## CHAPTER ONE

## 1.0 Introduction

Cardiovascular diseases are conditions that influence the constructions or capacity of your heart, for example, Abnormal heart rhythms/arrhythmias, Aorta infection, and Marfan disorder, Congenital coronary illness, Heart assault, Heart disappointment, Heart muscle sickness (cardiomyopathy) and Stroke, and so on (Steinbaum, 2019). Cardiovascular diseases are Non-Communicable Diseases (NCD) like cancers, chronic respiratory diseases, diabetes, and Mental Health Disorders. CVD is seen as “a silent epidemic” because they are at the forefront of diseases leading to death. These diseases share common risk factors namely; age, unhealthy diet sexual orientation, hypertension, diabetes mellitus, tobacco smoking, processed meat utilization, unnecessary liquor consumption, sugar consumption, family ancestry, weight, absence of exercise, psychosocial factors, and air contamination (Emakhu et al., 2020; WHO, 2019).

Over 36 million people die annually from Noncommunicable Diseases accounting for about 63% of all global deaths. More than threequarters of the global deaths from CVDs occur in developing nations. Low- and middle-income countries are known to bear 86% of the number of these premature deaths giving rise to estimated cumulative economic losses of US$7 trillion over the next 15 years (WHO, 2019).

According to the Federal Ministry of Health (FMoH), “cardiovascular disease is a major public health concern, accounting for 11% of the over 2 million NCD deaths in Nigeria each year.” People with Cardiovascular Disease are often unaware of their condition until a disaster such as a stroke, heart attack, or death occurs. (WHO, 2019).

The increase in the amount of health data gathered through the electronic health record (EHR) systems makes the use of strong analysis tools are necessary. Data Mining has proved itself to be very effective in data analysis and forecasting diverse scenarios for numerous fields (Sai Shekhar et al., 2021). Paring data mining with deep learning had led to the formation of many models to forecast specific scenarios for us to operate on them. The need of making accurate predictions of heart disease made the use of machine learning algorithms to point out predictions based on many factors. In the occurrences of many factors, the human mind cannot process that much and come up with estimation and can therefore provide incorrect feedback several times, leading to vital risk to the patient.

## 1.1 Background of the Study

Several research have been done in relation to heart disease in machine learning, some have work on predicting the early stage of heart disease (Almustafa, 2020; Gao et al., 2021; Usha Sri, 2020), increasing the efficiency of diagnosis and reduce the misclassification cost (Dandona et al., 2017) and on the relationship between certain factors and to the presence of heart disease (Muibideen & Prasad, 2020) as well as enhancing the selection of features(attributes) that are in causal relationship with presence of heart disease(Jothi Prakash & Karthikeyan, 2021).

Research have been done on the area of heart disease prediction using different machine learning techniques like logistic regression, Support Vector Classifier, Random Forest, Decision trees and with deep learning models e.g. Convolutional Neural Network (CNN). Muibideen & Prasad, (2020) in their research builds a probabilistic graphical model to understand the causal relationship among attributes of the Cleveland heart disease dataset from University of California Irvine (UCI) to predict heart disease. Emakhu et al., (2020) in their research used AdaBoost, Bagging, Random Forest, and Voting Ensemble (Decision Trees, Logistic Regression and Support Vector Machines) to analyze different heart disease prediction system models for designing an automated medical diagnosis.

Many factors or input features even to machine learning often make a predictive modeling task more challenging to model. This is more generally referred to as the curse of dimensionality. Feature engineering is important in increasing the efficiency and correctness of prediction on machine learning models. It entails changing raw data into characteristics that better describe the underlying problem to prediction models, resulting in enhanced model accuracy on previously unseen data. Feature engineering includes determining attributes that are relevant for your modeling assignment, estimating the utility of a feature, and extracting features. The automatic creation of new features from raw data, Feature Selection; the selection of features from a large number of features to a few that are helpful. (Shekhar, 2018). Also in all the above researches, 14 features(attributes) where used which brings the need for dimensionality reduction for computational efficiency and improvement on the accuracy of the analysis(Lidy & Rauber, 2008).

Bayesian networks (BNs) have received increasing research attention as it possesses potential significant benefit to the healthcare system(McLachlan et al., 2020). A Bayesian network represents the causal probabilistic relationship among a set of random variables, their conditional dependencies, and it provides a compact representation of a joint probability distribution. It is made up of two major components: a directed acyclic graph and a collection of conditional probability distributions. The directed acyclic graph is made up of nodes that represent random variables. A node might be a health domain in terms of health assessment, and the states of the node could be the possible reactions to that domain(Cheng et al., 2002).

The proposed method in this study is the use of wrapper feature selection technique as a dimension reduction technique for extracting important features and the use of Naïve Bayes, Bayesian Network, KNN, and Logistic Regression to make predictions. The performance of all the models is also measured to make a comparison.

## 1.2 Problem Statement

Medical dictation has traditionally been a high-maintenance sector in terms of time, accuracy, and cost. Human beings are prone to and capable of making mistakes. Cases of cardiovascular diseases are rising at an exponential pace, and that is very troubling and early detection is crucial in preventing the progression of the disease. Patients usually visits a hospital when the disease reaches an advanced stage. This leads to increase in the cost of treatment and also weakens the chances of recovery and often leads to loss of life. Cardiac disorders can be avoided if they are detected early. The importance of developing a constraint-free, reliable and timely prediction system has long been recognized by the health sector. Machine learning based prediction system can prove to be an effective tool in detecting the heart diseases in an early stage.

## 1.4 Aim and Objectives of the Study

The main aim of the study is to improve the performance of heart disease prediction through feature reduction. To achieve these, the work considered the following specific objectives:

* Study the best performing models in heart disease prediction.
* Perform Different feature reduction technique and measure performance on different models of each.
* Compare all the results obtained from the model.

## 1.5 Significance of the study

As heart disease has become is a silent epidemic, accurate and timely prediction of heart disease is important. This will help in reduction in the cost of treatment, loss of life and also increases the chances of recovery.

## 1.6 Scope and Limitation of The Study

**Scope**

This study covers the use of structured tabular data on supervised machine learning algorithms to accurately predict the risk of the presence of heart disease in a person.

**Limitation**

The Study does not include the use of heart image data to make prediction on the presence of heart disease in a person

# CHAPTER TWO: LITERATURE REVIEW

## Introduction

This chapter reviews the prominent prediction system in relation to this research. The chapter explain heart disease, its type and causes. It also explains the various symptoms and ways of manually diagnosing heart disease. An over view of machine learning as well as different tool and algorithms predominantly used in heart disease prediction is given. The Cleveland dataset as well as features are also explained. Various related works are also briefly outlined; the algorithm used and evaluation metrics used.

