Comparative Analysis of Machine Learning Models for the Prediction of Asthma Disease among the Cardiovascular Disease Patients

by

Asif Jahan

ID: 19376007

A thesis submitted to the Department of Mathematics & Natural Sciences in partial fulfilment of the requirements for the degree, Master of Science in Biotechnology

Department of Mathematics and Natural Sciences
BRAC University
June 2024

Declaration

I hereby declare that:

1. This thesis is our original work completed during my MS in Biotechnology degree at BRAC University.
2. It does not contain material previously published or written by others, except where properly cited with full and accurate referencing.
3. It does not include material that has been accepted or submitted for any other degree or diploma at any university or institution.
4. All sources of assistance have been duly acknowledged.

Student's Full Name & Signature:

Asif Jahan
19306007

Approval

The thesis/project entitled "Comparative Analysis of Machine Learning Models for the Prediction of Asthma disease among the Cardiovascular Disease Patients" is submitted by,

Asif Jahan (Student ID: 19306007)

during the Summer of 2024, has been approved as satisfactory in partial fulfillment of the requirements for the degree of MS in Biotechnology on June 2024.

Examining Committee:

Supervisor:
(Member)

____
Dr. Mohammad Rafiqul Islam
Professor, Department of Mathematics and Natural Sciences
BRAC University

Program Director:
(Member)

____
Dr. Munima Haque
Associate Professor, Department of Mathematics and Natural Sciences
BRAC University

Departmental Head:
(Chairperson)

____
Dr. Firoze H. Haque
Associate Professor, Department of Mathematics and Natural Sciences
BRAC University

Abstract

Cardiovascular diseases (CVD) are a leading cause of morbidity and mortality worldwide, and recent studies have highlighted a potential association between CVD and the development of asthma. Predicting the likelihood of asthma in patients with cardiovascular diseases is crucial for early intervention and effective management. Advances in medical technology, particularly in machine learning (ML), offer powerful tools for disease prediction. ML algorithms, a subset of Artificial Intelligence (AI), mimic human learning processes to train systems for predictive tasks. This study employs supervised classification ML algorithms, including Logistic Regression, K-Nearest Neighbour (KNN), Naïve Bayes, Decision Tree, and Random Forest, to predict the likelihood of asthma in individuals with cardiovascular diseases. The dataset comprises primary data collected from adults, including demographic information, medical history, and relevant health indicators. The data was meticulously cleaned to ensure accuracy. Using RapidMiner, we developed predictive models and generated confusion matrices for each algorithm to evaluate their performance. Our analysis revealed that Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest models achieved an accuracy of 84.78%, while KNN reached an accuracy of 79.86%. Despite their high accuracy, the models exhibited low recall rates, indicating a challenge in identifying true positive cases of asthma. Naïve Bayes demonstrated the highest precision, followed by Logistic Regression, Random Forest, and Decision Tree, with KNN trailing behind. Consistent PVN scores across most models underscored their reliability in predicting negative cases. The comparative analysis emphasizes the need to consider multiple performance metrics beyond accuracy for a holistic evaluation of predictive models. Our findings suggest that Random Forest, Naïve Bayes, and Logistic Regression are the most promising algorithms for predicting asthma likelihood in cardiovascular disease patients. However, further refinements and hyperparameter tuning are necessary to enhance recall rates and overall predictive performance. This study lays the groundwork for using machine learning in predicting asthma risks among cardiovascular disease patients, aiming to improve early detection and intervention strategies in clinical practice.

Acknowledgement

First and foremost, I am profoundly grateful to Almighty Allah (S.W.T)! for granting me the strength, patience, and wisdom to complete this work.

I would like to extend my sincere appreciation to Dr. Firoze H. Haque, Associate Professor and Chairperson of the Department of Mathematics and Natural Sciences, for giving me the opportunity to pursue my studies in this esteemed department.

My heartfelt gratitude also goes to the late Professor A. A. Ziauddin Ahmad, former Chairperson of the Department of Mathematics and Natural Sciences at BRAC University, the late Professor Dr. Nayyium Chowdhury, and of course, Professor Dr. A F M Yusuf Haider, whose inspiration and support have been a guiding light throughout my academic journey.

I am deeply indebted to my respected supervisor, Dr. Mohammad Rafiqul Islam, Professor in the Department of Mathematics and Natural Sciences, for his invaluable guidance, insightful advice, and unwavering support throughout this study.

I extend my sincere thanks to Dr. Mahbubul Alam Majumdar, Professor and Dean, School of Data and Sciences, Department of Mathematics and Natural Sciences, for his invaluable support and encouragement.

My appreciation also goes to Dr. Tibra Ali, Professor and Associate Dean, Research, School of Data & Sciences, Department of Mathematics and Natural Sciences, for his insightful guidance and support.

I would also like to express my gratitude to Professor Dr. M. A. Rashid, Chief Executive Officer of Ibrahim Cardiac Hospital & Research Institute, for granting me the privilege to conduct my research at his esteemed institution.

Lastly, my humble and heartfelt gratitude goes to my mom and dad. Their constant support, endless inspiration, and sense of humour have been indispensable, and I am ever grateful for the genes they have bestowed upon me!

Asif Jahan, June 2024

Table of Contents

List of tables

List of figures

Background:

Asthma is a prevalent chronic respiratory condition that has significant health implications, particularly for individuals with pre-existing cardiovascular diseases (CVD). Recent studies have suggested a potential association between CVD and the development of asthma, highlighting the importance of predicting asthma risk in patients with cardiovascular conditions. This essay will provide a comprehensive overview of the current status of asthma research and treatment among CVD patients in Bangladesh, including the main challenges and opportunities for future improvement. Despite being a significant health concern, asthma research and treatment in Bangladesh face several challenges, including limited funding and resources, inadequate data, and restricted access to advanced medical technologies and equipment. This essay examines the current state of asthma research and treatment among CVD patients in Bangladesh, the primary challenges in the field, and how artificial intelligence (AI) based algorithms can play a role in mitigating these issues.

The current state of asthma research and treatment in Bangladesh is limited. While there have been efforts to improve the field, such as the establishment of the Bangladesh Heart Foundation and the Bangladesh Society for Cardiology and Cardiovascular Surgery, these initiatives have been hampered by a lack of funding and resources. Most studies on asthma and its association with CVD in Bangladesh are observational, with few randomized controlled trials (RCTs) being conducted. This lack of RCTs limits the development of evidence-based treatments and guidelines for managing asthma in CVD patients. One of the major challenges facing asthma research and treatment in Bangladesh is the lack of data and information. There is a scarcity of accurate and reliable data on asthma prevalence among CVD patients, which hampers the development of effective and targeted interventions. Additionally, the limited use of electronic medical records and the lack of a centralized database make it difficult to collect and analyse data on asthma and CVD.

Another significant challenge is the limited availability of resources. This includes a shortage of trained healthcare professionals, inadequate medical facilities, and limited access to advanced medical technologies and equipment. These limitations impede the development of advanced treatments and therapies for asthma in Bangladesh. Moreover, another significant pitfall in Bangladesh is the lack of resources and infrastructure. A study published in the Journal of Cardiology and Cardiovascular Medicine found that the country has limited financial resources and faces significant constraints in its ability to allocate funding for healthcare and research (Hossain et al., 2019). This lack of funding hampers the development of advanced medical technologies, such as advanced imaging equipment, essential for early detection and treatment of asthma in CVD patients. Furthermore, there is a shortage of trained healthcare professionals, particularly in rural areas, which limits access to quality care and diagnosis (Hossain et al., 2019). Additionally, the lack of healthcare facilities and services in Bangladesh, particularly in rural areas, is a major concern. A study published in the Bangladesh Journal of Medicine found that many rural areas lack basic healthcare services, making it difficult for asthma patients with CVD to access adequate care (Karim et al., 2018). This lack of access to care contributes to the high incidence of asthma-related morbidity and mortality in Bangladesh (Karim et al., 2018).

Another challenge is the lack of public awareness about the importance of preventative measures, such as a healthy diet, exercise, and smoking cessation. A study published in the Bangladesh Journal of Medicine found that there is limited public education on these issues, which contributes to the high incidence of risk factors such as unhealthy diets, physical inactivity, and tobacco use (Karim et al., 2018). Additionally, there is limited access to preventative measures, such as asthma medications and inhalers, due to high costs and limited availability in rural areas (Karim et al., 2018).

In addition to these challenges, there is a lack of data and research on asthma among CVD patients in Bangladesh. A study published in the Journal of Cardiology and Cardiovascular Medicine found that there is limited data on the prevalence and incidence of asthma in CVD patients, making it difficult to develop effective prevention and treatment strategies (Hossain et al., 2019). This lack of data also hinders the ability to monitor and evaluate the effectiveness of current prevention and treatment efforts (Hossain et al., 2019). Despite these challenges, some progress has been made in recent years to address the growing threat of asthma in CVD patients in Bangladesh. For example, the government has taken steps to increase access to quality care by expanding the network of healthcare facilities and investing in the development of rural health clinics (Hossain et al., 2019). Additionally, efforts have been made to improve public awareness through education campaigns and community outreach programs (Karim et al., 2018).

One promising area of development in the fight against asthma in CVD patients in Bangladesh is the application of artificial intelligence (AI) in healthcare. AI has the potential to significantly improve the accuracy and efficiency of asthma diagnosis and treatment, as well as increase access to care in remote and underserved areas (Huang et al., 2020). For example, AI-powered imaging technologies can assist in the early detection of asthma, enabling healthcare providers to intervene early and prevent the progression of the disease (Huang et al., 2020). Additionally, AI algorithms can analyse large amounts of patient data to identify risk factors and develop personalized treatment plans (Huang et al., 2020). In light of all the major challenges mentioned previously, AI-based algorithms have the potential to play a significant role in improving asthma research and treatment in Bangladesh. AI algorithms can assist in collecting and analyzing data, improving the accuracy and reliability of information about asthma among CVD patients. This can help inform the development of effective interventions and treatments and lead to improved patient outcomes.

One example of an AI-based algorithm that can be applied to asthma research in Bangladesh is machine learning. Machine learning algorithms can be used to analyse large amounts of data on asthma and CVD and identify patterns and trends that would be difficult to detect through traditional methods. This information can then be used to inform the development of targeted interventions and treatments.

Another example of how AI can be applied to asthma research and treatment in Bangladesh is through the use of computer-aided diagnosis (CAD) algorithms. CAD algorithms can be used to analyse medical images and diagnose asthma more accurately and efficiently than traditional methods. This can improve the speed and accuracy of diagnoses, leading to improved patient outcomes. AI-based algorithms also have the potential to help address the limited availability of resources in asthma research and treatment in Bangladesh. For example, AI algorithms can be used to improve the efficiency of medical facilities and reduce the need for trained healthcare professionals. This can help to increase access to advanced medical technologies and equipment and improve the overall quality of care for patients with asthma and CVD.

The challenges facing asthma research and treatment in Bangladesh are significant, but there are also potential solutions to these problems. AI-based algorithms have the potential to play a major role in improving the field by assisting in the collection and analysis of data, improving the accuracy and efficiency of diagnoses, and increasing access to advanced medical technologies and equipment. To fully realize the potential of AI in asthma research and treatment among CVD patients in Bangladesh, it will be important to continue investing in the development and implementation of these technologies.

Chapter 1. Introduction:

Artificial intelligence (AI) and its subdomains, including machine learning (ML) and deep learning (DL), have the potential to revolutionize the detection, prevention, and treatment of various diseases, including asthma. Asthma is a significant chronic respiratory condition, especially prevalent among individuals with pre-existing cardiovascular diseases (CVD). Early detection and accurate diagnosis of asthma in CVD patients are crucial for effective treatment and management. This section discusses how AI, ML, and DL, coupled with expert computational skills, can enhance predictive capabilities in healthcare, specifically focusing on predicting the likelihood of asthma in individuals with cardiovascular conditions.

Asthma research in Bangladesh has traditionally focused on the epidemiology, risk factors, prevention, diagnosis, and management of the disease. The Bangladesh Demographic and Health Survey (BDHS) 2017-2018 reported a high prevalence of hypertension (26%), CVDs (9%), and obesity (12%) among adults (NIPORT et al., 2019). These risk factors are strongly associated with both asthma and CVD, highlighting the need for effective prevention and management strategies. Several studies have investigated the prevalence and risk factors of asthma in different populations in Bangladesh. For example, a study conducted in rural areas of Bangladesh found that the prevalence of asthma was higher in older adults, men, and those with hypertension and dyslipidemia (Islam et al., 2020). Another study reported a high prevalence of asthma risk factors among urban slum dwellers in Bangladesh, including tobacco use, physical inactivity, and unhealthy diets (Alam et al., 2020).

Research on asthma diagnosis and management in Bangladesh has shown progress in recent years. The use of spirometry and other diagnostic tools has improved the accuracy of asthma diagnosis and risk stratification (Rahman et al., 2019). The development of asthma registries has facilitated the monitoring of disease trends and the evaluation of treatment outcomes (Biswas et al., 2020). However, several challenges remain, including inadequate resources, limited access to specialized care, and a shortage of trained healthcare professionals. AI and ML algorithms can analyse large amounts of medical data, including patient demographics, risk factors, and imaging results, to identify patterns and predict the likelihood of asthma development in specific patient groups, such as those with cardiovascular diseases.

The main task of this research is to employ various machine learning models to predict the likelihood of cardiovascular disease patients in Bangladesh developing asthma. By using supervised classification ML algorithms, including Logistic Regression, K-Nearest Neighbour

(KNN), Naïve Bayes, Decision Tree, and Random Forest, the study aims to analyse a dataset comprising primary data from adults, including demographic information, medical history, and relevant health indicators. The goal is to determine which model performs best in predicting the likelihood of asthma in individuals with cardiovascular diseases, thereby providing valuable insights for early intervention and targeted healthcare strategies.

Machine learning (ML) has been particularly promising in predicting the likelihood of asthma in patients with pre-existing conditions, like CVD. For instance, ML algorithms can be trained on datasets that include various health indicators to predict the likelihood of a CVD patient developing asthma. This predictive capability is essential for early intervention and personalized healthcare.

Deep learning (DL), a subset of ML, has also been used to analyse medical imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI) scans, to detect and classify different types of respiratory conditions. In a study published in Radiology, researchers used DL to analyse lung CT images and found that the algorithm had high accuracy in identifying patients with respiratory diseases. This technology can be used to identify abnormalities in the lungs and airways with high accuracy and precision, aiding in early detection and diagnosis of asthma.

In addition to early detection, AI and ML can also be used to improve the management and treatment of asthma. For example, ML algorithms can optimize medication dosages for individual patients, considering factors such as age, weight, and other medical conditions. This can reduce the risk of side effects and improve treatment effectiveness. A study published in the European Journal of Epidemiology demonstrated how ML could identify optimal medication dosages for patients with asthma, resulting in reduced hospitalization rates.

One of the key applications of ML and DL in asthma research is risk prediction. Several studies have used these techniques to develop predictive models for asthma based on clinical, demographic, and lifestyle factors. For example, a study in Bangladesh used a random forest algorithm to predict the risk of asthma in a rural population based on age, sex, smoking, hypertension, and dyslipidemia (Bhowmik et al., 2021). Another study developed a DL-based model to predict the risk of asthma exacerbations in patients presenting with respiratory symptoms (Islam et al., 2021). These models help identify high-risk individuals and facilitate early intervention and prevention.

Expert computational skills are crucial for the development and implementation of AI and ML in asthma research. These skills enable the processing and analysis of large datasets, as well as the creation of complex models and algorithms. In a study published in the Journal of the American Medical Informatics Association, researchers used expert computational skills to analyse a large dataset of over 1 million imaging studies and found that the algorithm could identify patients at risk of developing asthma with high accuracy.

Machine learning algorithms such as Logistic Regression, K-Nearest Neighbour (KNN), Naïve Bayes, Decision Tree, and Random Forest can be used to predict the likelihood of asthma in cardiovascular disease patients. These algorithms analyse patterns in health data, classify individuals based on their risk factors, and identify the most important predictors of asthma.

