| | | |
|---|---|---|
| Syrian Arab Republic |  | الجمهورية العربية السورية |
| Ministry of Higher Education | | وزارة التعليم العالي |
| Syrian Virtual University | | الجامعة الافتراضية السورية |

# Prediction of Heart Disease Using Artificial Intelligence


By

**Reham Nofal / reham_168419**


A Thesis Submitted to Obtain

The Degree of Master in Bioinformatics

Under the supervision of:

## Dr. Abdulqader Abbady

***F22/2023***

## *Acknowledgement*

I would like to acknowledge **Dr. Abdulqader Abbady** for supervising this research. He has made a significant impact on our knowledge and skills during this master.

I want to thank **Syrian Virtual University** and **Program Manager Dr. Majd Aljamali** and all professors for providing the opportunity to study a master's degree in bioinformatics

## *Abstract*

Healthcare is one of the most important aspects of human life. Cardiovascular diseases (CVDs) or heart diseases are one of the most lethal diseases, affecting the lives of millions of people worldwide. 25% of people die suddenly without any prior symptoms. Correct diagnosis and treatment at an early stage will save people. Therefore, it is important to establish a system that can predict disease early. Artificial intelligence (AI) algorithms are the most advanced computer sciences that have a significant role in predicting diseases to help clinicians diagnose disease and optimize treatment processes.

In this study, we proposed an efficient and accurate model for early prediction of cardiovascular disease, based on 13 features that are important for physicians to diagnose like age, gender, chest pain type, blood pressure, cholesterol, blood glucose, also on ECG reading, and other investigations. The model is based on machine learning techniques and artificial neural networks by using three dataset related to California University (Cleveland, Statlog) and from kaggle heart predicted data. The model performed by 3 platforms (SPSS, WEKA, Python), then developed based on classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree, Random Forest, XGBoost.

The best artificial model when using Random forest, XGBoost with accuracy 98.25%, also many models have high accuracy and sensitivity like MLP, DT

Thus, the integration of machine learning and neural network is applied in predicting heart disease.AI methods have shown to be a promising tool in the field of medicine.



## **Keywords**

Artificial Intelligence, Artificial Neural Network, Machine learning, cardiovascular diseases, Heart dataset, Cleveland dataset, Statlog dataset, Heart-disease prediction, Multiple Perception, Logistic Regression, Support vector machine, decision tree, Random forest, K-NN, XGBoost.

# ***Contents***

# List of Figures

# List of Tables

# List of Abbreviations

1. CVDs: cardiovascular diseases
2. CHD: coronary heart disease
3. CAD: coronary artery disease
4. ACS: acute coronary syndrome
5. HF: Heart failure
6. MI: Myocardial infarction
7. SCD: Sudden Cardiac Death
8. ECG: Electrocardiogram
9. ICA: Invasive coronary angiography
10. AI: Artificial Intelligence
11. MLP: multilayer perceptron
12. WHO: World Health Organization
13. CNN: convolutional neural network
14. RNN: recurrent neural network
15. DBN: deep belief network
16. GAN: generative adversative network
17. EHRs: electronic health records
18. CMR: cardiac magnetic resonance
19. CT: computed tomography
20. IoT: internet of things
21. SPECT: single photon-emission computerized tomography.
22. ICU: intensive care units
23. CP: chest pain

# Introduction

## Cardiovascular Disease(CVD)

Cardiovascular diseases (CVDs) include heart and blood vessels disease like coronary heart disease (CHD) and coronary artery disease (CAD), and acute coronary syndrome (ACS).

coronary artery disease (CAD) is characterized by atherosclerosis in coronary arteries and can be asymptomatic, whereas ACS almost always presents with a symptom, such as unstable angina. CAD will result for chronic heart disease, acute coronary syndrome (ACS), Heart failure (HF), Ischemia and Myocardial infarction(MI) (1).

Coronary heart disease (CHD) is a slowly developing chronic disease that mainly results from a progressive narrowing of blood vessels that supply the myocardium with oxygenated blood, giving rise to ischemia at times of increased oxygen demands (2). As clinical result includes an inadequate ejection of blood from the heart (heart failure), irregular cardiac rhythms (arrhythmias), or acute coronary syndromes (ACSs) such as myocardial infarctions and unstable angina, which are often followed by sudden cardiac death.

Acute coronary syndrome (ACS) covers acute myocardial infarction and unstable angina pectoris. Symptoms of ACS are acute angina pectoris that is prolonged for >20 minutes, symptoms of nausea, fatigue, and dyspnea. ECG shows non-ST segment elevation myocardial infarction (NSTEMI) (2) .

Myocardial infarction (MI) is myocardial injury and necrosis due to myocardial ischemia with a subsequent elevation in cardiac troponin and ST-segment elevation(STEMI) (3) .

## Prevalence of CVD

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.

Prevalence of CVD (comprising CHD, HF, stroke, and hypertension) in adults ≥20 years of age was 48.6% overall (127.9 million in 2020) and increases with age in both males and females. HD and stroke currently claim more lives each year than cancer and chronic lower respiratory disease combined. • In 2020, 19.05 million deaths were estimated for CVD globally. The estimated annual incidence of myocardial infarction is 605 000 new attacks(4).

The World Health Organization(WHO) has declared coronary artery disease (CAD) is the common type of cardiovascular disease. More than 30% of deaths worldwide were due to CAD, which resulted in more than 17 million deaths in 2015. Additionally, more than 360,000 Americans have died from heart attacks. (5) and CHD is the leading cause of death in adults in the U.S (1).

## Pathophysiology of CAD

Atherosclerosis is a leading cause of artery disease. It is a progressive chronic inflammatory process of arterial wall thickening. Atherosclerosis promotes fatty streak formation, low-density lipoproteins (LDLs) under the endothelial layer of wall of vessel and activation of inflammatory response is mediated through macrophages, foam cells then fibrous plaque formation (6).

Chronic heart disease like Ischemia occurs when a plaque enlarges sufficiently to impair blood flow to meet tissue demand (usually >70% stenosis). Atherosclerosis plaque may stay stable leading to stenosis in arteries, but sometimes endothelial rupture leading to thrombus. That is most of the morbidity and mortality events from atherosclerosis, such as MI, stroke(7).

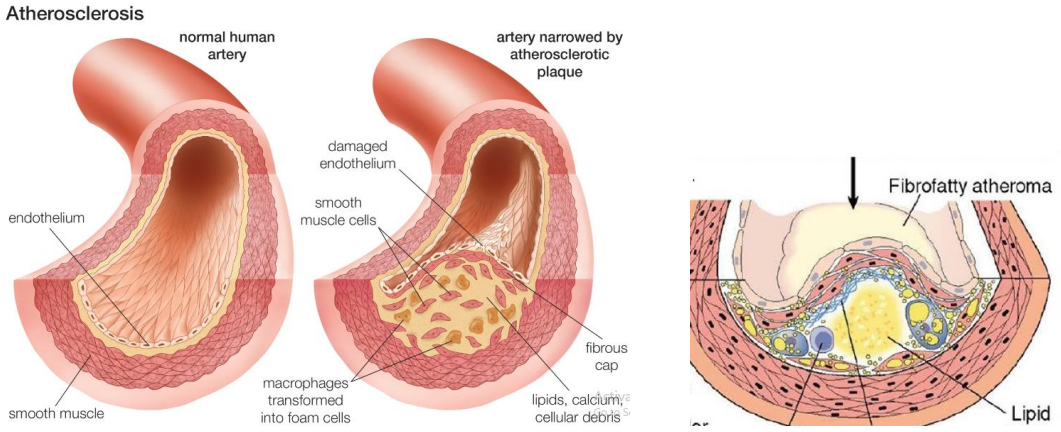



Figure (1): Atherosclerosis plaque causing stension in the artery (8)

## Risk Factors of Heart disease

There are several factors that increase the risk of heart disease, such as family history of heart disease, smoking habit, cholesterol level, obesity, nutrition, high blood pressure, and lack of physical exercise, diabetes mellitus, hypertension, hyperlipidemia, obesity, and psychosocial stress (6), beside role of hereditary like familial hypercholesterolemia (9,10).

## Diagnosis of CVD

- The diagnosis of heart disease is done by the analysis of the medical history of the patient, physical examination report. It is usually based on signs, symptoms and physical examination of the patient, that include confirmation of the causality for symptoms.
- Electrocardiogram (ECG) is considered the first line non-invasive diagnostic investigation for the evaluation of cardiovascular pathology.
- Exercise Electrocardiogram: The diagnostic endpoint of an exercise ECG test is ischemic. ECG changes defined as $\geq$ 1 mm horizontal or down-sloping ST-segment depression at peak exercise.
- Stress Echocardiography :exercise or pharmacological stress echocardiography is new or worsening of wall motion(11).
- Cardiac Magnetic Resonance (CMR).
- Coronary Computed Tomography Angiography (CTA).