## Cardiovascular Diseases

Cardiovascular diseases (CVDs) include any disorder, abnormality, or failure to function well, relating to the heart and blood vessels or the circulation.(Ike & Onyema, 2020). Cardiovascular diseases includes Abnormal heart rhythms/arrhythmias, Aorta infection, and Marfan disorder, Congenital coronary illness, Heart assault, Heart muscle sickness (cardiomyopathy) and Stroke, and so on (Steinbaum, 2019).

Cardiovascular diseases (CVDs) is not in itself an actual disease; it is rather refers to wide range of disorders affecting not only the heart but also blood vessels(Adedoyin & Adesoye, 2005). CVD is a cluster of injuries that affect the cardiovascular system (the heart and blood vessels). These are most commonly diseases of the heart and of the blood vessels of the heart and brain(Nason, 2007).

## Types Cardiovascular Diseases

CVD comprises many different types of condition. Some of these might develop at the same time or lead to other conditions or diseases within the group. Diseases and conditions that affect the heart include:

**Coronary heart disease**

Coronary heart disease (CHD) is caused by atheromatous plaques accumulating in the walls of the arteries that supply the myocardium. CHD damages the arteries that supply the cardiac muscle. It is the leading cause of sudden death.

**Angina**

Angina is a type of chest pain that can include arm, jaw, and other forms of discomfort. The word discomfort is selected over pain to describe angina since it varies greatly in nature and degree between individuals and most people do not regard angina as painful unless severe. Angina is a heart muscle cramp. Angina is a form of chest pain caused by reduced cardiac blood flow.

**Stroke**

A stroke occurs when the blood flow to a portion of the brain is cut off by an artery blockage or rupture (haemorrhage). A clogged or ruptured artery deprives the brain of oxygen, causing brain cells to damage or die (necrosis), limiting brain function. If left untreated, a stroke can cause lifelong brain damage or death.

Strokes are characterized as ischemic or hemorrhagic. Ischemia can be caused by thrombosis, embolism, or systemic hypoperfusion (reduction of the blood flow to all parts of the body).

**Rheumatic heart disease**

Rheumatic heart disease occurs when streptococcal rheumatic fever damages the heart valves. Rheumatic fever is an inflammatory condition that can damage the heart, joints, brain, and skin. Acute rheumatic fever affects anyone, but most commonly affects children aged 5 to 15. The resulting rheumatic heart disease can be fatal.

**Congenital heart disease**

Congenital heart disease is a wide term that describes a variety of cardiac problems caused by improper or disorganized heart development before birth. Congenital heart disease is a congenital defect in heart function or anatomy.

**Peripheral arterial disease**

A gradual buildup of fatty material within the artery walls causes the disease (atherosclerosis). A thrombus (blood clot) can form due to atheroma, completely blocking the artery. Peripheral arterial disease occurs when the arteries supplying the legs become narrowed or blocked. The artery usually narrows in the upper leg.

**Aortic aneurysm and dissection**

An aortic aneurysm is a balloon-like swelling of the aorta that can rupture causing large internal bleeding (dissection).

**Deep vein thrombosis**

In a deep vein, usually in the lower leg, develops a blood clot (thrombus). Deep vein thrombosis can cause leg pain and complications. The deep veins are surrounded by muscles in the leg's centre. A DVT is distinct from blood clots that form in the skin's superficial veins. The superficial thrombophlebitis is less serious.

**Other cardiovascular diseases**

These could include tumours of the heart, vascular tumours of the brain, disorders of the heart muscle (cardiomyopathy), heart valve diseases and disorders of the lining of the heart(Nason, 2007).

## Symptoms

Symptoms will vary depending on the specific condition. Some conditions, such as type 2 diabetes or hypertension, may initially cause no symptoms at all(Felman, 2019).However, typical symptoms of an underlying cardiovascular issue include:

* pain or pressure in the chest, which may indicate angina
* pain or discomfort in the arms, left shoulder, elbows, jaw, or back
* shortness of breath
* nausea and fatigue
* lightheadedness or dizziness
* cold sweats

Although these are the most common ones, CVD can cause symptoms anywhere in the body.

## Risk factors

Some risk factors are responsible for this rising trend of CVDs in Nigeria. These can be classified as modifiable risk factors, nonmodifiable risk factors, and emerging risk factors.

Worldwide, CVDs are largely driven by modifiable risk factors. In Nigeria, these modifiable risk factors were listed by the WHO (in 2016) in the order of occurrence. These lists include the harmful use of alcohol (22%), physical inactivity (22%), tobacco use (11%), hypertension/raised

blood pressure (18%), salt/sodium intake (8%), diabetes mellitus and dyslipidemia (4%), obesity (4%), ambient air pollution (1%), and household pollution (<1%).

On the other hand, the nonmodifiable risk factors include advancing age, male sex, family history of premature cardiovascular events, and race.

Emerging risk factors include elevated homocysteine, small, dense lipoprotein (Lpa), plasminogen activator inhibitor, inflammatory markers such as C‑reactive protein and infectious agents‑like chlamydia.(Ike & Onyema, 2020)

## Cardiovascular Disease in sub-Saharan Africa

CVD deaths in sub-Saharan Africa occur at younger ages on average than in the rest of the world(Moran et al., 2013). The most common underlying cause of heart disease in high-income countries is coronary artery disease while in sub-Saharan Africa (SSA), the predominant causes have traditionally been ascribed to rheumatic heart disease, hypertensive heart disease and cardiomyopathy(Gallagher et al., 2018). In Nigeria, hypertension and rheumatic heart disease are high on the list of causes of heart failure(Gallagher et al., 2018; Onwuchekwa & Asekomeh, 2009).

Poor health system response due to lack of funding have led to the increasing risk of CVDs in sub-Saharan Africa (SSA)(Hamid et al., 2019; Ike & Onyema, 2020). Nigeria’s high focus on communicable diseases leaves little room for addressing this rapidly increasing public health burden(Udeh et al., 2020).

However, in the global south, there is a significant rise in the incidence of CVDs resulting from epidemiological transition, increased urbanization(Gallagher et al., 2018; Hamid et al., 2019), adoption of the Western diet and lifestyle(Ike & Onyema, 2020) and poor knowledge of risk factors for CVDs(Babatunde et al., 2020; Olayinka Akinwusi et al., 2013). This has resulted in an increased mortalities among the working populations(Babatunde et al., 2020).

