
Chapter 12

A KNN and ANN model for predicting heart diseases

*Sulaiman Olaniyi Abdulsalam¹,
Micheal Olaolu Arowolo^{2,3}, Enobong Chidera Udofot²,
Ayodeji Matthew Sanni¹, Damilola David Popoola¹ and
Marion Olubunmi Adebisi^{2,4,5}*

Abstract

The heart is the single most important organ in the human body. Patients, professions, and medical systems are all bearing the brunt of heart failure's devastating effects on contemporary society. Since cardiac arrest may well be demonstrated as a better understanding or conceivably go unobserved, particularly in the vast population of clients that have other cardiovascular disorders, the true prevalence of heart failure is likely to be underestimated, accounting for only 1–4% of all hospitalized patients as test procedures in developed nations. A person with heart failure has a heart that is unable to circulate sufficient blood through the body, but the term “heart failure” does not explain why this happens. The clinical picture is confusing since there are several possible causes of heart problems, many of which are diseases in and of themselves. Many cases of heart failure can be avoided if the underlying medical conditions that cause them are identified and treated promptly. The study and prediction of cardiac conditions must be precise because numerous diseases have been connected to the cardiovascular system. The resolution of this problem requires intensive online research on the relevant topic. Since incorrect illness prognoses are a leading cause of death among heart patients, learning more about effective prediction algorithms is crucial.

This research utilizes K-nearest neighbor (KNN) and artificial neural network (ANN) to assess cardiovascular diseases using data collected from Kaggle. The highest accuracy (96%) was achieved by ANN trained with the standard scalar. Medical experts, specialists, and academics can all benefit greatly from this study. Based on the results of this study, cardiologists will be able to make more knowledgeable decisions about the inhibition, analysis, and handling of heart disease.

¹Department of Computer Science, Kwara State University, Nigeria

²Department of Computer Science, Landmark University, Nigeria

³Department of Electrical Engineering and Computer Science, University of Missouri Columbia, USA

⁴SDG 3 (Good Health and Well-being), Landmark University, Nigeria

⁵Covenant Applied Informatics and Communications-Africa Centre of Excellence (CAPIC-ACE), Covenant University, Nigeria

Keywords: Heart; Cardio; Disease; Machine learning; Prediction; KNN; ANN; CNN

12.1 Introduction

As the largest cause of death worldwide, cardiovascular disease (CVD) is a critical health concern no matter its form. CVD is not just one disease but a collection of disorders that affect the cardiovascular system (the heart and blood vessels). Diseases affecting the cardiovascular system and the central nervous system rank the highest in prevalence. The majority of persons who will acquire a kind of CVD may already have symptoms by the age of 35, according to a famous cardiologist [1]. These diseases typically afflict persons in the future (with prevalence drastically cumulative around the 30–44 age range).

When plaque shapes up in the veins and blocks plasma movement to the heart, this is called a heart attack. Excessive cholesterol, smoking, high blood pressure, extreme alcohol consumption, high blood sugar, insufficient physical activity, and a hypertensive heart are all risk factors for developing CVD [2].

Death from cardiovascular causes has surpassed all others in recent decades, both in high-income nations and in poorer ones. The mortality rate can be lowered through the early diagnosis of heart disorders and the constant monitoring of patients by medical professionals. Unfortunately, due to the additional intelligence, time, and expertise required, 24-hour medical consultation and the precise diagnosis of heart disorders are not currently options for patients [3].

To improve the accuracy with which heart illnesses may be predicted, this research makes use of machine learning methods. Essential features are selected using a standard scaler technique.

Since cardiac issues have such a high death rate, many lives can be spared if they are detected and treated promptly by utilizing the tools developed for heart ailment forecast. Data mining and machine learning are tools for sifting through mountains of information and distilling it into usable insights. The term “data mining” is used to describe the process of extracting non-trivial meaning from datasets by uncovering latent, previously unknown, and potentially useful information. Massive amounts of healthcare data are collected by the healthcare business every year, and these data can be mined for insights using a variety of data mining methods. This information can subsequently be put to use in a diversity of fields, in the medical, marketing, and financial sectors. In the medical field, treating cardiovascular illness is a crucial but challenging endeavor that calls for haste, efficiency, and the right kind of automation. Heart disease prediction systems have been examined, and it has been shown that varying amounts of medical characteristics and risk variables have been applied using various data mining approaches.

Removing valuable data from huge amounts of information is essential in the healthcare sector. Predicting outcomes and obtaining a deeper understanding of medical data are two areas where data mining and machine learning are rapidly emerging as crucial professions. The World Health Organization (WHO) intellects

that 17 million persons yearly lose their lives due to CVD. Accurate predictions of cardiovascular illnesses can be made utilizing data mining techniques. The prediction can aid doctors in making more informed diagnoses by providing answers to difficult questions about heart disease [4].

Medical experts, specialists, and researchers can learn a great deal from this study. Because of this research, cardiologists will be able to make more knowledgeable results on the inhibition, analysis, and handling of heart disease. Using deep learning strategies including ANN, CNN, and KNN, this research aims to detect and forecast heart problems. To do this, researchers compiled data from a wide range of sources and applied predictive analytics to cut down on cases of heart disease. This research creates a standardized scalar approach to data pre-processing using ANN, CNN, and KNN for predicting cardiac disorders.

12.2 Overview of the literature

12.2.1 Heart diseases

Coronary artery disease (CAD) is a complex process that narrows the coronary arteries, leading to heart disease (HD). The three most frequent kinds of IHD are angina, myocardial infarction (MI), and heart failure (HF), all of which are connected but have distinct clinical manifestations.

Approximately 31% of all fatalities worldwide in 2012 were attributed to CVDs, conferring to the WHO. It is projected that 7.4 million of these demises were brought on by IHD. The incidence of IHD was identified as the main cause of mortality worldwide in 2012, accounting for 13.2% of all deaths. In 2012, IHD was the main cause of mortality in both middle- and high-income nations, accounting for around 45% of demises in middle-income nations and 37% in high-income republics, respectively. Heart disease is a foremost source of death and disability worldwide. HD exerts a gigantic economic drain on patients, in addition to reducing their health-related quality of life (HRQOL) and life expectancy. Costs associated with HD patients' medical care were \$851 million in 2004, placing them third most expensive out of seven major disorders. "A state of total psychological, physical, and societal welfare, and not only the absenteeism of ailment," as stated by the WHO. Changes in the incidence and severity of diseases should be included when gauging health and the effects of health treatment. Improved HRQOL can be used as a proxy for overall well-being in an evaluation.

Many different conditions affecting the heart are collectively referred to as "heart disease." The most prevalent kind of heart ailment in the United States is CAD, which limits the blood supply to the heart. Heart Disease and Stroke Figures in 2021 appraise a statement from the American Heart Association, states that a decrease in blood movement can lead to a heart attack [5].

