A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Applied Machine Learning (CT046-3-M-AML-22-RESIT)

Assignment

19th of August 2022

Prof. Dr. Mandava Rajeswari

Applying Machine Learning Techniques on Prediction of Heart Disease

MARZIEH KHALILI CHACHAKI-TP067265

APUMF2112AI

Table of contents

Abstract

Heart attack infarction is among the most prevalent causes of death. Research into the causes, prevention, treatment, and cures of diseases such as heart attacks and strokes are dependent on data mining, comparisons, and other forms of data analysis on enormous datasets. Big Data analytics are used in the corporate sector for controlling, comparing, and managing large amounts of information, may be utilised to forecast, treat, manage, and cure cardiac illness. Massive datasets are mined for information using methodologies and technologies such as data mining, data visualisation, and Hadoop. Using machine learning and deep learning, one may turn data into usable knowledge. There are a number of disadvantages to adopting these algorithms, including resource consumption and extensive data pre-processing, but the proposed study presents a low-cost and dependable way for accurately and consistently forecasting heart attacks. This article describes how to identify and categorise risk factors for cardiovascular disease. To identify heart disease, researchers use Decision tree, ANN, and KNN. Heart disease data is utilised to compute the proposed method. The best approach is then evaluated and taught. This shows that the model utilised in this research is the most accurate for predicting heart attacks.

Keywords__ Machine learning, Heart disease, Decision tree, ANN, KNN

1.0 Introduction

During a heart attack, inadequate blood flow causes damage to the heart muscle. Clinical diagnostics, such as electrocardiograms and blood tests, may be used to detect early indicators of heart illness and aid in the prevention of sudden cardiac death (Waqar et al., 2021). In electrocardiography, or ECG, the electrical activity of the heart is monitored throughout time. Electrocardiograms may identify electrical activity abnormalities in the heart (ECG). In blood tests, CK-MB (local CK-MB) is an indicator of an imminent attack. In recent years, troponin levels have been used to help in the early identification of heart attacks (Takci, 2018).

Heart disease refers to the set of cardiovascular disorders that affect a variety of heart and circulatory system components. This article focuses on the condition known as "Heart Attack" and the risk factors that lead to it. Cardiomyopathy and cardiovascular disease are the most frequent manifestations of cardiac sickness (Sekar & Rao, 2012). The phrase "cardiovascular disease" encompasses conditions that affect the heart, blood vessels, and blood flow inside the body. Cardiovascular disease is the major cause of death and disability in the United States

(CVD). Most of the time, it is hard to predict cardiovascular disease (CVD) prior to a heart attack, stroke, angina, or heart disease. Therefore, it is essential to monitor any signs of cardiovascular illness and to seek medical help when required (Raza, n.d.). According to the World Health Organization, heart attacks and strokes annually kill 17.5 million people (World Health Organization). The leading cause of mortality is cardiovascular illness, with heart attacks and strokes accounting for 80 percent of all fatalities (Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam, 2018). Cardiovascular disease accounts for over one-third of all fatalities worldwide according to KaanUyar (2017).

If cardiac irregularities and cardiovascular disease prediction algorithms could be recognised early and utilised by clinicians to develop treatments that reduce the mortality rate associated with coronary artery disease, many lives may be saved. Coronary artery constriction causes cardiovascular disease by reducing the heart's blood and oxygen flow (CHD). The most prevalent symptoms of coronary artery disease are heart attacks and angina pectoris, which are both medical terminology for chest discomfort (CHD). If a blood clot grows and obstructs a coronary artery, a heart attack may ensue. When the heart muscle's blood supply is inadequate, chest symptoms ensue. Cardiovascular illness manifests itself in a variety of ways. Hypertension, coronary artery disease, arrhythmias, stroke, and rheumatic fever/rheumatoid fever are among examples (Sekar & Rao, 2012).

Machine learning is an area of mathematics concerned with how computers make judgments based on data. It was established at the nexus of statistics and computer science, both of which are concerned with discovering relationships in data. This convergence of mathematics and computer science is driven by the computational difficulties inherent to creating statistical models from massive data sets, such as billions or trillions of data points(Deo, 2015).

Machine learning seeks to find patterns and connections in data that were previously unknown. These patterns are also used in the development of other prediction models. Numerous functional units in several domains have been mechanised by technological breakthroughs. Utilizing a huge number of electronic devices, the health care business generates voluminous, interconnected data on hospitals, patients, and diseases. Despite the potential significance of raw data, caution must be taken while handling it. (Waqar et al., 2021).

Due to the rise of automated healthcare networks containing vast quantities of medical data, it is also feasible to develop predictive prototypes for heart diseases (i.e. Massive Data in Electronic Health Record System). Data mining, also known as machine learning, is the process of extracting meaningful information from massive amounts of data by analysing them from several perspectives. Important information is extracted from previously unrecognised, implicit data (Patel, J., Upadhyay, P. and Patel, 2016).

It is quite rare for healthcare organisations to generate vast quantities of data on medical illnesses, treatments, and other elements of patients' lives. Data mining employs a variety of techniques to identify patterns and similarities in vast quantities of data. Based on two publicly available datasets, this study presents a machine learning technique for predicting cardiovascular disease. Using machine learning algorithms, on the other hand, may assist anticipate future performance based on historical data (Vanisree K, 2011).

The subsequent sections of this work are as follows: In part 2 the researcher mentioned literature review of existing studies. The dataset should be categorised in part 3 of this document. In part 4, the consequences of using the suggested method are explained and analysed. The last of the work does so well in summarising the research's findings.

1.1 Problem Statement

The results of this data analysis will be beneficial. Machine learning is the most effective approach to accomplishing the aforementioned objective. In the medical field, machine learning and data mining have ushered in a new computer age. Various data mining approaches may include diagnostic procedures for identifying cardiovascular disease. (Amin and partners, 2019) To successfully deploy machine learning models, feature engineering requires considerable effort (El-Amir, 2020). Using this data, we are attempting to answer the following questions:

How unbalanced data affect machine learning algorithms?

How is the effectiveness of machine learning evaluated?

1.2 Research objective

This work attempts to develop a flawless model for analysis by means of optimization and fine-tuning in order to accurately forecast heart attacks. The following are the anticipated results, as indicated in the research plan for this study:

Evaluate their performance using three prediction models (Analyst Neural Network, Decision tree, K-nearest neighbour).

Determine if previous heart attack prognostications were correct.

1.3 Scope of the research

This investigation used three separate algorithms. While attempting to improve our rate of precision, we ran across a number of algorithmic difficulties. Even while some algorithms were highly precise, we learned that they had drawbacks and were time-consuming to build in our pursuit of the utmost degree of accuracy. We did an experiment to see which technique produced the greatest number of outcomes in the shortest length of time. We were unable of achieving our objective using any accepted means. Consequently, we devise a strategy that is both successful and feasible.

2.0 Literature review

In several medical centre research projects, various data mining approaches and machine learning algorithms have been applied to construct illness prediction systems.

Miled et al. (Zina Ben M, 2020) analysed electronic medical data for comprehensive diagnoses, prescriptions, and medical notes. They used machine learning methods to distinguish between people with and without dementia. Three EMR datasets were used in the development of Random Forest algorithms, which yielded an accuracy of 95.6%. (77.43 percent). Kar et al. used an ECG signal to monitor the heart state of a patient (N. Kar, 2020). When analysing the ECG data, they used both continuous and discrete wavelet adjustments. Using time intervals, statistical indicators, and irregular heartbeats, the data was categorised. Using K-NN and DT-CWT features, they were able to achieve an accuracy of classification of (98.92 percent).