Lifestyle factors such as alcohol consumption, smoking, and malnutrition(Gallagher et al., 2018; Ike & Onyema, 2020) may influence the burden of CVDs in the country, and these modifiable risk factors can be better managed to reduce CVD prevalence(Udeh et al., 2020). World Health Organization/International Society of Hypertension (WHO/ISH) has made a prediction charts which is cost-effective and assess the total cardiovascular risk through the integration of risk factors (age, sex, presence or absence of diabetes, smoking status, systolic blood pressure, total serum cholesterol)(Babatunde et al., 2020; Hamid et al., 2019).

The inability of people to describe the symptoms of the disease as most people in various communities are only aware of malaria to be a major cause of death has be problematic to the increase in death due to heart disease(Udeh et al., 2020). Other people have the belief that some types of disease like stroke and heart attack can be sent paranormally through occult incantations and charms from an enemy or an evil person (witch/wizard)(Udeh et al., 2020). Thus, exploring the knowledge, their risk factors, and predicting its presence from early symptoms may play an important role in the prevention and control of Death through CVDs in Nigeria. The purpose of this study is to develop a prediction system will few numbers of symptoms that will promptly predict the presence of CVD in a person.

## Predicting heart disease

## Data mining approach to heart diseases prediction

## Related Works

Research has been done in this field and people have produced methods to predict cardiovascular disease using supervised machine learning algorithms. Several research papers have been written on this topic. A short survey has been in which analyzes performance of various models based on machine learning algorithms and techniques is shown below:

In one of the works by (Prasad, 2020) , Bayesan network was used to predict heart disease. Bayesian networks apply Bayes’ Theorem (also known as Bayes’ rule or Bayes’ law). In Bayes’ theorem, a prior (unconditional) probability represents the likelihood that an input parameter will be in a particular state; the conditional probability calculates the likelihood of the state of a parameter given the states of input parameters affecting it; and the posterior probability is the likelihood that parameter will be in a particular state, given the input parameters, the conditional probabilities, and the rules governing how the probabilities combine. Bayesian networks (BN) have been used to build medical diagnostic systems as doesn't produce a range outcome but rather a probability for each potential predicted event.

In another paper by (Aniruddha Dutta, 2020) , Convolutional Neural Network (CNN) was used to make heart disease prediction. Convolutional Network is very useful in the field of image recognition. A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. Convolutional Neural Network (CNN) is known to provide better outcomes than other machine learning algorithms if tuned better and feed a good amount of data, it also poses some drawbacks ranging from being significantly slower due to an operation such as maxpool thereby making training process take a lot of time if the computer doesn’t consist of a good GPU, to the fact that it requires a large Dataset to process and train the neural network

In a work by (Mursal Furqan, 2020), algorithms like Logistic Regression Classifier, K-Nearest Neighbors Classifier, and Random Forest Classifier were individually used to predict heart disease with each resulted in its percentage accuracy. The accuracy gotten from all the models ranges from 72 -84 percent with logistic regression having the highest accuracy of 84%. Using method of running different algorithms is useful to know the performance of each model on a dataset for the major decision of which model prediction to be expected more accurate. Logistic regression which was the best performing model is one of the supervised Machine Learning algorithms used for classification i.e. to predict discrete valued outcome. Although logistic regression being **one of the simplest machine learning algorithms** and is easy to implement yet provides great training efficiency in some cases. It also possesses some drawbacks including model overfitting the training set on high dimensional datasets, it can best perform well when only important and relevant features are used to build the model otherwise the probabilistic predictions made by the model may be incorrect and the model's predictive value may degrade.

The table below shows comparative analysis with previous studies

## Literature Review

The table below shows some of the related literature and their weaknesses

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S/N** | **Title** | **Year** | **Author** | **Method Used** | **Evaluation Metrics** | **Weakness** |
|  | A fast algorithm for heart disease prediction using Bayesian network model | 2020 | Mistura Muibideen and Rajesh Prasad | Bayesian network (BN) | Accuracy, Precision, Recall, F1 Score | Other models apart from Bayesian models were not investigated |
|  | An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction | 2020 | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton | Convolutional Neural Network (CNN) | Confusion matrix  AUC  Recall  Specificity  Accuracy | The training process takes a lot of time on non-GPU systems. It also requires a large Dataset to process and train the **neural network** |
|  | Heart Disease Prediction using Machine Learning Algorithms | 2020 | Mursal Furqan, Hiba Rajput,Sanam Narejo | Logistic Regression Classifier, K-Nearest Neighbours  Classifier, and Random Forest Classifier | Accuracy | Other metrices were not used to ascertain the performance of the predictors. |
|  | A Novel Approach to the Diagnosis of Heart Disease using Machine Learning and Deep Neural Networks | 2020 | Sahithi Ankireddy | K-Nearest  Neighbors (KNN), Support Vector Machine (SVM), Random  Forest Classifier (RF), Naive-Bayes (NB) and Deep Neural  Network (DNN) | ROC-AUC  Accuracy  MCC | The performance of prediction was not improved |
|  | Improving the performance of heart disease prediction system using ensembling techniques | 2021 | Ekta Maini, and Bondu Venkateswarlu | Ensembling techniques (Naïve Bayes, SVM, Logistic Regression and and Multilayer Perceptron) | Accuracy | The merged ensemble method can be extremely **computationally expensive** due to Neural network used |
|  | Coronary Artery Disease Diagnosis; Ranking the Significant Features Using Random Trees Model | 2020 | Javad Hassannataj Joloudari , et al. | Random Trees, Decision Trees, support vector machine (SVM) | Accuracy, ROC curve, Gini, Gain and Confidence | Feature selection techniques could improve performance of prediction |
|  | Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means | 2021 | Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria | K-Means with PCA and  Hybrid Genetic Algorithm | Accuracy, Clustering Error, Recall, Precision,  F1 Score | Other feature selection techniques were not investigated |
|  | Heart Disease Prediction Using Hybrid Machine Learning Algorithms | 2020 | S.Raguvaran  R.Anandhi  A.Anbarasi  T.Megala | Logistic  Regression, Random Forests Classifier Algorithm, Neural network, KNN (K-Nearest Neighbour) | Accuracy  AUC  PR Curve | Feature selection techniques could improve performance of prediction |
|  | Heart disease prediction using machine learning techniques | 2021 | Apurv Garg, Bhartendu Sharma and Rijwan Khan | K-Nearest Neighbor (K-NN) and Random  Forest | Accuracy  Confusion matrix | Performance evaluation metrics eg AUC ROC could give more information about the model performances |
|  | Prediction System for Heart Disease Based on Ensemble  Classifiers | 2020 | Joshua Emakhu, Sujeet Shrestha and  Suzan Arslanturk | AdaBoost, Bagging, Random Forest, and Voting Ensemble (Decision Trees, Logistic Regression and Support Vector Machines) | Accuracy  Confusion matrix  Recall  Specificity | Feature reduction could improve performance and Performance evaluation metrics eg AUC ROC could give more information about the model performances |