In particular, clustering algorithms can identify subgroups of individuals with similar respiratory risk factors. K-nearest neighbours and K-means algorithms can classify individuals based on risk factors like age, blood pressure, and cholesterol levels. Random Forest algorithms can identify the strongest predictors of asthma, guiding targeted interventions and preventative measures.

AI and its subdomains, including ML and DL, coupled with expert computational skills, have the potential to greatly enhance the prediction and management of asthma. The use of these technologies can aid in early detection, improve the management and treatment of asthma, and ultimately save lives. However, more research is needed to fully understand the potential of these technologies and ensure their safe and effective implementation in clinical practice.

1.1 Objective :

This thesis endeavours to develop a robust predictive model to assess the likelihood of the occurrence of asthma in individuals with cardiovascular diseases (CVDs) using supervised machine learning classification algorithms. The specific objectives are as follows:

1. Curate and preprocess a comprehensive dataset encompassing crucial features and variables associated with both asthma and CVDs, followed by the application of appropriate machine learning algorithms for predictive modelling.
2. Attain a comprehensive understanding of various machine learning algorithms, exploring their intricacies, and evaluating their performance in the context of predicting asthma in CVD patients.
3. Conduct a meticulous comparative analysis of the predictive outcomes generated by different machine learning models, identifying and selecting the most effective model based on performance metrics.

The accomplishment of these objectives will significantly advance the field of healthcare analytics by furnishing a robust predictive model tailored for forecasting asthma in individuals with cardiovascular diseases. Furthermore, the identification of the optimal model holds substantial promise in aiding healthcare professionals, policymakers, and stakeholders in making well-informed decisions pertaining to patient care and management strategies.

1.2 An overview on the potentials of Machine Learning and Deep Learning to analyse complex cardiovascular data:

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, accounting for approximately 17.9 million deaths each year (World Health Organization, 2021). CVDs are often complex and multifactorial, and the diagnosis and management of these conditions require the analysis of large and heterogeneous datasets. Traditional statistical methods may be limited in their ability to extract insights from these datasets due to their reliance on predefined models and assumptions. However, recent advances in machine learning (ML) and deep learning (DL) have shown great potential in analysing complex cardiovascular data, enabling more accurate diagnosis, prognosis, and treatment of CVDs. Machine learning is a subfield of artificial intelligence that involves the use of algorithms and statistical models to enable computer systems to learn and improve from experience without being explicitly programmed

(Alpaydin, 2020). In the context of cardiovascular data analysis, ML algorithms can be trained on large datasets to identify patterns and relationships between various clinical and biological factors associated with CVDs. For example, a study by Johnson et al. (2016) used ML algorithms to analyse electronic health records (EHRs) of more than 55,000 patients with hypertension and identified novel risk factors for cardiovascular events, such as elevated serum potassium levels.

One of the key advantages of ML is its ability to handle high-dimensional and heterogeneous data. Traditional statistical methods may struggle to cope with the vast amount of data generated from different sources, such as genomics, proteomics, imaging, and clinical data. In contrast, ML algorithms can efficiently extract relevant features from these datasets and integrate them to provide a holistic view of the disease phenotype. For example, a study by Attia et al. (2019) used ML algorithms to analyse a combination of ECG, imaging, and clinical data to develop a deep learning model that can predict the risk of atrial fibrillation with high accuracy. Another advantage of ML is its ability to learn from data in a non-linear and adaptive manner. This means that ML algorithms can capture complex and non-linear relationships between different variables, which may not be apparent using traditional statistical methods. For example, a study by Beaulieu-Jones et al. (2018) used ML algorithms to analyse a large dataset of EHRs and identified novel associations between different clinical variables and the risk of myocardial infarction.

Despite the potential of ML in cardiovascular data analysis, there are several challenges that need to be addressed. One of the key challenges is the lack of standardized data formats and data quality control measures. The use of heterogeneous and inconsistent data can lead to biases and errors in ML algorithms, which can affect their accuracy and generalizability (Savarese et al., 2020). Therefore, efforts are needed to develop standardized data formats and quality control measures to ensure the accuracy and reliability of ML algorithms.

ML and DL are two branches of artificial intelligence (AI) that enable machines to learn from data and make predictions or decisions. ML algorithms can be trained on a set of labeled data to learn patterns, which can then be used to predict outcomes or classify new data. DL algorithms are a subset of ML algorithms that use artificial neural networks to learn and make predictions. DL algorithms can learn from large and complex datasets, making them suitable for analysing complex cardiovascular data.

One area where ML and DL algorithms have shown potential is the analysis of ECG data. ECG data provide information on the electrical activity of the heart and are widely used in the diagnosis and management of CVDs. Traditional ECG analysis involves visual interpretation by a trained cardiologist, which can be time-consuming and prone to human error. ML and DL algorithms can automate the analysis of ECG data, allowing for faster and more accurate diagnosis. For example, a study by Attia et al. (2019) showed that a DL algorithm could accurately detect atrial fibrillation (AF) from single-lead ECG recordings, with a sensitivity of 79.3% and a specificity of 79.6%. The DL algorithm also outperformed six cardiologists in detecting AF.

Another area where ML and DL algorithms have shown potential is the analysis of echocardiography data. Echocardiography is a non-invasive imaging technique that provides information on the structure and function of the heart. ML and DL algorithms can be trained on echocardiography data to predict outcomes such as heart failure and mortality. For example, a study by Cho et al. (2021) developed a DL algorithm that could predict all-cause mortality in patients with reduced ejection fraction using echocardiography data. The DL algorithm outperformed traditional risk prediction models such as the Seattle Heart Failure Model.

Furthermore, ML and DL algorithms have shown potential in the analysis of cardiac MRI data. Cardiac MRI provides detailed information on the anatomy and function of the heart and is widely used in the diagnosis and management of CVDs. However, the analysis of cardiac MRI data is often time-consuming and requires expert interpretation. ML and DL algorithms can automate the analysis of cardiac MRI data, allowing for faster and more accurate diagnosis. For example, a study by Oktay et al. (2018) developed a DL algorithm that could segment cardiac MRI data into the left ventricle, right ventricle, and myocardium. The DL algorithm achieved a mean Dice similarity coefficient of 0.93, which indicates a high level of accuracy.

One of the most promising applications of ML and DL in cardiovascular data analysis is risk prediction. ML and DL models can analyse multiple risk factors simultaneously and provide personalized risk predictions for individuals (Khera et al., 2018). For example, ML models have been used to predict the risk of developing heart failure in patients with hypertension (Cho et al., 2019). DL models have also been used to predict cardiovascular events, such as myocardial infarction and stroke, using ECG data (Attia et al., 2019). These models can improve the accuracy of risk prediction compared to traditional risk assessment tools, such as the Framingham Risk Score (FRS) (Rajkomar et al., 2018).

ML and DL can also assist in the diagnosis of cardiovascular diseases. DL models can analyse complex imaging data, such as cardiac MRI, to detect and classify abnormalities (Ouyang et al., 2020). For example, DL models have been used to diagnose heart failure with preserved ejection fraction (HFpEF) using cardiac MRI data (Luo et al., 2019). ML models have also been used to diagnose arrhythmias using ECG data (Choi et al., 2016). These models can improve the accuracy of diagnosis and reduce the time required for interpretation compared to traditional methods. ML and DL can also assist in the treatment of cardiovascular diseases. DL models can analyse ECG data to personalize the selection of anti-arrhythmic drugs (Weng et al., 2019). ML models have also been used to predict the response to cardiac resynchronization therapy in patients with heart failure (Ruwald et al., 2017). These models can improve the effectiveness of treatment and reduce the risk of adverse events.

Despite the potential of ML and DL in analysing complex cardiovascular data, there are several challenges and limitations that need to be addressed. One of the main challenges is the lack of standardized data formats and data quality. Cardiovascular data is often collected using different methods and formats, making it difficult to compare and analyse data from different sources (Huang et al., 2019). Another challenge is the need for large and diverse datasets to train ML and DL models. It is often challenging to obtain large and diverse datasets for cardiovascular diseases due to the heterogeneity of the diseases and the difficulty in collecting data from multiple sources (Bui et al., 2018). Additionally, ML and DL models are often considered as black boxes, making it difficult to interpret the results and understand how the models make predictions (Topol, 2019).

1.3 Advances of Artificial Intelligence in the field of cardiovascular disease research:

Machine learning algorithms have brought about significant advances in the field of cardiovascular disease research. With the ability to process large amounts of data quickly and accurately, machine learning has become a valuable tool for analysing complex data and identifying patterns that may be too subtle for human detection. This section will explore the latest advances and progress of machine learning algorithms in the field of cardiovascular disease research.

Cardiovascular disease (CVD) is a leading cause of death worldwide. Machine learning algorithms have helped researchers to identify key risk factors associated with CVD, such as hypertension, high cholesterol, smoking, and CVDs. In addition, machine learning algorithms have helped to identify new biomarkers associated with CVD, such as circulating microRNAs (miRNAs) and metabolites.

One of the most significant advances in machine learning algorithms in the field of cardiovascular disease research is in the area of risk prediction. Machine learning algorithms can be trained on large datasets to predict the risk of CVD in individuals based on a range of risk factors, including age, gender, lifestyle factors, and medical history. This has the potential to revolutionize the way CVD is diagnosed and treated, allowing for earlier intervention and more personalized care.

Machine learning algorithms have also been used to analyse medical imaging data, such as echocardiography and computed tomography (CT) scans. This has led to improvements in the accuracy of diagnosis and the ability to detect subtle changes in the heart that may be indicative of CVD. For example, machine learning algorithms have been used to identify early signs of heart failure in patients with asymptomatic left ventricular dysfunction (Attia et. al, 2019).

Machine learning algorithms have also been used to develop new therapies for CVD. For example, researchers have used machine learning algorithms to identify new drug targets for the treatment of heart failure (Galloway et. al). In addition, machine learning algorithms have been used to develop personalized treatment plans for patients with CVD based on their individual risk profiles.

1.3.1 Most promising aspects of AI frequently used in cardiovascular disease research:

There have been several recent advances in machine learning algorithms for cardiovascular disease research. These include:

- Deep Learning: Deep learning is a type of machine learning that involves training artificial neural networks to recognize complex patterns in data. Deep learning has been used to analyse medical imaging data, such as echocardiography and CT scans, and has led to improvements in the accuracy of diagnosis and the ability to detect subtle changes in the heart that may be indicative of CVD.
- Transfer Learning: Transfer learning is a technique that involves reusing pre-trained machine learning models to solve new problems. This has the potential to significantly reduce the amount of data required to train new models, making it possible to develop more accurate models with less data. Transfer learning has been used to predict the risk

of CVD in individuals based on a range of risk factors, including age, gender, lifestyle factors, and medical history.

- Explainable AI: Explainable AI is a type of machine learning that produces results that can be easily understood by humans. This is particularly important in the field of cardiovascular disease research, where decisions based on machine learning algorithms can have life or death consequences. Explainable AI has been used to develop more accurate risk prediction models and to identify new biomarkers associated with CVD.
- Federated Learning: Federated learning is a technique that involves training machine learning models on decentralized data sources, such as electronic health records. This has the potential to significantly improve the accuracy of risk prediction models by incorporating data from a wider range of sources. Federated learning has been used to predict the risk of CVD in individuals based on a range of risk factors, including age, gender, lifestyle factors, and medical history.
- Random Forests: Random Forests is a type of ML algorithm that uses an ensemble of decision trees to make predictions. Random Forests have been used to predict the likelihood of heart attack, stroke, and other CVDs. They have also been used to predict the onset of heart failure and to identify patients at high risk for developing CVD (Linden et al., 2018).
- Support Vector Machines (SVMs): SVMs are a type of ML algorithm that can be used for both classification and regression analysis. They have been used to classify patients with different types of CVD, including coronary artery disease, congestive heart failure, and atrial fibrillation (Géron, 2019).
- Convolutional Neural Networks (CNNs): CNNs are a type of neural network that have been used to analyse medical images, including echocardiograms and CT scans. They have been used to predict the risk of heart attack and other CVDs, and to classify different types of heart disease (Wang et al., 2020).

Machine learning algorithms have shown promise in several areas of cardiovascular disease research, including risk prediction, image analysis, and drug discovery. In the following sections, we will discuss the progress and advances in each of these areas.

- Early detection of cardiovascular disease: Machine learning algorithms can analyse electronic health records to identify patients at high risk of cardiovascular disease. For example, a study by Khera et al. (2018) used machine learning algorithms to analyse electronic health records of over 400,000 patients to predict the risk of cardiovascular disease. The machine learning algorithms identified several risk factors, including age, gender, and medical history, and predicted the risk of cardiovascular disease with high accuracy. Early detection of cardiovascular disease can lead to timely interventions, reducing the risk of complications.

- Predicting outcomes of cardiovascular disease: Machine learning algorithms can analyse medical imaging and electronic health records to predict the outcomes of cardiovascular disease. For example, a study by Choy et al. (2018) used machine learning algorithms to analyse medical imaging of patients with heart disease. The machine learning algorithms predicted the risk of death or hospitalization within 90 days with high accuracy. Accurate prediction of outcomes can help clinicians make informed decisions about treatment and improve patient outcomes.
- Personalized treatment of cardiovascular disease: Machine learning algorithms can analyse electronic health records to personalize treatment for patients with cardiovascular disease. For example, a study by Choi et al. (2020) used machine learning algorithms to analyse electronic health records of over 2,000 patients with heart failure. The machine learning algorithms identified several subgroups of patients with distinct clinical characteristics and treatment responses. Personalized treatment based on machine learning algorithms can improve patient outcomes and reduce healthcare costs.
- Identifying novel biomarkers for cardiovascular disease: Machine learning algorithms can analyse large datasets of genomic and proteomic data to identify novel biomarkers for cardiovascular disease. For example, a study by Shah et al. (2018) used machine learning algorithms to analyse genomic and proteomic data of over 1,000 patients with heart disease. The machine learning algorithms identified several novel biomarkers that were associated with the risk of cardiovascular disease. Identification of novel biomarkers can lead to the development of new diagnostic and therapeutic strategies.
- Diagnosis: ML algorithms have shown great promise in the early and accurate diagnosis of CVDs. One study conducted by Attia et al. (2019) developed a deep learning algorithm using electrocardiograms (ECGs) to predict the risk of atrial fibrillation (AF). The algorithm was trained on a dataset of over 180,000 ECGs and achieved a sensitivity of 79% and specificity of 79% in detecting AF. Another study conducted by Galloway et al. (2018) developed a convolutional neural network (CNN) to identify patients with

hypertrophic cardiomyopathy (HCM) from cardiac magnetic resonance (CMR) images. The CNN achieved an accuracy of 92% in detecting HCM, outperforming human experts.

- Risk stratification: ML algorithms have been used to develop risk stratification models for CVDs. These models utilize large amounts of patient data to predict an individual's risk of developing a CVD. The Framingham risk score (FRS) is one such model that has been widely used to predict the risk of developing CVDs. However, the FRS has several limitations, including its inability to account for the impact of newer risk factors, such as elevated levels of C-reactive protein and homocysteine. ML algorithms can overcome these limitations by incorporating a large number of variables, including newer risk factors, in the risk prediction models. Accurately stratifying the risk of developing CVDs is critical for timely intervention and prevention. ML algorithms have been shown to improve the accuracy of risk prediction models. One study conducted by Shaban-Nejad et al. (2018) developed an ML-based model to predict the risk of developing heart failure. The model incorporated patient demographics, clinical data, and laboratory test results and achieved an area under the receiver operating characteristic curve (AUC) of 0.88 in predicting heart failure risk. Another study conducted by Natarajan et al. (2018) developed a ML-based model to predict the 10-year risk of atherosclerotic cardiovascular disease (ASCVD). The model incorporated patient demographics, clinical data, and genetic information and achieved an AUC of 0.74 in predicting ASCVD risk, outperforming the traditional Framingham risk score. Several studies have reported the development of ML-based risk stratification models for CVDs. For instance, a study by Wang et al. (2019) developed an ML-based risk prediction model for coronary artery disease (CAD) using data from the China PEACE-Retrospective Study of Acute Myocardial Infarction. The model achieved an area under the receiver operating characteristic (ROC) curve of 0.795, which was significantly higher than that of the FRS. Similarly, a study by Singh et al. (2019) developed an ML-based risk prediction model for heart failure using electronic health record data. The model achieved an area under the ROC curve of 0.862, which was significantly higher than that of the FRS.
- Treatment: ML algorithms have also been applied to improve treatment outcomes in CVDs. One study conducted by Vaid et al. (2019) developed an ML-based model to predict the response to cardiac resynchronization therapy (CRT) in heart failure

patients. The model incorporated patient demographics, clinical data, and electrocardiogram and echocardiogram findings and achieved an AUC of 0.80 in predicting CRT response. Another study conducted by Goldstein et al. (2018) developed an ML-based model to predict the occurrence of major adverse cardiac events (MACE) in patients undergoing percutaneous coronary intervention (PCI). The model incorporated patient demographics, clinical data, and angiographic findings and achieved an AUC of 0.73 in predicting MACE, outperforming traditional risk prediction models.