To estimate the network's hyperparameters, the researchers utilised a trial-and-error method to develop an MLP network based on the Western Australian patient dataset. The hidden layer used a rectified linear activation function, whereas the output layer utilised a sigmoid activation function. The MLP-based approach was the most sensitive, most accurate, and had the greatest sensitivity (48.42 percent) and specificity (70.01 percent) (Awan, S.E., Bennamoun, M., Sohel, F., Sanfilippo, F.M., Dwivedi, 2019).

Mati et al. employed a range of techniques to analyse the 303 cases and 76 features of the Cleveland dataset (Matic, V., 2017). It was essential to create two 30 item 2D arrays, one for training and the other for testing. Two output characteristics were used to map the training data to the target data with a 96.67 percent degree of accuracy.

In an effort to eliminate heart disease misdiagnoses in a medical dataset, MLP and SVM were deployed (Olaniyi et al., 2015). SVM obtained an overall accuracy of 87.5 percent, whereas

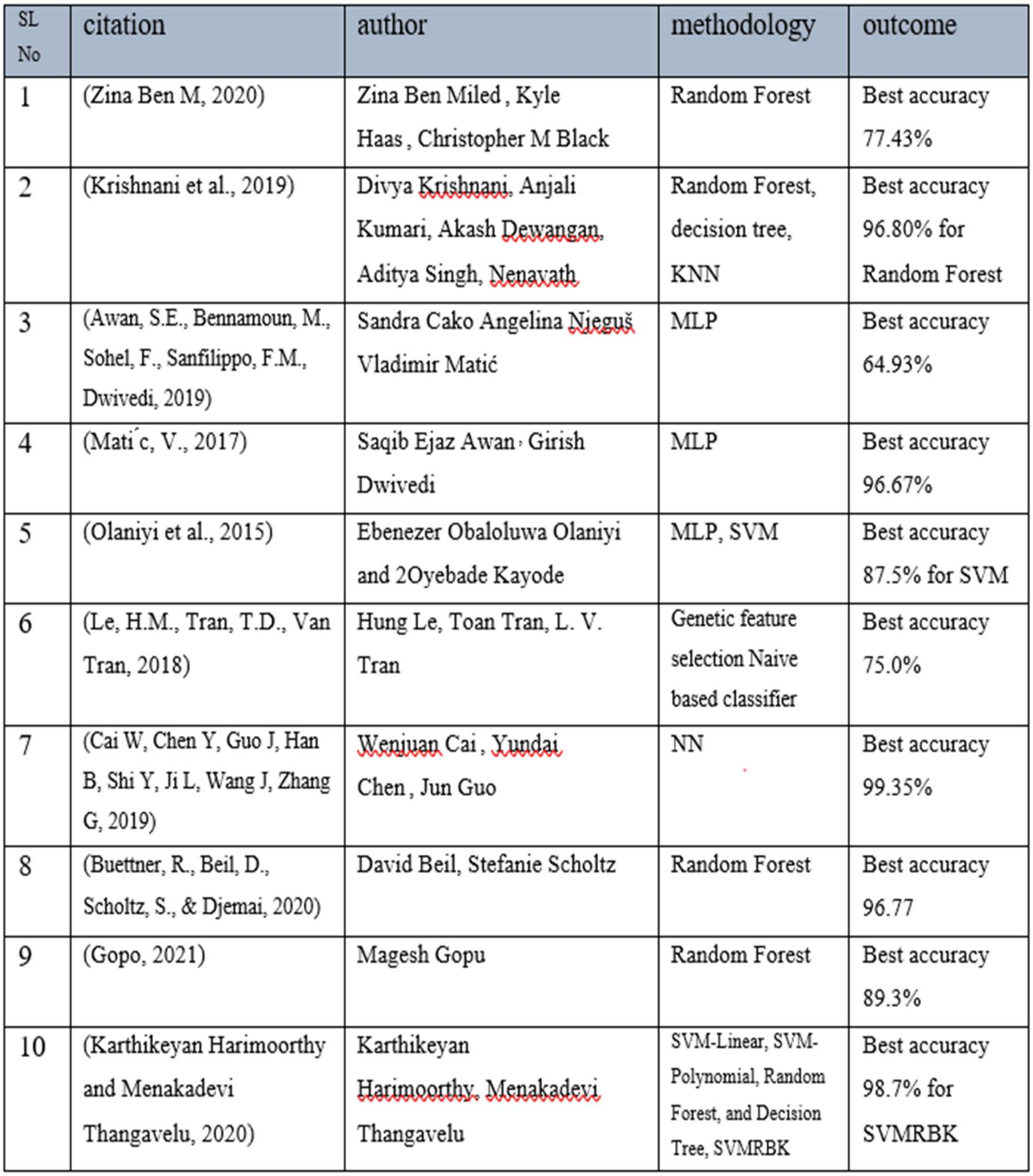
MLP earned an overall accuracy of 85.5 percent. They are accurate over seventy percent of the time. T.M. Le, T.D. Tran, and V. Tran (2018).

Cai et al. did researched on cardiac arrhythmias and 12-lead electrocardiograms (Cai W, Chen Y, Guo J, Han B, Shi Y, Ji L, Wang J, Zhang G, 2019) utilizing one-dimensional neural networks, ventricular fibrillation was identified. Author estimations place the test dataset's precision at 99.35, sensitivity at 99.19, and specificity at 99.44. Research by Buettner et al. evaluated the electroencephalograms (EEGs) of cardiac patients (Buettner, R., Beil, D., Scholtz, S., & Djemai, 2020). Classifiers based on machine learning were applied to explain the five granular divisions of EEG granularity. With a 96.77 percent accuracy rate, they were able to balance paranoid schizophrenics and non-schizophrenics using the Random Forest method.

Magesh and Swarnalatha investigated cardiovascular disease treatment facilities on behalf of the government. As a consequence of smoking, several risk factors and diseases associated with coronary artery disease were identified. Sample target level distributions were analysed using entropy, and patterns were discovered in the distributions. When using Random Forest to predict cardiac illness, cluster-based DT learning (89.30 percent accuracy) and non-cluster-based DT learning (76.70 percent accuracy) were discovered (Gopo, 2021).

Harimoorthy and his team(Karthikeyan Harimoorthy and Menakadevi Thangavelu, 2020), were able to improve the SVM Redial biaskernel by alleviating chronic renal failure-related symptoms. Compared to (SVM-Linear, SVM-Polynomial, RF, and DT), it was established that the accuracy of SVMRBK was (98.3 percent, (98.7%, and (89.9%).

Following table illustrates the use of a range of machine learning techniques in earlier research.