**TECHNIQUES USED**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S/N** | **Author and Year** | **LR** | **KNN** | **NB** | **SVM** | **BN** | **CNN** | **MLP** | **RF** | **DNN** | **RT** | **DT** | **Ada Boost** | **Bagging** | **Boosting** | **K-Means** | **Accuracy** |
|  | Mistura Muibideen and Rajesh Prasad **-2020** | √ | √ | √ |  | √ |  |  |  |  |  |  |  |  |  |  | **85%** |
|  | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton **-2020** | √ |  |  | √ |  | √ | √ | √ |  |  |  |  |  |  |  | **85.7%** |
|  | Mursal Furqan, Hiba Rajput, Sanam Narejo **-2020** | √ | √ |  |  |  |  |  | √ |  |  |  |  |  |  |  | **87%** |
|  | Sahithi Ankireddy **-2020** |  | √ | √ | √ |  |  |  | √ | √ |  |  |  |  |  |  | **92%** |
|  | Ekta Maini, and Bondu Venkateswarlu **-2021** | √ |  | √ | √ |  |  | √ |  |  |  |  |  | √ | √ |  | **87.5%** |
|  | Javad Hassannataj Joloudari , et al. **-2020** |  |  |  | √ |  |  |  |  |  | √ | √ |  |  |  |  | **93.4%** |
|  | Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria  **-2021** |  |  |  |  |  |  |  |  |  |  |  |  |  |  | √ | **94.06 %** |
|  | Apurv Garg, Bhartendu Sharma & Rijwan Khan -**2021** |  | √ |  |  |  |  |  | √ |  |  |  |  |  |  |  | **86.9%** |
|  | Joshua Emakhu and Sujeet Shrestha **-2020** | √ |  |  | √ |  |  |  | √ |  |  | √ | √ | √ |  |  | **87.04%** |

**EVALUATION METRICES USED**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S/N** | **Author and Year** | **Accuracy** | **Precision** | **Recall** | **F1- Score** | **AUC/ROC** | **Gini** | **Gain** | **Confidence** | **Cluster Error** | **Probability Mean Error** | **MCC** | **Kappa Statistic** | **Specificity** |
|  | Mistura Muibideen and Rajesh Prasad **-2020** | √ | √ | √ | √ |  |  |  |  |  |  |  |  |  |
|  | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton **-2020** |  |  |  |  | √ |  |  |  |  |  |  |  |  |
|  | Mursal Furqan, Hiba Rajput, Sanam Narejo -**2020** | √ |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Sahithi Ankireddy **-2020** |  |  |  |  | √ |  |  |  |  |  |  |  |  |
|  | Ekta Maini, and Bondu Venkateswarlu **-2021** | √ |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Javad Hassannataj Joloudari , et al **-2020** | √ |  |  |  | √ | √ | √ | √ |  |  |  |  |  |
|  | Md. Touhidul Islam, Sanjida Reza Rafa, Md. Golam Kibria -**2021** | √ | √ | √ | √ |  |  |  |  | √ |  |  |  |  |
|  | Apurv Garg, Bhartendu Sharma & Rijwan Khan -**2021** | √ |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Joshua Emakhu and Sujeet Shrestha **-2020** | √ |  | √ |  |  |  |  |  |  |  |  |  | √ |

# CHAPTER THREE: SYSTEM ANALYSIS AND DESIGN

## Introduction

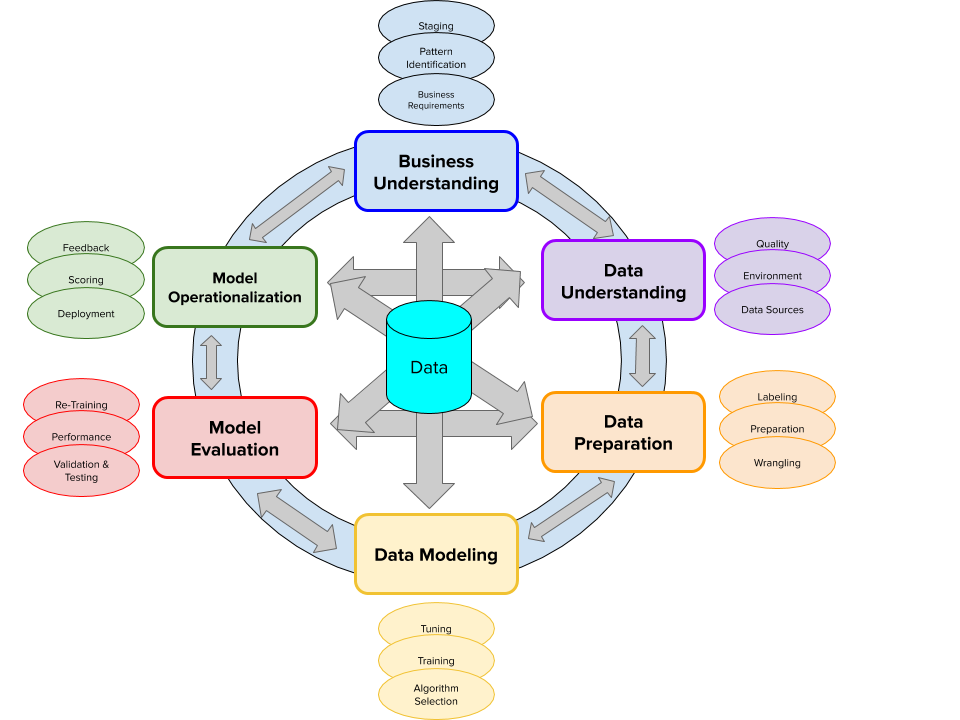
The study aimed to improve performance of heart disease prediction through feature reduction.

The first step was to design the heart disease performance prediction model and then improve the performance. To achieve this a data mining process called Cognitive Project Management for Artificial Intelligence (CPMAI)(Cognilytica, 2021) was followed to build the heart disease classifier models. These models were compared to determine that which is the best suited for the type of data used in the present study.