- Invasive coronary angiography (ICA) has remained the gold standard in diagnosis. It provides anatomical evaluation of the coronary artery anatomy, including presence and severity of atherosclerosis by determining lesion location, luminal obstruction, lesion length (2,12).

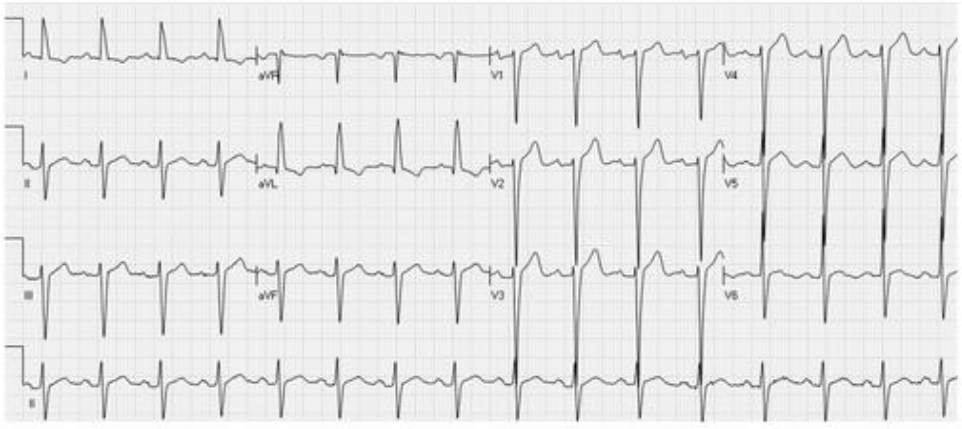


Figure (2): ECG example for diagnosis CVD. (13)

**Therapeutics**

- Antiplatelet drugs have been used in CAD treatment like aspirin
- Nitrates in the form of sublingual nitroglycerin have been used for immediate relief of angina by relieving the symptoms alone by increasing the myocardial oxygen supply and decreasing the myocardial oxygen demand (9).
- Treatment using angioplasty and stent placement, also known as percutaneous coronary intervention is a procedure used during a heart attack to quickly open a blocked artery and reduce the amount of damage to the heart.

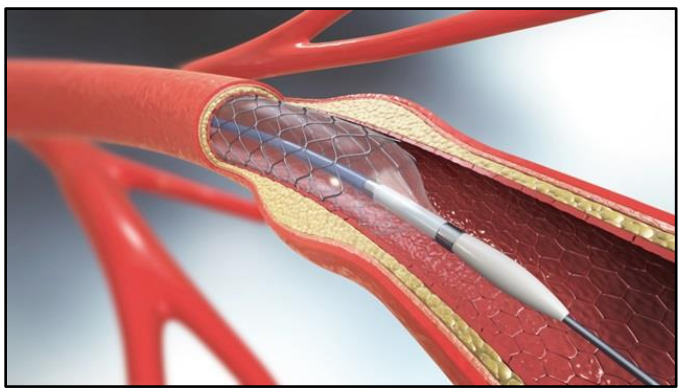


Figure (3): Angioplasty and stent procedure (14)

## Artificial Intelligence (AI)

The invention of artificial intelligence was a revolutionary breakthrough for humanity, which opened the gateway to a different world. Artificial intelligence (AI) was first described in 1950 by Alan Turing (15).

AI is accomplished by studying how the human brain thinks, and how humans learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems (16,17). AI is a combination of multiple disciplines, such as logistics, biology, linguistics, computer science, mathematics, engineering, and psychology. It has achieved extraordinary results in the field of speech and facial recognition, natural language processing, intelligent robots, and image recognition (18).

In this AI technique, computers learn from previous experiences and data. The amount of data is increasing rapidly, so there is a need to efficiently handle the data. Sometimes, it becomes quite difficult for humans to manually extract useful information from raw data due to their inconsistency and uncertainty. This is where machine learning is useful (19).

The goal is to identify hidden patterns in the data and predict new data. AI is able to use very complex nonparametric models from a vast amount of data in comparison to simple parametric models requiring a suitable-sized data set used in statistics (20).

AI can improve many aspects of patient care like medical imaging quality, early diagnosis, prognosis prediction, risk stratification, patient data analysis, personalized treatments and more (15).

There are many subfields in AI, such as machine learning (ML), deep learning (DL), and computer vision (21).

## Deep learning

DL is one of those remarkable advancements. Deep learning is the branch of machine learning that was named in 2006. It was inspired by the structure of the human brain, which contains neural networks (20). It is a data-processing method that uses a multiple-layer technique. The working of the layers can be considered to be a layer receiving weighted input, transforming it into mostly nonlinear functions, and then sending the output to the next layer(19).

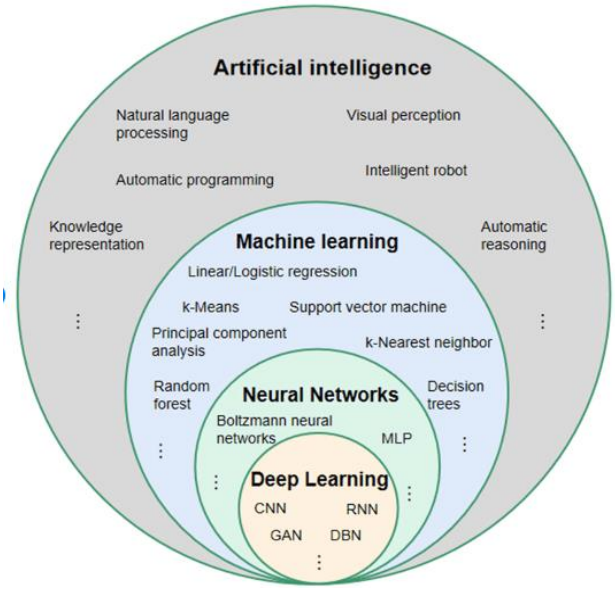



Figure (4): Relationship between artificial intelligence, machine learning, neural network, and deep learning (22).

## Machine Learning

Machine learning (ML) is one of the tools or pathways to artificial intelligence, where a computer model is enabled to learn new skills and information to provide useful tasks.The core principle of ML is to learn from data in order to forecast or make decisions depending on the assigned task. Once the ML algorithm is trained with data, the ML model will be provided with an input. The output will be a predictive model, based on the data that trained the mode (20,23).

ML focuses on the learning by developing algorithms in which can be explicitly coded using known features. In ML, there are four commonly used learning methods, each useful for solving different tasks: supervised, unsupervised, semi supervised, and reinforcement learning (15,18).

- Supervised ML: Most commonly used supervised learning tasks include classification (identification of the group a new measurement belongs to) and regression (prediction of a continuous value of a new observation).(20). It is useful if the task at hand requires the input data to be sorted into predetermined classes or making predictions(15) . The basic steps of supervised machine learning are:
    1. acquire a dataset and split it into separate training, and test datasets
    2. use the training datasets to inform a model of the relationship between features and target
    3. evaluate the model via the test dataset to determine how well it predicts(18).
- Unsupervised learning: aims to detect patterns in a dataset and categorize individual instances in the dataset to said categories. Some of the most common unsupervised learning tasks are clustering, association, and anomaly detection (20)
- Semi Supervised Learning: Semi supervised learning can be thought of as the "happy medium" between supervised and unsupervised learning and is particularly useful for datasets that contain both labeled and unlabeled data
- Reinforcement learning: is the technique of training an algorithm for a specific task where no single answer is correct, but an overall outcome is desired (18).

## Description for algorithms

### support vector machine (SVM)

It is used for classification and regression; it is a popular ML approach. SVM was introduced by Vapnik in the late twentieth century. For unlabeled data, supervised ML algorithms are unable to perform. Using a hyperplane to find the clustering among the data, SVM can categorize unlabeled data. However, SVM output is not nonlinearly separable.

To overcome such problems, selecting appropriate kernel and parameters is two key factors when applying SVM in data analysis (23)

## Logistic Regression:

Logistic Regression: is a classification algorithm where the goal is to find a relationship between features and the probability of a particular outcome.

It is the best regression analysis to use when the dependent variable or response variable is binary. It uses a sigmoidal curve to estimate class probability (18).

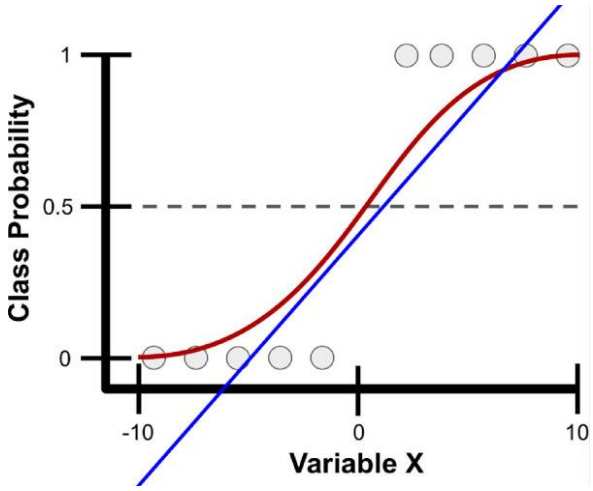



Figure (5) : illustrate logistic regression (18)

It works by combining the input variable (X) in a linear form and using coefficients to predict an output variable (Y) which is a binary value of 0 or 1. The logistic regression technique models the chance of an outcome based on the individual characteristics or input variables (X). It is represented mathematically as follows:

$$\log_{10}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

where π indicates the probability of an event, β represents estimated parameter values or regression coefficients associated with the variables via maximum likelihood estimation, and x indicates the parameter variables.