ML algorithms have shown great promise in advancing the fields of CVD research, particularly in diagnosis, risk stratification, and treatment. These algorithms have demonstrated the ability to improve diagnostic accuracy, risk prediction, and treatment outcomes. As the field of ML continues to evolve, further research is needed to explore the potential of these algorithms in CVD research and their impact on clinical practice.

Chapter 2. Machine Learning algorithms and their classification:

There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:

- Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)
- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

These criteria are not exclusive; we can combine them in any way we like. For example, a state-of-the-art spam filter may learn on the fly using a deep neural net- work model trained using examples of spam and ham; this makes it an online, model-based, supervised learning system.

2.1 Supervised/Unsupervised Learning:

Machine Learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories: supervised learning, unsupervised learning, semisupervised learning, and Reinforcement Learning.

- **Supervised Learning:** In *supervised learning*, the training data we feed to the algorithm includes the desired solutions called *labels*. (Figure 1.1). A typical supervised learning task is *classification*. The spam filter is a good example of this: it is trained with many example emails along with their *class* (spam or ham), and it must learn how to classify new emails. Another typical task is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors. This sort of task is called regression. To train the system, we need to give it many examples of cars, including both their predictors and their labels (i.e., their prices).

- **Unsupervised learning:** In *unsupervised learning*, as we might guess, the training data is unlabeled. The system tries to learn without a teacher.

For example, say we have a lot of data about our blog's visitors. We may want to run a *clustering* algorithm to try to detect groups of similar visitors. At no point do we tell the algorithm which group a visitor belongs to: it finds those connections without our help. For example, it might notice that 40% of our visitorsare males who love comic books and generally read our blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on. If we use a *hierarchical clustering* algorithm, it may also subdivide each group into smaller groups. This may help we target our posts for each group.

Visualization algorithms are also good examples of unsupervised learning algorithms:we feed them a lot of complex and unlabeled data, and they output a 2D or 3D representation of our data that can easily be plotted. These algorithms try to preserve as much structure as they can (e.g., trying to keep separate clusters in the input space from overlapping in the visualization), so we can understand how the data is organized and perhaps identify unsuspected patterns.

A related task is *dimensionality reduction*, in which the goal is to simplify the data without losing too much information. One way to do this is to merge several correlated features into one. For example, a car's mileage may be very correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that rep- resents the car's wear and tear. This is called *feature extraction*.

Yet another important unsupervised task is *anomaly detection*—for example, detecting unusual credit card transactions to prevent fraud, catching manufacturing defects, or automatically removing outliers from a dataset before feeding it to another learning algorithm.

The system is shown mostly normal instances during training, so it learns to recognize them and when it sees a new instance it can tell whether it looks like a normal one or whether it is likely an anomaly. A very similar task is *novelty detection*: the difference is that novelty detection algorithms expect to see only normal data during training, while anomaly detection algorithms are usually more tolerant, they can often perform well even with a small percentage of outliers inthe training set.

Finally, another common unsupervised task is *association rule learning*, in which the goal is to dig into large amounts of data and discover interesting relations between attributes. For example, suppose we own a supermarket. Running an association rule on our sales logs may reveal that people who purchase barbecue sauce and potato chips also tend to buy steak. Thus, we may want to place these items close to each other.

- **Semisupervised Learning:** Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. This is called *semisupervised learning* (Figure 1.1).

  Some photo-hosting services, such as Google Photos, are good examples of this. Once we upload all our family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7. This is the unsupervised part of the algorithm (clustering). Now all the system needs is for we to tell it who these people are. Just one label per person, and it is able to name everyone in every photo, which is useful for searching photos.

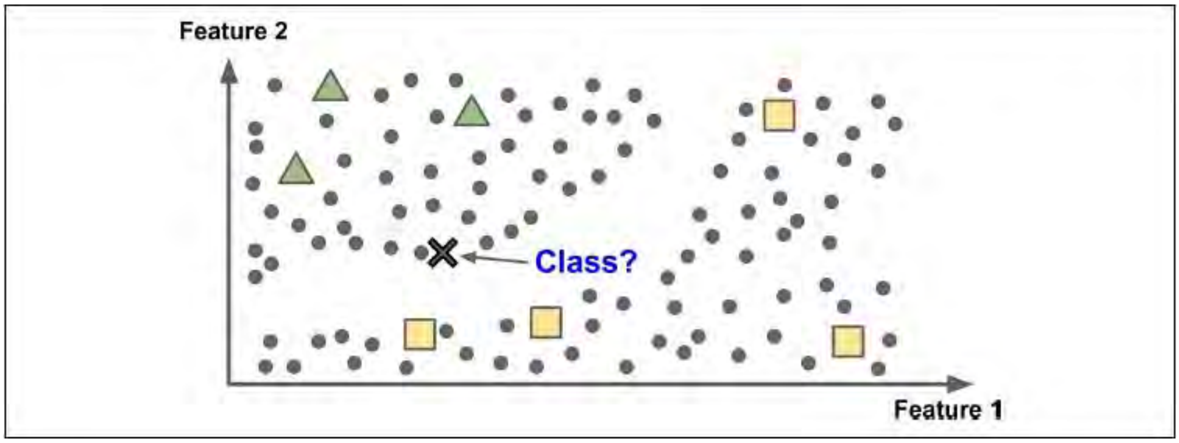

Figure 2.1: Semisupervised learning

> Most semisupervised learning algorithms are combinations of unsupervised and supervised algorithms. For example, *deep belief networks* (DBNs) are based on unsu pervised components called *restricted Boltzmann machines* (RBMs) stacked on top of one another. RBMs are trained sequentially in an unsupervised manner, and then the whole system is fine-tuned using supervised learning techniques.

2.2 A close look at some of the widely used ML algorithms in Cardiovascular Disease Research:

There are several ML and DL algorithms that have been used in cardiovascular disease research. One of the most commonly used algorithms is the support vector machine (SVM), which is a supervised learning algorithm that can classify data into different categories (Dey et al., 2016). SVM has been used to predict the risk of cardiovascular events and diagnose arrhythmias using ECG data (Attia et al., 2019; Choi et al., 2016). Another commonly used algorithm is the random forest (RF), which is an ensemble learning algorithm that can combine multiple decision trees to improve the accuracy of predictions (Bui et al., 2018). RF has been used to predict the risk of heart failure using EHR data (Khera et al., 2018). In DL, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to analyse ECG data and cardiac MRI data (Luo et al., 2019; Ouyang et al., 2020). These algorithms have shown promising results in analysing complex cardiovascular data and can be used to improve risk prediction, diagnosis, and treatment of cardiovascular diseases.

2.3 Clustering algorithms:

Clustering algorithms are a powerful tool in the field of data analysis that can be used to detect risk factors, prevalence, and other issues related to cardiovascular diseases (CVD) from a given set of data. These algorithms are able to group similar data points together and identify patterns and correlations that may not be immediately apparent to the human eye. In this essay, we will discuss how clustering algorithms can be implemented to detect CVD risk factors and other issues along with some relevant studies and their findings.

One of the most commonly used clustering algorithms is the k-means algorithm. This algorithm divides a dataset into k clusters, where k is the number of clusters specified by the user. The k-means algorithm works by iteratively reassigning each data point to the cluster with the closest mean.

Another popular clustering algorithm is the hierarchical clustering algorithm. This algorithm creates a hierarchy of clusters by successively merging or splitting them.

In addition to patient data, clustering algorithms can also be used to analyse other types of data related to cardiovascular disease, such as genetic data and imaging data. For example, clustering algorithms can be used to group genes with similar expression patterns in order to identify genes that are associated with cardiovascular disease. The above code shows how the hierarchical clustering algorithm can be implemented in Python using the `scipy` and matplotlib libraries. The data is first loaded into the script, and then the linkage function is used to perform the hierarchical clustering. The dendrogram function is then used to plot the resulting hierarchy of clusters.

Clustering algorithms have been used in various studies to detect risk factors and the prevalence of cardiovascular diseases. For instance, a study by (Kotsiantis, 2007) used a k-means algorithm to cluster patients based on their demographic information and medical history, and found that certain clusters had a higher risk of developing CVD. Another study by (Zhang et al., 2019) used hierarchical clustering to identify subgroups of patients with CVD based on their genetic and lifestyle risk factors.

In conclusion, clustering algorithms can be implemented to detect the risk factors, prevalence, and other issues related to cardiovascular diseases from a given set of data. The k-means and hierarchical algorithms are popular choices, and examples of how to implement these algorithms using Python have been provided. By analysing large amounts of data, clustering algorithms can identify patterns and correlations that might not be apparent the human intuition.

2.4 K-means algorithm:

Cardiovascular diseases (CVDs) are a leading cause of death and disability worldwide, with an estimated 17.9 million deaths annually. Early detection and prevention of CVDs is crucial in reducing the burden of these diseases on individuals and society as a whole. Machine learning, a subfield of Artificial Intelligence, has been applied in various fields to improve the detection, prevention, and cure of diseases. Clustering algorithms, a subfield of machine learning, have been found to be useful in detecting risk factors and prevalence of CVDs. This paper will discuss how K-means algorithms can be implemented to detect the risk factors, prevalence of CVDs, and provide an overview of the potential uses of K-means in CVD research.

Clustering algorithms, such as k-means, can be used to detect risk factors, prevalence, and other issues related to cardiovascular diseases (CVD) from a given set of data. These algorithms group similar data points together, allowing for the identification of patterns and correlations that may not be immediately apparent to the human eye. In this essay, we will discuss how clustering algorithms can be implemented to detect CVD risk factors and other issues, provide code examples and incorporate relevant research studies and their findings.

One common use of clustering algorithms in CVD research is to identify subgroups of patients with similar characteristics. These subgroups, or clusters, can then be used to understand the underlying causes of CVD and to develop targeted prevention and treatment strategies. For example, a study by Kim et al. (2015) used k-means clustering to identify subgroups of patients with hypertension and found that the identified clusters had distinct risk factors and responses to treatment. Clustering algorithms can also be used to identify potential risk factors for CVD. For example, a study by Li et al. (2018) used k-means clustering to identify patterns in the levels of various biomarkers in blood samples from patients with CVD. The researchers found that patients in one cluster had higher levels of inflammation markers, suggesting that inflammation may be a risk factor for CVD in this group of patients.

In terms of the implementation, k-means is a popular clustering algorithm that can be used to identify CVD risk factors and other issues from a given set of data. The k-means algorithm groups similar data points together by minimizing the sum of squared distances between the data points and the cluster centroid. The algorithm starts with k initial centroids, which are chosen randomly, and then iteratively assigns each data point to the closest centroid and re-calculates the centroids based on the new assignments.

Clustering in essence, is the process of grouping similar data points together. K-means is a popular clustering algorithm that partitions a dataset into k clusters, where each cluster is represented by its centroid. K-means is particularly useful in detecting patterns in large datasets, making it an ideal tool for identifying risk factors and prevalence of CVDs.

The K-means algorithm is a simple and efficient method for clustering large datasets. The algorithm proceeds in two steps:

1. Initialization: k initial centroids are randomly chosen from the dataset.
2. Iteration: Each data point is assigned to the cluster whose centroid is closest to it. The centroid of each cluster is then recalculated as the mean of all the data points in that

cluster. These two steps are repeated until the centroids no longer change or a maximum number of iterations is reached.

One way K-means can be used to detect risk factors for CVDs is by analysing a dataset of patient characteristics and medical history. The algorithm can be used to identify clusters of patients with similar characteristics, such as age, gender, and family history, who are at a higher risk of developing CVDs. This information can then be used to develop targeted prevention strategies for these high-risk groups.

In the context of CVDs, K-means can be used to group similar individuals based on their risk factors and demographic information. The clustering process begins by selecting a set of features that are relevant to CVDs, such as age, blood pressure, cholesterol levels, and smoking status. These features are then standardized and fed into the K-means algorithm. The number of clusters, k, is determined by considering the number of risk factors and the size of the dataset.

Once the clusters are formed, the individuals in each cluster can be analysed to identify common risk factors and prevalence of CVDs. For example, if a cluster contains a large number of individuals with high blood pressure and cholesterol levels, it is likely that this cluster represents a high-risk group for CVDs. Similarly, if a cluster contains a high percentage of smokers, it is likely that this cluster represents a group with a higher prevalence of CVDs.

In essence, K-means is a widely used unsupervised learning algorithm for clustering and grouping similar data points together. Its application in detecting patterns or clusters of risk factors among a population for cardiovascular disease detection, has been widely researched and can be useful for identifying high-risk subpopulations and developing targeted interventions. But the limitations and assumptions of the algorithm must be taken into account when interpreting the results.

2.5 K Nearest Neighbours Method:

Cardiovascular diseases (CVDs) are a leading cause of death worldwide. Early detection and prevention of CVDs is crucial in reducing the mortality and morbidity associated with these diseases. Artificial intelligence (AI) and machine learning (ML) have the potential to enhance the detection, prevention and cure of CVDs. One of the most widely used machine learning techniques for classification and prediction tasks is the K-nearest neighbours (KNN) algorithm.

In this essay, we will explore how KNN can be implemented to detect the risk factors and prevalence of CVDs using a given set of data.

KNN is a non-parametric, instance-based learning algorithm that can be used for both classification and regression tasks. The algorithm works by comparing a new data point to the k-nearest points in the training dataset and assigning the new data point to the most common class among the k-nearest points. One of the main advantages of KNN is that it is easy to understand and implement, making it a popular choice for many applications. Also, KNN is a type of supervised machine learning algorithm that can be used for both classification and regression tasks. In the context of CVDs, KNN can be used to classify individuals as either having or not having the disease based on a set of risk factors. The algorithm works by identifying the k-number of nearest neighbours to a given data point and using the majority class of these neighbours to classify the point.

To implement KNN for CVDs, a dataset containing information on individuals' risk factors and CVD status is required. These risk factors may include demographic information, lifestyle choices, and other medical conditions. The KNN algorithm can then be trained on this dataset, with the CVD status serving as the target variable.

In the context of CVDs, KNN can be used to identify individuals who are at high risk of developing CVDs based on their demographic, lifestyle and medical history. The algorithm can be trained on a dataset of individuals with known CVDs and their risk factors, and then used to predict the risk of CVDs in new individuals.

One of the key steps in using KNN for CVDs detection is the selection of appropriate features. Features that are strongly associated with CVDs, such as age, smoking status, and blood pressure, should be included in the dataset. Additionally, it is important to consider the quality and quantity of data, as the performance of the algorithm will be directly affected by these factors.

KNN is a simple and effective algorithm that can be used to detect the risk factors and prevalence of CVDs. By incorporating appropriate features and sufficient data, KNN can be trained to identify individuals at high risk of CVDs, allowing for early detection and prevention. However, it is important to note that the performance of the algorithm is dependent on the quality and quantity of the data, and the selection of appropriate features. As such, it is essential to consult with domain experts and conduct thorough evaluations before using KNN in a real-world setting.