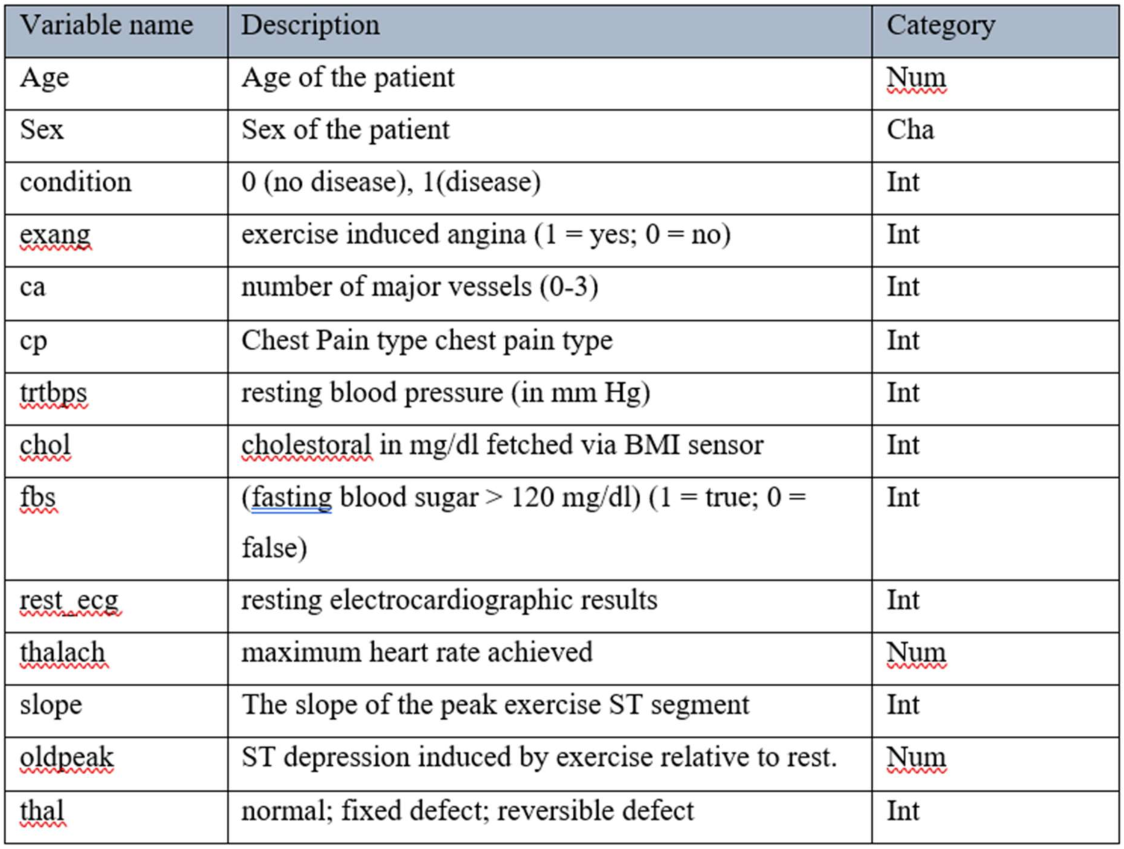


SL No	citation	author	methodology	outcome
1	(Zina Ben M, 2020)	Zina Ben Miled , Kyle Haas , Christopher M Black	Random Forest	Best accuracy 77.43%
2	(Krishnani et al., 2019)	Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath	Random Forest, decision tree, KNN	Best accuracy 96.80% for Random Forest
3	(Awan, S.E., Bennamoun, M., Sohel, F., Sanfilippo, F.M., Dwivedi, 2019)	Sandra Cako Angelina Njeguš Vladimir Matić	MLP	Best accuracy 64.93%
4	(Mati´c, V., 2017)	Saqib Ejaz Awan · Girish Dwivedi	MLP	Best accuracy 96.67%
5	(Olaniyi et al., 2015)	Ebenezer Obaloluwa Olaniyi and 2Oyebade Kayode	MLP, SVM	Best accuracy 87.5% for SVM
6	(Le, H.M., Tran, T.D., Van Tran, 2018)	Hung Le, Toan Tran, L. V. Tran	Genetic feature selection Naive based classifier	Best accuracy 75.0%
7	(Cai W, Chen Y, Guo J, Han B, Shi Y, Ji L, Wang J, Zhang G, 2019)	Wenjuan Cai , Yundai Chen , Jun Guo	NN	Best accuracy 99.35%
8	(Buettner, R., Beil, D., Scholtz, S., & Djemai, 2020)	David Beil, Stefanie Scholtz	Random Forest	Best accuracy 96.77
9	(Gopo, 2021)	Magesh Gopu	Random Forest	Best accuracy 89.3%
10	(Karthikeyan Harimoorthy and Menakadevi Thangavelu, 2020)	Karthikeyan Harimoorthy, Menakadevi Thangavelu	SVM-Linear, SVM-Polynomial, Random Forest, and Decision Tree, SVMRBK	Best accuracy 98.7% for SVMRBK

3.0 Dataset

This inquiry used data from an ongoing study on cardiovascular disease. This test aims to predict whether or not a patient has a high chance of acquiring heart disease in the future. In all, the collection includes 297 items and 14 patient features. The website of Kaggle may be reached at the following address:

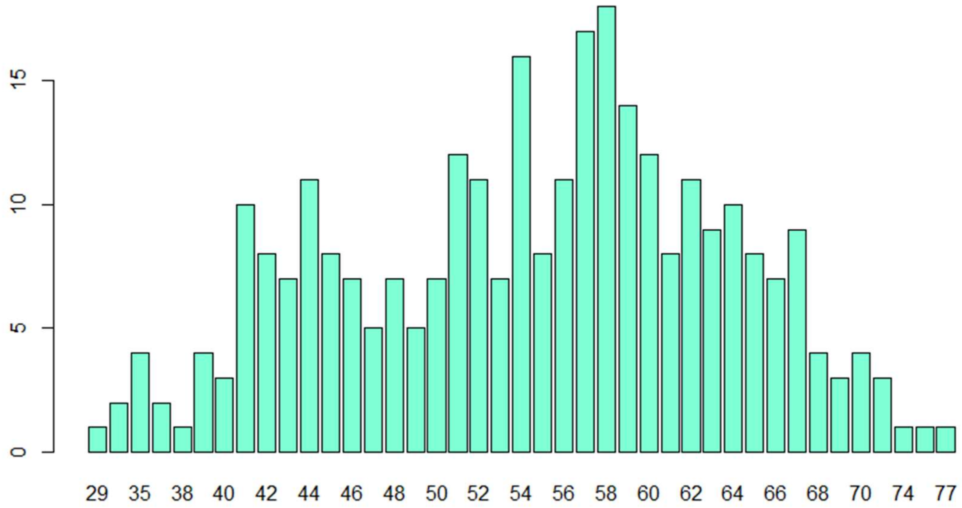
https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci



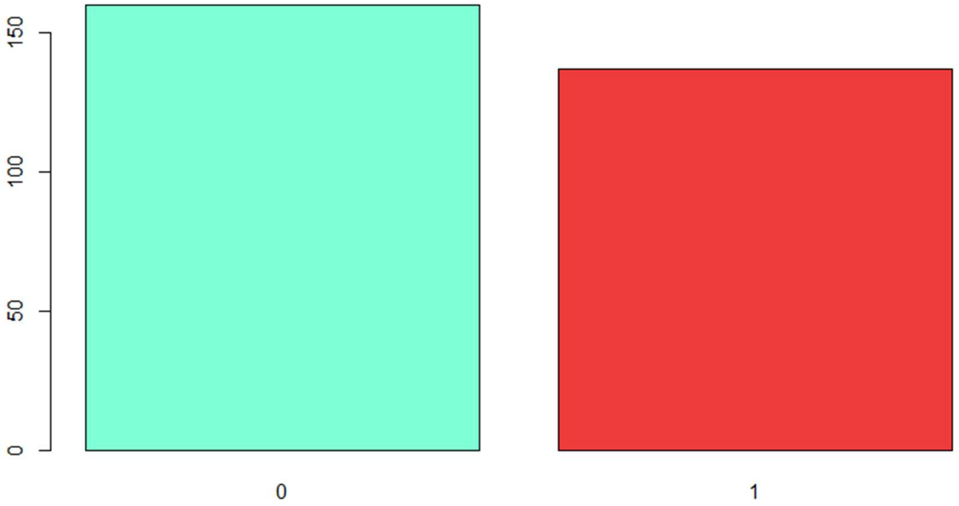
Variable name	Description	Category
Age	Age of the patient	Num
Sex	Sex of the patient	Cha
condition	0 (no disease), 1(disease)	Int
exang	exercise induced angina (1 = yes; 0 = no)	Int
ca	number of major vessels (0-3)	Int
cp	Chest Pain type chest pain type	Int
trtbps	resting blood pressure (in mm Hg)	Int
chol	cholestoral in mg/dl fetched via BMI sensor	Int
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Int
rest_ecg	resting electrocardiographic results	Int
thalach	maximum heart rate achieved	Num
slope	The slope of the peak exercise ST segment	Int
oldpeak	ST depression induced by exercise relative to rest.	Num
thal	normal; fixed defect; reversible defect	Int

3.1 Plots of Dataset

There are two plots related to dataset's different categories. In the Following plot, distribution of age variable is shown.