Someone in the United States suffers a heart attack every 40 sec on an average. The average rate of death in the United States due to heart disease is higher than one person every 60 sec. There are an estimated 720,000 annual new heart attacks and 335,000 annual recurrences in the United States, according to the study. On an

average, a man will have his first heart attack at the age of 65.6 years old, while a woman will be 72.0 years old [6].

A human heart will beat 2.5 billion times during its lifespan, distributing millions of gallons of blood to all parts of the body. Oxygen, fuel, hormones, and other substances, along with a wide variety of cells, are carried by this constant flow. It also helps get rid of metabolic waste. When the heart stops beating, life support systems collapse almost instantly. Incredibly, the heart can stay beating for so long and for so many people, given how hard it works all the time. However, it can falter for a variety of reasons [7], including a poor diet and lack of exercise, smoking, illness, and unfortunate genes.

The development of atherosclerosis is a serious health concern. Atherosclerosis is the hardening and narrowing of arteries due to a buildup of cholesterol-rich plaque. The coronary blood vessel, which delivers plasma to the heart, and other arteries in the body can be impeded by these plaque pockets. When a portion of the plaque breaks off, it can cause a heart attack or a stroke [8].

A stroke occurs when a clot in a cerebral artery grazes off body fluid supply to the brain. Heart disease occurs when the heart is incapable to pump sufficient blood to the body's tissues and organs. Initial symptoms can include a rapid heart rate, difficulty breathing, discomfort in the chest, abrupt confusion, nausea, swollen feet, and a cold sweat. Cardiovascular ailment is a broad term relating to the variability of situations disturbing the heart and blood containers, and it is quite common as individuals' age. However, it is not inevitable. When adopted at an early age, a healthy lifestyle greatly reduces the probability of emerging vascular ailment. High blood pressure and high cholesterol are both perilous features of CVD, but they can be prevented with healthy lifestyle choices and medication. Drugs, treatments, and devices can help keep a damaged heart functioning. Those who suffer from heart disease have an abnormality involving the cardiovascular system (such as coronary heart disease, arrhythmia, or heart-valve defect).

A heart valve ailment happens when the heart's valves do not function normally. Naturally, occurring heart valves are delicate and smooth structures. They control how the heart's blood is pumped from one chamber to the next. Furthermore, they stop blood from re-entering the heart from the atria.

CAD, strokes, temporary ischemic attacks (TIAs), peripheral artery ailment, and aortal ailment are the key types of cardiac ailment.

12.2.2 Machine learning

Algorithms developed for use on computers that can learn from their own experience and incorporate new information into their operation are the focus of machine learning research. They classify it under AI. The field of machine learning focuses on the study of algorithms that may be programmed to learn and improve on their user data and previous examples. They classify it under AI. In data science, it refers to a technique that uses programming to construct analytical models automatically. It is a subfield of AI that seeks to automate as much of the learning, pattern recognition, and decision-making processes as possible. Applications, where it would be

impractical or impossible to build traditional statistics which is focused on making predictions using computers, include e-mail filtering and computer vision, two areas where machine learning methods are widely employed. Results for huge datasets can be predicted using machine learning methods. Prediction systems can benefit greatly from the application of machine learning, an AI method.

Supervised, unsupervised, and reinforcement learnings are types of machine learning algorithms.

Supervised learning: Supervised learning requires labeled or known data for training. Learning is supervised or geared toward productive action because the data is already known. Once the data is prepared, the machine learning method helps train the model. Large amounts of data are needed to adequately grasp the patterns and enable effective prediction of test data. By comparing the training outputs to the actual ones and making adjustments to the algorithms based on the discrepancies, the method can be refined [9].

Supervised machine learning is the evolving algorithms that can use the information provided by an outside source to form general hypotheses and then predict new data. A compact model of the circulation of class labels in terms of predictor attributes is the goal of supervised learning. The resulting classifier is applied to testing instances where predictor feature values are identified but class label values are unidentified [10].

An algorithm is used to discover the association among a set of input (x) and output (Y) variables ($Y=f(X)$)

The objective is to get close enough to the mapping function so that new data may be used to reliably predict the output variables (Y) (X). Supervised learning is the procedure of teaching an algorithm how to learn from a given dataset; in this case, the training data set. As the algorithm iteratively processes the training data, it makes predictions that are then corrected by the instructor, who knows the right answers. The learning process concludes when the algorithm performs to the desired standard. Once the model has been trained on identified data, new results can be generated by feeding in unknown input [11].

Today, the most popular supervised learning algorithms include random forest, polynomial regression, K-nearest neighbors (kNNs), linear regression, decision tree, Naive Bayes, and logistic regression [12].

In classification, a supervised learning strategy is used to assign labels to new observations based on the labels assigned to the training data. The software learns from a collection of observations and then uses that knowledge to categorize new observations into one of many groups. Predicting a class label is an example of supervised learning. The detection of spam in electronic messages is a shining illustration of the success of categorization machine learning. There are primarily different types of categorization, and these are [13]; linear classification methods include, for instance, logistic regression and support vector machine (SVM).

Naive Bayes, decision tree, and random forest classification are all forms of non-linear classification [14].

Mathematical methods known as regression allow data scientists to make predictions about a continuous outcome (y) given the values of one or more

predictor variables (x). The reason why linear regression is so popular is that it can be easily implemented in the context of prediction and forecasting. Predicting a number label is an example of supervised learning. Linear, logistic, polynomial, support vector, decision tree, random forest, ridge, and Lasso Regressions are just a few of the many forms of regression that can be performed [15].

In unsupervised learning, the data used for training purposes has never been seen by a human being and is therefore unlabeled. The term “unsupervised” refers to the inability to guide the algorithm’s input with the benefit of prior knowledge. The machine learning algorithm is fed this information to train a model. For the machine learning method to work, the model must be trained using the information that will eventually be used as input [16]. A trained model is one that actively searches for patterns and then takes action based on what it finds.

In contrast to supervised methods, unsupervised machine learning can process data that has not been labeled. Since no human effort is needed to make the dataset machine-readable, significantly larger datasets can be processed by the program. Relational data modeling is a type of challenge wherein a model is used to either define or extract relationships within a dataset. Due to the lack of labels, it is unable to build anything transparent. Without any guidance from humans, the program can abstractly perceive relationships between data pieces. Partial least squares, singular value decomposition, Fuzzy means, K-means clustering, hierarchical clustering, a priori, and principal component analysis are currently utilized as unsupervised learning techniques [17].

