Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Early Detection of Cardiovascular Disease Utilizing Machine Learning Techniques: Evaluating the Predictive Capabilities of Seven Algorithms

Maad M. Mijwil[1]*, Alaa K. Faieq[2], Mohammad Aljanabi[3,4]

[1]Computer Techniques Engineering Department, Baghdad College of Economic Sciences University, Baghdad, IRAQ
[2]College of Engineering, University of Information Technology and Communications, Baghdad, IRAQ
[3]Department of Computer, College of Education, Aliraqia University, Baghdad, IRAQ
[4] Imam Ja'afar Al-Sadiq University, Iraq

*Corresponding Author: Maad M. Mijwil

DOI: https://doi.org/10.52866/ijcsm.2024.05.01.018
Received May 2023 ; Accepted August 2023 ; Available online February 2024

ABSTRACT: Heart disease is the leading cause of death in developed countries, as it causes many deaths annually. Despite the availability of effective treatments, heart disease remains a significant challenge to public health, so early detection is essential in enhancing patient outcomes and reducing mortality. Artificial intelligence seeks to help physicians make the right decisions about a patient's health condition. In this regard, the authors decided to utilize machine learning techniques (k-nearest neighbor, decision tree, linear regression, support vector machine, naïve bayes, multilayer perceptron, random forest) to contribute to the classification of the heart disease dataset, where it is determined whether a person is suffering or not. After that, the execution of all techniques will be measured, and the accuracy of each technique will be compared to determine the most suitable performer. The public dataset is organized from the UC Irvine machine learning repository and have significantly different characteristics. The dataset will be divided such that 80% of the data is designated for training and 20% is designated for testing. This article concluded that the adequate performance is for the multilayer perceptron technique, as it gained an accuracy of more than 88%, while the poor performance is for the decision tree technique, as it gained an accuracy of more than 79%.

Keywords: Artificial intelligence, Heart, Cardiovascular disease, Machine learning, UC Irvine, Accuracy.

1. INTRODUCTION

Today, cardiovascular diseases occupy first place among the most dangerous diseases in this universe as well as the death rate is rising at an alarming rate [1-3]. The prevalence of heart disease is caused by harmful behaviors sustained over extended periods such as lack of sports activity, malnutrition, and the use of toxic substances such as drugs and tobacco (smoking) [4-6]. Despite widespread awareness efforts, the habits mentioned earlier are widespread in our society and result in a high mortality rate. Older adults are more susceptible to cardiovascular disease because age plays a major role in weakening the role of the cardiovascular system [7][8]. Therefore, these diseases are getting more severe with the advancing individuals' age. In 2019 [9], the American Heart Association conducted a study on heart disease and stroke and found that the incidence of these conditions was highest in older age groups. People between the ages of 40 to 60 had an average incidence of 35-40%. For those between 60 to 80, the average incidence was 75-78%. And for those over 80 years old, the incidence was alarmingly high, at over 80%. The literature have shown that there is a marked disparity between the sexes in terms of the occurrence of heart disease [10-13]. This is due to the impact of sex hormones and the higher occurrence of metabolic syndrome in women. The website "*Our World in Data*" has published statistics from the years 1990 → 2019 indicating that cardiovascular diseases are the most prevalent type of illness (as shown in Figure 1) [14].

The most typical heart diseases are aneurysms, heart ischemia, arrhythmias, and heart failure. Changes in the regular pattern or sequence of electrical impulses generated by the heart indicate abnormal behavior or a defect in the heart, where an irregular heartbeat occurs, and the patient may feel that the heartbeat is accelerating or slowing down [15][16]. This disease can be classified into two types; the first type poses a threat to the patient's life and requires medical intervention from healthcare workers. The second type does not pose a danger but rather requires conducting a set of diagnostics to find out the leading cause of the disease, following the physician's recommendations, and adhering to treatment and psychological comfort. An electrocardiogram (ECG) is a widely utilized test to detect heart problems

and irregular heartbeats in patients, as it records the timing and strength of electrical signals. Through this technique, the heart's electrical activity is demonstrated through a group of electrodes placed on the patient's skin [17][18]. Marks showing the presence of some arrhythmia may appear rarely, as it requires monitoring the patient's heart activity for long periods of time and making an appropriate diagnosis by experts. Tracking the patient is accomplished through wearable devices and sensors that allow them to be monitored without disrupting their daily routine. For instance, Nuubo created a compact ECG patient monitor (see Figure 2) that adjusts to each individual and captures heart electrical activity without any adhesive or wires [19]. This approach is significantly improved compared to traditional methods.