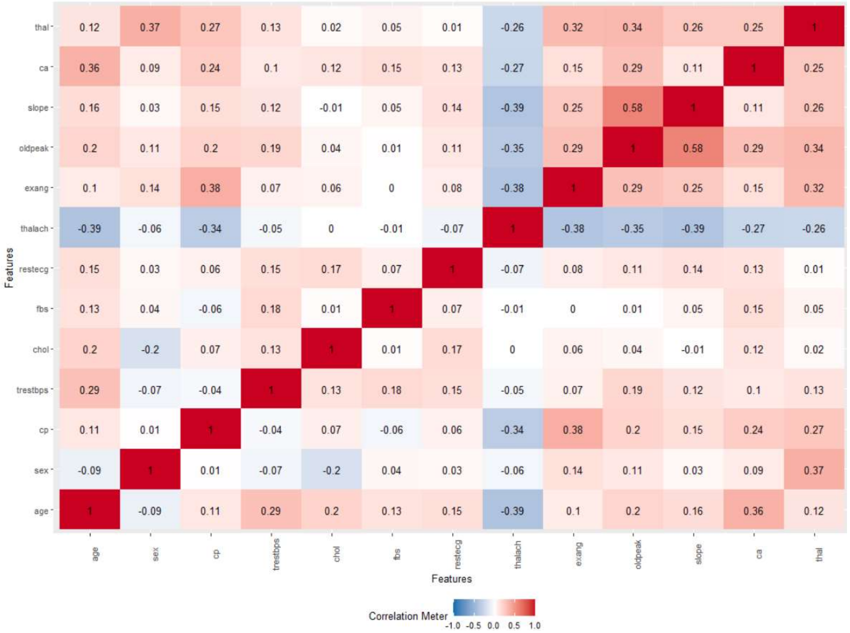


In the next plot we have condition variable that 0 means No disease and 1 means Disease.



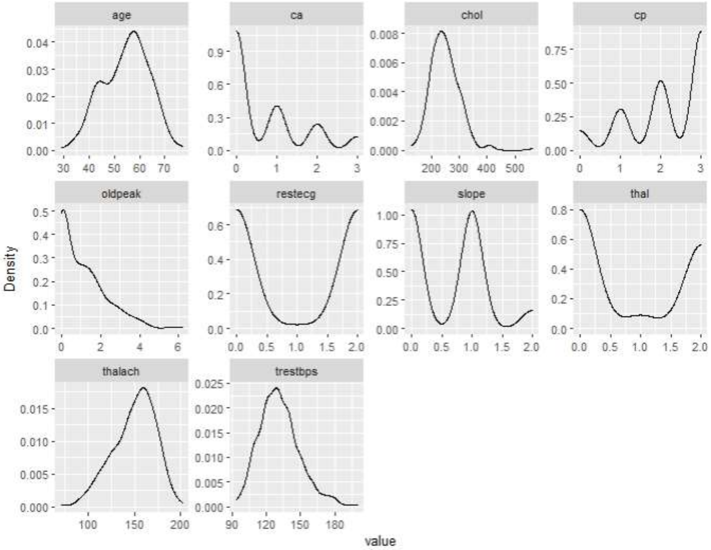
3.2 Correlation between independent variables

According to following plot, the correlation among independent variables is shown. Correlation is an important statistical measurement. The degree of relationship between two variables can be quantified by a single number. The range of correlation is from -1 to 1. Negative indicates that as one variable increase and the positive shows decreases.



3.3 Gaussian test

We plot density estimates for each continuous feature using the plot density function to determine if the density plot is Gaussian. Following figure demonstrates that Guassanese tests may be used to investigate variables.



4.0 Machine learning techniques

The purpose of this study was to demonstrate a machine's capacity to learn from medical data. Learning is the process of obtaining new information via the accumulation of old information. When we are bored, we must immediately begin acquiring information. We will investigate how we get knowledge from our own experiences. When a machine self-learns and trains, this is known as "machine learning." In machine learning, statistics, psychology, and brain modelling are all included. For ML to function, a collection of records or instances with distinct characteristics is required. They are Two forms of machine learning: Supervised Learning and Unsupervised. In supervised learning, labelled data are often utilised to identify mapping functions that translate input variables into output variables. Unsupervised learning models are applied to each sample of unsupervised procedures, as opposed to modelling the output data.

4.1 Decision Tree

Due to its origins in the decision-making process of humans, decision trees are intuitive. Whether the provided information is discrete or continuous, it is capable of addressing the issue. Using deterministic decision tree approaches, the traits are separated into "Best" subclasses. Each branch's partitioned data must be as PURE as is practically possible, given that the criteria for splitting must be same. Using this model, predict what will occur in the case of an occurrence (Patel & Prajapati, 2018). For this project, four various kinds of DT are designed.

4.2 Artificial Neural Network

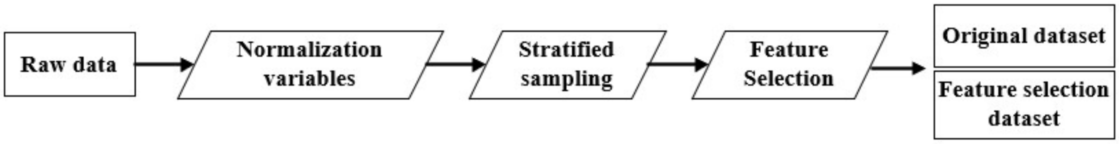
Artificial intelligence-based models of neural networks try to imitate the structure and behaviour of biological brain networks. The core of a neural network consists of mathematical functions or artificial neurons. The three concepts of addition, multiplication, and activation govern these models. At the entry of the artificial neuron, input values are multiplied by weights. All inputs and biases are summed in an artificial neuron's deepest layer. During the creation of an artificial neuron, an activation function (transfer function) is utilised to manage the sum of previously weighted inputs and bias (Dwivedi, 2018).

4.3 K-nearest Neighbour

KNN extends the nearest neighbour technique to unknown categories by choosing k-closest neighbours from the training samples and then counting the number of samples inside each of these k-closest neighbours. For instance, it is used to forecast how individuals will classify items. KNN classifies incoming input data by a comparison to previously learnt data. Each piece of data is classified based on its closest familial relationships. Despite its extensive usage, KNN has a number of shortcomings.