How humans learn from data in their daily lives is a major inspiration for the field of reinforcement learning. It has a learning mechanism based on trial and error that can adapt to new situations. Outputs that are deemed unfavorable are “punished,” while those that are deemed desirable are bolstered. Through the application of the psychological concept of conditioning, reinforcement learning places the algorithm in a working setting with an interpreter and a reward system. Every time the algorithm completes an iteration, the result is sent to an interpreter who determines whether or not the result was useful. In this class of challenges, an agent is placed in its natural environment and tasked with figuring out how to best interact with it. Positive and negative reinforcement are two distinct types of reinforcement learning [18].

CNN: An example of an artificial neural network (ANN) is a CNN. To accomplish both generative and descriptive tasks, they rely on deep learning, making them a subfield of artificial intelligence (AI) [19].

CNN is a type of deep learning model used to evaluate data with a grid form, such as photographs. With inspiration from the visual brain of animals, CNN is programmed to robotically and adaptively learn longitudinal orders of qualities, beginning with basic features and progressing to more complex ones. The three categories of layers (or “structure blocks”) that make up a typical CNN are convolution, pooling, and fully linked [20].

ANN: Synonymous with the word “neural network,” an ANN is a mathematical model that is biologically inspired and consists of a collection of artificial neurons that are connected. The architecture has three layers: an input layer, a transition layer(s), and an output layer(s). For computation, it employs a connectionist model.

This method is a highly advanced analytical technique, capable of modeling exceedingly complex non-linear functions. Multilayer perceptron (MLP) is a famous ANN architecture (MLP). The input layer, the output layer, and the hidden layer are its three components. It takes data from the output layer and feeds it into the input layer. As a result, we may tell it to add as many hidden layers as we like to the model. The MLP is widely recognized as a highly effective function estimate for classification and forecasting issues. Given the right parameters, MLP is capable of efficiently learning non-linear functions of arbitrary complexity and precision. Nonlinear neurons (perceptrons) are the building blocks of the MLP, which consists of many layers of neurons coupled via a feed-forward architecture.

Interesting patterns and new insights can be mined from massive datasets using data mining techniques. For the resolution of investigation and decision-making, medical data mining techniques have been increasingly applied. In poor and middle-income countries, CVD was responsible for almost 80% of deaths. Big, unstructured data sets on heart disease are generated by the healthcare industry every year.

12.2.3 *Related work*

A CNNs model was proposed as part of the Cardio-Help system [21] to predict CVD in persons. For early HF prediction, the suggested method uses CNN and is concerned with temporal data modeling. The heart disease dataset was developed, and the results were encouraging when compared to cutting-edge approaches. Experiment results achieve better results in terms of evaluation metrics. There is a 97% success rate with the proposed method.

It was suggested [22] that researchers compile and report on results from analyses of different categorizations of learning models used for predicting cardiovascular disease. The review focuses on three main areas: CVD classification approaches. Performance measures, datasets, and tools used to report accuracy, prediction, and category of these approaches are also gathered and stated.

An innovative deep learning model using 1D CNN for classification between healthy and non-healthy individuals with a balanced dataset was suggested [23], which aims to overcome the restrictions of conventional machine learning methods. Multiple scientific parameters are utilized to determine a patient's risk profile, which is useful for making an early prognosis. To prevent overfitting, the proposed model employs several regularization procedures. The proposed model obtains over 97% training accuracy and over 96% test accuracy on the dataset. The effectiveness of the proposed model is demonstrated through in-depth comparisons to other machine learning techniques utilizing multiple performance metrics.

Prognosis prediction using recurrent neural networks (PPRNNs) [24] suggest the use of deep recurrent neural networks (RNNs) to predict high-risk prognosis from patient diagnostic histories, like that of language models. To learn from patient diagnostic code sequences and predict the existence of high-risk illnesses, the proposed PP-RNN uses several RNNs. Our findings also imply that our proposed technique has the potential to improve upon previous efforts.

It looked into several different methods for silent heart attack predicting [25], including machine learning algorithms and deep neural networks, but none of them yielded satisfactory results. Deep learning techniques, in particular RNN, are proposed as part of a heart attack prediction architecture for determining the severity of a patient's cardiovascular illness. Due to the findings of this study, the author has decided to implement RNN and GRU to enhance the system's ability to detect silent heart attacks and provide the user with timely notifications. This method has increased the accuracy of heart attack prediction to 92%, making it a reliable tool for predicting even silent heart attacks.

Most of the article [26] is devoted to determining which types of patients are more likely to acquire heart disease based on a variety of factors. To anticipate whether patients are diagnosed with heart diseases based on the patient's medical history. Using several machine learning procedures, including logistic regression and KNN, for classifying and predicting cardiac patients. A very useful method was applied to regulate how the model may be utilized to progress the accuracy of heart attack prediction in individuals. The proposed model's strength was seen in its ability to predict symptoms of heart disease in an individual, utilizing KNN and logistic regression, with respectable accuracy after comparing with other classifiers.

The diagnosis of CVD by the use of data mining techniques was presented [27], with only 14 parameters such as gender, age, BMI, down sloping, sugar, and fat being used as examples.

Data from the UCI machine learning for heart disease were analyzed and compared using a variety of machine and deep learning methods. The accuracy target was met with 94.2% [28].

Strong pre-processing models, such as feature selection and clustering with DBSCAN, were recommended in addition to HDPM via SMOTE and ENN algorithm [29]. The XGBoost algorithm was also utilized for cardiac illness prognosis. In some cases, the deployed model can achieve an accuracy of 99.4%.

The use of a deep neural network in a self-operating diagnostic model for the detection of cardiac problem disease [30]. The classification of healthy and unwell persons is greatly improved by the use of machine learning methods. A method for predicting a patient's risk profile based on aspects of their clinical data has been created in this study. The proposed model is built utilizing both deep neural networks and the 2-statistical model. Lack of a proper fit or an excessive one is no longer concern. The training and test data model performed exceptionally well. Using DNN and ANN, analyze the performance of a model that can accurately predict if a person has heart disease.

Based on a trained recurrent fuzzy neural networks (RFNNs), a genetic approach for cardiac illness diagnosis was proposed [31]. This research proposes utilizing genetically based trained RFNNs for cardiac disease diagnosis. The Cleveland Heart Disease dataset from the University of California, Irvine (UCI) is used for this investigation. Only 45 are used for the actual testing, while the remaining 252 are used for training. The outcomes presented an accuracy of the testing set was 98% with some other metrics that can be used to evaluate a test's efficacy. Upon inspection, it was decided that the outcomes were adequate.

A high-quality heart disease prediction model that uses the unsupervised K-means clustering method to identify outliers in healthcare data was proposed [32]. Most existing methods for anomaly detection focus on building profiles of typical occurrences. But these models need a large enough sample of typical people to be convincing. Our proposed model uses the Silhouette method to determine the best possible value for K. As a next step, it intends to identify outliers that are more than a predetermined distance from their respective clusters. Five of the most common categorization methods KNN, random forest, SVM, Naive Bayes, and logistic regression were adopted. The effectiveness of the proposed strategy is proved on a benchmark dataset for CVD.