In addition to finding the best model, finding the optimal feature subset was also considered. In this chapter, five phases of the six CRISP-DM phases are discussed: (i)

1. business understanding; (ii)
2. data understanding; (iii)
3. data preparation; (iv)
4. data modeling
5. Model evaluation and
6. Model Operationalization

The chapter discusses how data was successfully used to build the classifier models. The chapter concludes with a discussion of the metrics that were used in the evaluation of the performance of the classifier models and, further, points to the sixth phase, indicating that it would make use of the discovered knowledge to design and implement the mobile academic performance prediction system.



In order to accurately, easily and effectively make prediction for cardiovascular disease, the research plan to the proposed method is developing an effective model to predicting heart disease. To achieve this, the following has to be considered:

1. The dataset to be used
2. Data Retrieval
3. Data preprocessing
4. Feature Selection
5. Target Class Transformation and
6. the model to be used for the study.
7. **Dataset**

This research is proposed to use the heart disease dataset from the UCI machine learning repository called Cleveland Heart Disease Data set. Cleveland Heart Disease Dataset is a publicly available supervised dataset provided by the Cleveland Clinic Foundation was used for the ML model. This data set contains 14 total attributes of patient medical information for 303 patients. The figure below shows the chosen attributes, and its information. These attributes have been selected by other researchers and healthcare professionals because they are known to be the best determining factors of heart disease.

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Description** | **Type of Attribute** | **Attribute Value Range** |
| Age | Age in years | Numeric | 29 to 77 |
| Sex | Gender | Nominal | 0 = female,  1 = male |
| cp | Chest pain type | Nominal | 1 = typical angina,  2 = atypical angina,  3 = non-angina pain,  4 = asymptomatic |
| trestbps | Resting blood pressure in mm Hg on admission to the hospital | Numeric | 94 to 200 |
| chol | Serum cholesterol in mg/dL | Numeric | 126 to 564 |
| fbs | Fasting blood sugar > 120 mg/dL | Nominal | 0 = false,  1 = true |
| restecg | Resting electrocardiographic results | Nominal | 0 = normal,  1 = ST-T wave abnormality,  2 = definite left ventricular hypertrophy by Estes’ criteria |
| thalach | Maximum heart rate achieved | Numeric | 71 to 202 |
| exang | Exercise induces angina | Nominal | 0 = no  1 = yes |
| oldpeak | ST depression induced by exercise relative to rest | Numeric | 0 to 6.2 |
| slope | The slope of the peak exercise ST segment | Nominal | 1 = upsloping,  2 = flat,  3 = down sloping |
| ca | Number of major vessels colored by fluoroscopy | Nominal | 0–3 |
| thal | The heart status | Nominal | 3 = normal,  6 = fixed defect,  7 = reversible defect |
| num | Prediction attribute | Nominal | 0= Unlikely to obtain heart disease  1 = Likely to obtain heart disease |

1. **Data Retrieval**

Data retrieval is usually the first step. Data can be gotten from various sources. It can be as easy as someone handing over a file on a drive for you to analyze them directly. Or you need to download it. Our dataset is to be downloaded from the UCI machine repository.

1. **Data Preprocessing**

Data preprocessing is also known as cleaning data. It is one of the most important steps to achieve the best from the dataset. This is a process whereby data inconsistencies such as missing values, out of range values, unformatted data, and noise are removed from the data. The process is usually time-consuming because it involves a lot of experimentation trying out various data analysis tools. Our preprocessing would involve data retrieval, handling missing values, target class transformation and data discretization

Handling Missing Values

Missing data values is a common problem faced by analysts. This occurs due to different reasons such as incomplete extraction, corrupt data, failure to load the information, etc. This is a great challenge that must be fixed because good models are generated when you make the right decisions on how to fix it. These are 5 ways of handling missing data:

* 1. Deleting Rows
  2. Replacing with mean/median/mode
  3. Assigning a unique category
  4. Predicting the missing values
  5. Using algorithms which supports missing values

We would adopt the method of handling missing values that proves best.

1. **Feature Selection**

In order to select the best features for our model. Recursive feature elimination is proposed. RFE is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

1. **Target Class Transformation**

As stated in the data set description, the target class contains values (0, 1, 2, 3, 4). Where 0 means healthy (no heart disease) and (1, 2, 3, 4) means the presence of sickness of varying degrees. Interest is in the absence or presence of heart disease, so the need to limit the class to (0, 1). Level (1, 2, 3, 4) was converted to 1.

1. **Proposed Model**

The proposed models comprise of Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. This is because the models proved to have high performance on predicting heart disease from previous studies.

1. **Performance Metrics**

Performance metrics are used to evaluate how different algorithms perform based on various criteria such as accuracy, precision, recall, etc. They are discussed below.

Accuracy

Accuracy is the ratio of the number of correctly classified instances to all the cases. It is the sum of TP and TN divided by the total number of instances. Accuracy = (TP + TN) / (TP+TN+FP+FN)

Precision

Precision is the proportion of true positive instances that are classified as positive. It shows how near the projected values are to each other. Precision = TP/(TP+FP)

Recall/ Sensitivity

The recall is the proportion of positive instances that are correctly classified as positive. Recall is known as sensitivity. Sensitivity = TP / (TP +FN)

F1 Score

F1 score combines both precision and recall and finds a balance between both. In other words, it computes the harmonic mean of precision and recall. F-measure = (2\*Precision + Recall) / (precision + Recall)

MCC

MCC known as Mathew correlation coefficient. Matthew’s correlation coefficient is a contingency matrix method of calculating the Pearson product-moment correlation coefficient between actual and predicted values. It ranges in the interval [ −1, +1], with extreme values –1 and +1 reached in case of perfect misclassification and perfect classification, respectively.

AUROC curve

It's a graph depicting the ratio of false positives to real positives. The area assesses discrimination, or the classifier's ability to accurately classify the test data.

Kappa Statistics

The kappa measure of agreement is the ratio K = P(A) - P(E) 1 - P(E)

Where P(A) denotes the percentage of times the k raters agree, i.e., the percentage agreement between the classifier and the ground truth, and P(E) is the proportion of times the k raters are expected to agree by chance alone i.e., the chance agreement. K=1 indicates perfect agreement and K=0 indicates chance agreement. The value greater than 0 means classifier is doing better. The classifier's result improves as the kappa statistic value rises.

# CHAPTER FOUR: RESULTS AND DISCUSSION

In total for machine learning classifiers, 303 records were used each with 9 total attributes. 8 out of the 9 features played a part in determining the last feature or the heart disease diagnosis. The evaluation involved different classification methods, i.e., Naïve Bayes, Bayesian Network, KNN, and Logistic Regression. In this study, WEKA was used to classify the Cleveland dataset. This is done using different proportions of the dataset for training and testing and the accuracy measure is taken as shown in Table 1. In Table 1, Splitting the training and testing data into 80:20 ration has the highest accuracy.