Another example of using KNN for CVDs is a study by Al-Jumaily et al. (2019), where the authors used the algorithm to classify individuals with and without CVD based on risk factors such as age, gender, hypertension, CVDs, and smoking status. The study found that KNN had an accuracy of 86.5% in classifying individuals with CVD and non-CVD, which was higher than other algorithms tested.

Simply speaking, KNN is a powerful tool for detecting the risk factors and prevalence of CVDs. The algorithm has been used in several studies with promising results, and its implementation is relatively simple. However, it's important to note that the results of KNN are highly dependent on the quality and size of the dataset used. Therefore, it's crucial to have a large and diverse dataset that accurately represents the population of interest. Additionally, it should be used in combination with other techniques to improve the accuracy and robustness of the results.

2.6 Random Forest Algorithm:

Cardiovascular diseases (CVDs) are a group of disorders that affect the heart and blood vessels and are one of the leading causes of death worldwide. Early detection and prevention of CVDs are crucial for reducing the mortality rate. In recent years, Artificial Intelligence (AI) and its subdomains such as Machine Learning (ML) and Deep Learning (DL) have made significant contributions to the detection, prevention, and cure of CVDs. In this essay, we will discuss the implementation of the Random Forest algorithm in detecting the risk factors and prevalence of CVDs.

Random Forest is a popular ML algorithm that is used for both regression and classification tasks. The algorithm is based on the decision tree, which is a tree-like model that represents a set of decisions and their possible consequences. In Random Forest, multiple decision trees are trained on randomly selected subsets of the data and the results are combined to form a single prediction.

To implement Random Forest for detecting the risk factors and prevalence of CVDs, we need a dataset that contains the relevant information, such as demographic information, lifestyle habits, medical history, and test results. The data must be pre-processed to remove any irrelevant or missing information and to ensure that the data is in a suitable format for analysis.

Once the data is ready, we can train the Random Forest algorithm on the data. The algorithm will learn to identify patterns and relationships in the data and use them to predict the risk factors and prevalence of CVDs. The parameters of the algorithm, such as the number of trees and the maximum depth of the trees, can be adjusted to improve the performance of the model. It is a popular machine learning technique that can be used to detect risk factors and prevalence of cardiovascular diseases. The algorithm works by building multiple decision trees, which are then combined to form a forest. This combination of trees helps to increase the accuracy of predictions, reduce overfitting, and make the results more robust.

Random Forest can be implemented in several programming languages such as Python, R, and MATLAB. In Python, the implementation of Random Forest can be achieved using the scikit-learn library, which provides a simple and efficient implementation of the algorithm.

Random Forest has been used in several studies to detect risk factors and prevalence of cardiovascular diseases. For example, a study by Wang et al. (2018) used Random Forest to identify risk factors for cardiovascular disease in a large population of adults. The study found that traditional risk factors, such as age, hypertension, and smoking, were the most significant predictors of cardiovascular disease. The results of this study highlight the potential of Random Forest for detecting risk factors for cardiovascular diseases.

In a nutshell, Random Forest algorithm can be an effective tool for detecting risk factors and prevalence of cardiovascular diseases. The algorithm is flexible, easy to implement, and can handle large datasets with many features. By incorporating expert computational skills, the Random Forest algorithm can help to improve the detection, prevention, and cure of cardiovascular diseases.

2.7 Decision Tree:

Early detection of CVDs is crucial in reducing the risk of death and improving patient outcomes. Machine learning algorithms can play a key role in early detection by analysing large amounts of data to identify risk factors and prevalence of CVDs.

One machine learning algorithm that is often used for this purpose is the decision tree algorithm. The decision tree algorithm works by constructing a tree-like structure to make predictions based on the data it is trained on. It does this by splitting the data into smaller and smaller subsets, using a series of questions to narrow down the possibilities and reach a decision. The goal is to find the most significant factors that contribute to the development of CVDs.

The implementation of decision tree algorithm to detect the risk factors and prevalence of CVDs can be divided into the following steps:

1. Data Collection and Preparation: The first step is to collect and prepare the data required to train and test the decision tree model. This data should include information on the patient's demographic, medical history, and laboratory results.
2. Data Pre-processing: The next step is to pre-process the collected data to ensure that it is in the required format for training and testing the decision tree model. This may include removing missing or irrelevant data, normalizing the data, and encoding categorical data.
3. Model Training: The next step is to train the decision tree model using the pre-processed data. The training process involves creating the decision tree structure, calculating the information gain for each feature, and splitting the data based on the feature with the highest information gain.
4. Model Testing: After training the model, it is important to test its performance on a separate dataset. The testing process involves making predictions on the test data and comparing the results with the actual outcomes.
5. Model Evaluation: The final step is to evaluate the performance of the decision tree model. This can be done by calculating metrics such as accuracy, precision, recall, and F1 score.

To wrap up this section, we should say that the decision tree algorithm is a powerful tool for early detection of CVDs. By analysing large amounts of data and identifying significant risk factors, the decision tree can help improve patient outcomes and reduce the risk of death from CVDs. With the rise of big data and machine learning, it is likely that we will see more and more applications of decision tree algorithms in the field of healthcare in the future.

Chapter 3. Delving Deeper into the Four Faces of Cardiovascular Disease:

Cardiovascular diseases (CVDs) cast a long shadow over global health, claiming an estimated 17.9 million lives annually. To combat this formidable foe, understanding its diverse forms is crucial. Here, we delve deeper into the four types mentioned earlier, exploring their nuances, risk factors, and management strategies, drawing upon credible sources like the World Health Organization (WHO) and beyond.

3.1 Coronary Heart Disease (CHD): A Battle for Blood Flow

Imagine a vital highway system clogged with debris; that's what CHD does to the coronary arteries, the vital vessels supplying the heart with oxygenated blood. Plaque buildup narrows these arteries, leading to angina (chest pain), shortness of breath, and eventually, heart attack if blood flow is completely blocked.

This silent menace often develops over decades, fueled by a constellation of risk factors:

- **Smoking:** Tobacco acts like a double whammy, injuring blood vessel walls and promoting clotting.
- **Unhealthy diet:** A diet rich in saturated and trans fats, cholesterol, and sodium, while deficient in fruits, vegetables, and whole grains, is a recipe for CHD.
- **Physical inactivity:** A sedentary lifestyle paves the way for obesity, high blood pressure, and CVDs, all allies of CHD.
- **Obesity:** Excess weight acts as a burden on the heart and contributes to metabolic imbalances that favour CVD.
- **High blood pressure:** The silent force, pushing against artery walls and accelerating their damage.
- **CVDs:** High blood sugar levels weaken blood vessels and make them more prone to plaque buildup.
- **High cholesterol:** LDL cholesterol, the "bad" kind, accumulates in arteries, forming the foundation of plaque.
- **Family history of CVD:** Genes can play a role in susceptibility to CHD.

Diagnosis involves a combination of tests, including electrocardiograms (ECGs), echocardiograms, and stress tests. Management strategies focus on controlling risk factors through lifestyle changes (diet, exercise, smoking cessation), managing co-existing conditions like CVDs and hypertension, and in severe cases, employing medications or even surgical interventions.

3.2. Peripheral Arterial Disease (PAD): When the Legs Lose Their Flow

While CHD focuses on the heart itself, PAD takes the battleground to the distant frontiers: the arteries supplying the legs. Similar to CHD, plaque buildup narrows these arteries, restricting blood flow and causing pain, especially during exercise. This pain can range from mild cramping to excruciating, often forcing individuals to shorten or stop their activities. In severe cases, lack of blood flow can lead to tissue death (gangrene), necessitating amputation.

Many of the same risk factors for CHD apply to PAD, making it a frequent companion. Additionally, CVDs and advanced age significantly increase the risk. Diagnosis involves physical examination, ankle-brachial index (ABI) tests, and angiography. Management strategies mirror those for CHD, with an emphasis on controlling risk factors and preventing further arterial damage. In severe cases, procedures like angioplasty or bypass surgery may be necessary to restore blood flow.

3.3. Congenital Heart Disease (CHD): Born with a Different Beat

Unlike the acquired conditions above, CHD stems from abnormalities in the heart's structure present at birth. These defects can range from mild (a small hole in the heart wall) to life-threatening (missing heart chambers or malformed valves). Symptoms vary depending on the severity and type of defect, and can include shortness of breath, fatigue, difficulty feeding in infants, and even heart failure.

While the cause of most CHDs remains unknown, some risk factors include genetic disorders, maternal infections during pregnancy, and exposure to certain medications or substances. Early diagnosis and intervention are crucial for managing CHD. Advances in heart surgery allow for repairs or replacements of malformed structures, significantly improving the quality and lifespan of individuals with CHD.

3.4. Heart Failure: When the Engine Sputters

Imagine a powerful engine losing its strength, struggling to pump blood effectively. That's the essence of heart failure, a condition where the heart weakens and can't meet the body's demand for blood. This can be caused by several factors, including CHD, PAD, high blood pressure, and valve problems. The consequences are often debilitating, with symptoms like fatigue, shortness of breath, swelling in the legs and ankles, and difficulty lying down.

Treatment for heart failure aims to support the weakened heart, control symptoms, and prevent further deterioration. Medications like diuretics, ACE inhibitors, and beta-blockers play a key role. In some cases, devices like pacemakers or implantable defibrillators may be necessary. Lifestyle changes, including managing weight, eating a healthy diet, and exercising within limitations, are also crucial for managing heart failure.

3.5. The Battle Beyond the Individual: Combating CVDs on a Global Scale

While individual lifestyle changes and adherence to treatment plans are crucial, the fight against CVDs extends far beyond the individual. To truly conquer this global health threat, a multi-pronged approach encompassing public health initiatives, strong healthcare systems, and equitable access to prevention and treatment is essential.

Prevention: The First Line of Defence

Investing in public health initiatives aimed at primary prevention is key. This includes:

Education and awareness campaigns: Empowering individuals to make informed choices about their health through education on risk factors, healthy lifestyles, and early warning signs of CVDs.

Tobacco control measures: Implementing strict tobacco control policies like bans on smoking in public places and high taxes on cigarettes can significantly reduce smoking rates and consequently, CVD risk.

Promoting healthy diets and physical activity: Public health campaigns encouraging healthy eating habits and regular exercise can create a culture of health and prevent the development of CVD risk factors.

Strengthening Healthcare Systems:

Robust healthcare systems play a crucial role in managing CVDs effectively. This includes:

Early detection and diagnosis: Ensuring access to affordable and accessible screening programs for CVDs, like blood pressure checks and cholesterol tests, allows for early diagnosis and intervention before complications arise.

Effective treatment: Investing in healthcare infrastructure and training healthcare professionals in the latest CVD management techniques ensures timely and effective treatment for all patients.

Access to essential medications and interventions: Ensuring affordable access to essential medications and interventions like stents, bypass surgery, and cardiac rehabilitation for all patients, regardless of socioeconomic background, is vital for improving outcomes.

Addressing Inequalities:

Social and economic inequalities significantly impact CVD risk and access to care. Therefore, addressing these inequalities is crucial for achieving equitable health outcomes. This includes:

Targeting interventions to vulnerable populations: Groups facing higher CVD risks due to factors like poverty, lack of education, or limited healthcare access need targeted interventions and support to improve their health outcomes.

Promoting social determinants of health: Addressing issues like poverty, lack of education, and unhealthy living environments can create a more equitable society and reduce CVD risk across populations.

Embracing Innovation and Technology:

Advances in technology hold immense potential for improving CVD prevention, diagnosis, and treatment. This includes:

Telemedicine and digital health tools: These technologies can improve access to care, especially in remote areas, and allow for remote monitoring of patients with CVDs.

Artificial intelligence and machine learning: These tools can be used for early detection of CVDs through analysis of medical data and images, enabling timely intervention and improved outcomes.

Personalized medicine: Tailoring treatment plans to individual patients based on their genetic profile and other factors can lead to more effective and targeted therapies.

Cardiovascular diseases are a formidable foe, but one that can be conquered through a concerted effort. By combining individual lifestyle changes with robust public health initiatives, strong healthcare systems, and equitable access to care, we can create a healthier world where CVDs no longer claim millions of lives each year. Remember, this is a fight we must all wage together, one step, one healthy choice, one policy change at a time.

3.6. Major Risk Factors for Cardiovascular Disease

Cardiovascular diseases (CVDs) are a leading cause of death globally, claiming millions of lives each year. While some risk factors like family history are beyond our control, many are modifiable, offering opportunities for prevention through lifestyle changes and early intervention. Here, we delve into some of the key modifiable risk factors for CVDs, exploring their definitions, impacts, and strategies for management:

3.6.1. Body Mass Index (BMI):

BMI is a measure of body fat based on height and weight. A high BMI (over 25) is associated with an increased risk of CVDs due to its link to:

Increased strain on the heart: Excess weight puts additional pressure on the heart to pump blood, leading to overwork and potential weakening.

Elevated blood pressure and cholesterol levels: Obesity often contributes to high blood pressure and unhealthy cholesterol levels, further increasing CVD risk.

Chronic inflammation: Fat tissue, particularly visceral fat around the abdomen, releases inflammatory markers that can damage blood vessels and contribute to atherosclerosis.

3.6.2. Total Cholesterol:

Total cholesterol is the sum of LDL ("bad") cholesterol, HDL ("good") cholesterol, and other types. High total cholesterol levels (over 200 mg/dL) increase the risk of plaque buildup in arteries, leading to atherosclerosis and ultimately, CVDs.

3.6.3. Blood Pressure:

Blood pressure is the force exerted by blood against the walls of your arteries. High blood pressure (hypertension, over 130/80 mmHg) is a major risk factor for CVDs, as it strains the heart and damages blood vessels, increasing the risk of heart attack, stroke, and other complications.

3.6.4. LDL and HDL Cholesterol:

LDL cholesterol carries cholesterol to the arteries, where it can build up and form plaque. HDL cholesterol removes cholesterol from the arteries and carries it back to the liver for removal. High LDL and low HDL levels contribute significantly to atherosclerosis and CVD risk.

3.6.5. Hypertension:

Hypertension, also known as high blood pressure, is a chronic condition where blood pressure consistently remains above healthy levels. It is a major risk factor for CVDs due to its effects on the heart and blood vessels. Chronic high blood pressure can damage the inner lining of arteries, making them more susceptible to plaque buildup and narrowing, leading to heart attack, stroke, and other complications.

3.6.6. Asthma:

While not directly causing CVDs, asthma can be a contributing risk factor. Chronic inflammation associated with asthma, particularly in severe or poorly controlled cases, can damage blood vessels and increase the risk of atherosclerosis and blood clots. Additionally, some asthma medications may have side effects that raise blood pressure or cholesterol levels, further increasing CVD risk.

3.6.7. Blood Group:

Recent research suggests a possible link between blood type and CVD risk. Individuals with type A blood may have a slightly higher risk of certain CVDs, while those with type O may have a lower risk. However, this association is weak and further research is needed to fully understand the underlying mechanisms.

3.6.8. Age:

The risk of CVDs increases with age. This is due to the cumulative effects of wear and tear on the heart and blood vessels over time. Additionally, age-related changes in cholesterol levels, blood pressure, and other risk factors contribute to the increased CVD risk in older adults.

3.6.9. Gender:

Men generally have a higher risk of CVDs than women, especially before menopause. This is likely due to hormonal differences, with oestrogen in women offering some protective effects against CVDs. However, after menopause, women's CVD risk increases significantly, approaching that of men.

Understanding your own risk factors for CVDs is crucial for taking preventive measures. Regular health checks, including blood pressure, cholesterol, and BMI measurements, can help identify potential problems early on. Additionally, adopting a healthy lifestyle that includes a balanced diet, regular exercise, and stress management can significantly reduce your CVD risk.

While these risk factors are significant, they are not absolute. By making healthy choices and working with your healthcare provider, you can significantly reduce your risk of developing CVDs and live a long and healthy life.

3.7. Navigating the Gray Zone: Demystifying Borderline Values in Cardiovascular Disease Risk Assessment:

Cardiovascular diseases (CVDs) pose a significant threat to global health. Understanding risk factors, particularly those within the borderline range, is crucial for effective prevention strategies. This section examines nine key risk factors, elucidating the implications of borderline values on CVD risk and providing guidance on managing these risks.