Using a genetic algorithm for feature selection, researchers at the University of Toronto were able to create a heart disease prediction model based on 303 samples from the Cleveland dataset. Through this process, they were able to collect seven features that were then utilized to train models for heart prediction with four different machine learning strategies: SVM, MLP, jackknife (J48), and kernel (KNN). The framework was further assessed by comparing its findings to those of models built using the traditional feature selection techniques and by employing a 10-fold cross-validation procedure. When the SVM was used in conjunction with the genetic algorithm, it was shown to have an accuracy of 88.34%, up from 83.76% when using the original dataset alone.

An evaluation and categorization of machine learning procedures for the analysis of cardiac disease were proposed [33], using four different datasets from diverse locations. Several different classification methods, including KNN, decision tree NB, J48, SVM, JRip, AB, decision tree, and stochastic gradient descent classifiers, were applied to establish the classification and prediction of cases of heart disease. Results showed that using various classification algorithms for heart disease classification improves accuracy, with KNN ($N=1$), JRip, and Decision tree J48, all achieving above 98% accuracy in their respective classifications. Extraction of features using the classifiers subset evaluator enhanced the performance of KNN ($N=1$) and decision table classifiers. The feature selection technique for CVD prediction improved when only 4 out of 13 features were included.

There was a suggestion [34] to compare the efficacy of six data mining tools such as Weka, Orange, Knime, RapidMiner, MATLAB, and Scikit-Learn to classify heart disease with the Cleveland data comprising 297 annotations and 13 features. In total, there are 164 healthy individuals and 139 individuals with CVD in the dataset used for this study. Using these three performance indicators, a comparison of the accuracy, sensitivity, and specificity of the procedures available in each instrument was evaluated. Based on the results, it was clear that Matlab's Artificial Neural Network model was the most efficient approach. After generating a Matlab Receiver Operating Characteristic Curve, we made many suggestions for the most appropriate tool to utilize according to the customers' level of experience with data mining.

Changing mislaid values with the mean through pre-processing was presented as a method to determine cardiac disease presence or absence [35] using Naïve Bayes, SVM (linear and radial basis function), and KNN classifiers. A better preprocessing

was used to increase the precision of predictions about cardiovascular illness. It helps medical professionals establish if a patient has cardiovascular disease and, if so, what stage of the condition they may be in. To prove the dependability of the outcomes, several machine learning algorithms were tested utilizing accuracy, precision, f1-score, and recall as performance measures. This research analyzed machine learning algorithms based on a variety of performance metrics to improve their accuracy. In the preprocessing phase, where data is missing, the average is substituted. The results show how well the mean work as a stand-in for missing variables. Using SVMs and a linear kernel, 86.8% accuracy in scoring was attained.

In the medical field, data science is used to foresee cardiac issues [36]. Numerous studies have been done on the topic, but more research is needed to improve forecasting precision. This study examines numerous heart disease datasets to analyze experiments and show how accuracy can be improved. Algorithms and techniques for selecting features are then discussed. The use of the Rapid Miner instrument allows for the use of the decision tree, logistic regression, SVM, Naive Bayes, and random forest algorithms for feature selection and enhancement. Accuracy for several models such as the decision tree, random forest, logistic regression, Naive Bayes, and the logistic regression support vector machine is 82.22%, 82.56%, 84.17%, and 84.85%, respectively. Naive Bayes and logistic regression (SVM) are two methods utilized in this paper, and the results show that the accuracy of both methods has improved.

12.3 Materials and methods

Feature extraction and classification are two of the main methods used in this study’s dataset. In this research, we load data from an open source, process it with a feature extraction method (standard scalar), which extracts the necessary data from the dataset, and then classify it using deep learning algorithms (ANNs and Kernel Naive Bayes). Finally, the results are compared using ANN and KNN algorithms using performance evaluation. Figure 12.1 shows the proposed workflow.

Dataset was downloaded from the Kaggle repository (<https://www.kaggle.com/johnsmith88/heart-disease-dataset/version/2>) for use in the study. Developers and other technical specialists can use Kaggle, an online data science platform that hosts a wide range of crowdsourced datasets and frameworks, to collaborate and advance their work. Using the search term “heart diseases data set,” we were able to locate the dataset.

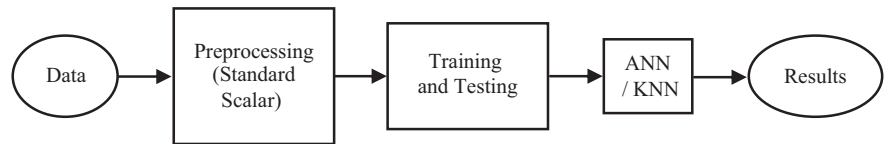


Figure 12.1 Workflow

The Cleveland, Hungary, Switzerland, and Long Beach V databases make up this 1988 bundle of information. It has 76 properties, including the anticipated attribute, but only 14 have been used in any of the published trials. If the patient has a heart illness, that fact will be included in the “target” field. The value ranges from 0 (no disease) to 1 (severe disease). Attributes include age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0, 1, 2), maximum heart rate achieved exercise-induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, the number of major vessels (0–3) colored by fluoroscopy, thal: 0 = normal. Recently, real patient information was scrubbed from the database and replaced with fictitious data.

The existence of the cardiac disease in the patient is what the “target” field is tracking. It takes on the form of an integer, with 0 denoting the absence of disease and 1 indicating the presence of illness. I plan to employ random forest, naive Bayes, and SVMs to evaluate this dataset of people with heart disease.

12.3.1 *Standard scalar*

To standardize the range of functionality of the input dataset, many machine learning models execute a standard scalar as a preprocessing step [37]. It standardizes values by converting variance to a fixed value of 1.0 and disregarding the mean. To calculate the standard score, z , for training sample x , we divide x by the mean, μ , of the training samples, which is zero if mean=False, and by the standard deviation, s , which is one if std=False.

Each feature undergoes its centering and scaling by estimating the appropriate statistics on the samples in the training set. When data is converted, the mean and standard deviation are retained for analysis.

Many machine learning estimators need that a dataset is standardized; otherwise, they may not perform as expected if the underlying features do not roughly conform to the assumptions of normal distribution (e.g. Gaussian with 0 mean and unit variance).

The RBF kernel of SVMs and the L1 and L2 regularizers of LMs are just illustrations of components used in the optimal solution of a learning algorithm that assume all features are symmetrical around 0 and have the same variance. For example, if one feature’s variance is hundreds of times bigger than the rest, it could overwhelm the objective function and prevent the estimator from learning as predicted from the other features. To preserve the sparsity structure of sparse CSR or CSC matrices, this scalar can be applied to them with the with mean=False argument.