Table 1. Test proportion to determine the best ration for highest model accuracy (rounded to 4 digits)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ALGORITHMS** | **TRAINING/TESTING ACCURACY RATIO** | | | | | | | | |
| 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| Naïve Bayes | 82.4176 | 83.8843 | 85.3774 | 84.0659 | 86.0927 | 87.6033 | 86.8132 | 86.8852 | 86.6667 |
| Bayesian Network | 81.685 | 83.4711 | 86.3208 | 85.1648 | 87.4172 | 87.6033 | 86.8132 | 88.5246 | 86.6667 |
| KNN | 79.8535 | 79.3388 | 79.717 | 80.2198 | 82.1192 | 80.1653 | 82.4176 | 80.3279 | 86.6667 |
| Logistic Regression | 68.8645 | 65.2893 | 75.000 | 82.4176 | 86.0927 | 83.4711 | 81.3187 | 85.2459 | 83.3333 |

In determining which ML algorithm to use, accuracy scores were produced for each of the common machine learning algorithms. Figure 1 shows the accuracy scores of the K-Nearest Neighbors, Logistic regression Bayesian network, and Naïve Bayes algorithm Algorithms, and proves how the Bayesian Network had the highest accuracy score on 80:20 training/testing split.

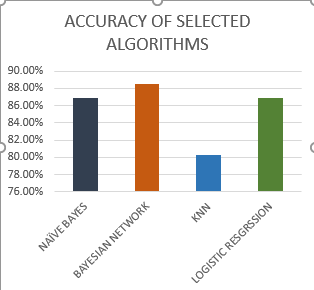


Figure 1: Accuracy Comparison of the models

ROC graphs were also established for the Bayesian Network (Fig 4.), Naïve Bayes (Fig 5), Logistic Regression (Fig 6), and KNN (Fig 7). Bayesian network and Naïve Bayes had the highest a score of approximately 92%. Logistic regression had a score of approximately 91% with KNN having a Score of 87.3%

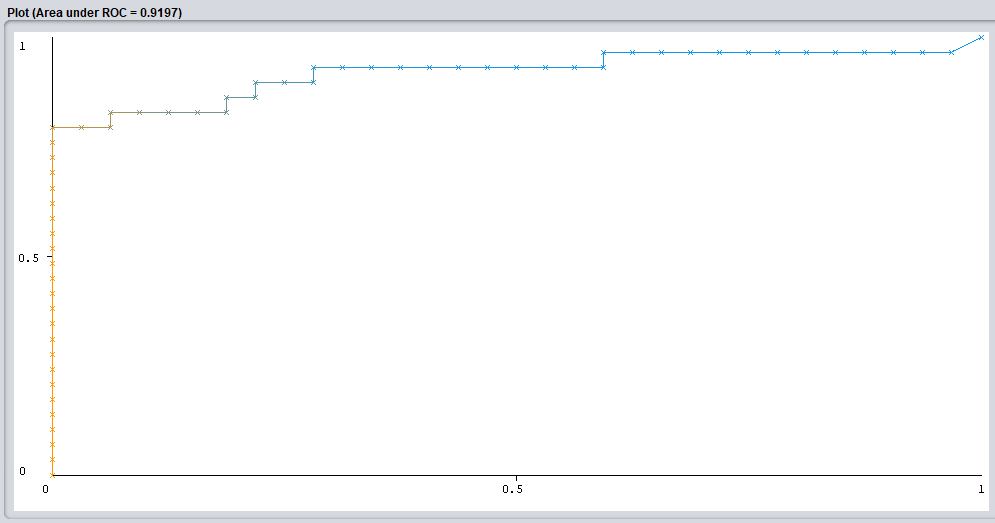
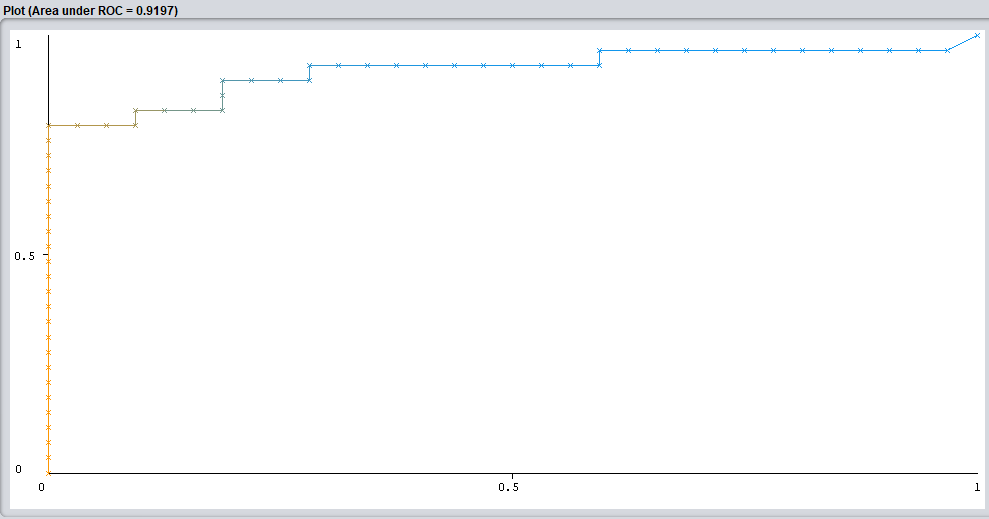
 

