86.5%
	Sensitivity	99.15%	99.89%	92.11%	93.3%	99.25%	99.07%
	Specificity	7.18%	1.1%	9.62%	20.79%	7.73%	3.85%

Each of the algorithm applied in each particular dataset, the best performance was observed for dataset 3 having 11 features. Thus, the results indicated that for predicting cardiovascular disease in the human body, features used in the dataset 3 is most likely to be the best recommended attributes. The features considered in Dataset 3 are Age, Sex, Chest Pain Type, Resting Blood Pressure, Smoking Year, Fasting Blood Sugar, Diabetes History, Family History Coronary, ECG, Pulse Rate and Presence of Disease.

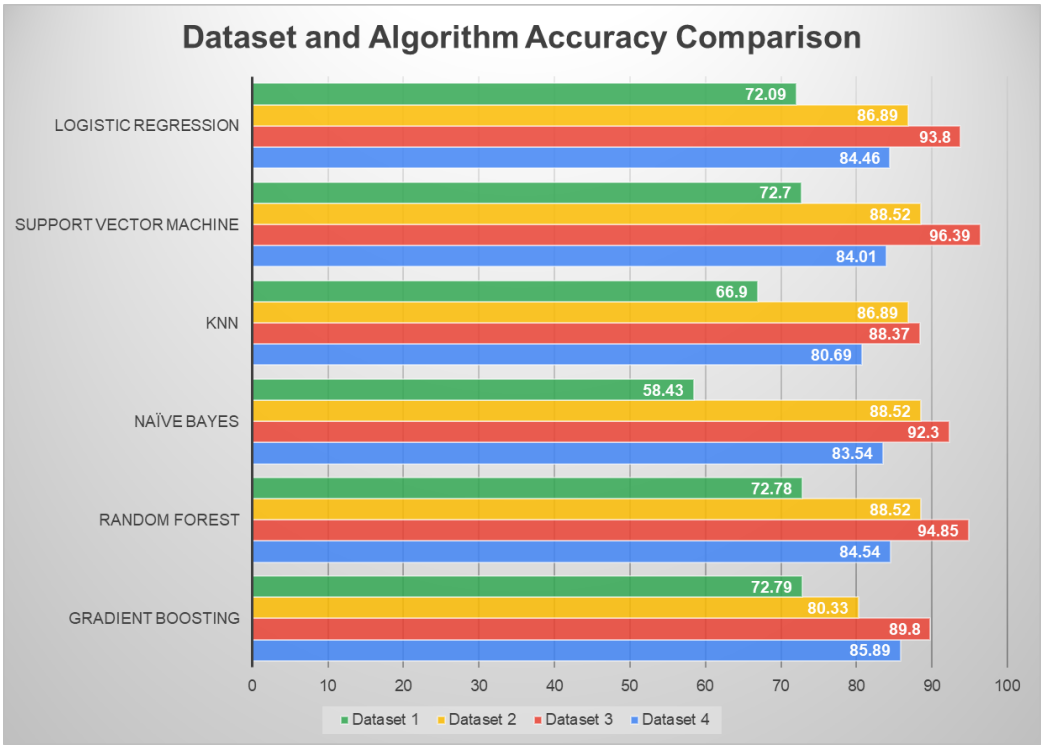

Fig. 4.1. Accuracy comparison

CHAPTER 5

CONCLUSION

5.1 Synopsis of this Chapter

This Chapter summarizes the overall work of this study. This chapter also addresses the impact of this thesis. We will also discuss the limitations and future works of this study.

5.2 Conclusion

Cardiovascular diseases have been one of the note-worthy reasons for mortality all over the world. According to the reports of WHO, every year more than 18 million people die because of cardiovascular diseases which covers almost 31% of global death. Damages in parts or all of the heart, coronary artery, or inadequate supply of nutrients and oxygen to this organ result in cardiovascular disease. Several lifestyle choices can increase the risk of heart disease that include, for example, high blood pressure and cholesterol, smoking, overweight and obesity, and diabetes. Disease detection is generally dependent on the experience and expertise of doctors. Though the decision support system could be a more feasible choice in the diagnosis of cardiovascular diseases through prediction. Healthcare organizations and hospitals collect data from patients regarding various health-related issues all around the world. The collected

set of data can be utilized using various machine learning classification techniques to draw effective insights that are overwhelming for human minds to comprehend. To rectify the hospital errors, prevention, easy detection of diseases along with better health policy-making, and preventable hospital deaths data mining applications can be used. In the vein, prediction of cardiovascular disease using machine learning can efficiently assist medical professionals.

The symptoms related to this disease may increase slowly but can go unpredicted for a very long time. Moreover, some people may become reluctant to immediately act even when they find out that something is wrong with them. Though it is increasing at an alarming rate, most people are unaware of this serious problem. There are a number of researches that show that an early detection might help in slowing down the deterioration process and help one in avoiding the later irreversible state. In today's age, Machine learning algorithms can play a significant role in diagnosing such conditions in less time and a price-friendly way.