12.3.2 *ANNs*

Computable algorithms known as ANNs have been shown to mimic the learning and problem-solving abilities of the human brain. The idea was to model the functioning of biological systems made up of “neurons.” The inspiration for ANNs comes from the structure and function of the central nervous systems of animals. The ability to learn new things and recognize patterns are two of its many talents [25].

For the sake of visualization, an ANN is best depicted as a directed graph with artificial neurons serving as the nodes and the weights. Directional edges with weights represent the connection between neuron outputs and neuron inputs. The ANN takes in data from the outside world in the form of a vector representing a pattern or an image. For each set of n inputs, a corresponding mathematical notation $x(n)$ is written [38].

Then, each input is multiplied by its weight (these weights are the details utilized by the ANNs to solve a specific problem). Commonly, the strength of the connections between neurons within the ANN is described by these weights. A compilation of all the weights used in the calculation is stored in the computer system [39].

A set of transfer functions, or the activation function, is applied to some input to produce a specific output. While there are a wide variety of activation functions, most can be placed into two main categories: linear and non-linear sets of operations. Some of the most common activation functions are the binary, linear, and tan hyperbolic sigmoidal sets.

12.3.3 *K-nearest neighbor*

K-nearest neighbor (kNN) is an abbreviation for this. This algorithm is used in supervised machine learning. Statements of classification and regression problems can both be solved using the procedure. With K, we signify the number of neighbors of a new unknown variable that needs to be predicted or categorized [40].

The kNN method is widely used since it does not rely on the data having a normal distribution or being homoscedastic. As a form of supervised machine learning, kNN can be used for classification and modeling in numerous forest mapping applications. It is also not too difficult to put together. To forecast the value of an unknown pixel, we use the weighted observations of k plots established from the training sample that is furthest from the projected pixel in the feature space [41,42].

12.3.4 *Performance metrics*

Researching this topic requires a computer with at least 8 GB of RAM and a 2.30 GHz processor, running Microsoft Windows 10 (64-bit edition) or a later version. For this study, we also used data from <https://www.kaggle.com/john-smith88/heart-disease-dataset/version/2>.

Accuracy: A model's accuracy is measured by how well it predicts future values for a target variable given a set of historical values for those variables (the "training data"). The accuracy of a classification algorithm is typically used as a benchmark for its performance. $\text{Accuracy} = ((\text{TP} + \text{TN})) / ((\text{FP} + \text{FN}))$.

The degree of specificity is defined as the percentage of actual negatives that matched the projection of a negative outcome. Which is true (or in the latter case), more predicted false negatives. A consequence of this is the potential for unintended "positives," or false positives. False positive rate is a word for this phenomenon. $\text{Specificity} = \text{TN} / ((\text{FP} + \text{TN}))$.

Specifically, precision is the fraction of correct predictions as a percentage of all correct guesses (i.e., the number of true positives plus the number of false positives). $\text{Precision} = \text{TP} / ((\text{TP} + \text{FP}))$.

Sensitivity is a measure of how effectively a model can predict the real positive results of the model for each available category. To put it another way, specificity measures how well a model can predict the true negative of each available category. We can calculate the sensitivity as $\text{TP} / (\text{TP} + \text{FN})$ [43].

12.4 Results and discussions

Discussion of the outcomes of the models' implementations and evaluations form the bulk of this chapter. The evaluation findings justify the study's completion because they show that the aims and objectives were met. The data was preprocessed using a regular scalar technique. In this part, we provide the findings from our investigation using the proposed model. By running it through a typical scalar preprocessor, we were able to normalize the data. This necessitated separating the information into a training set and a test set. In this study, we use deep learning techniques, including KNN and neural networks, to simulate the effects of cardiac disease.

For pandas to be able to read the dataset, it must be imported into the environment. `Df = PD.read CSV("/content/heart.csv")` is the code to use for accessing the data. The complete dataset is displayed in Figure 12.2.

Information is used to describe datasets in the software. The `describe()` function can be used to determine the mean, median, and standard deviation of a data frame's numerical values. In this data set, we estimated the number of occurrences, the average, the standard deviation, and the minimum percentile (25th, 50th, 75th). In machine learning, it is common practice to separate data into train and test sets.

[5] data.head(10)

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluoro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
5	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
6	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
7	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
8	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
9	63	0	4	150	407	0	2	154	0	4.0	2	3	7	Presence

Figure 12.2 Dataset

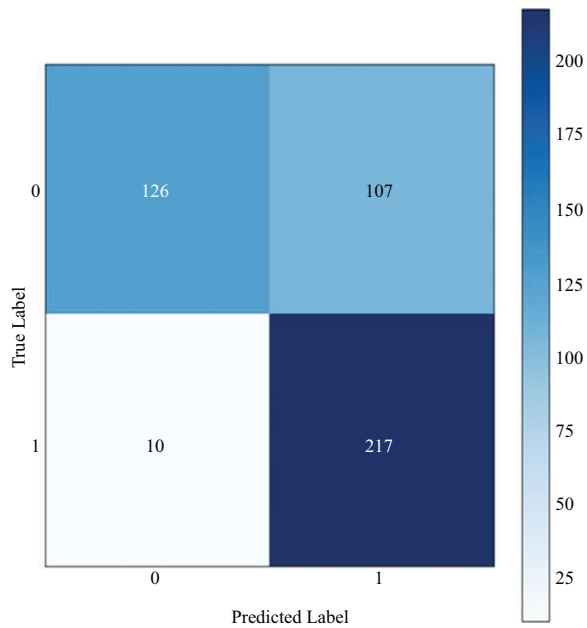


Figure 12.3 Confusion matrix for KNN model without standard scaler ($TP=126$; $TN=217$; $FP=107$; $FN=10$)

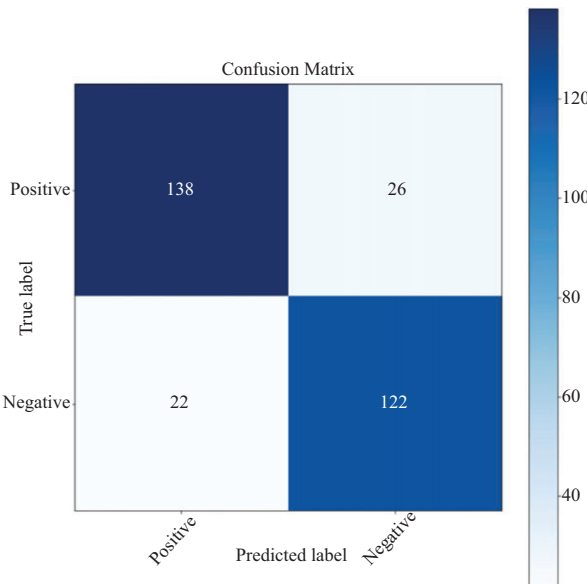


Figure 12.4 Confusion matrix for KNN with standard scaler model ($TP=138$; $TN=122$; $FP=26$; $FN=22$)