## Naive Bayes

A Naive Bayes classifier is a simple probabilistic classifier. Based on a given record or data point, it forecasts membership probability for each class. The most probable class is the one having the greatest probability. (23)

Naive Bayes classifier can be trained very efficiently in the context of supervised learning. The Bayesian rule is given in the following equation

$$P(H|X) = \frac{P(X|A)P(H)}{P(H)}.$$

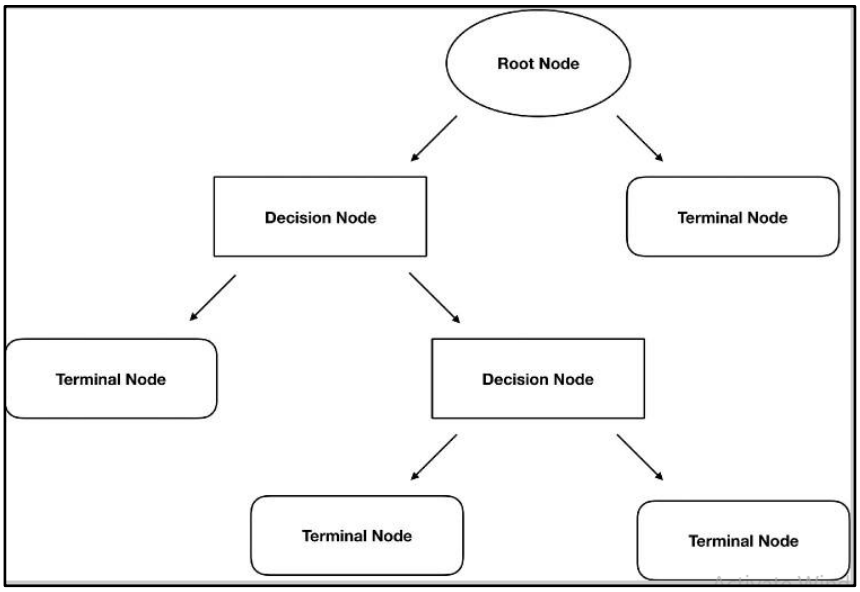
P(H|X) is a conditional probability, that is, the likelihood of event H occurring given X is true. P(X) and P(H) are the probabilities of observing X and H independently of each other.

## Decision Trees

A decision tree is a supervised learning technique, primarily used for classification tasks. In DT models, the attribute may take on various values known as classification trees; leaves indicate distinct classes, whereas branches reflect the combination of characteristics that result in those class labels. It begins with a root node, the first decision point for splitting the dataset, and contains a single feature that best splits the data into their respective classes. Each split has an edge that connects either to a new decision node that contains another feature to further split the data into homogenous groups(23).

## Random forest

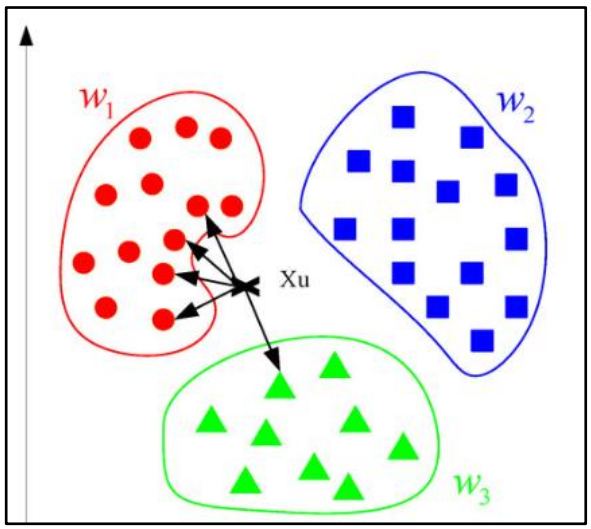
Random forest is a supervised machine learning algorithm that constructs several decision trees. The final decision is made based on the majority of the decision tree. It suffers from low bias and high variance. RF converts high variance to low variance, Random forest is an extension of this DT, known as an ensemble method, that produces multiple decision trees. Rather than using every feature to create every decision tree in a random forest, a subsample of features is used to create each decision tree. Trees then predict a class outcome, and the majority vote among trees is used as the model's final class prediction (24).




Figure(6): Decision tree method (18).

## K-Nearest Neighbour

K-NN is a simple, supervised machine learning (ML) algorithm., non-parametric approach used widely for both pattern recognition and classification applications. Predictions are made based on k numbers of frequently used neighbours for a new object, and a different distance metric for finding the K-NN is used. K-NN classifies new training data points based on similarity measurements. Data points are classified by considering the majority of votes from its neighbours. This works effectively for small dimensional data sets. K-NN does not require extra training for classification if a new data point is added to the existing data set. It is an inefficient algorithm for large data sets and requires more memory space for computation and longer model testing times because of the need to compute the distance between training data set and testing data set during each test. The K-NN algorithm needs to calculate the distance between the forecasted data point and the known data point, so as to select the nearest *k* labeled data ( fig:7) (25,26)




Figure(7) :K-NN logarithm map (25)

**XGBOOST (Extreme gradient Boosting)**

XGBoost is an implementation of the ensemble learning algorithm boosting. The fundamental principle of the XGBoost is to train the model using residuals. It is used to extract the numerical features of patients, which are an integrated learning method proposed by Tianqi Chen based on GBDT. The outcome of the most recent tree training is utilized as the input for the subsequent iteration, and the error is progressively decreased over numerous serial iterations (27)

## Artificial Neural networks

An artificial neural network (ANN) is a machine learning algorithm inspired by biological neural networks (human brain). Each ANN contains nodes (analogous to cell bodies) that communicate with other nodes via connections (analogous to axons and dendrites). Much in the way synapses between neurons are strengthened when their neurons have correlated outputs in a biological neural network. (18)

NNs contain a layer of input nodes, a layer of output nodes, and a number of “hidden layers” between the two. In simple ANNs, there exists an input layer between zero and three hidden layers and an output layer, whereas deep neural networks contain tens or even hundreds of hidden layers. ANNs feed information forward. This is known as a

feedforward neural network, meaning information from each node in the previous layer is passed to each node in the next layer, transformed, and passed forward to each node in the next layer. Each layer in an ANN can contain any number of nodes, the number of nodes in the output layer typically corresponds to the number of classes being predicted if the goal is multiclass classification (20).

ANN are classified depending on their structure, data flow, neurons used and their density. The most important types of neural networks involve:

1. Feed Forward Neural Network
2. Multilayer Perceptron
3. Radial Basis Function Neural Network
4. Recurrent Neural Network
5. Modular Neural Network
6. Convolutional Neural Network

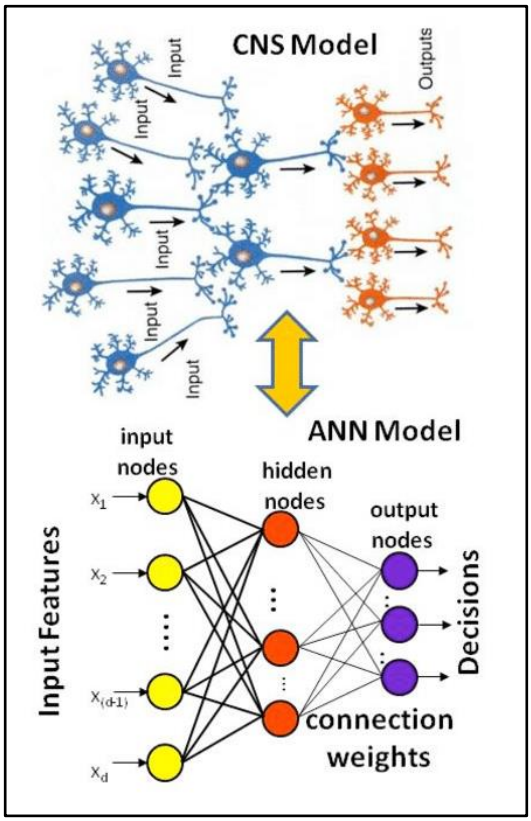



Figure (8): Comparison of central nervous system and artificial neural network.