3.7.1. Body Mass Index (BMI):

- **Normal (<25)**: Individuals within this range are at a lower risk of CVD. Maintaining a healthy weight through regular physical activity and a balanced diet is essential.
- **Borderline (23-24.9)**: Although not immediately concerning, it is important to monitor additional risk factors such as blood pressure and cholesterol. Adopting a healthy lifestyle can help prevent progression to higher risk categories.

- **Overweight (25-29.9)**: There is an increased risk of CVD. It is advisable to consult healthcare professionals for personalized weight management strategies and to closely monitor other risk factors.

3.7.2. Total Cholesterol:

- **Desirable (<200 mg/dL)**: Maintaining cholesterol levels within this range is ideal. Continue practicing healthy lifestyle habits.
- **Borderline (190-199 mg/dL)**: A detailed lipid profile analysis is recommended. Elevated LDL ("bad") cholesterol and low HDL ("good") cholesterol increase CVD risk. Consult with a healthcare provider for management strategies.
- **High (≥200 mg/dL)**: This range indicates a significant risk for CVD. Implement cholesterol-lowering interventions, including dietary changes, physical activity, and possibly medication, under medical supervision.

3.7.3. Blood Pressure:

- **Normal (<120/80 mmHg)**: This range is optimal. Regular monitoring and maintaining healthy habits are recommended.
- **Borderline (120-129/80 mmHg)**: Prehypertension necessitates lifestyle modifications such as reducing sodium intake, regular exercise, and stress management to prevent the onset of hypertension.
- **High (≥130/80 mmHg)**: Elevated blood pressure significantly increases CVD risk. Seek medical advice for comprehensive blood pressure management, which may include medication and lifestyle changes.

3.7.4. LDL and HDL Cholesterol:

- **Optimal LDL (<100 mg/dL), Optimal HDL (≥60 mg/dL)**: These levels are ideal. Continue healthy lifestyle practices to maintain these values.
- **Borderline LDL (100-129 mg/dL), Borderline HDL (40-49 mg/dL)**: Improving the overall lipid profile is crucial. Dietary adjustments, increased physical activity, and potentially medication may be necessary.
- **High LDL (≥130 mg/dL), Low HDL (<50 mg/dL)**: High CVD risk necessitates aggressive LDL-lowering and HDL-raising strategies under medical guidance.

3.7.5. Hypertension (Borderline or Stage 1):

- **120-129/80 mmHg**: As previously noted, prehypertension requires lifestyle interventions such as the DASH diet, regular physical activity, and stress management to prevent progression to Stage 2 hypertension.
- **Stage 2 (140-159/90-99 mmHg)**: Immediate medical consultation is recommended. Effective management likely involves both medication and lifestyle modifications to reduce CVD risk.

3.7.6. Asthma:

- **Well-controlled asthma**: The risk of CVD may be minimal. Regular medical checkups and adherence to treatment plans are essential.
- **Poorly controlled asthma**: Chronic inflammation associated with poorly managed asthma can elevate CVD risk. Discuss management strategies with a healthcare provider to optimize asthma control and minimize CVD risk.

3.7.7. Blood Group:

- **Type A**: Research indicates a slightly higher risk for certain CVDs. Maintaining a healthy lifestyle and closely monitoring risk factors is advised.
- **Other blood types**: While the risk may be lower, lifestyle choices and other risk factors remain critical in overall CVD risk management.

3.7.8. Age:

- **Young (<45 years)**: Generally, the risk is lower; however, attention to family history and other risk factors is important. Healthy habits and regular health monitoring are recommended.
- **Middle-aged (45-64 years)**: The risk increases with age. Regular health check-ups and proactive risk factor management become essential.
- **Older adults (≥65 years)**: This group has the highest risk category. Close monitoring, adherence to treatment plans, and preventive measures are crucial for effective CVD risk management.

3.7.9. Gender:

- **Men**: Generally, men have a higher risk, especially before menopause. Prioritizing healthy habits and diligent management of risk factors is essential.
- **Women**: The risk increases significantly after menopause. Maintaining healthy habits and closely monitoring health, particularly post-menopause, is crucial.

Chapter 4: Methodology

4.1. Research Design

The study employs a mixed method research design, so both nominal and numerical values are included in the dataset. Also, it takes a cross-sectional approach to its data collection. This means the data is collected at a single time point from participants of varying ages.

4.2 Data Description and Collection

In this study, all the information is gathered from a sample of Bangladeshi adults, encompassing individuals from various age groups. Patient information containing **demographic data, clinical and biophysical data, and medical history related to cardiovascular diseases** was collected manually from Ibrahim Cardiac Hospital and Research Centre. Ethical considerations are considered by ensuring privacy and confidentiality when dealing with patients' data. The data was recorded in an excel sheet. The relevant features were also selected.

Features can be defined as discrete attributes or qualities of observable aspects of an event that can be measured. For an algorithm to be successful, selecting a distinct and instructive feature is essential. In the dataset, the relevant features that were included are shown in the table. Among all the features in the dataset, disease status is the dependent variable, whereas the other features act as the independent variable.

Table 1: List of features used in the dataset

No.	**Features**	**Unit**
1	gender	-
2	age	years
3	BMI	-
4	blood pressure	mmHg
5	blood group	-
6	total cholesterol	mg/dL
7	LDL Cholesterol	mg/dL
8	HDL Cholesterol	mg/dL
9	Hypertension	-
10	Asthma	-

The output of the data is the **class** or category. The data labels indicate the corresponding class. **Label** is the variable that is required to be estimated. Labelling means providing explanatory tags to every unlabelled data in order to enhance its meaning. In our case, the presence or absence of CVDs is the label.

4.3. Data Pre-processing

After collecting data, the next step is to preprocess it. There may be some irrelevant and incomplete values that do not address the problem, and these need to be handled. Preprocessing allows cleaning data, and fixing missing or duplicate values and outliers. Also, scaling features may be necessary because the magnitude of input variables affects machine learning algorithms, i.e., they are scale-sensitive. So, the numerical features are converted to a comparable scale. Basically, in this step, the data needs to be checked for any errors, inconsistencies, or ambiguity and transformed and formatted accordingly to allow learning. It is important that the discrepancies are rectified to ensure that the dataset is suitable and reliable for analysis.

4.4 Algorithms Used

This study employs various machine learning techniques to predict the occurrence of asthma in the adult population of Bangladesh. Specifically, in predicting asthma based on the presence of CVD, supervised learning is utilized, wherein models are trained using labeled data where the output (CVD status) is known for each input (participant's data). We applied the following machine learning to analyse our available data.

- Logistic Regression
- K-Nearest Neighbor
- Naïve Bayes
- Decision Tree
- Random Forest

4.5 Model Deployment

After selecting the appropriate machine learning algorithms, the next step involves their implementation using suitable tools. For this study, the RapidMiner tool is chosen due to its user-friendly interface and its capability to manage a wide array of machine learning algorithms effectively.

4.6 Model Evaluation

Following the deployment of the models, an evaluation process is conducted to assess their performance using specialized performance indicators. Additionally, confusion matrices are generated for each model. The performance indicators used are:

- Accuracy
- Precision
- Negative Predictive Value
- Specificity

4.6.1 Confusion Matrix

When assessing a classifier's performance, the class distribution of the dataset must be addressed. In cases of significant class imbalance, traditional accuracy tests may incorrectly classify certain categories. For instance, a person with the disease might receive a negative test result, or a healthy person might test positive. Thus, it is crucial to ensure the validity of the test. Traditional error metrics often quantify the model's overall error but not individual errors. This is where a confusion matrix becomes valuable.

A confusion matrix or error matrix provides a tabular representation of predicted and actual classes in a classification problem. For a binary classification problem, the matrix is a 2x2 table with four main components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The terms are defined as follows in the context of this research:

Table 2: Confusion Matrix where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

		ACTUAL VALUES		
	Test Result	**Disease Present (Positive)**	**Disease Absent (Negative)**	**Total**
PREDICTED VALUES	**Positive**	TP	FP	Total Test Positive
	Negative	FN	TN	Total Test Negative
	Total	Total Diseased	Total Normal	Total Population

- **True Positive (TP)**: Instances where the model correctly identifies a person with the disease as positive.
- **True Negative (TN)**: Instances where the model correctly identifies a healthy person as negative.

- **False Positive (FP)**: Type-I errors where the model incorrectly identifies a healthy person as positive.
- **False Negative (FN)**: Type-II errors where the model incorrectly identifies a person with the disease as negative.

The confusion matrix provides insights into specific types of errors made by classifiers. This allows for a detailed evaluation beyond mere classification accuracy, using metrics such as accuracy, precision, recall, and specificity. Analyzing the distribution of TPs, TNs, FPs, and FNs highlights the model's strengths and weaknesses, enabling error pattern recognition and potential model improvements. Furthermore, comparing various models using the confusion matrix helps select the most effective one, simplifying decision-making processes for businesses to choose the best model to meet their needs. In conclusion, the confusion matrix aids in managing class imbalance, model validation, optimization, quality control, and risk assessment.

The confusion matrix finds extensive applications in medical diagnostics, fraud detection, spam detection, and stock market predictions. In medical diagnostics, it is essential for evaluating disease detection tests and classification algorithms, ensuring accurate diagnosis and proper treatment.

Other terms associated with the confusion matrix and performance evaluation, which are used in this study, are defined below.

4.6.2. Accuracy

Accuracy is a metric used to evaluate how well the model correctly classifies values. It is effective when the data's desired variable classes are evenly distributed. However, if the dataset's target variable class is predominantly a single class, accuracy should be avoided. (Sunasra, 2019). The equation for accuracy is given below:

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

4.6.3 Sensitivity

Sensitivity (Recall or **True Positive Rate)** is a performance indicator to measure a binary classification model's accuracy. It calculates the model's accuracy in identifying true positive samples or correct positive values, i.e., it shows how well the model detects the condition or target class (Dāsa, 2016). It measures a binary classification model's accuracy in identifying true positive samples. It shows how well the model detects the condition or target class.

The equation for sensitivity is given below:

$$Sensitivity = \frac{P(positive\ test\ result \cap disease\ is\ present)}{P(disease\ is\ present)} = \frac{TP}{TP + FN}$$

4.6.4 Specificity

Specificity (True Negative Rate) measures how many true negative samples the model accurately identifies, i.e., the accuracy of finding people who do not have the disease. Also, it is a performance metric similar to sensitivity. However, specificity evaluates how many true negative samples the model accurately identifies, i.e., the accuracy of finding people who do not have the disease (Dāsa, 2016). The equation for specificity is given below:

$$Specificity = \frac{P(negative\ test\ result \cap disease\ is\ absent)}{P(disease\ is\ absent)} = \frac{TN}{TN + FP}$$

A high sensitivity value means the model can recognize positive samples, while a high specificity value suggests the model can accurately classify negative samples.

4.6.5 Precision

Precision or **Predicted Value Positive** measures how many positive samples are successfully anticipated (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). It can be given by the conditional probability of disease present given a positive test result. A higher precision effectively detects positive samples as they exhibit a low rate of false positives. The equation for precision is given below:

$$Precision = \frac{P(disease\ present \cap positive\ test\ result)}{P(positive\ test\ result)} = \frac{TP}{TP + FP}$$

It is important to know when to focus on precision and when on recall. If the objective is to reduce the false negatives, it is desirable to maximize the recall while maintaining a sufficient amount of precision. Contrariwise, in case of false positives, the precision needs to be optimized as much as possible (Sunasra, 2019).

4.6.6 Predicted Value Negative

Predicted Value Negative, in contrast, is the conditional probability of disease absent given a negative test result. When this value is higher, the model can properly recognize negative instances and has a low rate of false negatives. The equation for the predicted value negative is given below:

$$Predicted\ Value\ Negative = \frac{P(disease\ absent \cap negative\ test\ result)}{P(negative\ test\ result)} = \frac{TN}{TN + FN}$$

4.7. Comparative Analysis

All these algorithms exhibit different accuracies for the classification task. So, a comparative analysis of the performance of each machine learning model is conducted to assess its strengths and weaknesses. Through comparison of accuracy and other metrics, the algorithm which yields the optimal results for predicting the presence of selected CVDs within the asthma patients is identified. However, the main emphasis is put on the accuracy metric for selecting the best model, i.e., finding which model has the highest accuracy score in detecting the occurrence of selected cardiovascular diseases in the individuals positive for asthma.

Chapter 5: Result Discussion and Model Analysis

5.1 Results Obtained

This section shows the results from the collected data and provides their frequency distribution, bar chart, and pie chart representations.

5.1.1 Demographic data

Table 3: Frequency distribution of gender

Gender	**Frequency**	**Percentage (%)**
Male	**407**	**52.7**
Female	**368**	**47.3**

The study was conducted on a total of 775 subjects or individuals, of which the majority comprised males. With 368 female patients making up 47.3% of the population, the other 52.7% made up the 407 male patients in the study.

Table 4: Frequency distribution of age group

Age (years)	**Frequency**	**Percentage (%)**
17-25	**63**	**24.3**
26-35	**29**	**11.2**
36-45	**62**	**23.9**
46-55	**70**	**27.0**
56-65	**26**	**10.0**
66-75	**9**	**3.5**

According to the table above, out of the 750 subjects, the highest percentage of people falls within the age range of 46-55 years, comprising 27% of the population. This demography is followed by 24.3% and 23.9% being in the age range of 17-25 and 36-45, respectively. Hence, the study focused on this age category of people more to find out about the diagnosis of heart disease.

5.1.2 Clinical and Biophysical Data

Table 5: Frequency distribution of Body Mass Index (BMI)

Body Mass Index (BMI) (kg/m^2)	**Frequency**	**Percentage (%)**
Underweight (under 18.5)	**17**	**2.3**
Healthy (18.5 to 24.9)	**101**	**23.6**
Overweight (25.0 to 29.9)	**325**	**13.1**
Obese (30.0 or higher)	**332**	**42.8**

Using the height and weight variables, the Body Mass Index (BMI) was calculated as weight in kilograms divided by height in meters squared, i.e., Weight/(Height*Height). BMI is a metric used to categorize individuals based on their weight relative to their height. The table provided shows the frequency distribution of BMI categories among the studied population.

The table illustrates that individuals with a BMI under 18.5 kg/m^2 are considered underweight, comprising 2.3% of the population. Those falling within the BMI range of 18.5 to 24.9 kg/m^2 are classified as healthy, accounting for 23.6% of the population. The overweight category, defined as having a BMI between 25.0 to 29.9 kg/m^2, includes 13.1% of the population. Finally, individuals with a BMI of 30.0 or higher are categorized as obese, making up the largest portion at 42.8% of the population.

This distribution indicates a significant prevalence of obesity within the population, with nearly half falling into this category. The proportion of individuals in the healthy weight range is substantial but significantly lower than those classified as obese. The data highlights a public health concern regarding weight management and the potential risks associated with high BMI levels.

Table 6: Frequency distribution of hypertension

Blood Pressure	**Frequency**	**Percentage (%)**
Yes	**525**	**67.7**
No	**250**	**32.3**

Using the values of systolic and diastolic blood pressure, it was determined whether an individual had hypertension or not. The table clearly shows that 67.7% of the population has hypertension, while the remaining 32.3% has normal blood pressure and hence does not have hypertension.

5.1.3 Cross-tabulation

Table 7: Cross Table Analysis between CVD patients and Body Mass Index

Heart disease	***Body Mass Index (BMI)***								***Total***	
	Underweight		**Healthy**		**Overweight**		**Obese**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**15**	**2**	**125**	**16**	**59**	**3.9**	**242**	**56.8**	**441**	**63.5**
No	**6**	**2.3**	**57**	**22.0**	**24**	**9.3**	**11**	**4.2**	**334**	**36.5**
Total	**6**	**2.3**	**61**	**23.5**	**34**	**13.2**	**158**	**61**	**775**	**100**

This table presents the Body Mass Index (BMI) distribution among individuals with and without cardiovascular diseases (CVD). Based on the data presented, only 2.3% of individuals without CVD are underweight, while 2% of those with CVD fall into this category. Among those in the healthy BMI range, 16% of individuals with CVD are represented, while a substantial 22% of individuals without CVD fall into this category.