roc_auc_score for KNN: 0.8443428184281843

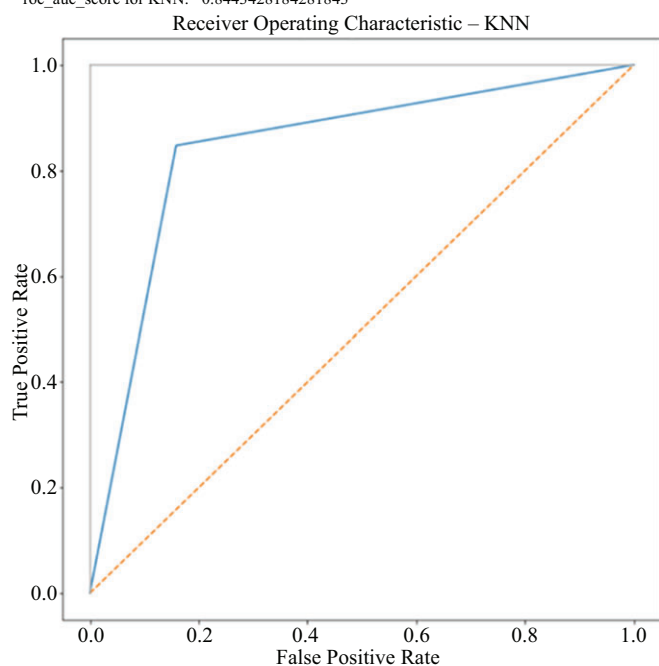


Figure 12.5 ROC curve for the KNN model

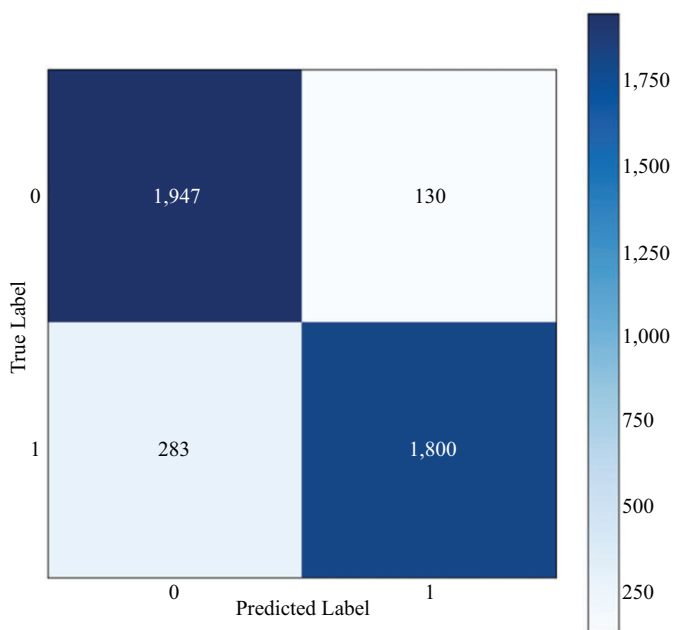


Figure 12.6 Confusion matrix for ANN model without standard scaler ($TP=157$; $TN=140$; $FP=4$; $FN=7$)

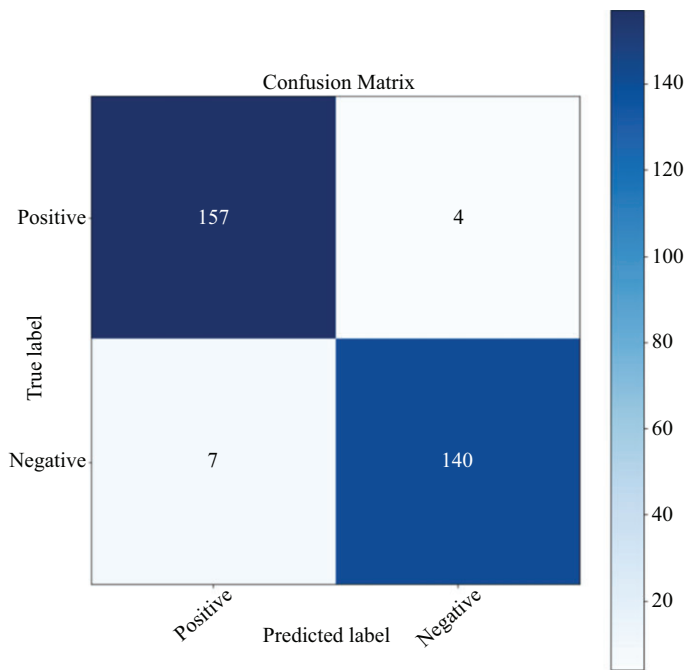


Figure 12.7 Confusion matrix for the ANN with standard scalar model (TP=157; TN=140; FP=4; FN=7)

It was necessary to split the data into training and testing sets for each method. Model fitting and testing were both performed using the training set as the foundation for comparison.

After data is partitioned into a training set and a testing set, the former is used to teach the latter. If you want to visually assess the efficacy of a classification model, plot the confusion matrix. Here you can see the KNN model's confusion matrix, roc curve, and accuracy. Figures 12.3–12.8 show the confusion matrix and ROC curve of the developed model.

KNN and ANN use a common scalar to classify the data in this analysis. Standard scaler is also given into the classifiers along with the data.

The resultant confusion matrices were evaluated using evaluation methods such as sensitivity, specificity, precision, accuracy, and F1 score. Metrics comparing each classifier's performance with and without the addition of supplementary features. Table 12.1 shows the performance evaluation of the developed models.

12.4.1 Comparison with previous work

This study included multiple experiments, the outcomes of which are displayed in Table 12.2. With an accuracy of 96%, ANN easily beat out the other models.