## Multilayer Perceptron(MLP)

MLPNN is one of the most significant models in artificial neural networks. It is one of the powerful supervised types of classifiers for efficient prediction. It contains three nodes: input node, output node, and hidden node. The input nodes pass values to the first hidden layer, and then nodes of the first hidden layer pass values to the second and so on till producing outputs (26,28). It functions like a human brain; such that information passes from an input node to an output node via hidden nodes in a forward direction. An MLP is the combination of a number of neural units known as perceptron. Each layer contains a number of weights via which perceptron are connected with each other (Fig8)

In Multilayer Perceptron (MLP), input data travels through various layers of artificial neurons. It is a fully connected neural network, as all nodes are connected to all the neurons in the next layer. Input and output layers and multiple hidden layers (three or more) are present, and propagation is bi-directional (forward and backward). MLP is used in speech recognition, machine translation and complex classification (20). MLP network is trained using the backpropagation which uses data to adjust the network's weights and thresholds to minimize the error in its predictions on the training set. It computes the total weighted input:

$$X_j = \sum y_i w_{ij},$$

where yi is the activity level of the j-th unit in the previous layer and wij is the weight of the connection between the i-th and the j-th unit. Next, the unit calculates the activity yj using the sigmoid function (29)**.**

## Performance Evaluation

To maximize the chance of generalizability to the performance of the algorithm on unseen data, the training dataset is usually split into a slightly smaller training dataset and a separate validation dataset. Performance of a learned model can be evaluated in a number

of ways, but is most commonly evaluated based on prediction accuracy (classification) or error (regression)

Model performance is monitored via some form of accuracy on the training and testing datasets during. So long as the accuracy of the model on the training set (*X*%) and validation set (*Y*%) are increasing and converging after each training iteration, the model is considered to be learning

**Accuracy (Acc)**

The accuracy denotes total correctly identifying instances among all of the instances.as Eq:

$$ACC = \frac{T_p + T_N}{T_p + T_N + F_p + F_N}$$

**Receiver Operating Characteristic Curve(ROC)**

ROC curve is a plot of the sensitivity versus 1 – specificity. It is used for performance of a classification model, that means TPR versus false positive rate (FPR), where TPR is on the y-axis and FPR on the x-axis(30).

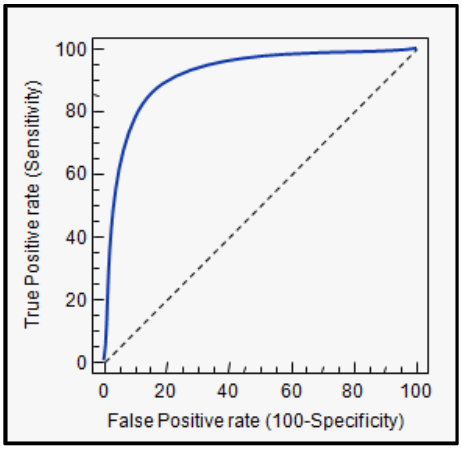



Figure (9): ROC curve.

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

Eq: 1

Precision is measured as the proportion of precisely predicted to all expected positive observations. Eq:2

$$Precision = \frac{TP}{TP + FP}$$

False Positive Rate(FPR): defined as

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$FPR = 1 - \text{Specificity}$$

$$= \frac{FP}{TN + FP}$$

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision}$$

## Recall

Represents the rate of values that measures positive records that the classifier correctly predicted. Moreover, it is called true positive rate (TPR) or sensitivity. Thus, recall is calculated as shown in Eq1. Precision is the ratio of TP records to the total of positive predicted records, as shown in Eq 2(19).

**Confusion Matrix**

It is a summary of predictions that are correct and incorrect per class. It shows the ways in which your classification model is confused when it makes predictions. (31)

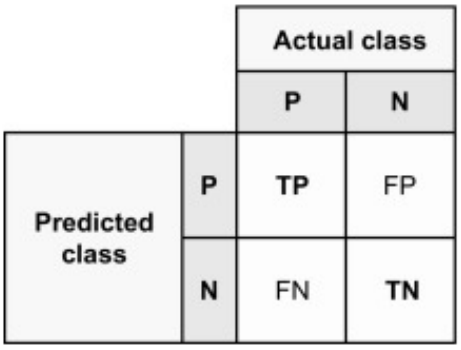



Figure (10): box of confusion matrix

## AI and Cardiology

With the development of new medical devices, new knowledge can be gained in the field of disease diagnosis. One of the best ways to quickly diagnose diseases is to use computer-assisted decision making, i.e., machine learning to extract knowledge from data. Advances in diagnostic uses of artificial intelligence (AI) for cardiac diseases, like interpreting echocardiograms to identify heart rhythms or left ventricular dysfunction from the ECG. in order to emphasize expected benefits to both patients and healthcare specialists. (32,33).

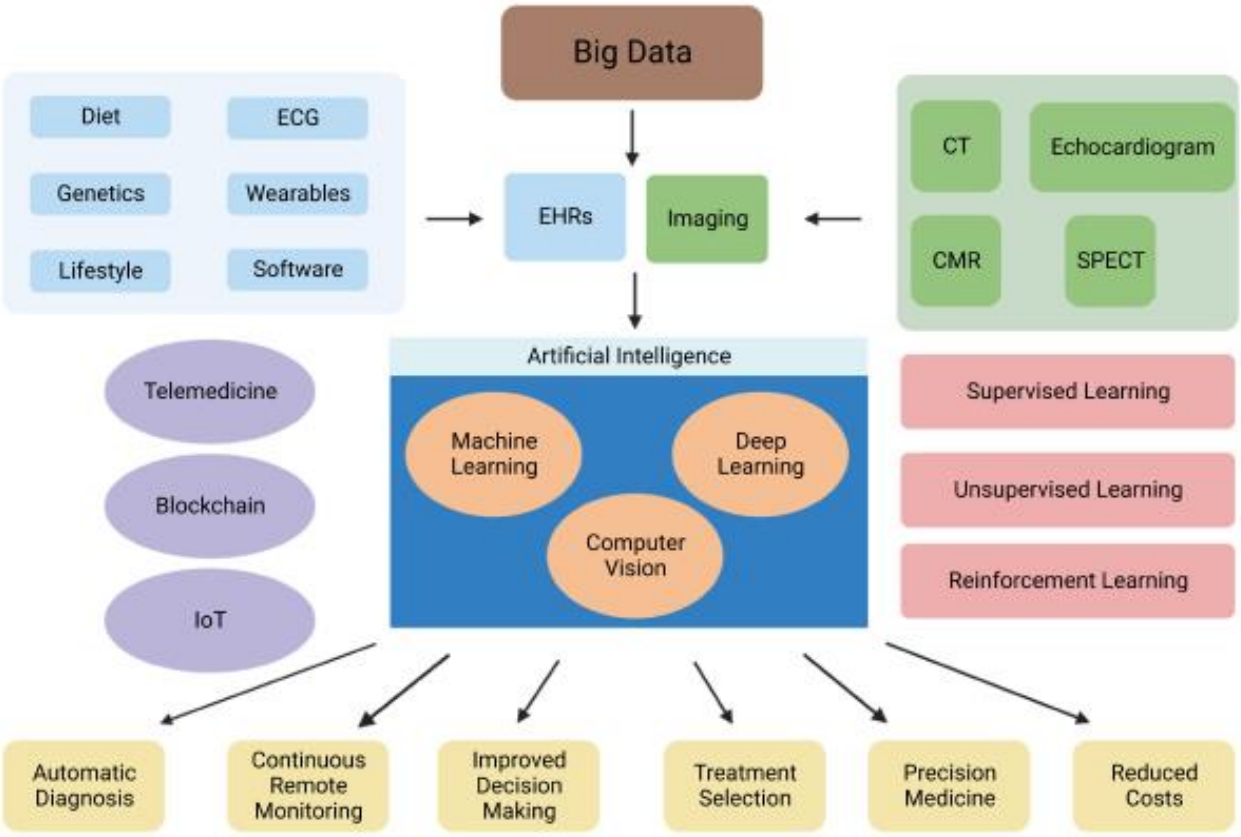



Figure (11) : Role of AI in cardiovascular disease (20)

# Problems & Aims

## Problem

Cardiovascular diseases (CVDs) or heart diseases are the leading international cause of human death. According to a report for World Health Organization (WHO) in 2019, approximately 18 million people died due to CVDs, representing 32% of total deaths. Among them, 85% were caused by heart failure and stroke.

Diagnosing heart disease not only takes a lot of time and effort, but also demands many investigations to diagnose these diseases early and correctly.

A significant number of patients with acute coronary syndrome(ACS) have been admitted every day to intensive care units (ICU) in hospitals, which require rapid and accurate intervention for the patient's survival. There is no prognostic knowledge about the development of the infraction, so either the patient's condition will stabilize and recover, or he will develop other complications such as ischemic heart attack or death, approximately half of myocardial infarctions (MIs) and strokes will occur in people suddenly without any prior symptoms Therefore, it is important to establish a system that can predict disease early.

A total of 25% of people die suddenly without any prior symptoms of heart disease. Therefore, it is important to establish a system that can predict heart diseases at an early stage.