5.0 Data preparation

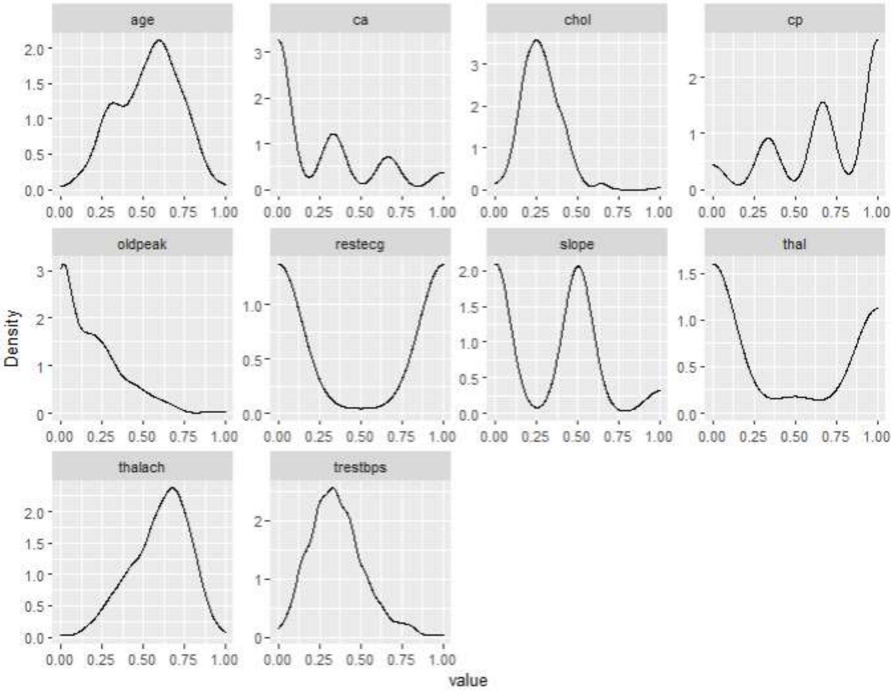
Data preparation is the act of changing raw data into a format that used by machine learning algorithms to uncover new insights or predict future events. The data must be cleansed in order to accelerate data processing and analysis and increase the accuracy of the prediction. R studio's given tools and packages are utilised to accomplish this operation. Using the programmes listed the pre-processing, visualisation, and analysis of data were investigated.



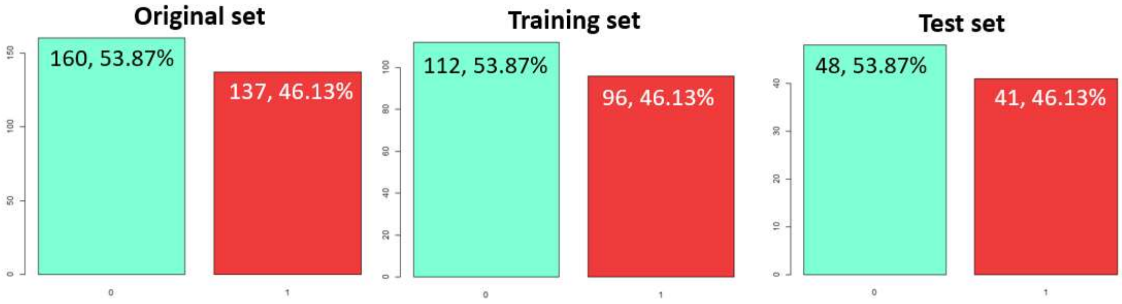
The researcher conducts normalisation, stratified sampling, and feature selection, and then creates original dataset and feature selection dataset.

5.1 Normalization variables

Following figure depicts the variable distributions following Min-Max normalisation, which we utilised to increase the accuracy of our predictions by normalising numeric variables.

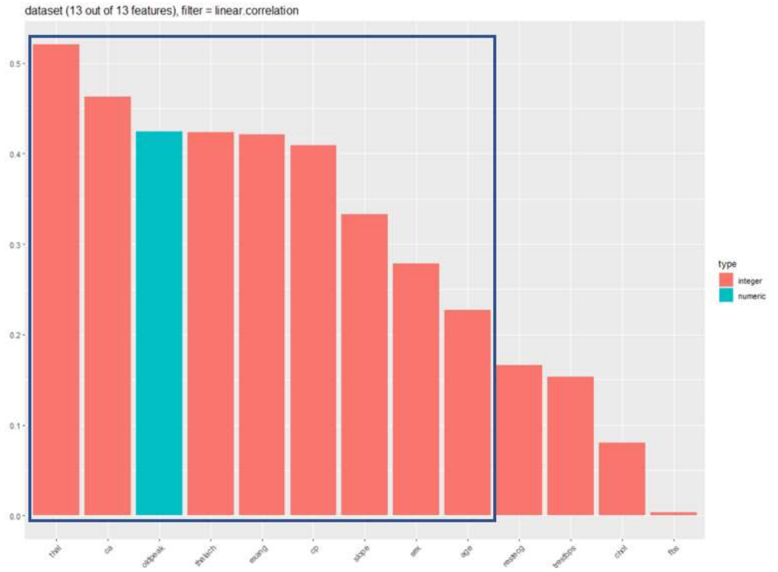


5.2 Stratified sampling

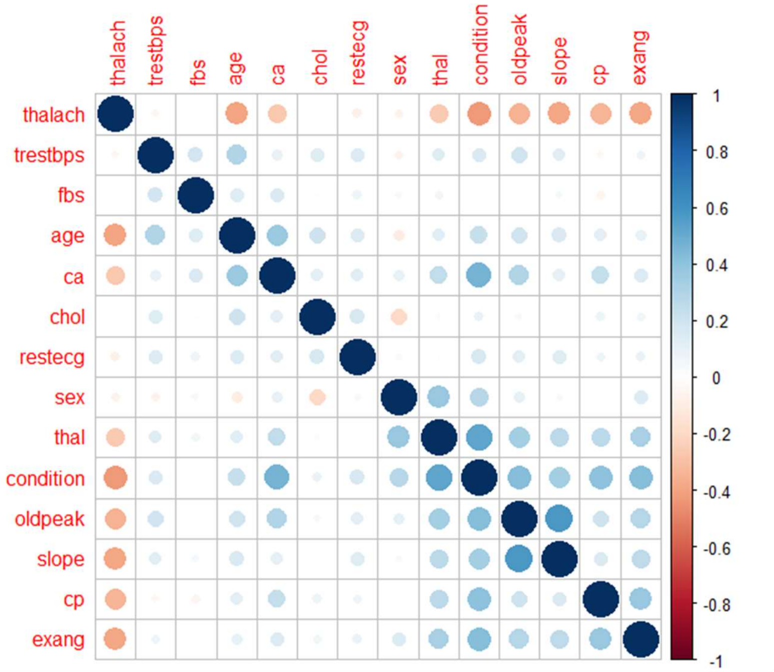


For training or validation purposes, you may partition a dataset into two distinct datasets using random or stratified sampling. Using stratified sampling, the dataset was divided into training and testing subsets. In stratified sampling, since samples are sorted prior to selecting sample categories, the sample dataset has the same proportion as the original dataset.

5.3 Feature Selection



We use the mlr programme to exclude all but the most essential characteristics from the list of possible options. This figure, which is generated using the correlation coefficient approach, displays the p-value of the correlation coefficient and the correlation coefficient's correlation coefficient. Correlation coefficients are shown in Figure below. We generate a new dataset that excludes the trestbps, chol, fbs, and restecg characteristics due to their low relevance values.