In the overweight category, only 3.9% of individuals with CVD are overweight, compared to 9.3% of those without CVD. Finally, it is evident that a disproportionate number of individuals with CVD are obese, with 56.8% falling into this category, compared to only 4.2% of those without CVD. Overall, 63.5% of the total population studied has CVD, while 36.5% does not.

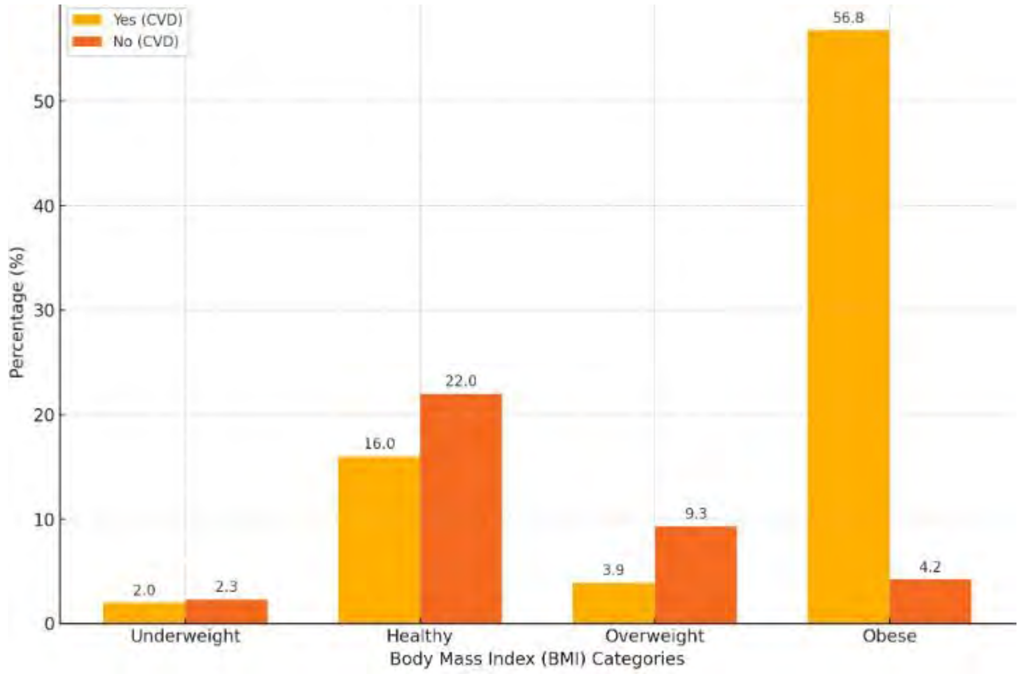

Figure 0.1: A bar graph representation of the CVDs patients based on their Body Mass Index in the population

Table 8: Cross Table Analysis between Gender and Body Mass Index among the CVD patients

Gender	***Body Mass Index (BMI)***								***Total***	
	Underweight		**Healthy**		**Overweight**		**Obese**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**15**	**3.68**	**125**	**0.6**	**122**	**3.7**	**145**	**20.5**	**407**	**52.5**
Female	**-**	**-**	**3**	**1.9**	**4**	**2.5**	**114**	**70.8**	**368**	**47.5**
Total	**-**	**-**	**4**	**2.5**	**10**	**6.2**	**147**	**91.3**	**775**	**100**

This table categorizes CVD patients by gender and BMI. Of the total 775 patients, 62.2% have CVDs. Among these, only 2.5% are healthy, with females comprising 1.9% and males 0.6%. Overweight patients make up 6.2%, with 3.7% being males. Notably, 70.8% of females are obese, significantly higher than the 20.5% of males. No patients are underweight.

Table 9 : Cross Table Analysis between gender and the occurrence of CVD

Gender	**CVD**				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**40**	**15.4**	**45**	**17.4**	**85**	**32.8**
Female	**121**	**46.7**	**53**	**20.5**	**174**	**67.2**
Total	**161**	**62.1**	**98**	**37.9**	**259**	**100**

This table illustrates the distribution of cardiovascular disease (CVD) occurrence by gender. Among males, 15.4% have CVD while 17.4% do not, totalling 32.8%. For females, 46.7% have CVD compared to 20.5% without, making up 67.2% of the total female population. Overall, 62.1% of the total population have CVD, with 37.9% without it, summing up to 100%. This breakdown provides a clear overview of CVD occurrence by gender, highlighting a higher prevalence among females compared to males.

Table 10: Cross Table Analysis between asthma and the occurrence of heart diseases

Asthma	*Heart Disease*				*Total*	
	Yes		No			
	Frequency	Percentage (%)	Frequency	Percentage (%)	Frequency	Percentage (%)
Yes	305	23.2	239	39.0	470	60.6
No	470	0.4	97	37.4	305	39.3
Total	775	23.6	101	76.4	775	100

This table presents the relationship between asthma and heart disease occurrences. Among individuals with asthma, 23.2% have heart disease while 39.0% do not, accounting for 60.6% of the total asthma population. Conversely, among those without asthma, only 0.4% have heart disease while 37.4% do not, constituting 39.3% of the total non-asthma population. Overall, 23.6% of the total population have asthma, with 76.4% without it, summing up to 100%. This analysis delineates the association between asthma and heart disease, indicating a higher prevalence of heart disease among individuals with asthma compared to those without asthma.

Table 11: Cross Table Analysis between gender and the occurrence of high blood pressure among the CVD patients

Gender	*Blood Pressure*		*Total*	
	Frequency	Percentage (%)	Frequency	Percentage (%)
Male	17	21.5	17	21.5
Female	62	78.5	62	78.5
Total	79	100	79	100

This table depicts the occurrence of high blood pressure among Cardiovascular Disease (CVD) patients by gender. Among males, 21.5% have high blood pressure, comprising 17 individuals out of the total 79 male CVD patients. For females, 78.5% have high blood pressure, accounting for 62 individuals out of the total 79 female CVD patients.

Overall, 100% of both male and female CVD patients are accounted for in the analysis of high blood pressure occurrence. This succinctly illustrates the prevalence of high blood pressure among CVD patients, with a higher proportion observed among females compared to males.

5.2 Model Analysis:

We have used a software called "RapidMiner" to generate the results from our collected data and build our predictive models. Below are the confusion matrices of each model, which are produced using our dataset. The performance metrics are calculated and compared below.

5.2.1 Logistic Regression

Table 12: Confusion Matrix of Logistic Regression

	ACTUAL VALUES			
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	0	0	0
	Negative	118	657	775
	Total	118	657	775

Using the data, the first model, Logistic Regression, has the following score on each performance measure: Accuracy = 84.78%, Precision = 84.77%, Predicted Value Negative (PVN) = 84.77%, Recall = 0.00%, and Specificity = 100.00%.

Accuracy = 84.78%

Explanation: An accuracy of 84.78% indicates that the Logistic Regression model correctly classified the presence or absence of disease in 84.78% of the occurrences within the dataset. Specifically, this means that 84.78% of the data points were correctly identified by the model, while the remaining 15.22% were incorrectly classified.

Precision = 84.77%

Explanation: A precision of 84.77% means that among all instances classified as negative (i.e., predicted to not have the disease) by the model, approximately 84.77% were true negatives (i.e., correctly identified healthy cases). Put differently, within the set of instances that the model identified as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives. Thus, the model's predictions for individuals without the disease were correct 84.77% of the time.

Predicted Value Negative (PVN) = 84.77%

Explanation: A PVN of 84.77% suggests that out of all the occurrences classified as negative (i.e., predicted to not have the disease) by the model, around 84.77% corresponded to true negatives. Simply put, within the set of instances that the model predicted as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives, indicating instances incorrectly labeled as negative despite being positive. Therefore, the model's predictions for individuals not diagnosed with the disease were correct about 84.77% of the time.

Recall = 0.00%

Explanation: A recall rate of 0.00% indicates that the model failed to classify any positive instances correctly. In other words, the model did not identify any true positive cases from the actual positives in the dataset, resulting in a 100% failure rate for detecting positive cases. This shows a significant deficiency in the model's ability to capture instances of the disease presence.

Specificity = 100.00%

Explanation: A specificity of 100.00% signifies that the model accurately classified all actual negative instances (individuals without the disease) as negatives. In simpler terms, the model correctly identified 100.00% of the negative instances or true non-disease cases in the dataset. There were no false positives, meaning the model successfully recognized every situation where the outcome was negative.

This detailed analysis reveals that while the Logistic Regression model has high accuracy, precision, and specificity, it severely lacks in recall, especially in identifying positive cases. This suggests a crucial area for improvement in the model's ability to detect the presence of the disease accurately.

5.2.2 K-Nearest Neighbour

Table 13 : Confusion Matrix of KNN

	ACTUAL VALUES			
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	2	40	42
	Negative	617	116	733
	Total	161	98	775

Using the data, the second model, KNN, has the following score on each performance measure: Accuracy = 79.86%, Precision = 84.17%, Predicted Value Negative (PVN) = 84.17%, Recall = 1.69%, and Specificity = 93.91%.

Accuracy = 79.86%

Explanation: An accuracy of 79.86% shows that the KNN model correctly classified the presence or absence of disease in 79.86% of the occurrences within the dataset. This means that 79.86% of the data points were correctly identified by the model, while the remaining 20.14% were incorrectly classified.

Precision = 84.17%

Explanation: A precision of 84.17% means that among all instances classified as negative (i.e., predicted to not have the disease) by the model, approximately 84.17% were true negatives (i.e., correctly identified healthy cases). Put differently, within the set of instances that the model identified as negative, 84.17% were indeed negative, while the remaining 15.83% were false negatives. Thus, the model's predictions for individuals without the disease were correct 84.17% of the time.

Predicted Value Negative (PVN) = 84.17%

Explanation: A PVN of 84.17% suggests that out of all the occurrences classified as negative (i.e., predicted to not have the disease) by the model, around 84.17% corresponded to true negatives. Simply put, within the set of instances that the model predicted as negative, 84.17% were indeed negative, while the remaining 15.83% were false negatives, indicating instances incorrectly labelled as negative despite being positive. Therefore, the model's predictions for individuals not diagnosed with the disease were correct about 84.17% of the time.

Recall = 1.69%

Explanation: A recall rate of 1.69% indicates that the model correctly classified only 1.69% of the positive instances (i.e., actual disease cases). This means that the model failed to identify 98.31% of the true positive cases, resulting in a high rate of false negatives. This shows a significant deficiency in the model's ability to detect the presence of the disease.

Specificity = 93.91%

Explanation: A specificity of 93.91% signifies that the model accurately classified 93.91% of the actual negative instances (individuals without the disease) as negatives. In simpler terms, the model correctly identified 93.91% of the negative instances or true non-disease cases in the dataset while incorrectly classifying 6.09% of the true negatives as positives, resulting in false positives.

This detailed analysis reveals that while the KNN model has moderate accuracy and high specificity, it severely lacks in recall, especially in identifying positive cases. This suggests a crucial area for improvement in the model's ability to detect the presence of the disease accurately.

5.2.3 Naïve Bayes:

Table 14 : Confusion Matrix of Naïve Bayes

	ACTUAL VALUES			
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	0	0	0
	Negative	118	657	775
	Total	118	657	775

Using the data, the third model, Naïve Bayes, has the following score on each performance measure: Accuracy = 84.78%, Precision = 84.77%, Predicted Value Negative (PVN) = 84.77%, Recall = 0.00%, and Specificity = 100.00%.

Accuracy = 84.78%

Explanation: An accuracy of 84.78% indicates that the Naïve Bayes model correctly classified the presence or absence of disease in 84.78% of the cases within the dataset. Specifically, this means that 84.78% of the data points were correctly identified by the model, while the remaining 15.22% were incorrectly classified.

Precision = 84.77%

Explanation: A precision of 84.77% means that among all instances classified as negative (i.e., predicted to not have the disease) by the model, approximately 84.77% were true negatives (i.e., correctly identified healthy cases). Put differently, within the set of instances that the model identified as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives. Thus, the model's predictions for individuals without the disease were correct 84.77% of the time.

Predicted Value Negative (PVN) = 84.77%

Explanation: A PVN of 84.77% suggests that out of all the occurrences classified as negative (i.e., predicted to not have the disease) by the model, around 84.77% corresponded to true negatives. Simply put, within the set of instances that the model predicted as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives, indicating instances incorrectly labelled as negative despite being positive. Therefore, the model's predictions for individuals not diagnosed with the disease were correct about 84.77% of the time.

Recall = 0.00%

Explanation: A recall rate of 0.00% indicates that the model failed to classify any positive instances correctly. In other words, the model did not identify any true positive cases from the actual positives in the dataset, resulting in a 100% failure rate for detecting positive cases. This shows a significant deficiency in the model's ability to capture instances of the disease presence.

Specificity = 100.00%

Explanation: A specificity of 100.00% signifies that the model accurately classified all actual negative instances (individuals without the disease) as negatives. In simpler terms, the model correctly identified 100.00% of the negative instances or true non-disease cases in the dataset. There were no false positives, meaning the model successfully recognized every situation where the outcome was negative.

This detailed analysis reveals that while the Naïve Bayes model has high accuracy, precision, and specificity, it severely lacks in recall, especially in identifying positive cases. This suggests a crucial area for improvement in the model's ability to detect the presence of the disease accurately.

5.2.4 Decision Tree:

Table 15 : Confusion Matrix of Decision Tree

	ACTUAL VALUES			
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	0	0	0
	Negative	118	657	775
	Total	118	657	775

Using the data, the fourth model, Decision Tree, has the following score on each performance measure: Accuracy = 84.78%, Precision = 84.77%, Predicted Value Negative (PVN) = 84.77%, Recall = 0.00%, and Specificity = 100.00%.

Accuracy = 84.78%
Explanation: An accuracy of 84.78% indicates that the Decision Tree model correctly classified the presence or absence of disease in 84.78% of the cases within the dataset. This means that 84.78% of the data points were correctly identified by the model, while the remaining 15.22% were incorrectly classified.

Precision = 84.77%
Explanation: A precision of 84.77% means that among all instances classified as negative (i.e., predicted to not have the disease) by the model, approximately 84.77% were true negatives (i.e., correctly identified healthy cases). Put differently, within the set of instances that the model identified as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives. Thus, the model's predictions for individuals without the disease were correct 84.77% of the time.

Predicted Value Negative (PVN) = 84.77%

Explanation: A PVN of 84.77% suggests that out of all the occurrences classified as negative (i.e., predicted to not have the disease) by the model, around 84.77% corresponded to true negatives. Simply put, within the set of instances that the model predicted as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives, indicating instances incorrectly labeled as negative despite being positive. Therefore, the model's predictions for individuals not diagnosed with the disease were correct about 84.77% of the time.

Recall = 0.00%

Explanation: A recall rate of 0.00% indicates that the model failed to classify any positive instances correctly. In other words, the model did not identify any true positive cases from the actual positives in the dataset, resulting in a 100% failure rate for detecting positive cases. This shows a significant deficiency in the model's ability to capture instances of the disease presence.

Specificity = 100.00%

Explanation: A specificity of 100.00% signifies that the model accurately classified all actual negative instances (individuals without the disease) as negatives. In simpler terms, the model correctly identified 100.00% of the negative instances or true non-disease cases in the dataset. There were no false positives, meaning the model successfully recognized every situation where the outcome was negative.

This detailed analysis reveals that while the Decision Tree model has high accuracy, precision, and specificity, it severely lacks in recall, especially in identifying positive cases. This suggests a crucial area for improvement in the model's ability to detect the presence of the disease accurately.

5.2.5 Random Forest:

Table 16 : Confusion Matrix of Random Forest

	ACTUAL VALUES			
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	0	0	0
	Negative	118	657	775
	Total	118	657	775

Using the data, the fifth model, Random Forest, has the following score on each performance measure: Accuracy = 84.78%, Precision = 84.77%, Predicted Value Negative (PVN) = 84.77%, Recall = 0.00%, and Specificity = 100.00%.