Traditional diagnosis processes are costly, time-consuming, and often require human intervention. It requires medical tests, which low-income people often find expensive and difficult to afford.

Artificial Intelligence (AI) methods handle complex data and provide accurate risk-prediction models at the individual level**.** With the development of artificial intelligence, it was necessary to find an expert model capable of early prediction of these diseases and thus avoid these complications. That suggests that AI tool can be used as a clinical tool in the detection of CVD and will be particularly useful for physicians

## Aims:

- To apply advances in diagnostic uses of AI in cardiac diseases.
- Perform an intelligent model to predict CVD based on the given variable that related to risk factor, blood tests and cardiology markers and investigation.
- Perform a model by using three platforms (SPSS, Weka, Python) and different algorithms like MLP, SVM, LR, DT, ANN, KNN to get the best accuracy in predicting heart disease.
- To improve accuracy compared with other models from other studies.
- Determine the best features for early detection of disease.

# Methodology

## Datasets Source

- Data collection, different types of data mainly structured, collected from various sources.
- First one from IUC Machine Learning Repository. .https://archive.ics.uci.edu/ which has 4 databases but Cleveland Clinic Foundation dataset database is the most that has been used by ML researchers to this date. http://archive.ics.uci.edu/dataset/45/heart+disease.
- Second data is Statlog Data from Kaggle. https://www.kaggle.com/datasets/shubamsumbria/statlog-heart-data-set
- Third data:collection of Cleveland data and statlog data because they have the same features.
- Fourth data:Heart predicts data from Kaggle.
  https://www.kaggle.com/code/rafaelsakuma/heart-disease-

### Datasets Description

- Cleveland and Statlog dataset containing 14 variables: 13 features and 1 label. These features in relation to heart disease.
- They are: age, gender, Cp: chest pain type, trestbps: resting blood pressure, chol: serum cholesterol in mg/dl, fbs: fasting blood sugar > 120 mg/dl, restecg: resting electrocardiographic results, Thalach: maximum heart rate achieved, exang: exercise induced angina, oldpeak: ST depression induced by exercise relative to rest, slope: the slope of the peak exercise ST segment, ca: number of major vessels (0-3) colored by fluoroscopy, thal: thallium stress test result and finally diagnosis of heart disease or not.
- Cleveland data has 303 records in total, In the dataset 6 subjects have missing values, 4 values of ‘number of major vessels’ and 2 values of ‘thalassemia, so they are deleted from records to get more accurate data.
- Statlog data has 270 records in total and no missing values.

- We formed another data from Cleveland and Statlog to become 567 records.
- Heart predicted dataset 918 records but it has 11 features in common with cleveland and statlog dataset except the last two (ca.thal).

Table (1): Description for each feature for all datasets.

| N | | features | description | values |
|---|---|---|---|---|
| 1 | age | Age | Ages of patients taken in years | |
| 2 | gender | gender | 0 for female, 1 for male | 0, 1 |
| 3 | cp | Chest pain type | There are four types—1 for angina, 2 for atypical angina, 3 for non-angina pain, and 4 for asymptomatic angina. | 1, 2, 3, 4 |
| 4 | trestbps | Resting blood pressure | Blood pressure of the patient when at rest in mmHg. | |
| 5 | chol | Serum cholesterol | the amount of cholesterol in the blood in mg/dL. | |
| 6 | fbs | Fasting blood sugar | Amount of sugar present at fasting. 0 for false—fasting blood sugar is not above 120 mg/dL; 1 for true—fasting blood sugar is above 120 mg/dL. | 0, 1 |
| 7 | restecg | Resting electrocardiograph | Values produced by electrocardiography at rest. 0 is normal; 1 is having ST-T wave abnormality; 2 for showing probable or definite left ventricular hypertrophy. | 0, 1, 2 |

| | | | | |
|---|---|---|---|---|
| 8 | thalach | Maximum heart rate | Maximum heart rate of patients. | |
| 9 | exang | Exercise-induced angina | Whether or not the patient gets angina when exercise is performed. They are 0 for no and 1 for yes. | 0, 1 |
| 10 | old peak | ST depression | Finding on an electrocardiogram wherein the ST segment is abnormally below the baseline. | 1 to 3 |
| 11 | slope | Slope | the slope of the ST segment. 1 for up sloping, 2 for flat, and 3 for down sloping. | 1, 2, 3 |
| 12 | ca | Number of vessels | Number of vessels colored by fluoroscopy. | 0 to 3 |
| 13 | thal | Thallium stress test result | Thallium stress test is an imaging study of the flow of blood to the heart through coronary arteries. 3 is normal, 6 is a fixed defect, and 7 is a reversible defect. | 3, 6, 7 |
| 14 | | target(heart disease) | Predicted attribute that contains values 0: no presence or 1 presence of heart disease | 0,1 |

- **Method:**

Three platforms have been done for each type of data to compare which model achieves the best prediction of heart disease.

**First**: **Multilayer Perceptron (MLP) by SPSS Statistics 25**

It is used to build the neural network model and test its accuracy. After splitting data, training, all factors, covariates were normalized, target was determined, The result shows the percentage of correct and incorrect prediction for training and testing data Classification model, important features, Roc curve.

**Second: Decision tree (j48) logarithms used by Weka3.9.5**

Four measures were implemented to assess the performance of the classification models: accuracy, recall, precision, receiver operating characteristic (ROC), and area under the ROC curve (AUC). Accuracy represents the rate of correctness of a classifier. Therefore, we take the sum of true positive (TP) records and true negative (TN) records and then divide by the total number of records which represents the sum of TN, TP, false negative (FN) and false positive (FP); thus, accuracy denotes the ratio of the number of correctly predicted records to the total

**Third: Machine learning algorithms by python**

It involved working on Google Colab, a software platform used to write scientific code in Python to build various models. Three experiments were done by python.

1. Splitting data by 80% training ,20% testing for all datasets.
2. Splitting data by 70% training,30% testing for Statlog and sum of ( cleveland+statlog )
3. Perform accuracy for (Statlog and Cleveland) after deletion 2 features (Ca, thal) to compare it with heart data 918.

After the data segregation: the data are fed into various machine learning algorithms like Logistic Regression (LR) Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and K-nearest neighbor (KNN), XGBoost, Artificial Neural Network. (ANN). The learned model is tested using test data to check its accuracy.

# Results & Discussions

## Results:

Experiment by SPSS: Results of model by using MLP show model classification and Roc curve, sensitivity for diagnosis of disease. Also, it presents percentage of important features that are related to heart disease.

Table (2): Comparative performance of the training and testing accuracy using MLP.

| | **percentage** | **Training** | **Testing** |
|---|---|---|---|
| **Cleveland data 297** | **percentage** | 68.4 | 31.6 |
| | **Incorrect prediction%** | 12.8 | 14.9 |
| | **classification%** | 87.2 | 85.10 |
| | **ROC** | 0.93 | |
| **Statlog data 270** | **percentage** | 71.9 | 28.1 |
| | **Incorrect prediction%** | 14.4 | 7.9 |
| | **classification%** | 85.6 | 92.1 |
| | **ROC** | 0.93 | |
| **Cleveland Statlog 567** | **percentage** | 70.2 | 29.8 |
| | **Incorrect prediction%** | 8.8 | 12.4 |
| | **classification%** | 91.2 | 87.6 |
| | **ROC** | 0.96 | |
| **Heart 918** | **percentage** | 74.6 | 25.4 |
| | **Incorrect prediction%** | 14.9 | 14.4 |
| | **classification%** | 85.1 | 85.6 |
| | **ROC** | 0.91 | 0.91 |