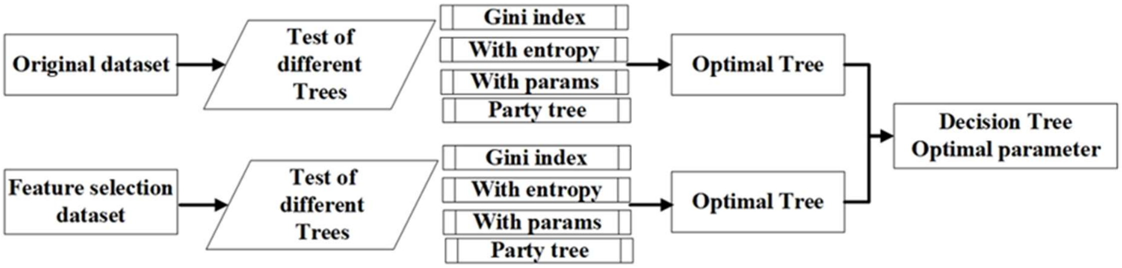
5.4 Two types of datasets

After the preceding data pre-processing, there are two types of datasets as original dataset and feature selection, which will be used and analysed individually in machine learning methods.

6.0 Model explanation and validation

This research utilises two kinds of datasets: the original dataset and the feature selection dataset. These two datasets and the three ML techniques which assist us in predicting cardiac disease.

6.1 Decision Tree

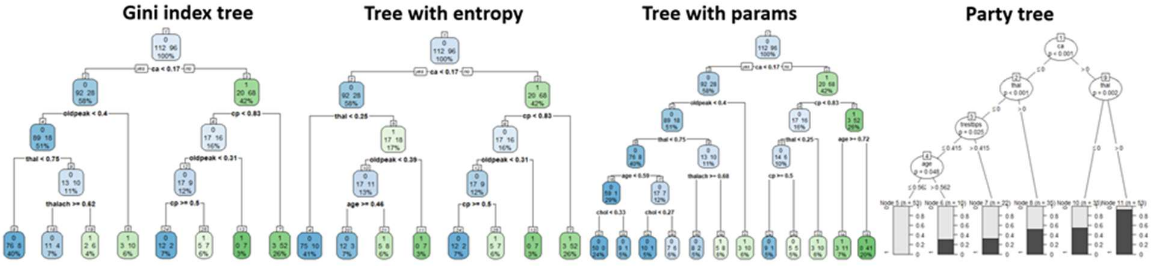


We employ several trees to get the ideal decision tree utilising the original dataset and the feature selection dataset separately.

6.1.1 The results from the original dataset

Model	Training Test	Test set
Gini index	Accuracy0.8702 Sensitivity0.8839 Specificity0.8542	Accuracy0.7865 Sensitivity0.7917 Specificity0.7805
entropy	Accuracy0.8654 Sensitivity0.8839 Specificity0.8438	Accuracy0.764 Sensitivity0.8125 Specificity0.7073
params	Accuracy0.8558 Sensitivity0.8750 Specificity0.8333	Accuracy0.8539 Sensitivity0.8542 Specificity0.8537
Party Tree	Accuracy0.775 Sensitivity0.6696 Specificity0.8958	Accuracy0.7303 Sensitivity0.5625 Specificity0.9268

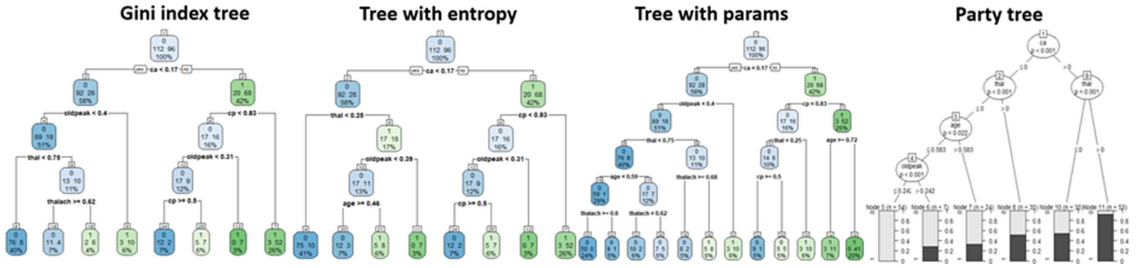
On the training set, a tree of gini index performs the best, but a tree of params performs the best on the test set. Consequently, there is overfitting in the tree structure of the gini index. Following Figure demonstrates that the initial dataset produces several tree designs.



6.1.2 The results from the feature selection dataset

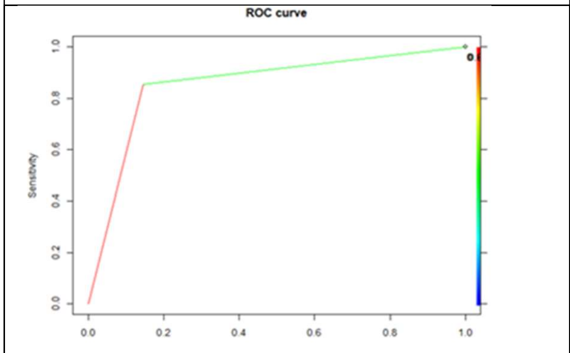
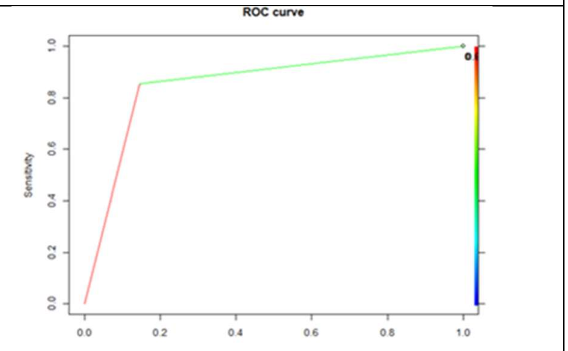
Model	Training Test	Test set
Gini index	Accuracy 0.8702 Sensitivity 0.8839 Specificity 0.8542	Accuracy 0.7865 Sensitivity 0.7917 Specificity 0.7805
entropy	Accuracy 0.8654 Sensitivity 0.8839 Specificity 0.8438	Accuracy 0.764 Sensitivity 0.8125 Specificity 0.7073
params	Accuracy 0.8558 Sensitivity 0.8750 Specificity 0.8333	Accuracy 0.8539 Sensitivity 0.8542 Specificity 0.8537
Party tree	Accuracy 0.774 Sensitivity 0.6696 Specificity 0.8958	Accuracy 0.7303 Sensitivity 0.5625 Specificity 0.9268

The results of DT using feature selection data demonstrates this point: The gini index tree has the best training performance, whereas the tree with parameters has the best testing performance. When utilising a feature selection dataset, the same results are produced as when using the original dataset. Following figure demonstrates that the initial dataset produces several tree designs.

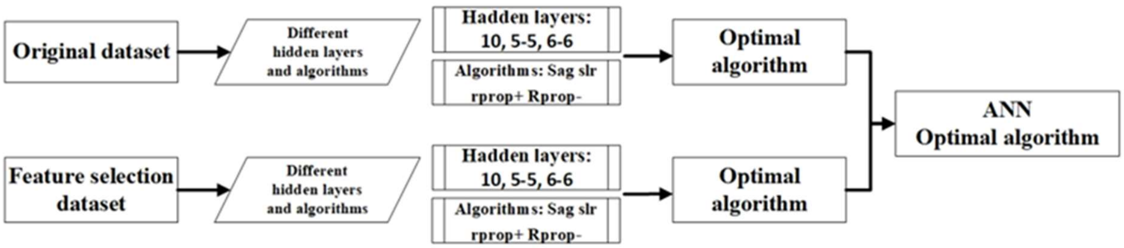


6.1.3 Optimal model of DT models

Two set of datasets in decision three algorithm gave us accuracy 0.8539 for original dataset and 0.8539 for feature selection dataset. So, based on these results, Decision tree is not sensitive to these two different datasets.