In this research, for early and accurate detection of cardiovascular diseases we used six different approaches. We applied those approaches on four different datasets from the UCI machine learning repository and kaggle which consisted of different numbers of instances each holding its own significance. This raw dataset was transformed into an understandable format and was split into 70% and 30% for training and testing purposes. The result showed that 11 features of dataset 3 are the most efficient features while the Random Forest showed the best accuracy for most of the datasets with different set of features.

While the existing work have primarily focused on the implementation of several algorithms in a particular dataset and then compared their performance; This study demonstrated performance comparison among multiple datasets having different set of features along with evaluating multiple machine learning algorithms on them.

5.3 Limitations and future work

There are some limitations that can be mentioned. One of the limitations of this work is, this study considers only the traditional and ensemble ML algorithms. Different insights can be brought by either incorporating more algorithms or hybrid ensemble models. An app could have been built considering the convenience of the users and patients. These disease rely hugely on the geography of the patient. If individual prediction system can be incorporated for each geographical location, the benefited user group can grow even larger.

In future, an app or tool can be developed using ML algorithms to detect cardiovascular disease. Besides, algorithms can also be applied on new datasets to validate and generalize the outcomes of this research related to the best features to predict cardiovascular diseases.

REFERENCES

[1] A. Timmis, N. Townsend, C. Gale, R. Grobbee, N. Maniadakis, M. Flather, E. Wilkins, L. Wright, R. Vos, J. Bax, *et al.*, "European society of cardiology: cardiovascular disease statistics 2017," *European heart journal*, vol. 39, no. 7, pp. 508–579, 2018.

[2] G. Santulli, "Epidemiology of cardiovascular disease in the 21st century: Updated updated numbers and updated facts," *Journal of Cardiovascular Disease Research*, vol. 1, no. 1, 2013.

[3] E. G. Nabel, "Cardiovascular disease," *New England Journal of Medicine*, vol. 349, no. 1, pp. 60–72, 2003.

[4] M. I. Szerlip and H. M. Szerlip, "Identification of cardiovascular risk factors in homeless adults," *The American journal of the medical sciences*, vol. 324, no. 5, pp. 243–246, 2002.

[5] D. A. Falb and M. A. Gimbrone Jr, "Compositions and methods for treatment and diagnosis of cardiovascular disease," Sept. 26 2000. US Patent 6,124,433.

[6] D. Williams and J. Hill, "Machine learning," May 19 2005. US Patent App. 10/939,288.

[7] T. T. Inan, M. B. R. Samia, I. T. Tulin, and M. N. Islam, "A decision support

model to predict icu readmission through data mining approach.," in *PACIS*, p. 218, 2018.

[8] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[9] X. J. Zhu, "Semi-supervised learning literature survey," 2005.

[10] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[11] P. Rajyalakshmi, G. S. Reddy, K. G. Priyanka, V. L. B. S. Sai, and D. Anveshini, "Prediction of cardiovascular disease using machine learning," *entropy*, vol. 23, p. 24.

[12] S. K. Sen, "Predicting and diagnosing of heart disease using machine learning algorithms," *Int. J. Eng. Comput. Sci*, vol. 6, no. 6, 2017.

[13] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1–7, IEEE, 2018.

[14] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," in *Information and Communication Technology for Competitive Strategies*, pp. 447–454, Springer, 2019.

[15] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.

[16] A. K. Dwivedi, "Performance evaluation of different machine learning techniques

for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018.

[17] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," *Heart Disease*, vol. 7, no. 1, pp. 129–137, 2015.

[18] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, pp. 204–207, IEEE, 2017.

[19] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[20] E. I. Georga, N. S. Tachos, A. I. Sakellarios, V. I. Kigka, T. P. Exarchos, G. Pelosi, O. Parodi, L. K. Michalis, and D. I. Fotiadis, "Artificial intelligence and data mining methods for cardiovascular risk prediction," in *Cardiovascular Computing—Methodologies and Clinical Applications*, pp. 279–301, Springer, 2019.

[21] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[22] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. 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, p. e0213653, 2019.

[23] A. Frank, A. Asuncion, *et al.*, "Uci machine learning repository, 2010," *URL http://archive. ics. uci. edu/ml*, vol. 15, p. 22, 2011.

[24] P. Vasant, I. Zelinka, and G.-W. Weber, eds., *Intelligent Computing and Optimization*. Springer International Publishing, 2020.

[25] I. Krak, O. Barmak, E. Manziuk, and A. Kulias, "Data classification based on the features reduction and piecewise linear separation," in *International Conference on Intelligent Computing & Optimization*, pp. 282–289, Springer, 2019.

[26] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review, gests international transactions on computer science and engineering 30 (2006) 25–36," *Synthetic Oversampling of Instances Using Clustering*.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[28] Y. Miah, C. N. E. Prima, S. J. Seema, M. Mahmud, and M. S. Kaiser, "Performance comparison of machine learning techniques in identifying dementia from open access clinical datasets," in *Advances on Smart and Soft Computing*, pp. 79–89, Springer, 2020.

[29] G. R. Kumar, G. Ramachandra, and K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques," *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 2, no. 4, p. 139, 2013.

[30] M. Tahmooresi, A. Afshar, B. B. Rad, K. Nowshath, and M. Bamiah, "Early detection of breast cancer using machine learning techniques," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 3-2, pp. 21–27, 2018.