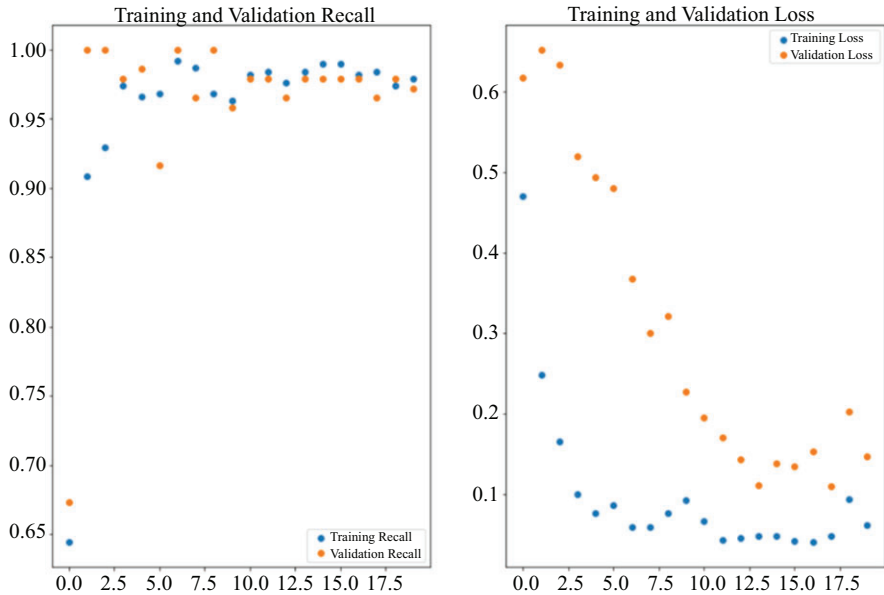


Figure 12.8 Scattered plot for the ANN model

Table 12.1 Performance metrics

Performance measures	KNN	ANN	ANN+SC	KNN+SC
Sensitivity	86.25	87.31	95.73	92.65
Specificity	82.43	93.26	97.22	66.98
Precision	84.15	93.74	97.52	54.08
Accuracy	84.42	90.07	96.43	74.57
F1 score	85.19	90.41	96.62	68.29

Table 12.2 Comparison of the findings

Reference	Technique	Results
[44]	SVM	83.7%
[25]	RNN	92%
[32]	SVM	88%
[36]	Random forest	82%

Results from different works are compared in Table 12.2.

Data mining models SC with ANN were shown to be the most accurate predictors of cardiovascular illnesses, as shown in Table 12.1. Existing works such as Nashif were used to inform the development of a heart disease prediction system that

achieved 95.8% accuracy in experiments. Standard scalar was used to find the missing values in the dataset, and those missing values were then replaced with the most suitable filter values. The results of previous research are shown in Table 12.2 and compared to the built model, demonstrating that a neural network has better accuracy than other data mining techniques. Heart disease is fatal because it can cause heart attacks and other potentially fatal complications. The suggested approach aims to reduce this by identifying the most effective model across the two datasets where gold-standard methods for cardiac disease prognosis have been tested.

12.5 Conclusions

Machine learning techniques such as KNN and ANN were used to diagnose heart conditions in this research. If you had like access to the dataset used for this study, you may get it here: <https://www.kaggle.com/johnsmith88/heart-disease-dataset/version/2>. With heart disease being the field's "goal," it is clear that this patient is suffering from a cardiac condition. If the value is 0, no disease is present; if it is 1, the disease is present. After running the data through a normal scalar, we fed it into KNN and ANN, and we got accuracy rates of 84%, 96%, and 88%, respectively, which we compared to those of comparable research.

The heart is an essential organ. Heart conditions necessitate higher precision and accuracy in diagnosis and analysis. More study is needed to determine whether or not they can be detected in real-time. Using a dataset of heart disorders, this work presents a reliable and early prediction of these conditions. Several machine learning algorithms are needed to implement the proposed methodology. Kaggle data is used in this endeavor. The model was constructed using an advanced learning algorithm, like ANN or KNN, and fed data using the standard scalar algorithm. The purpose of this study is to provide a summary of the methods and results used.

It is crucial to stress that the goal of the methods used in creating a model for predicting cardiac disease is to enhance it. The scope of this model might be expanded by using a more comprehensive dataset, which is why I propose doing just that in future research. Among these algorithms, the ANN algorithm has the highest accuracy, hence it should be used.

References

- [1] Chen Z and Xiu D. On generalized residual network for deep learning of unknown dynamical systems. *J Comput Phys.* 2021;438: 110362. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0021999121002576>
- [2] Buttar HS, Li T, and Ravi N. Prevention of cardiovascular diseases: role of exercise, dietary interventions, obesity and smoking cessation. *Exp Clin Cardiol.* 2005;10(4):229–49. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19641674>

- [3] Amini M, Zayeri F, and Salehi M. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. *BMC Public Health*. 2021;21(1):401. Available from: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-10429-0>
- [4] Shafique U, Majeed F, Qaiser H, and Mustafa IU. Data mining in healthcare for heart diseases. *Int J Innov Appl Stud*. 2016;10(January):1312.
- [5] Sakellarios A, Correia J, Kyriakidis S, *et al*. A cloud-based platform for the non-invasive management of coronary artery disease. *Enterp Inf Syst*. 2020;14(8):1102–23. Available from: <https://www.tandfonline.com/doi/full/10.1080/17517575.2020.1746975>
- [6] Adegoke O, Awolola NA, and Ajuluchukwu JN. Prevalence and pattern of cardiovascular-related causes of out-of-hospital deaths in Lagos, Nigeria. *Afr Health Sci*. 2018;18(4):942. Available from: <https://www.ajol.info/index.php/ahs/article/view/180231>
- [7] O'Callaghan J. How does the heart work? Wikimedia Commons. 2012. Available from: <http://www.howitworksdaily.com/whats-inside/inside>.
- [8] Dvorkin N, Clark MW, and Hamkalo BA. Ultrastructural localization of nucleic acid sequences in *Saccharomyces cerevisiae* nucleoli. *Chromosoma*. 1991;100(8):519–23. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1764970>
- [9] Nartey OT, Yang G, Wu J, and Asare SK. Semi-supervised learning for fine-grained classification with self-training. *IEEE Access*. 2020;8:2109–21. Available from: <https://ieeexplore.ieee.org/document/8943213/>
- [10] Kopper AE and Apelian D. Predicting quality of castings via supervised learning method. *Int J Met*. 2022;16(1):93–105. Available from: <https://link.springer.com/10.1007/s40962-021-00606-7>
- [11] Sakhrawi Z, Sellami A, and Bouassida N. Software enhancement effort prediction using machine-learning techniques: a systematic mapping study. *SN Comput Sci*. 2021;2(6):468. Available from: <https://link.springer.com/10.1007/s42979-021-00872-6>
- [12] Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, and Akinjobi J. Supervised machine learning algorithms: classification and comparison. *Int J Comput Trends Technol*. 2017;48(3):128–38. Available from: <http://www.ijctjournal.org/archives/ijctt-v48p126>
- [13] Arumugam K, Swathi Y, Sanchez DT, *et al*. Towards applicability of machine learning techniques in agriculture and energy sector. *Mater Today Proc*. 2022;51:2260–3. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2214785321074186>
- [14] Pranckevičius T and Marcinkevičius V. Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt J Mod Comput*. 2017;5(2). Available from: <https://doi.org/10.22364/bjmc.2017.5.2.05>