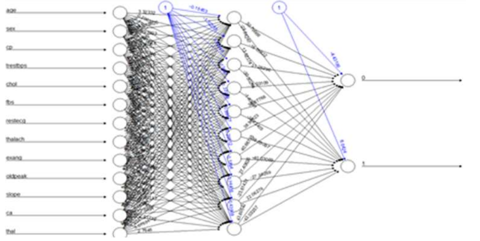
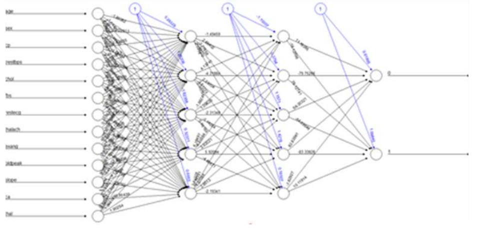
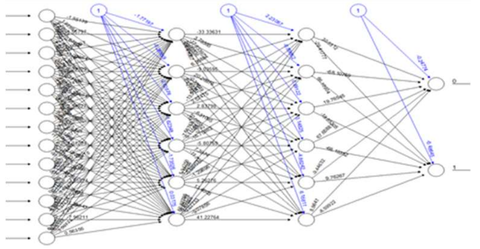
Original dataset	Feature selection dataset
Accuracy 0.8539 Sensitivity 0.8542 Specificity 0.8537	Accuracy 0.8539 Sensitivity 0.8542 Specificity 0.8537
 AUC: 0.854	 AUC: 0.854

6.2 ANN

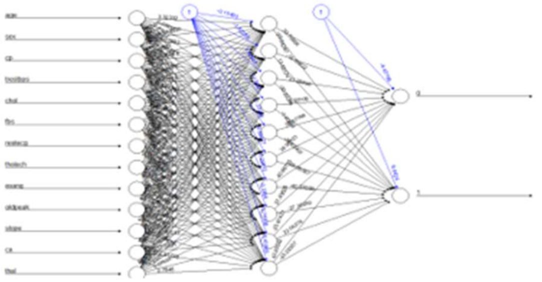

Original dataset
Different hidden layers and algorithms
Hadden layers: 10, 5-5, 6-6
Algorithms: Sag slr rprop+ Rprop-
Optimal algorithm
Feature selection dataset
Different hidden layers and algorithms
Hadden layers: 10, 5-5, 6-6
Algorithms: Sag slr rprop+ Rprop-
Optimal algorithm
ANN Optimal algorithm

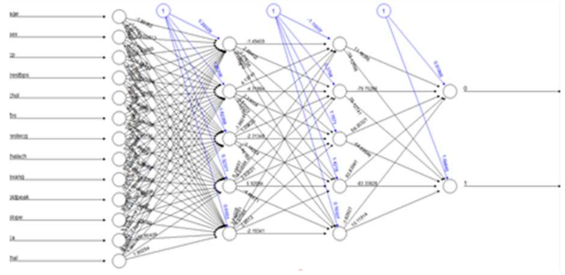
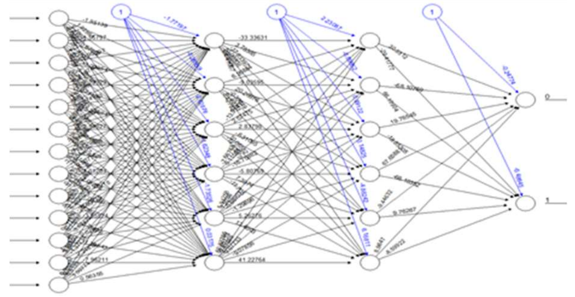
Various ANN models are developed using a variety of techniques and hidden layers, such as the 5-5 and 6-6 layers.

6.2.1 The results from the original dataset

Hidden layers	Evaluation		ANN structure
10	training set	accuracy1 sensitivity1 specificity1	
	test set	accuracy0.7528 sensitivity0.7500 specificity0.7561	
5-5	training set	accuracy1 sensitivity1 specificity1	
	test set	accuracy0.7978 sensitivity0.7917 specificity0.8049	
6-6	training set	accuracy1 sensitivity1 specificity1	
	test set	accuracy0.8202 sensitivity0.8750 specificity0.7561	

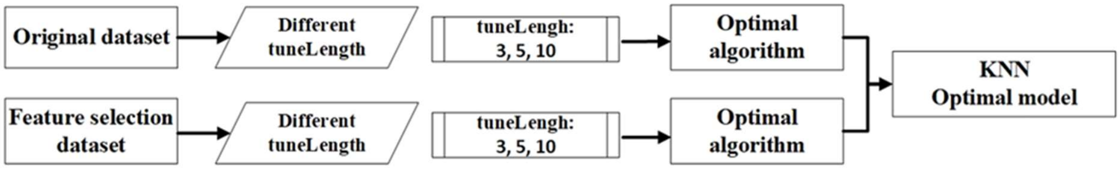
6.2.2 The results from the feature selection dataset

Hidden layers	Evaluation		ANN structure
10	training set	accuracy1 sensitivity1 specificity1	
	test set	accuracy0.7978 sensitivity0.7708 specificity0.8293	
5-5	training set	accuracy1 sensitivity1 specificity1	

	test set	accuracy0.7416 sensitivity0.7917 specificity0.6829	
6-6	training set	accuracy1 sensitivity1 specificity1	
	test set	accuracy0.7865 sensitivity0.8333 specificity0.7317	

According to results, the best performance of hidden layer is 10.

6.3 K-Nearest Neighbour



This flowchart displays the K-Nearest Neighbor algorithm, in which multiple lunLength parameters are used to generate separate models.

6.3.1 The results from the original dataset

tuneLength	Training set	Test set
3	Accuracy 0.8269 Sensitivity 0.8167 Specificity 0.8409	Accuracy 0.7978 Sensitivity 0.8571 Specificity 0.7447
5	Accuracy 0.8317 Sensitivity 0.8080 Specificity 0.8675	Accuracy 0.7978 Sensitivity 0.8125 Specificity 0.7805
10	Accuracy 0.8125 Sensitivity 0.8017	Accuracy 0.7978 Sensitivity 0.8261

	Specificity 0.8276	Specificity 0.7674

As shown in Table above, a range of tuneLength parameters (3, 5, 10) were used to develop a model and test it against the original data set, with the best result yielding an accuracy of 0.7979.