Figure 2: Bayesian Network Figure 3: Naive Bayes

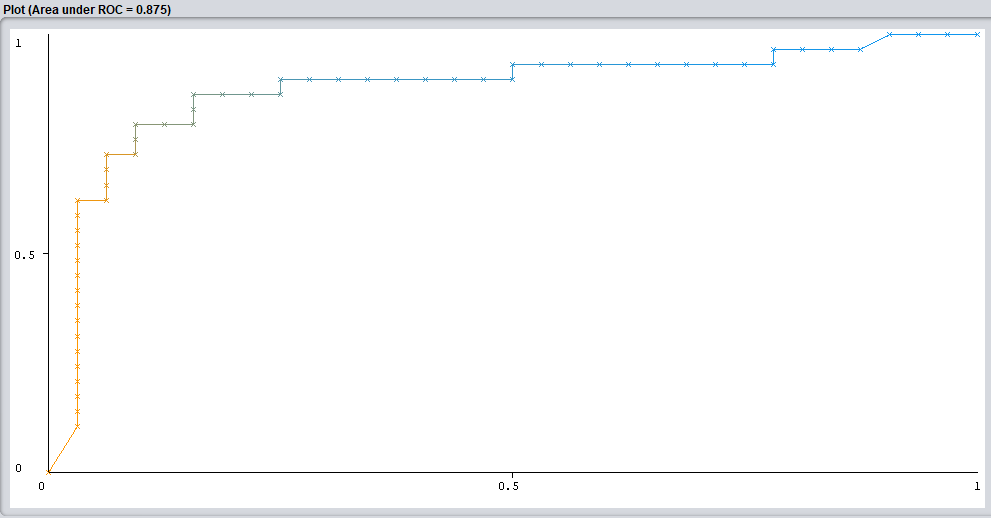
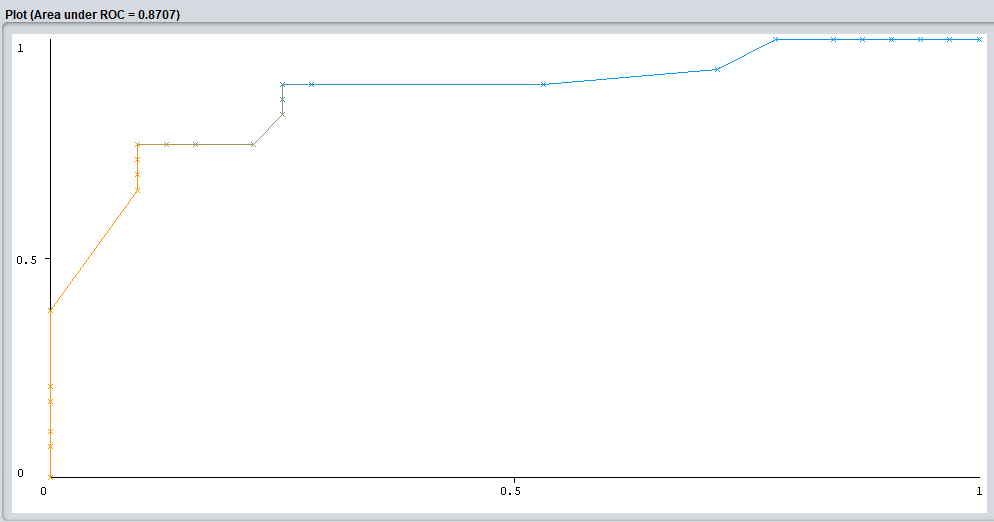
 

Figure 4: Logistic Regression Figure 5: KNN

Comparison of the models in relation to Matthews Correlation Coefficient (MCC) chart is shown in Figure 8 below. The Chart shows Bayesian network having the highest score of 77% followed by Naïve Bayes and Logistic regression with similar score of about 74%.KNN has the least MCC score of about 61%.

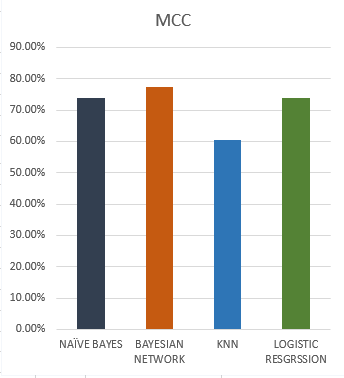


Figure 6: MCC chart of the Models

Table 2: Metrices Comparison between Naive Bayes, Bayesian Network, Logistic Regression and KNN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F-Measure** | **MCC** | **ROC Area** | **Kappa statistic** |
| **Naïve Bayes** | 86.8852 | 0.870 | 0.869 | 0.868 | 0.738 | 0.920 | 0.7362 |
| **Bayesian Network** | 88.5246 | 0.888 | 0.885 | 0.885 | 0.773 | 0.920 | 0.7688 |
| **Logistic Regression** | 86.8852 | 0.870 | 0.869 | 0.868 | 0.738 | 0.908 | 0.7362 |
| **KNN** | 80.3279 | 0.804 | 0.803 | 0.803 | 0.606 | 0.873 | 0.6043 |

Looking at the performance of the various models, each model had proved to have a good predictive power to heart disease However, looking at the MCC score and accuracy score Bayesian Network, despite having a similar AUC score with Naive Bayes was the overall winner. Bayesian Network proved to be the better overall model as seen by its much higher performance when looking at accuracy score and Matthews Correlation Coefficient. Thus, it can be deemed that it was the best algorithm out of the 4 tested.

Table 3: Comparison Various Approaches with our Proposed Approach

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S/N | Author and Year | Method | Dataset | Metrics |
|  | Mistura Muibideen & Rajesh Prasad, 2020 | Bayesian Network | Cleveland dataset: 14 Attributes | Accuracy: 85%  Precision: 86%  Recall: 85%  F1- Score: 85% |
|  | Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton **-**2020 | 2-layer CNN | NHANES dataset:  7 attributes | Accuracy: 81.78%  Recall: 77.3%  Specificity: 81.8 %  AUC: 76.78 % |
|  | Sahithi Ankireddy -2020 | Deep Neural  Network (DNN) | Cleveland dataset: 14 Attributes | Accuracy: 85.60% |
|  | Ekta Maini, and Bondu Venkateswarlu **-**2021(Maini & Venkateswarlu, 2021) | Ensembling techniques (Naïve Bayes, SVM, Logistic Regression and and Multilayer Perceptron) | Cleveland dataset: 14 Attributes | Accuracy: 87.5% |
|  | **Our proposed approach** | Bayesian Network with Wrapper subset evaluation (For feature selection) | Cleveland dataset: 8 Attributes. Namely: age, sex, cp, exang, oldpeak, slope, ca, thal | Accuracy: **88.53%**  Precision: **88.8** %  Recall: **88.5**%  F1- Score: **88.5** %  ROC Area**: 92.0%**  MCC: **77.3%**  Kappa statistic:**76.88%** |

# CHAPTER FIVE: SUMMARY, CONCLUTION AND RECOMMENDATION

Conclusion

It is worth researching much of what is required to forecast and diagnose any disease using machine learning effectively. Heart disease is considered one of the major threats to life and now it is a critical challenge to predict heart disease at an early stage in the area of clinical data analysis in order to minimize the death rate. Our method reduces the dimensionality of the dataset using WEKA wrapper method of data selection to select the best subset on Cleveland dataset features for better accuracy and efficiency predicting heart disease. The selected features are 8 in numbers and they include: age, sex, cp, exang, oldpeak, slope, ca, thal . The proposed method in the study has been evaluated with various metrics, and its performance results are compared with explores different machine learning algorithms.