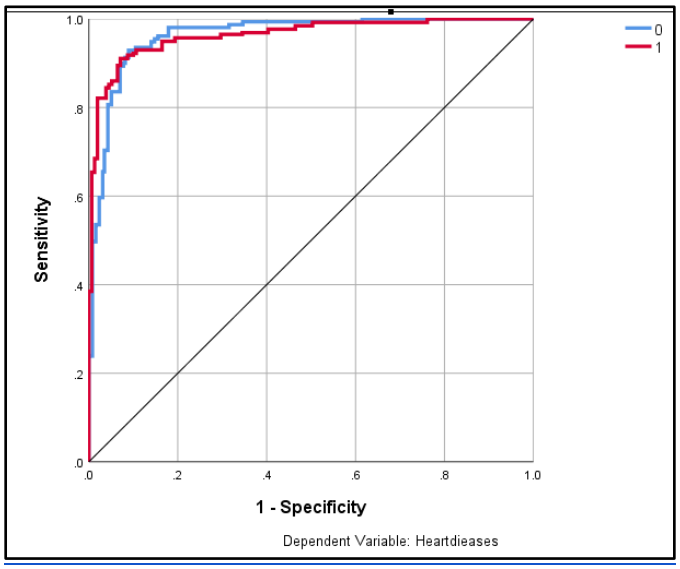



Figure (12): Roc curve for sum of cleveland and Statlog datasets

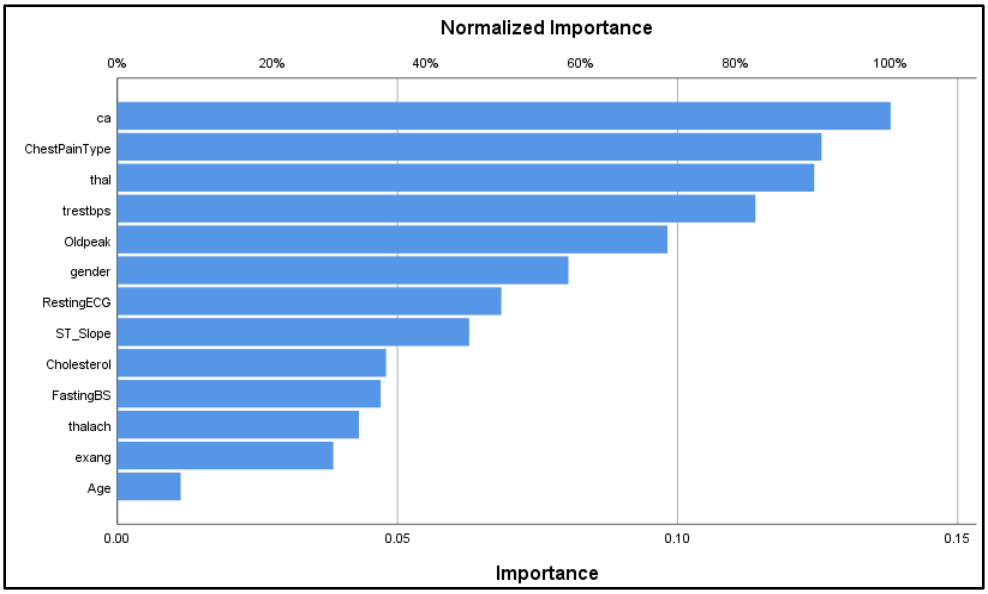



Figure (13): Important features of heart disease in Cleveland+statlog data by MLP

**Experience by Weka:**

Model was built by decision tree (J48) it show the correct and incorrect percentage, also many evaluation performances like, TP, TV, sensitivity, specificity, Re call, ROC, AUV, and confusion matrix(Tab.3)

Table (3): Results of model classification using Decision tree by Weka.

| | **Classification%** | | | | |
|---|---|---|---|---|---|
| | **correct** | **incorrect** | **Precision %** | **Recall%** | **Roc%** |
| **Cleveland data 297** | 77.7 | 22.2 | 77 | 77 | 76 |
| **Statlog data 270** | 76.6 | 23.3 | 76 | 76 | 74 |
| **Cleveland + Statlog 567** | 88.7 | 11.2 | 88 | 88 | 94 |
| **Heart data 918** | 84.4 | 15.5 | 84 | 84 | 85 |

Table (4): confusion matrix for CVD prediction using DT model for cleveland

| | Cleveland dataset | Predicted class | |
|---|---|---|---|
| | | Without CVD | With CVD |
| True class | Without CVD | 130 | 30 |
| | With CVD | 36 | 101 |

Table (5): confusion matrix for CVD prediction using DT model for Statlog data

| | Statlog dataset | Predicted class | |
|---|---|---|---|
| | | Without CVD | With CVD |
| True class | Without CVD | 88 | 32 |
| | With CVD | 31 | 119 |

Table (6): confusion matrix for CVD prediction using DT model for Cleveland and statlog

| | Cleveland + Statlog | Predicted class | |
|---|---|---|---|
| | | Without CVD | With CVD |
| True class | Without CVD | 281 | 29 |
| | With CVD | 35 | 446 |

Table (7): confusion matrix for CVD prediction using DT model for heart

| | Heart data918 | Predicted class | |
|---|---|---|---|
| | | Without CVD | With CVD |
| True class | Without CVD | 329 | 81 |
| | With CVD | 62 | 446 |

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         503               88.7125 %
Incorrectly Classified Instances        64               11.2875 %
Kappa statistic                          0.7718
Mean absolute error                      0.1227
Root mean squared error                  0.3071
Relative absolute error                 24.756  %
Root relative squared error             61.6891 %
Total Number of Instances              567

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.136     0.889      0.906    0.898      0.772    0.949     0.937     A
                 0.864    0.094     0.884      0.864    0.874      0.772    0.949     0.949     B
Weighted Avg.    0.887    0.117     0.887      0.887    0.887      0.772    0.949     0.942