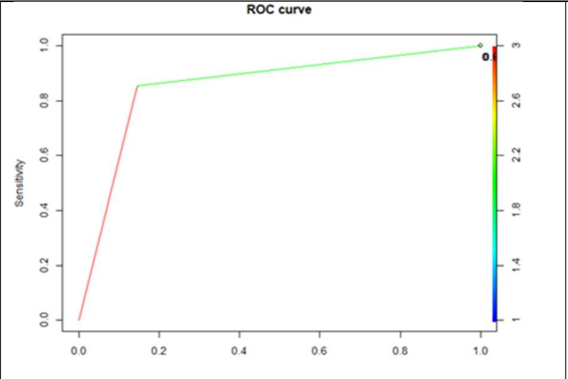
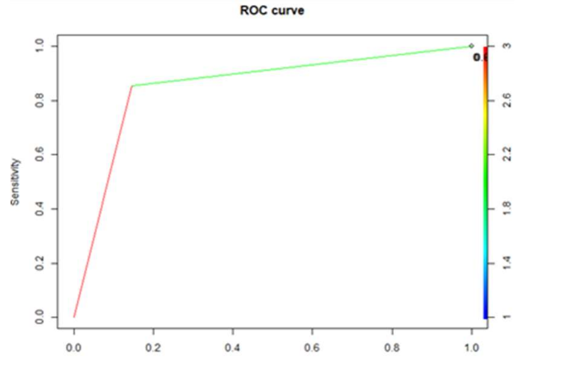
6.3.2 The results from the feature selection dataset

tuneLength	Training set	Test set
1	Accuracy 0.875 Sensitivity 0.8909 Specificity 0.8571	Accuracy 0.8315 Sensitivity 0.8367 Specificity 0.8250
5	Accuracy 0.8317 Sensitivity 0.8291 Specificity 0.8352	Accuracy 0.8202 Sensitivity 0.8333 Specificity 0.8049
10	Accuracy 0.8317 Sensitivity 0.8291 Specificity 0.8352	Accuracy 0.8202 Sensitivity 0.8333 Specificity 0.8049

As shown in table of 6.3.2-part, researcher constructed a model and assessed it using a feature selection dataset with different tuneLength parameter values (1, 5, 10). The model's highest degree of precision was 0.8315.

6.3.3 Optimal model

Original dataset		Feature selection dataset	
TuneLengh	5	tuneLengh	1
Accuracy0.7978 Sensitivity0.8125 Specificity0.7805		Accuracy0.8315 Sensitivity0.8367 Specificity0.8250	

	
AUC: 0.853	AUC: 0.854

Above table demonstrates that, with the feature selection dataset and the tuneLength parameter set to 1, the optimal KNN model achieves the specified accuracy of 0.8315.

7.0 DISCUSSION OF RESULTS

ML Algorithm	Dataset	Best Model Validation (Test set)	Parameters (best model)
Decision Tree	Original dataset	Accuracy 0.8539 Sensitivity 0.8542 Specificity 0.8537	Tree with params
ANN	Original dataset	Accuracy 0.8202 Sensitivity 0.8750 Specificity 0.7561	Hidden layer: 6-6 Algorithm: default
KNN	Feature selection dataset	Accuracy 0.8315 Sensitivity 0.8367 Specificity 0.8250	tuneLengh: 1

After evaluating a variety of machine learning methods, we've established the most precise way; the findings are shown in the table above. In the next stage, we will analyse and evaluate the data to assess their applicability.

(I) Validation of decision tree

The accuracy of the decision tree is 0.8539, and its sensitivity (0.8542) is greater than its specificity (0.8539). (0.8541). (0.8537). Using parameters, this decision-making model is as precise as feasible. Because the dataset is limited, decision trees' capacity to properly manage small data sets is advantageous.

(II) Validation of KNN

Despite the notion that sensitivity outweighs specificity in this scenario, KNN's performance is mediocre at 0.8315 (0.8367). (0.8250). Using a feature selection dataset and a parameter value of 1 for TuneLength, this model is more effective. In contrast to the original dataset, the feature selection dataset has nine separate variables (13). The optimal method for determining the parameter tuneLength, whose desired value is 1, is to use a small data set and independent components.

(III) Validation of ANN

The low performance of the ANN is owing to the absence of a large dataset. The precision of ANN is 0.8202, while its specificity is 0.7561.

7.1 Compare and contrast the references with this actual situation.

(I) Matching between machine learning algorithms and size of dataset

Exemplary reference models, neural network and ensemble models are well suited for large-scale data. Due to the limits of our technology, which hinders the performance of the ANN, we are restricted to selecting only small-scale data.

(II) Insufficient modelling scope

This task emphasises algorithm and parameter manipulation, however the parameter's scope is too limited. As an example, we have tested the hidden layer with the following parameters: The numbers 10, 5, 5, and 6 are significant, but there is a significantly greater range of choices to investigate.

(III) Lack of excessive sampling and feature selection

There is just a single dataset for feature selection, and there is no over-examination. We investigate an insufficiently small portion of machine learning techniques.

8.0 CONCLUSION

This research examined the predictive accuracy of multiple ML algorithms for heart disease. The selection of the optimal method was based on the construction and tuning of models using a variety of machine learning techniques. The following explanations explain why this was the case. Using normalisation variables and feature selection, we generate two datasets for data pre-processing: the original dataset and the feature selection dataset. Three machine learning algorithms were utilised to construct models from two kinds of information. We were unable to provide high-quality conclusions due to a lack of data and adjusting scope. Future advancements in machine learning are anticipated as we experiment with larger datasets and expand the breadth of our tuning.

9.0 References

Awan, S.E., Bennamoun, M., Sohel, F., Sanfilippo, F.M., Dwivedi, G. . (2019). *Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC heart failure 2019;6(2):428–435.*

Buettner, R., Beil, D., Scholtz, S., & Djemai, A. (2020). Development of a machine learning based algorithm to accurately detect schizophrenia based on one-minute EEG recordings. *In Proceedings of the 53rd Hawaii International Conference*.

Cai W, Chen Y, Guo J, Han B, Shi Y, Ji L, Wang J, Zhang G, L. J. (2019). *Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. Comput Biol Med. ;116:103378. doi: 10.1016/j.compbiomed.2019.103378. Epub 2. P*.

Castillo E. (1997). Expert systems and probabilistic network models. *Springer, Berlin*.

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*(10), 685–693. https://doi.org/10.1007/s00521-016-2604-1

Gopo, M. (2021). *ttps://www.researchgate.net/publication/338330612_Optimal_feature_selection_through_a_cluster-based_DT_learning_CDTL_in_heart_disease_prediction*.

Karthikeyan Harimoorthy and Menakadevi Thangavelu. (2020). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intell*.

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. *IEEE Region 10*

Annual International Conference, Proceedings/TENCON, *2019-Octob*, 367–372. https://doi.org/10.1109/TENCON.2019.8929434

Le, H.M., Tran, T.D., Van Tran, L. . (2018). *Automatic heart disease prediction using feature selection and data mining technique. Journal of Computer Science and Cybernetics 2018;34(1):33–48. le*.

Mati´c, V., et al. (2017). Effective diagnosis of heart disease presence using artificial neural networks. In: *Sinteza 2017- International Scientific Conference on Information Technology and Data Related Research. Singidunum University; 2017,* 3–8.8.e.

Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart Diseases Diagnosis Using Neural Networks Arbitration. *International Journal of Intelligent Systems and Applications*, *7*(12), 75–82. https://doi.org/10.5815/ijisa.2015.12.08

Patel, H. H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74–78. https://doi.org/10.26438/ijcse/v6i10.7478

Yildirim O, Baloglu UB, Tan RS, Ciaccio EJ, A. U. (2019). *A new approach for arrhythmia classification using deep coded features and LSTM networks. Comput Methods Programs Biomed. 2019 Jul;176:121-133. doi: 10.1016/j.cmpb.2019.05.004. Epub 2019 May*.

Zina Ben M. (2020). *No TiBen Miled Z, Haas K, Black CM, Khandker RK, Chandrasekaran V, Lipton R, Boustani MA. Predicting dementia with routine care EMR data. Artif Intell Med. 2020 Jan;102:101771. doi: 10.1016/j.artmed.2019.101771. Epub 2019 Dec 5. PMID: 31980108.tle*.