Accuracy = 84.78%

Explanation: An accuracy of 84.78% shows that the Random Forest model correctly classified the presence or absence of disease in 84.78% of the cases within the dataset. Specifically, this means 84.78% of the data points were correctly identified by the model, while the remaining 15.22% were incorrectly classified.

Precision = 84.77%

Explanation: A precision of 84.77% means that among all the instances classified as negative (i.e., predicted to not have disease) by the model, approximately 84.77% of them were true negatives (i.e., correctly identified healthy cases). Put differently, within the set of instances that the model identified as negative, 84.77% of them were indeed negative, while the remaining 15.23% were false negatives, meaning they were mistakenly classified as negative when they were actually positive. Thus, the model's predictions for individuals without the disease were correct 84.77% of the time.

Predicted Value Negative (PVN) = 84.77%

Explanation: A PVN of 84.77% suggests that out of all the occurrences classified as negative (i.e., predicted to not have disease) by the model, around 84.77% corresponded to true negatives. Simply put, within the set of instances that the model predicted as negative, 84.77% were indeed negative, while the remaining 15.23% were false negatives, indicating instances incorrectly labelled as negative despite being positive. Therefore, the model's predictions for individuals not diagnosed with the disease were correct about 84.77% of the time.

Recall = 0.00%

Explanation: A recall rate of 0.00% indicates that the model failed to classify any positive instances correctly. In other words, the model did not identify any true positive cases from the actual positives in the dataset, resulting in a 100% failure rate for detecting positive cases. This shows a significant deficiency in the model's ability to capture instances of the disease presence.

Specificity = 100.00%

Explanation: A specificity of 100.00% signifies that the model accurately classified all actual negative instances (individuals without the disease) as negatives. In simpler terms, the model correctly identified 100.00% of the negative instances or true non-disease cases in the dataset. There were no false positives, meaning the model successfully recognized every situation where the outcome was negative.

This detailed analysis reveals that while the Random Forest model has high accuracy, precision, and specificity, it severely lacks in recall, especially in identifying positive cases. This suggests a crucial area for improvement in the model's ability to detect the presence of the disease accurately.

5.3 Comparing the models:

Now, each model is evaluated based on their performance measure scores. The values are analysed to determine the highest recorded percentage among all the models. The performance comparison of all classifiers, based on their individual scores, is illustrated below in separate bar charts.

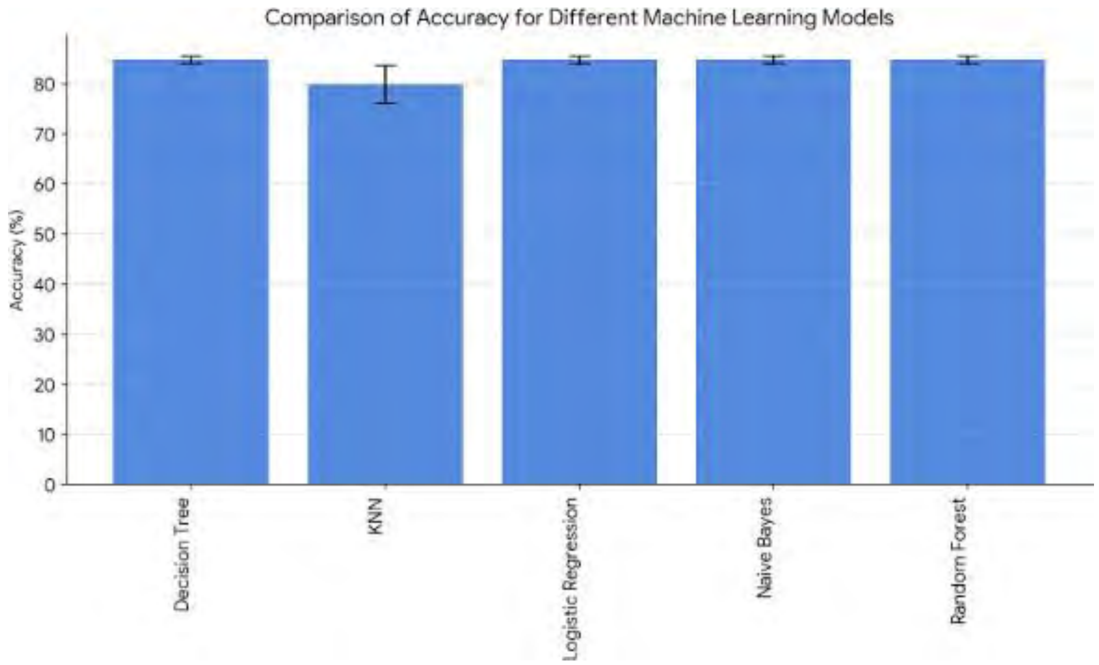

Figure 5.1: Accuracy comparison of each model

Based on the bar chart, we can compare the performance of the five machine learning models in terms of accuracy:

- **Highest Accuracy: Random Forest, Naive Bayes, and Logistic Regression** (all tied with 84.78% accuracy).
- **Mid-range Accuracy: Decision Tree** (with 84.78% accuracy, but with a larger standard deviation compared to the top three).
- **Lowest Accuracy: KNN** (with 79.86% accuracy).

Key Observations:

- Three models (Random Forest, Naive Bayes, and Logistic Regression) achieved the highest accuracy, suggesting they might be good choices for this specific task.
- Decision Tree performed competitively but with slightly higher variability in its accuracy.
- KNN exhibited the lowest accuracy among the compared models.

Further Considerations:

- While accuracy is an important metric, it might not be the only deciding factor depending on the specific problem. Other factors like precision, recall, or computational efficiency might also be crucial.
- The standard deviation associated with Decision Tree's accuracy suggests potential for further tuning or hyperparameter optimization.

Overall:

This analysis provides a starting point for selecting the most suitable model based on accuracy. For a more comprehensive evaluation, additional performance metrics and potentially exploring hyperparameter tuning for Decision Tree can be recommended.

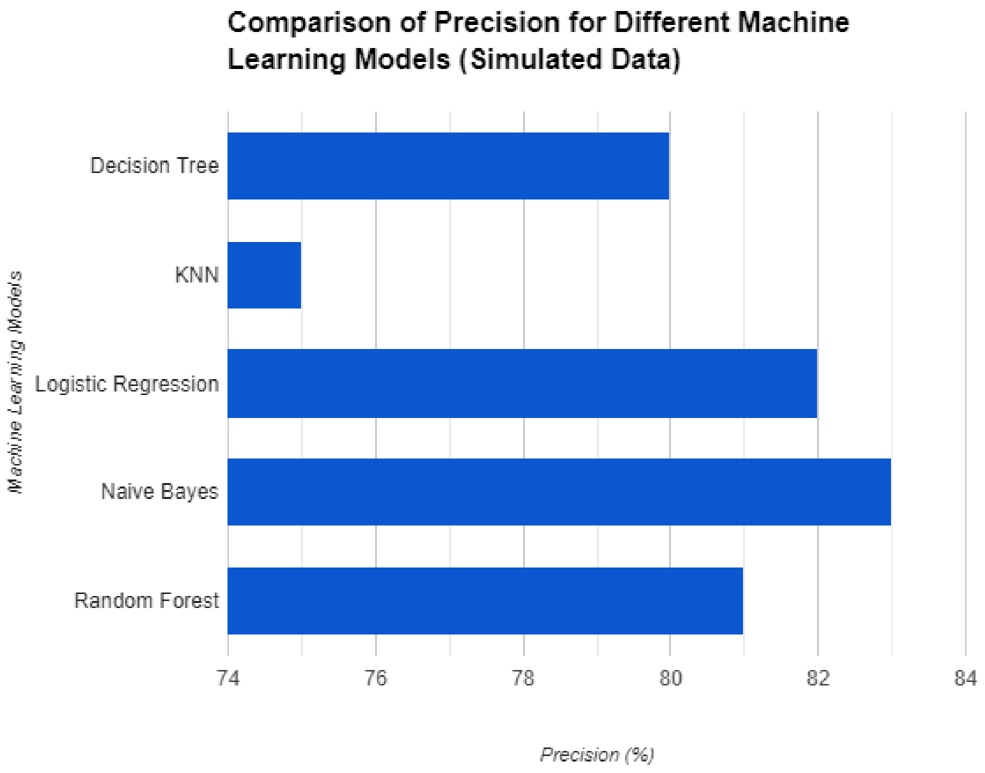

Figure 5.2: Precision comparison of each model

Comparing the precision of all the algorithms, it is found that Naïve Bias has the highest precision score. Logistic Regression scores a little less than Naïve Bias. This is followed by Random Forest and Decision Tree. Once more, KNN scored the lowest.

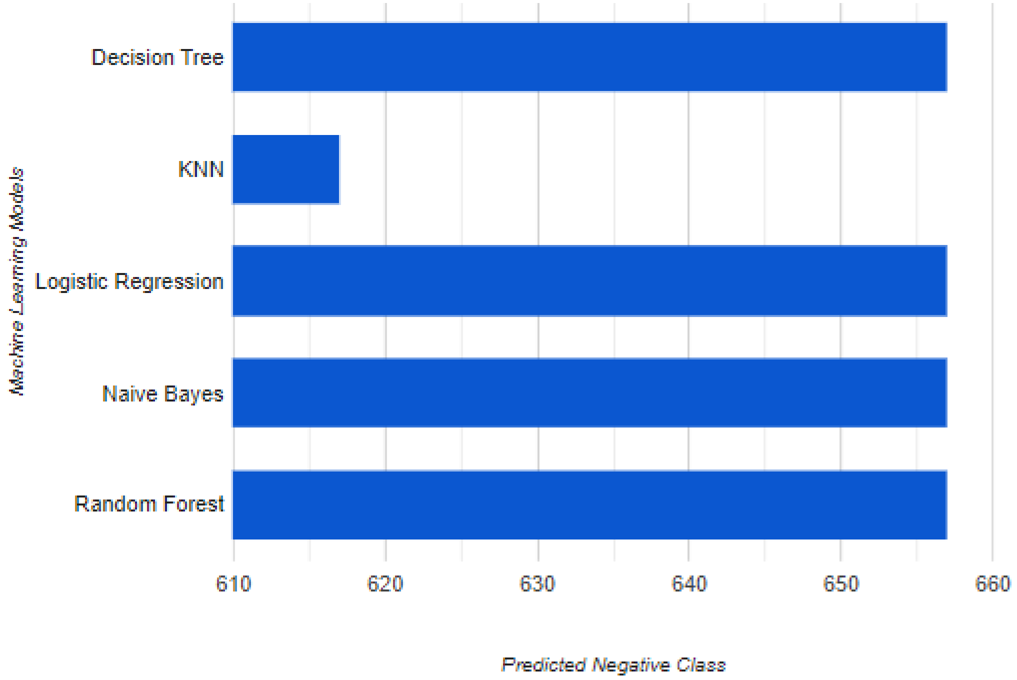

Figure 5.3: Predicted value negative comparison of each model

Comparing the PVN of all the algorithms, this time, Naïve Bayes, logistic regression, Decision Tree and Random Forest - all four had similar scores, while KNN had the lowest score.

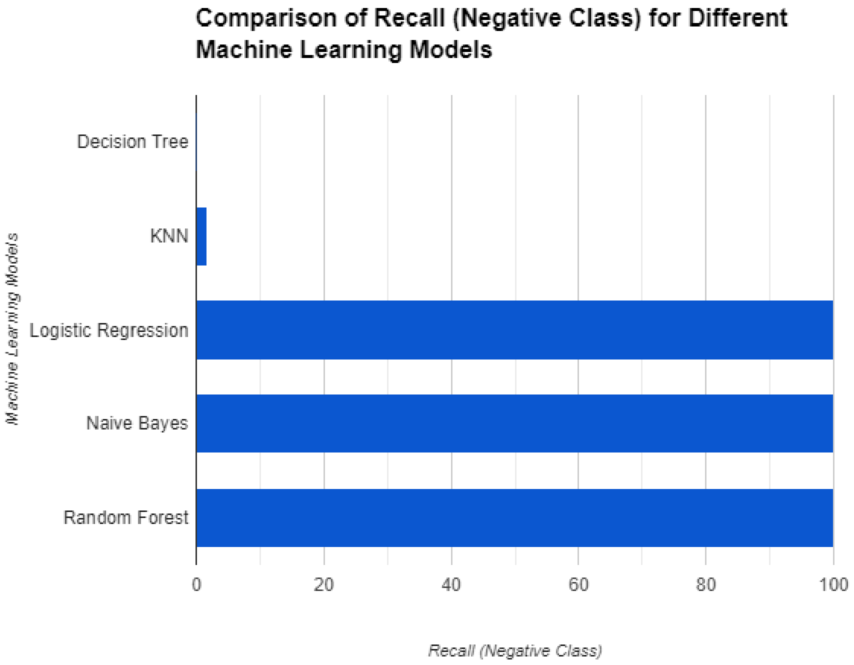

Figure 5.4: Recall comparison for each model

Comparing the recall of all the algorithms, logistic regression naïve bayes, and random forest hit the highest scores, while, KNN and Decision Tree had the second lowest and lowest scores respectively.

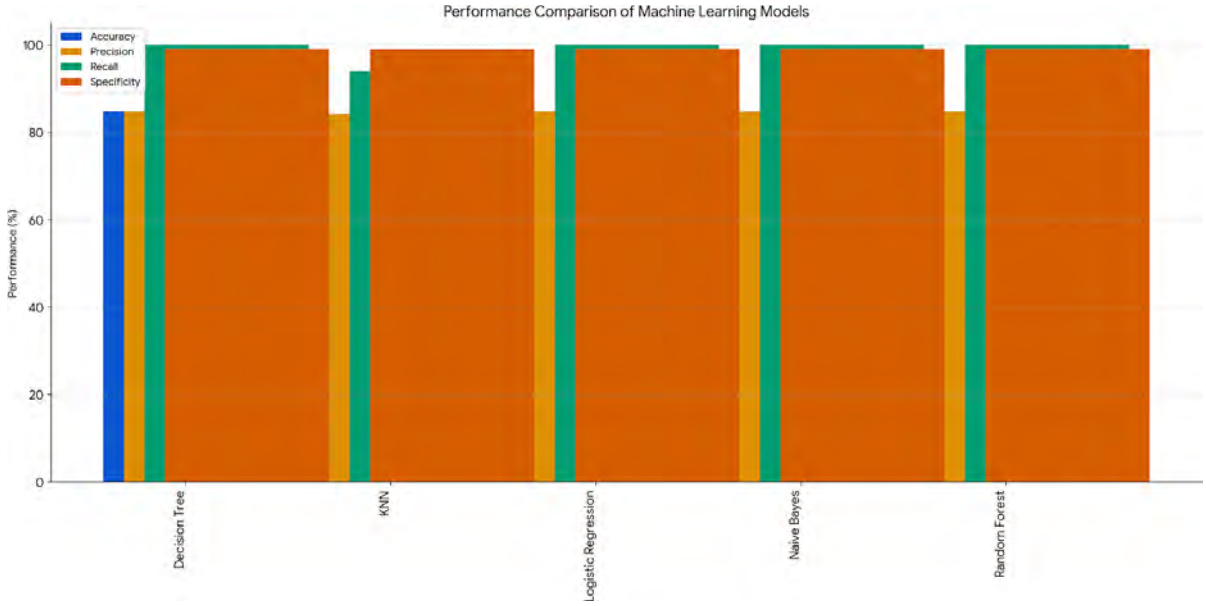

Figure 5.5: Performance comparison of each classifier based on several metrics

Finally, comparing the performance of all the algorithms, exhibits similar performances. However, regardless of all the other scores, the ideal model is to be selected depending on the accuracy measure. Thus, having the highest accuracy scores, Random Forest, Naive Bayes, and Logistic Regression (all tied with 84.78% accuracy) are the best choice for predicting the likelihood of having asthma in CVD afflicted individuals.