=== Confusion Matrix ===

   a   b   <-- classified as
 281  29 |   a = A
  35 222 |   b = B
```

Figure (14): Performance evaluation of DT model for sum (Cleveland+ Statlog) data.

### Experiment by python:

Table (8): Accuracy for algorithms used in Machine learning by python

| logarithm | Cleveland data | Statlog data | Cleveland+ Statlog data | Heart data918 | Cleveland+ Statlog data after deletion |
|---|---|---|---|---|---|
| **Logistic Regression** | 81.67 | 83.33 | 84.21 | 83.15 | 78.9 |
| **Naive Bayes** | 85 | 74.07 | 85.09 | 83.15 | 78.07 |
| **Support Vector Machine** | 83.33 | 77.78 | 81.58 | 84.24 | 74.56 |
| **K-Nearest Neighbors** | 63.3 | 68.52 | 68.42 | 69.57 | 65.79 |
| **Decision Tree** | 75 | 85.19 | 96.49 | 78.8 | 96.49 |
| **Random Forest** | 83.33 | 85.19 | 98.25 | 88.04 | 98.25 |
| **XGBoost** | 73.33 | 74.07 | 98.25 | 84.24 | 96.49 |
| **Artificial Neural Network** | 81.67 | 81.48 | 83.33 | 85.33 | 75.44 |

Table (9): accuracy when splitting data 70%training and 30% testing

| logarithm | Statlog dataset | Statlog+Cleveland |
|---|---|---|
| **Logistic Regression** | 85.19 | 82.46 |
| **Naive Bayes** | 80.25 | 87.13 |
| **Support Vector Machine** | 82.7 | 82.46 |
| **K-Nearest Neighbors** | 67.9 | 70.76 |
| **Decision Tree** | 82.72 | 95.91 |
| **Random Forest** | 86.42 | 92.98 |
| **XGBoost** | 77.78 | 91.81 |
| **Artificial Neural Network** | 77.78 | 80.12 |

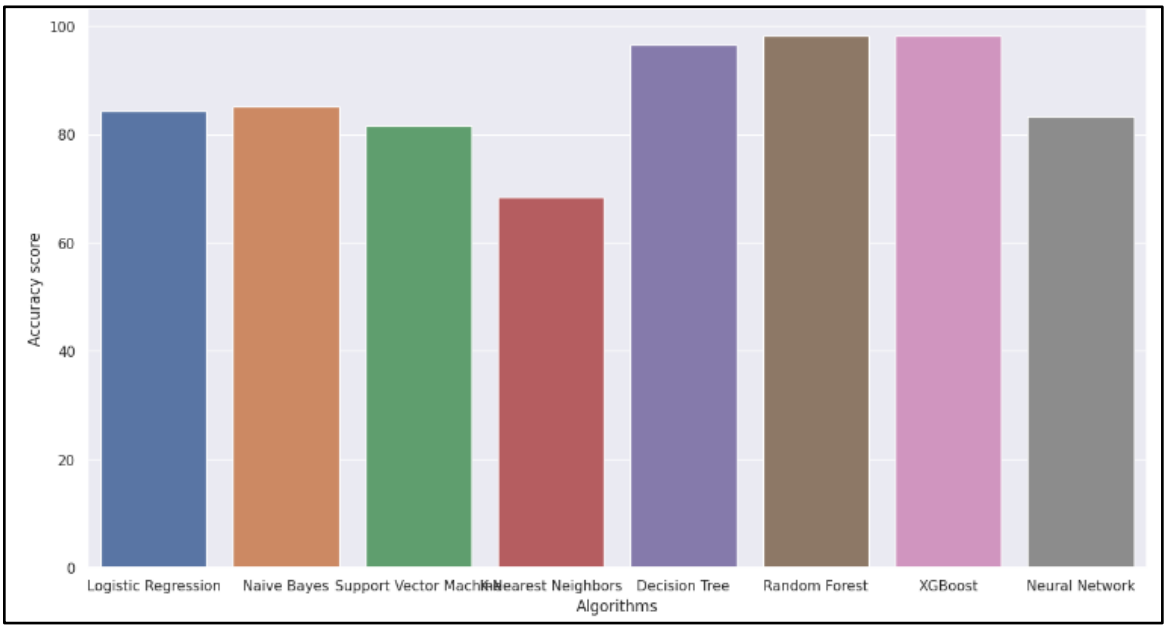



Figure (16): Model accuracy for prediction of heart disease in Cleveland and Statlog dataset

Experiment of model when splitting data 70% for training and 30% for testing.

Table (10): Accuracy of model for Statlog data when testing 30%

| | **Statlog data** |
|---|---|
| **logarithm** | **train70%test 30 %** |
| **Logistic Regression** | 85.19 |
| **Naive Bayes** | 80.25 |
| **Support Vector Machine** | 82.7 |
| **K-Nearest Neighbors** | 67.9 |
| **Decision Tree** | 82.72 |
| **Random Forest** | 86.42 |
| **XGBoost** | 77.78 |
| **Artificial Neural Network** | 77.78 |

Table (11): comparing the accuracy of the model in Cleveland and Statlog when changing the percentage of splitting.

| | **Cleveland +Statlog data** | |
|---|---|---|
| **logarithm** | **train80/test 20%** | **train70%/test30%** |
| **Logistic Regression** | 84.21 | 82.46 |
| **Naive Bayes** | 85.09 | 87.13 |
| **Support Vector Machine** | 81.58 | 82.46 |
| **K-Nearest Neighbors** | 68.42 | 70.76 |
| **Decision Tree** | 96.49 | 95.91 |
| **Random Forest** | 98.25 | 92.98 |
| **XGBoost** | 98.25 | 91.81 |
| **Artificial Neural Network** | 83.33 | 80.12 |

# Discussion

Nowadays, heart diseases are significantly contributing to deaths all over the world. Thus, heart-disease prediction has garnered considerable attention in the medical domain globally. Accordingly, Artificial Intelligence algorithms for the early prediction of heart diseases were developed in several studies to help physicians design medical procedures. Machine learning is a type of artificial intelligence that makes machines learn. There are various methods that can be adopted to predict the disease from training data and make predictions on the test data based on the learned data. The basic idea behind the ML is to find the patterns among the data and make the prediction (34).

In this study, 13 features are very important related to diagnosing heart disease. It includes laboratory tests like fasting glucose, cholesterol that are risk factors for CVD, in addition to gender and age.

Other kinds of medical data obtained from electrocardiography (ECG), echocardiography, or coronary angiography are known as evidence for cardiovascular diagnosis. That is why we use changes in ECG like ST-slope, Old peak as a feature.

Many models by different algorithms by 3 platforms have been done to get the best model and best data. The performance of the model is validated via evaluation metrics, accuracy, specificity, sensitivity, and Roc curve using two datasets from the University of California, namely, Cleveland and Statlog.

Firstly, we use the Cleveland dataset because it is used in many studies, so to compare our model with another study's model. When performing MLP algorithm, the accuracy of test data by MLP is 85.10% and training data 87.2%. Incorrect prediction is 14.9% for testing data and 12.8% for training data. Model can distinguish between healthy and heart disease about 93% by ROC curve. Compared to another study done by (Madhumita Pal,2022) using MLP the accuracy was 82.47% ROC curve 86.4%. That refers to our model being better in classification. (26)

By Decision tree(J48) algorithm, the model can classify 77.7% correctly and 22.2% incorrect. The accuracy is 77% and ROC 76%. Confusion matrix showed that the model can correctly predict 130 healthy from 160 and 30 incorrect. Also it can predict 101 as patient from 137 and 36 as healthy. When using google colab for coding by python language. The best model achieved using Naive Bayes 85% then 83.33% by SVM and RF. Uma Maheswari study by using ANN, accuracy is 84% higher than our model, but when we perform MLP accuracy 85% and sensitivity 94% is higher than Uma Maheswari study (34). Cleveland data shows that the MLP and Naive Bayes model predicts CVD more accurately.

(Pal M) study, two reliable machine learning techniques, multi-layer perceptron (MLP), and *K*-nearest neighbor (K-NN) have been employed for CVD detection using publicly available University of California Irvine repository data. Experimental-based results demonstrate that a higher accuracy in detection of 82.47% are obtained using the MLP model, unlike the K-NN model. The proposed MLP model is more efficient in CVD prediction when compared to other ML algorithms. the K-NN model with higher accuracy than our model. (26)

Fazl-e-Rabbi study: used multiple classifiers to predict heart diseases. The Cleveland dataset from UCI repository was used This study also used only 13 attributes of the dataset. Three different classifiers were used for the prediction of heart diseases i.e. SVM, ANN, and k-nearest neighbor. The classification accuracy with SVM was 85.18%. The accuracy value with KNN is 80.74%. Accuracy with the ANN was 73.33%. (35) compared with our study, our model's accuracy is higher in SVM.

Javad Hassannataj Joloudari study: this study uses the same data but differs in splitting data by 90% training and 10% testing , the accuracy was 91%,67% for RF,SVM respectively(5). Their accuracy is higher than our model in this data, that is related to splitting 90% and that may be unacceptable because overfitting may occur. So we tried to train a model for 90% and the accuracy is 100%, that is inaccurate.

Kartik Budholiya study work on XGBoost algorithms for Cleveland data, they have higher accuracy by his model 91% via 83% in our study, that related to using another technique (One-Hot (OH) encoding technique) to increase the accuracy. (27)

Table (13): Comparison of models of cleveland data with another studies

| Reference | Algorithms | Accuracy % | Our Study % |
|---|---|---|---|
| (26) | MLP | 82.4 | 85. |
| (26) | K-NN | 73.7 | 63.6 |
| (34) | ANN | 84 | 81.3 |
| (35) | SVM | 85.15 | 83.33 |
| (35) | K-NN | 80.7 | 63.33 |
| (35) | ANN | 70.3 | 81.67 |
| (5) | RF | 91 | 83.3 |
| (5) | SVM | 69.7 | 63.3 |
| (27) | XGBoost | 91 | 83 |

Second dataset is Statlog dataset: There are 270 records, by MLP the model accuracy in classification for testing 92.1% and it predicts 7.9% incorrectly and accuracy for training data is 85.6% and 14.4% in correct prediction. Evaluation performance by Roc curve is able to predict correctly 96%. Decision tree: the model classifies data 76.6% correctly and 23.3% in correct. Accuracy is 76% and Roc curve 74%.

By Python platform: The best model achieved by using decision tree and random forest (RF) 85.19% % then 81.48 by ANN. When the model trained 80%.

Another method for Statlog dataset the data were split for 70% training and 30%. The primary reason behind using this distribution was to compare our approach with those in other researches on the same dataset. We performed the same algorithms Random Forest, the accuracy is 86.42%, but another study by Mohamed G. El-Shafiey 87.7%. It is a little better than our approach that is related for more details in algorithms (36).

Comparing Statlog data accuracy between 20% and 30% for testing, the model accuracy differs between each kind of algorithm. It is better in (DT, RF, XGBoost, ANN) when using 20% for testing, but the accuracy in (LR, NB, SVM) is better when using 30% testing(Table.11). The best accuracy for (RF, XGBoost) for sum of two data (Cleveland+Statlog)

Table (14): comparing accuracy of model with other Statlog study

| | Algorithms | accuracy% |
|---|---|---|
| Statlog in our study | Random Forest | 86.42 |
| (36) | Random Forest | 87.4 |

**Third data: Total of cleveland +Statlog:**

Our study is the first study that sums two datasets, because they have the same features, so we sum them to get more records and study the effect of big data on the accuracy of the model. Indeed, it gets the best results for accuracy and sensitivity.

By MLP algorithm the model classification is the best model built by MLP. It can classify 87.6% for test data and 91.2% for training data and ROC curve 96% can predict if one is healthy or patient. When using decision trees by Weka also the best model is 88% the accuracy and ROC. It predicted 11.2% from testing data incorrectly and 88.7 % correctly. As in the confusion matrix the model can predict 281 healthy and 29 patients from 310 healthy and 222 patients as correct from 267 and 35 wrong predictions.

By python also the best data achieved this accuracy. Prediction accuracy is 98.25% and 98.25% for both Random forest and XGBoost. And 96.49% for the decision tree. And in all algorithms the accuracy is better than other datasets (Fig. 15)