A very detailed, useful, and highly preferable Machine Learning based model in this paper that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification window. The paper used Naïve Bayes, Bayesian Network, KNN, and Logistic Regression on the reduced features. The same features are used to both train and test the dataset. The outcome reveals that these data mining techniques can predict heart disease early with an accuracy of approximately 89%.

# REFERENCES

Adedoyin, R. A., & Adesoye, A. T. (2005). Incidence and pattern of cardiovascular disease in a Nigerian teaching hospital. *Tropical Doctor*, *35*(2), 104. https://doi.org/10.1258/0049475054037075

Almustafa, K. M. (2020). Prediction of heart disease and classifiers’ sensitivity analysis. *BMC Bioinformatics*, *21*(1), 1–18. https://doi.org/10.1186/s12859-020-03626-y

Babatunde, O. A., Olarewaju, S. O., Adeomi, A. A., Akande, J. O., Bashorun, A., Umeokonkwo, C. D., & Bamidele, J. O. (2020). 10-year risk for cardiovascular diseases using WHO prediction chart: Findings from the civil servants in South-western Nigeria. *BMC Cardiovascular Disorders*, *20*(1), 1–10. https://doi.org/10.1186/s12872-020-01438-9

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, *137*(1–2), 43–90. https://doi.org/10.1016/S0004-3702(02)00191-1

Cognilytica. (2021). *CPMAI Methodology - Cognilytica*. https://www.cognilytica.com/cpmai/

Dandona, L., Dandona, R., Kumar, G. A., Shukla, D. K., Paul, V. K., Balakrishnan, K., Prabhakaran, D., Tandon, N., Salvi, S., Dash, A. P., Nandakumar, A., Patel, V., Agarwal, S. K., Gupta, P. C., Dhaliwal, R. S., Mathur, P., Laxmaiah, A., Dhillon, P. K., Dey, S., … Swaminathan, S. (2017). Nations within a nation: variations in epidemiological transition across the states of India, 1990–2016 in the Global Burden of Disease Study. *The Lancet*, *390*(10111), 2437–2460. https://doi.org/10.1016/S0140-6736(17)32804-0

Emakhu, J., Shrestha, S., & Arslanturk, S. (2020). Prediction system for heart disease based on ensemble classifiers. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, *August*, 2337–2347.

Felman, A. (2019). *Cardiovascular disease: Types, symptoms, prevention, and causes*. Medicalnewstoday. https://www.medicalnewstoday.com/articles/257484

Gallagher, J., McDonald, K., Ledwidge, M., & Watson, C. J. (2018). Heart Failure in Sub-Saharan Africa. *Cardiac Failure Review*, *4*(1), 1. https://doi.org/10.15420/cfr.2018:4:1

Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity*, *2021*. https://doi.org/10.1155/2021/6663455

Hamid, S., Groot, W., & Pavlova, M. (2019). Trends in cardiovascular diseases and associated risks in sub-Saharan Africa: a review of the evidence for Ghana, Nigeria, South Africa, Sudan and Tanzania. *Aging Male*, *22*(3), 169–176. https://doi.org/10.1080/13685538.2019.1582621

Ike, S., & Onyema, C. (2020). Cardiovascular diseases in Nigeria: What has happened in the past 20 years? *Nigerian Journal of Cardiology*, *17*(1), 21. https://doi.org/10.4103/njc.njc\_33\_19

Jothi Prakash, V., & Karthikeyan, N. K. (2021). Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction. *Interdisciplinary Sciences: Computational Life Sciences*, *0123456789*. https://doi.org/10.1007/s12539-021-00430-x

Lidy, T., & Rauber, A. (2008). Machine Learning Techniques for Multimedia. *Machine Learning Techniques for Multimedia*, *11*, 249–285. https://doi.org/10.1007/978-3-540-75171-7

Maini, E., & Venkateswarlu, B. (2021). Improving the performance of heart disease prediction system using ensembling techniques. *AIP Conference Proceedings*, *2316*(February). https://doi.org/10.1063/5.0036478

McLachlan, S., Dube, K., Hitman, G. A., Fenton, N. E., & Kyrimi, E. (2020). Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, *107*(April), 101912. https://doi.org/10.1016/j.artmed.2020.101912

Moran, A., Forouzanfar, M., Sampson, U., Chugh, S., Feigin, V., & Mensah, G. (2013). The epidemiology of cardiovascular diseases in sub-saharan Africa: The global burden of diseases, injuries and risk factors 2010 study. *Progress in Cardiovascular Diseases*, *56*(3), 234–239. https://doi.org/10.1016/j.pcad.2013.09.019

Muibideen, M., & Prasad, R. (2020). *A Fast Algorithm for Heart Disease Prediction using Bayesian Network Model*. 1–11. http://arxiv.org/abs/2012.09429

Nason, E. (2007). *An overview of cardiovascular disease and research*.

Olayinka Akinwusi, P., Olarewaju Peter, J., Tolulope Oyedeji, A., & Oluwadamilola Odeyemi, A. (2013). The new face of rheumatic heart disease in South West Nigeria. *International Journal of General Medicine*, *6*, 375–381. https://doi.org/10.2147/IJGM.S44289

Onwuchekwa, A. C., & Asekomeh, G. E. (2009). Pattern of heart failure in a Nigerian teaching hospital. *Vascular Health and Risk Management*, *5*, 745–750. https://doi.org/10.2147/vhrm.s6804

Sai Shekhar, M., Mani Chand, Y., & Mary Gladence, L. (2021). Heart Disease Prediction Using Machine Learning. *Lecture Notes in Electrical Engineering*, *708*(Inccst 20), 603–609. https://doi.org/10.1007/978-981-15-8685-9\_63

Shekhar, A. (2018). *What Is Feature Engineering for Machine Learning?* Medium. https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a

Steinbaum, S. R. (2019). *Cardiovascular (Heart) Diseases: Types and Treatments*. https://www.webmd.com/heart-disease/guide/diseases-cardiovascular

Udeh, K., Nwakasi, C., & Fulton, J. (2020). ‘I think one of the ways they will help is to create awareness’: primary school teachers’ perceptions of cardiovascular diseases in Nigeria. *Global Health Promotion*, *27*(4), 24–31. https://doi.org/10.1177/1757975919893981

Usha Sri, B. (2020). Effective Heart Disease Prediction Model Through Voting Technique. *International Journal of Engineering Technology and Management Sciences*, *4*(5), 10–13. https://doi.org/10.46647/ijetms.2020.v04i05.003

WHO. (2019). *WHO and Nigerian Government move to curb cardiovascular diseases | WHO | Regional Office for Africa*. https://www.afro.who.int/news/who-and-nigerian-government-move-curb-cardiovascular-diseases