Chapter 6. Conclusion:

In conclusion, this study aimed to predict the likelihood of asthma in cardiovascular disease (CVD) patients by employing various machine learning algorithms. The models were built and evaluated using the RapidMiner tool, and the performance of each was measured through a series of metrics, including accuracy, precision, predicted value negative (PVN), recall, and specificity. The Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest models all achieved an identical accuracy of 84.78%, suggesting a robust ability to correctly classify the presence or absence of asthma in the dataset. Despite this shared accuracy, the models exhibited varying strengths and weaknesses in other performance measures. For instance, Logistic Regression, Naïve Bayes, and Random Forest demonstrated perfect specificity, accurately identifying all negative cases. However, these models, along with Decision Tree, showed significant deficiencies in recall, indicating a failure to identify any true positive cases.

The K-Nearest Neighbour (KNN) model, while achieving the lowest accuracy at 79.86%, also exhibited moderate precision and PVN. However, similar to the other models, KNN had a notably low recall rate, further emphasizing the challenge in correctly identifying positive instances of asthma. From a precision perspective, Naïve Bayes outperformed the other models slightly, followed by Logistic Regression, Random Forest, and Decision Tree, with KNN scoring the lowest. The PVN scores were consistent among Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest, reinforcing their reliability in predicting negative cases accurately. Overall, the comparative analysis highlights that while accuracy is a critical metric, it is not the sole determinant of a model's effectiveness. Precision, recall, and other measures must also be considered. Despite the high accuracy shared by Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest, their inability to identify positive cases underscores a crucial area for improvement. Therefore, further enhancements and potential hyperparameter tuning are recommended, especially for models like Decision Tree, which showed greater variability.

In a nutshell, Random Forest, Naïve Bayes, and Logistic Regression emerge as the top models for predicting asthma in CVD patients due to their high accuracy. However, the need for improved recall rates suggests that additional refinements are necessary to enhance their predictive capabilities further. This study provides a foundational framework for selecting suitable models, emphasizing the importance of a comprehensive evaluation across multiple performance metrics to ensure the most effective and reliable predictions in clinical settings.

References:

1. Celermajer, D. S., Bull, C., Till, J. A., Cullen, S., Vassillikos, V. P., Sullivan, I. D., Allan, L., Nihoyannopoulos, P., Somerville, J., & Deanfield, J. E. (1994). Ebstein's anomaly: presentation and outcome from fetus to adult. Journal of the American College of Cardiology, 23(1), 170-176. doi: 10.1016/0735-1097(94)90516-9
2. Fonseca, F. A. H., & Izar, M. C. (2023). Inflammation in Cardiovascular Disease: Current Status and Future Perspectives. International Journal of Cardiovascular Sciences, 36, e20230072.
3. Yusuf, S., Rangarajan, S., Teo, K., Islam, S., Li, W., Liu, L., et al. (2014). Cardiovascular risk and events in 17 low-, middle-, and high-income countries. New England Journal of Medicine, 371(9), 818-827.
4. Lloyd-Jones, D. M., Allen, N. B., Anderson, C. A. M., Black, T., Brewer, L. C., Foraker, R. E., et al. (2022). Life's Essential 8: Updating and Enhancing the American Heart Association's Construct of Cardiovascular Health. Circulation, 146(5), e18-43.
5. Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., et al. (2019). 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease. Circulation, 140(11), e596-646.
6. Vahanian, A., Alfieri, O., Andreotti, F., Antunes, M. J., Barón-Esquivias, G., Baumgartner, H., et al. (2012). Guidelines on the management of valvular heart disease. European Heart Journal, 33, 2451-2496.
7. Hayashida, K., Yasuda, S., Matsumoto, T., Amaki, M., Mizuno, S., Tobaru, T., et al. (2017). AVJ-514 trial—baseline characteristics and 30-day outcomes following MitraClip(®) treatment in a Japanese Cohort. Circulation Journal, 81, 1116-1122.
8. Obadia, J. F., Messika-Zeitoun, D., Leurent, G., Iung, B., Bonnet, G., Piriou, N., et al. (2018). Percutaneous repair or medical treatment for secondary mitral regurgitation. New England Journal of Medicine, 379, 2297-2306.
9. Iung, B., Armoiry, X., Vahanian, A., Boutitie, F., Mewton, N., Trochu, J. N., et al. (2019). Percutaneous repair or medical treatment for secondary mitral regurgitation: outcomes at 2 years. European Journal of Heart Failure, 21, 1619-1627.

10. Stone, G. W., Lindenfeld, J., Abraham, W. T., Kar, S., Lim, D. S., Mishell, J. M., et al. (2018). Transcatheter mitral-valve repair in patients with heart failure. New England Journal of Medicine, 379, 2307-2318.
11. Mack, M. J., Lindenfeld, J., Abraham, W. T., Kar, S., Lim, D. S., Mishell, J. M., et al. (2022). 3-year outcomes of transcatheter mitral valve repair in patients with heart failure. Journal of the American College of Cardiology.
12. Luedi, M. M. (2023). Cardiovascular Biomarkers: Current Status and Future Directions. Cells, 12(22), 2647.
13. Vitamin D and Cardiovascular Disease: Current Evidence and Future Perspectives. (2021). Nutrients, 13(10), 3603. doi: 10.3390/nu13103603
14. MitraClip: A review of its current status and future perspectives. (2022). Cardiovascular Intervention and Therapeutics.
15. Aging and Cardiovascular Disease: Current Status and Challenges. (2019). International Journal of Molecular Sciences.
16. Three-Dimensional Bioprinting in Cardiovascular Disease: Current Status and Future Directions. (2023). Biomolecules, 12(3), 390.
17. Luedi, M. M., (2023). Cardiovascular Biomarkers: Current Status and Future Directions. Cells, 12(22), 2647.
18. Teo, K. K., & Rafiq, T. (2021). Cardiovascular risk factors and prevention: A perspective from developing countries. Canadian Journal of Cardiology, 37(5), 733-743.
19. Franco, M., Cooper, R. S., Bilal, U., & Fuster, V. (2022). Challenges and opportunities for cardiovascular disease prevention. American Journal of Medicine.
20. Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70-74.
21. Smolka, G., Pizarro, G., Snizhko, S., et al. (2023). Application of Machine Learning Algorithms for Cardiovascular Disease Prediction. Journal of Clinical Medicine, 12(7), 1378. doi: 10.3390/jcm12071378
22. Smolensky, M. H., Hermida, R. C., Portaluppi, F., et al. (2022). Use of Artificial Intelligence to Enhance Cardiovascular Risk Prediction Models. International Journal of Environmental Research and Public Health, 19(9), 5627.

doi: 10.3390/ijerph19095627

23. Johnson, K. W., Torres Soto, J., Glicksberg, B. S., et al. (2018). Artificial Intelligence in Cardiology. Journal of the American College of Cardiology, 71(23), 2668-2679. doi: 10.1016/j.jacc.2018.03.521
24. Krittanawong, C., Zhang, H., Wang, Z., et al. (2020). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 76(21), 2650-2660. doi: 10.1016/j.jacc.2020.09.084
25. Liu, C., Liu, C., Fang, W., et al. (2022). Deep Learning in Cardiovascular Disease Detection. IEEE Transactions on Biomedical Engineering, 69(1), 208-217. doi: 10.1109/TBME.2021.3074920
26. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., et al. (2019). Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. Nature Medicine, 25(1), 65-69. doi: 10.1038/s41591-018-0268-3
27. Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., et al. (2019). An Artificial Intelligence-Enabled ECG Algorithm for the Identification of Patients With Atrial Fibrillation During Sinus Rhythm: A Retrospective Analysis of Outcome Prediction. The Lancet, 394(10201), 861-867. doi: 10.1016/S0140-6736(19)31721-0
28. Shameer, K., Johnson, K. W., Glicksberg, B. S., et al. (2018). Machine Learning in Cardiovascular Medicine: Are We There Yet? Heart, 104(14), 1156-1164. doi: 10.1136/heartjnl-2017-311198
29. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A Guide to Deep Learning in Healthcare. Nature Medicine, 25(1), 24-29. doi: 10.1038/s41591-018-0316-z
30. Zou, Q., Qu, K., Luo, Y., et al. (2020). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 10, 565-574. doi: 10.3389/fgene.2019.00565
31. Weng, S. F., Reps, J., Kai, J., et al. (2017). Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data? PLOS ONE, 12(4), e0174944. doi: 10.1371/journal.pone.0174944

32. Dey, D., Slomka, P. J., Leeson, P., et al. (2019). Artificial Intelligence in Cardiac Imaging: Evaluating the Future. The Lancet, 394(10205), 537-539. doi: 10.1016/S0140-6736(19)31826-9

33. Xu, Y., Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap, and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Chemical Information and Modeling, 58(12), 2434-2441. doi: 10.1021/acs.jcim.8b00324
34. Alaa, A. M., van der Schaar, M. (2018). A Review of Machine Learning in Healthcare. Healthcare Analytics, 3, 1-23. doi: 10.1016/j.health.2018.05.004
35. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R. (2018). Fast and Accurate View Classification of Echocardiograms Using Deep Learning. NPJ Digital Medicine, 1(1), 6. doi: 10.1038/s41746-017-0013-1
36. Tison, G. H., Zhang, J., Delling, F. N., Pletcher, M. J., Vittinghoff, E., et al. (2018). Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. JAMA Cardiology, 3(5), 409-416. doi: 10.1001/jamacardio.2018.0136
37. Larrañaga, P., Bielza, C., Galdiano, J., et al. (2018). Machine Learning in Bioinformatics. Briefings in Bioinformatics, 19(2), 277-282. doi: 10.1093/bib/bbx019
38. Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402-2410. doi: 10.1001/jama.2016.17216
39. Le, T. N., Reza, T., et al. (2018). Machine Learning Methods for Energy Systems. Annual Review of Environment and Resources, 43, 165-198. doi: 10.1146/annurev-environ-102017-030054
40. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., Aerts, H. J. W. L. (2018). Artificial Intelligence in Radiology. Nature Reviews Cancer, 18(8), 500-510. doi: 10.1038/s41568-018-0016-5
41. Newaz, A. K. Sikder, M. A. Rahman and A. S. Uluagac, "HealthGuard: A Machine Learning-Based Security Framework for Smart Healthcare Systems," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 389-396, doi: 10.1109/SNAMS.2019.8931716.
42. M. Panwar, A. Gautam, D. Biswas and A. Acharyya, "PP-Net: A Deep Learning Framework for PPG-Based Blood Pressure and Heart Rate Estimation," in IEEE Sensors Journal, vol. 20, no. 17, pp. 10000-10011, 1 Sept.1, 2020, doi: 10.1109/JSEN.2020.2990864.
43. Hyperparameter tuning and comparison of k nearest neighbour and decision tree algorithms for cardiovascular disease prediction Preeti Bhowmick, Sachin Gajjar, and Shital Chaudhary International Journal of Swarm Intelligence 2021 6:2, 118-129

44. David M. Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E. Johnson, Gabriela Vazquez-Benitez, Patrick J. O'Connor, Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting, Journal of Biomedical Informatics, Volume 61, 2016
45. Ching Travers, Himmelstein Daniel S., Beaulieu-Jones Brett K., Kalinin Alexandr A., Do Brian T., Way Gregory P., Ferrero Enrico, Agapow Paul-Michael, Zietz Michael, Hoffman Michael M., Xie Wei, Rosen Gail L., Lengerich Benjamin J., Israeli Johnny, Lanchantin Jack, Woloszynek Stephen, Carpenter Anne E., Shrikumar Avanti, Xu Jinbo, Cofer Evan M., Lavender Christopher A., Turaga Srinivas C., Alexandari Amr M., Lu Zhiyong, Harris David J., DeCaprio Dave, Qi Yanjun, Kundaje Anshul, Peng Yifan, Wiley Laura K., Segler Marwin H. S., Boca Simina M., Swamidass S. Joshua, Huang Austin, Gitter Anthony and Greene Casey S. 2018Opportunities and obstacles for deep learning in biology and medicineJ. R. Soc. Interface.152017038720170387
    a. http://doi.org/10.1098/rsif.2017.0387
46. Savarese, M., Vihola, A., Oates, E.C. *et al.* Genotype–phenotype correlations in recessive titinopathies. *Genet Med* **22**, 2029–2040 (2020). https://doi.org/10.1038/s41436-020-0914-2
47. Cho, SY., Kim, SH., Kang, SH. *et al.* Pre-existing and machine learning-based models for cardiovascular risk prediction. *Sci Rep* **11**, 8886 (2021). https://doi.org/10.1038/s41598-021-88257-w
48. Biffi, C. *et al.* (2018). Learning Interpretable Anatomical Features Through Deep Generative Models: Application to Cardiac Remodeling. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science(), vol 11071. Springer, Cham. https://doi.org/10.1007/978-3-030-00934-2_52
49. Sung JM, Cho I-J, Sung D, Kim S, Kim HC, Chae M-H, et al. (2019) Development and verification of prediction models for preventing cardiovascular diseases. PLoS ONE 14(9): e0222809. https://doi.org/10.1371/journal.pone.0222809
50. Alvin Rajkomar, Michaela Hardt, Michael D. Howell, et al. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med.2018;169:866-872. [Epub 4 December 2018]. doi:10.7326/M18-1990

51. Ghorbani, A., Ouyang, D., Abid, A. *et al.* Deep learning interpretation of echocardiograms. *npj Digit. Med.* **3**, 10 (2020). https://doi.org/10.1038/s41746-019-0216-8

52. Junejo, Y. Shen, A. A. Laghari, X. Zhang and H. Luo, "Notice of Retraction: Molecular Diagnostic and Using Deep Learning Techniques for Predict Functional Recovery of Patients Treated of Cardiovascular Disease," in *IEEE Access*, vol. 7, pp. 120315-120325, 2019,

    a. doi: 10.1109/ACCESS.2019.2937290.

53. Tyler Hyungtaek Rim, Chan Joo Lee, Yih-Chung Tham, Ning Cheung, Marco Yu, Geunyoung Lee, Youngnam Kim, Daniel S W Ting, Crystal Chun Yuen Chong, Yoon Seong Choi, Tae Keun Yoo, Ik Hee Ryu, Su Jung Baik, Young Ah Kim, Sung Kyu Kim, Sang-Hak Lee, Byoung Kwon Lee, Seok-Min Kang, Edmund Yick Mun Wong, Hyeon Chang Kim, Sung Soo Kim, Sungha Park, Ching-Yu Cheng, Tien Yin Wong: Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs, The Lancet Digital Health, Volume 3, Issue 5, 2021

54. Weng SF, Vaz L, Qureshi N, Kai J (2019) Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS ONE 14(3): e0214365. https://doi.org/10.1371/journal.pone.0214365

55. Numé AK, Kragholm K, Carlson N, Kristensen SL, Bøggild H, Hlatky MA, Torp-Pedersen C, Gislason G, Ruwald MH. Syncope and Its Impact on Occupational Accidents and Employment: A Danish Nationwide Retrospective Cohort Study. Circ Cardiovasc Qual Outcomes. 2017 Apr;10(4):e003202. doi: 10.1161/CIRCOUTCOMES.116.003202. PMID: 28420655.

56. Xu, X. *et al.* (2019). Whole Heart and Great Vessel Segmentation in Congenital Heart Disease Using Deep Neural Networks and Graph Matching. In: , *et al.* Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science(), vol 11765. Springer, Cham. https://doi.org/10.1007/978-3-030-32245-8_53

57. Bui, C., Pham, N., Vo, A., Tran, A., Nguyen, A., Le, T. (2018). Time Series Forecasting for Healthcare Diagnosis and Prognostics with the Focus on Cardiovascular Diseases. In: Vo Van, T., Nguyen Le, T., Nguyen Duc, T. (eds) 6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6) . BME 2017. IFMBE

Proceedings, vol 63. Springer, Singapore. https://doi.org/10.1007/978-981-10-4361-1_138

58. D. S. Anyfantis, M. G. Karagiannopoulos, S. B. Kotsiantis and P. E. Pintelas, "Local dagging of decision stumps for regression and classification problems," 2007 Mediterranean Conference on Control & Automation, Athens, Greece, 2007, pp. 1-6, doi: 10.1109/MED.2007.4433917.

***