- [15] Weinberg GH and Schumaker JA. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. In: *Stat an Intuitive Approach*. Baltimore, MD: Brooks, 2009, pp. 227–245.
- [16] Bengio Y. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2013, pp. 1–37. Available from: https://doi.org/10.1007/978-3-642-39593-2_1
- [17] Adadi A. A survey on data-efficient algorithms in big data era. *J Big Data*. 2021;8(1):24. Available from: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00419-9>
- [18] Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–30. Available from: <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.001593>
- [19] Yamashita R, Nishio M, Do RKG, and Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9(4):611–29. Available from: <https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9>
- [20] Lindsay GW. Convolutional neural networks as a model of the visual system: past, present, and future. *J Cogn Neurosci*. 2021;33(10):2017–31.
- [21] Mehmood A, Iqbal M, Mehmood Z, *et al*. Prediction of heart disease using deep convolutional neural networks. *Arab J Sci Eng*. 2021;46(4):3409–22. Available from: <http://link.springer.com/10.1007/s13369-020-05105-1>
- [22] Swathy M and Saruladha K. A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques. *ICT Express*. 2022;8(1):109–16. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2405959521001119>
- [23] Hussain S, Nanda SK, Barigidad S, Akhtar S, Suaib M, and Ray NK. Novel deep learning architecture for predicting heart disease using CNN. In: *2021 19th OITS International Conference on Information Technology (OCIT)*. New York, NY: IEEE, 2021, pp. 353–357. Available from: <https://ieeexplore.ieee.org/document/9719326/>
- [24] Pillai NSR and Bee KK. Prediction of heart disease using Rnn algorithm. *Int Res J Eng Technol*. 2019;06(03):4452–8.
- [25] Pandey AK, Basandrai AK, Basandrai D, *et al*. Field-relevant new sources of resistance to anthracnose caused by *Colletotrichum truncatum* in a mung-bean mini-core collection. *Plant Dis*. 2021;105(7):2001–10. Available from: <https://apsjournals.apsnet.org/doi/10.1094/PDIS-12-20-2722-RE>
- [26] Jindal H, Agrawal S, Khera R, Jain R, and Nagrath P. Heart disease prediction using machine learning algorithms. *IOP Conf Ser Mater Sci Eng*. 2021;1022(1):012072. Available from: <https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072>
- [27] Dhara Mehta NV. Comparative analysis of data mining classification techniques for heart disease prediction. *Int Res J Eng Technol*. 2018;5(12):206–10.

- [28] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, and Singh P. Prediction of heart disease using a combination of machine learning and deep learning. *Comput Intell Neurosci*. 2021;2021:1–11. Available from: <https://www.hindawi.com/journals/cin/2021/8387680/>
- [29] Alfian G, Syafrudin M, Fitriyani NL, *et al*. Deep neural network for predicting diabetic retinopathy from risk factors. *Mathematics*. 2020;8(9):1620. Available from: <https://www.mdpi.com/2227-7390/8/9/1620>
- [30] Ramprakash P, Sarumathi R, Mowriya R, and Nithyavishnupriya S. Heart disease prediction using deep neural network. In: *2020 International Conference on Inventive Computation Technologies (ICICT)*. New York, NY: IEEE, 2020, pp. 666–70. Available from: <https://ieeexplore.ieee.org/document/9112443/>
- [31] Uyar K and Ilhan A. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Comput Sci*. 2017;120:588–93.
- [32] Ripan RC, Sarker IH, Hasan Furhad M, Musfique Anwar M, and Hoque MM. An effective heart disease prediction model based on machine learning techniques. *Adv Intell Syst Comput*. 2021;1375 AIST:280–8.
- [33] Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinform*. 2020;21(1):278. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03626-y>
- [34] Tougui I, Jilbab A, and El Mhamdi J. Heart disease classification using data mining tools and machine learning techniques. *Health Technol (Berl)*. 2020;10(5):1137–44. Available from: <https://link.springer.com/10.1007/s12553-020-00438-1>
- [35] Louridi N, Amar M, and Ouahidi B.El. Identification of cardiovascular diseases using machine learning. In: *2019 7th Mediterranean Congress of Telecommunications (CMT)*. New York, NY: IEEE, 2019, pp. 1–6. Available from: <https://ieeexplore.ieee.org/document/8931411/>
- [36] Bashir S, Khan ZS, Hassan Khan F, Anjum A, and Bashir K. Improving heart disease prediction using feature selection approaches. In: *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. New York, NY: IEEE, 2019, pp. 619–23. Available from: <https://ieeexplore.ieee.org/document/8667106/>
- [37] Singh D and Singh B. Investigating the impact of data normalization on classification performance. *Appl Soft Comput*. 2020;97:105524. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1568494619302947>
- [38] Montesinos López OA, Montesinos López A, and Crossa J. Fundamentals of artificial neural networks and deep learning. In: *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer International Publishing, 2022, pp. 379–425. Available from: https://link.springer.com/10.1007/978-3-030-89010-0_10
- [39] Al-Hroob A, Imam AT, and Al-Heisa R. The use of artificial neural networks for extracting actions and actors from requirements document. *Inf*

- Softw Technol.* 2018;101:1–15. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0950584918300752>
- [40] Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, and Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput.* 2020;58(5):991–1002. Available from: <http://link.springer.com/10.1007/s11517-020-02132-w>
 - [41] McRoberts RE, Næsset E, and Gobakken T. Optimizing the k-nearest neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. *Remote Sens Environ.* 2015;163:13–22. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0034425715000851>
 - [42] Arowolo MO, Adebisi MO, Adebisi AA, and Olugbara O. Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. *J Big Data.* 2021;8(1):29. Available from: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00415-z>
 - [43] Afolayan JO, Adebisi MO, Arowolo MO, Chakraborty C, and Adebisi AA. Breast cancer detection using particle swarm optimization and decision tree machine learning technique. In: *Intelligent Healthcare*. Singapore: Springer Nature Singapore, 2022, pp. 61–83. Available from: https://link.springer.com/10.1007/978-981-16-8150-9_4
 - [44] Gokulnath CB and Shantharajah SP. An optimized feature selection based on genetic approach and support vector machine for heart disease. *Cluster Comput.* 2019;22(S6):14777–87. Available from: <http://link.springer.com/10.1007/s10586-018-2416-4>
 - [45] Bashir S, Qamar U, Khan FH, and Javed MY. MV5: a clinical decision support framework for heart disease prediction using majority vote based classifier ensemble. *Arab J Sci Eng.* 2014;39(11):7771–83. Available from: <http://link.springer.com/10.1007/s13369-014-1315-0>
 - [46] Tiwari V, Garg B, and Sharma UP. Significant impact of improved machine learning algorithm in the processes of large data sets. *Int J Sci Res Comput Sci Eng Inf Technol.* 2020;458–67. Available from: <https://doi.org/10.32628/cseit206133>
 - [47] Aldahiri A, Alrashed B, and Hussain W. Trends in using IoT with machine learning in health prediction system. *Forecasting.* 2021;3(1):181–206. Available from: <https://www.mdpi.com/2571-9394/3/1/12>