From Kaggle which is a web for datasets, we choose dataset with 918 record because it is big data comparing with cleveland and statlog, but there is no two feature (Ca, thal), to know impact these two feature on model, by MLP the accuracy of classification is about 85% for testing and training. Roc curve 91%.

When using Decision Tree by Weka the accuracy was 84% with 15.5% incorrect classification.

By applying python algorithms, the best model 88.04% built by random forest, then 85.33% by ANN, 84.24%, 83.15% for SVM and Naive Bayes consequently. We compared this data with sum Cleveland and statlog after deletion the same feature, the model is better than the data 918

The accuracy is 98.25% by RF and 96.49 by XGBoost and DT. That means the (statlog data + Cleveland) also the best data for building models.

The important approach has done is identifying the best important features for early prediction, in Cleveland are ca, cp, ST-slope, thal, gender old peak , whereas cp, ca, thal, slop, gender in Statlog data, when we sum two dataset (Cleveland and Statlog) the best attribute are on consequently Ca, CP, thal, tresbps, Old peak...(Tab 15) In general they are in common between data, and with other studies like (S.Chellammal,2019) on Cleveland data whose purpose is to choose correlation between features to choose the best and they are the same features of our study and the same of Cleveland. (37)

As shown about important features in data 918 the order of features is not logical in medical science, for example the last feature is ECG in reality it is a very important tool to diagnose heart disease. That means this data is not good quality. Where other features order in other dataset is important for physicians to diagnose.

Table (15): Descending Order for features

| order | Cleveland | Statlog | Cleveland and Statlog | Heart data |
|---|---|---|---|---|
| 1 | Ca | Chest pain | Ca | ST-slope |
| 2 | Chest pain | Ca | Chest pain | cholesterol |
| 3 | St-slope | thal | thal | exangina |
| 4 | thal | ST-slope | trestbps | maxHR |
| 5 | gender | gender | oldpeak | Resting BP |
| 6 | oldpeak | Old peak | gender | Old peak |
| 7 | Rest ECG | trestbbps | Rest ECG | age |
| 8 | exang | Rest ECG | ST-slop | Chest pain |
| 9 | trestbps | exang | cholesterol | Fasting BS |
| 10 | Fasting BS | age | Fasting BS | gender |
| 11 | Age | thalah | thalach | Resting ECG |
| 12 | thalach | Fasting BS | exang | |
| 13 | cholesterol | cholesterol | age | |

The experimental results confirm that the Random forest approach attained the high heart-disease-prediction accuracies of 98.25% on sum the Cleveland and Statlog datasets. It is also important to note that predictions and classifications made by AI models are only as strong as the dataset quality and selected feature reliability

AI can become extremely useful for healthcare in the near future. It is technology that should be an additional tool assisting doctors. AI end goal should be to provide better quality of healthcare, reduce hospital and administration costs to make getting medical help cheaper and more accessible, for everyone, everywhere.

# Conclusion

The prediction of heart disease at an early stage can prevent many complications. The use of an efficient algorithm can help physicians in detecting the possible presence of heart disease before it manifests. This research focused on using much data for the early detection of heart diseases. Initially, the dataset collected 14 features, and different records. The preprocessed data were used with many algorithms MLP, DT, NB, SVM, KNN, RF, XGBoost on SPSS, Weka and Google Collab for python. The proposed model was evaluated regarding the performance metrics of accuracy, precision, Roc curve. For each data used there is a better model, MLP 85.10% for Cleveland data, MLP 92% for Statlog data, and many models for both Cleveland and Statlog like firstly RF and XGBoost, then DT. Also Important features for diagnosis were determined by MLP, they are the number of vessels(ca) and Chest pain(cp) and others.AI methods have shown to be a great interest and promising tool in healthcare. Recent studies have demonstrated that these methods can develop effective diagnostic and predictive tools to identify various diseases.

# Limitations

- Small data: Small data impact on the accuracy of mode. Machine learning models are best trained and have higher accuracy when using big data. datasets that could be much larger and much more detailed would result in better overall performance
- real-world data is not often available for global research purposes.
- Noisy data: Frequently, the clinical data contains noise or missing values; therefore, such kind of data give less accuracy
- Overfitting is one of the most important issues that need to be addressed when building an AI model. With overfitting, the model tries to fit the training data, so the accuracy is not reliable.

# References

Bibliography


1. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Ann Transl Med. 2016 Jul;4(13):256.
2. Jensen RV, Hjortbak MV, Bøtker HE. Ischemic heart disease: an update. Semin Nucl Med. 2020 May;50(3):195–207.
3. Lu L, Liu M, Sun R, Zheng Y, Zhang P. Myocardial infarction: symptoms and treatments. Cell Biochem Biophys. 2015 Jul;72(3):865–7.
4. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. Circulation. 2023 Feb 21;147(8):e93–621.
5. Joloudari JH, Joloudari EH, Saadatfar H, GhasemiGol M, Razavi SM, Mosavi A, et al. Coronary artery disease diagnosis; ranking the significant features using a random trees model. Int J Environ Res Public Health. 2020 Jan 23;17(3).
6. Wirtz PH, von Känel R. Psychological stress, inflammation, and coronary heart disease. Curr Cardiol Rep. 2017 Sep 20;19(11):111.
7. Jebari-Benslaiman S, Galicia-García U, Larrea-Sebal A, Olaetxea JR, Alloza I, Vandenbroeck K, et al. Pathophysiology of Atherosclerosis. Int J Mol Sci. 2022 Mar 20;23(6).
8. Rafieian-Kopaei M, Setorki M, Doudi M, Baradaran A, Nasri H. Atherosclerosis: process, indicators, risk factors and new hopes. Int J Prev Med. 2014 Aug;5(8):927–46.
9. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. J Cell Physiol. 2019 Aug;234(10):16812–23.
10. Holmes MV, Asselbergs FW, Palmer TM, Drenos F, Lanktree MB, Nelson CP, et al. Mendelian randomization of blood lipids for coronary heart disease. Eur Heart J. 2015 Mar 1;36(9):539–50.
11. Hanson MA, Fareed MT, Argenio SL, Agunwamba AO, Hanson TR. Coronary artery disease. Prim Care. 2013 Mar;40(1):1–16.
12. Nelson AJ, Ardissino M, Psaltis PJ. Current approach to the diagnosis of atherosclerotic coronary artery disease: more questions than answers. Ther Adv Chronic Dis. 2019 Nov 1;10:2040622319884819.

13. Birnbaum Y, Wilson JM, Fiol M, de Luna AB, Eskola M, Nikus K. ECG diagnosis and classification of acute coronary syndromes. Ann Noninvasive Electrocardiol. 2014 Jan;19(1):4–14.

14. Coronary angioplasty and stents (PCI) - BHF [Internet]. [cited 2023 Jul 5]. Available from: https://www.bhf.org.uk/informationsupport/treatments/coronary-angioplasty-and-stents

15. Uzun Ozsahin D, Ozgocmen C, Balcioglu O, Ozsahin I, Uzun B. Diagnostic AI and cardiac diseases. Diagnostics (Basel). 2022 Nov 22;12(12).

16. Zhang C, Lu Y. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration. 2021 Sep;23:100224.

17. Yan Y, Zhang J-W, Zang G-Y, Pu J. The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine? J Geriatr Cardiol. 2019 Aug;16(8):585–91.

18. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14.

19. Arooj S, Rehman SU, Imran A, Almuhaimeed A, Alzahrani AK, Alzahrani A. A deep convolutional neural network for the early detection of heart disease. Biomedicines. 2022 Nov 3;10(11).

20. Karatzia L, Aung N, Aksentijevic D. Artificial intelligence in cardiology: Hope for the future and power for the present. Front Cardiovasc Med. 2022 Oct 13;9:945726.

21. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020 Oct;92(4):807–12.

22. Relationship between artificial intelligence, machine learning, neural... | Download Scientific Diagram [Internet]. [cited 2023 Jun 27]. Available from: https://www.researchgate.net/figure/Relationship-between-artificial-intelligence-machine-learning-neural-network-and-deep_fig3_354124420

23. Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3).

24. Pal M, Parija S. Prediction of Heart Diseases using Random Forest. J Phys: Conf Ser. 2021 Mar 1;1817(1):012009.

25. Fan G-F, Guo Y-H, Zheng J-M, Hong W-C. Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting. Energies. 2019

Mar 9;12(5):916.

26. Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022 Jun 17;17(1):1100–13.

27. Budholiya K, Shrivastava SK, Sharma V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University - Computer and Information Sciences. 2022 Jul;34(7):4514–23.

28. Singh P, Singh S, Pandi-Jain GS. Effective heart disease prediction system using data mining techniques. Int J Nanomedicine. 2018 Mar 15;13(T-NANO 2014 Abstracts):121–4.

29. Owusu E, Boakye-Sekyerehene P, Appati JK, Ludu JY. Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine. Comput Intell Neurosci. 2021 Dec 23;2021:3152618.

30. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013;4(2):627–35.

31. Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. Data Democracy. Elsevier; 2020. p. 83–106.

32. Koulaouzidis G, Jadczyk T, Iakovidis DK, Koulaouzidis A, Bisnaire M, Charisopoulou D. Artificial Intelligence in Cardiology-A Narrative Review of Current Status. J Clin Med. 2022 Jul 5;11(13).

33. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc. 2020 May;95(5):1015–39.

34. Maheswari Mrs. Neural Network based Heart Disease Prediction.

35. Rabbi Md, Uddin Md, Ali Md, Kibria Md, Afjal M, Islam Md, et al. Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction.

36. El-Shafiey MG, Hagag A, El-Dahshan E-SA, Ismail MA. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. Multimed Tools Appl. 2022 May;81(13):18155–79.

37. Recommendation of Attributes for Heart Disease Prediction using Correlation Measure. ijrte. 2019 Aug 10;8(2S3):870–5.