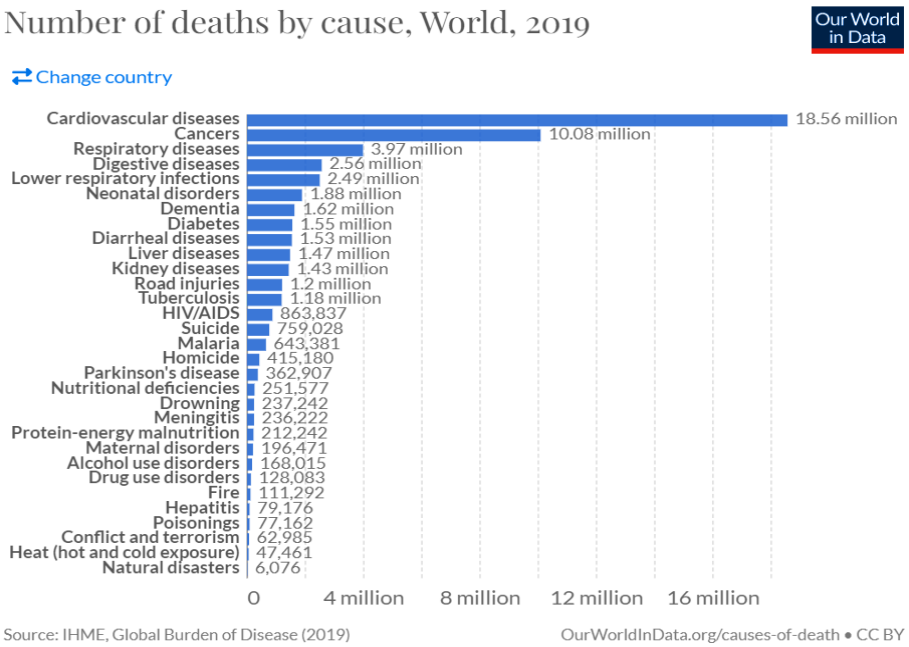


FIGURE 1. – Disease statistics from 1990 to 2019

Machine learning is a crucial component of artificial intelligence and involves the use of algorithms and statistical models for performing tasks such as prediction and decision-making based on data [20-22]. Artificial intelligence of this type seeks to play a significant role in the healthcare industry to improve patient outcomes, streamline operations, and help physicians and healthcare workers in making informed decisions [23][24]. Diagnostics is a critical application of machine learning in healthcare. The algorithms of machine learning strive to identify new patterns in medical images or data through training, assisting in the diagnosis of diseases like heart disease [25], viral pneumonia [26], and cancer [27]. By utilizing these algorithms, a prediction can be made, the progression of the disease can be determined, the most appropriate treatment can be selected for the patient, and physicians can be aided in making decisions and determining the patient's health status. Physicians and healthcare professionals are seeking to take advantage of the capabilities of machine learning to personalize treatments according to the unique genetic profile of each patient. These algorithms possess strong abilities in processing vast amounts of genomic data to identify genetic markers for various diseases and create personalized treatments, as demonstrated during the COVID-19 pandemic when artificial intelligence was utilized by specialized companies in the development of vaccines. Arshadi et al. [28] highlighted recent advancements in the use of artificial intelligence techniques for the creation of drugs and vaccines for COVID-19, emphasizing the employment and training of deep learning-based models. They also emphasized the acceleration of the discovery of effective treatments for the virus through the identification of its multiple molecular targets. Bagabir et al. [29] performed an investigation exploring the use of AI techniques for determining the viral genomic sequence and its application in developing vaccines for COVID-19 patients. Therefore, it is evident that machine learning plays a role in the discovery and production of medicines and vaccines, where machine learning algorithms can be trained to anticipate the effects of these treatments based on partial structure and other features. Besides, machine learning is being operated to process vast amounts of clinical trial data to discover treatments, improve patient outcomes, and plan precise clinical trials. Other usefulness of machine learning in healthcare includes optimizing hospital resource allocation, reducing wait times, and improving appointment scheduling efficiency. Furthermore, machine learning techniques are involved in analyzing medical and electronic records and finding patterns that can be utilized in patient treatment to achieve better results [30-32]. The main contribution of this research study is as follows:

1- The medical data employed for training algorithms are obtained from the UC Irvine machine learning repository, which is a trustworthy source of information.
2- During the pre-processing stage, a procedure is executed to eliminate duplicates from the obtained data.

3- Seven algorithms (K-NN, DT, LR, SVM, NB, MLP, and RF) are utilized for both training and testing on the medical data and data separation.
4- The results are analyzed utilizing different evaluation metrics, and the execution of the seven algorithms is evaluated based on their ability to predict patient and non-patient data.

FIGURE 2. - Nuubo System Overview [downloaded from Google].

The remainder of this article is framed as follows: In Section 2, the author examines relevant literature from other authors. In Section 3, they give a brief explanation of the data set used and highlight the significance of the algorithms employed. The results of the algorithms are presented and evaluated in Section 4. Finally, the article concludes by summarizing the conclusions and proposing future directions in Section 5.

2. LITERATURE SURVEY

This section of the article reviews a collection of literature that applies machine learning techniques to predict heart disease and discusses the findings achieved by these literature. Accuracy metric is one of the most critical aspects that is primarily focused on in current studies. This metric evaluates the performance and practices of classification algorithms and identifies the most adequate algorithm. Accuracy is the number of predictions that the algorithm got correctly, that is, the number of proper predictions. The starting point would be in a study conducted by Hagan et al. [33]; they involved various machine learning techniques such as support vector machines, multi-layer perception Neural Networks, random forest, extra trees, gradient boosting, and bagging to classify cardiovascular disease from datasets obtained from datasets collected from UCI Machine Learning repository and Kaggle platform. This study found that the most practical performance technique is the random forest, which gained an accuracy of 95% with UCI data and 74% with Kaggle data. In another analysis [34], Arunachalam and Rekha used a combination of machine learning techniques such as ensemble X-boost, k-Nearest Neighbor, random subspace, Adaboost, and linear support vector feature measure to predict heart disease from a dataset sourced from the UCI Machine Learning repository. This study, executed within the MATLAB environment, found that the linear support vector feature measure (SVFM) technique performed the most practically, earning an accuracy of 96% and 93% of Matthew's correlation coefficients. Saboor et al. [35] evaluated various machine learning techniques such as AB, LR, ET, MNB, CART, SVM, LDA, RF, and XGB to predict human heart disease using three datasets (UCI, StatLog, and Z-Alizadeh Sani). The effects of the tests revealed that the support vector machine (SVM) technique performed the most suitable, achieving an accuracy metric of over 96%. Sajeev et al. [36] executed a study in which they utilized deep learning techniques (two linear and two nonlinear) to predict heart disease using a dataset from the UCI repository, which comprises more than 560 patients. The study achieved an accuracy of 94% and an AUC score of 96%. This research demonstrates that the performance of deep learning and nonlinear machine learning models is superior to linear machine learning models. Likewise, Akella and Akella [37] have applied six machine learning techniques (generalized linear, decision tree, random forest, support vector machine, neural network, and k-nearest neighbor) in detecting coronary artery disease using data from the UCI repository and predicting the presence of the disease. The authors concluded that the most acceptable technique in terms of performance was the neural network, which achieved an accuracy of more than 93% but with an AUC score of 79.6%. In another study accomplished by Shehzadi et al. [38] about the diagnosis of chronic ischemic heart disease using three machine learning techniques (naïve bayes, logistic regression, and random forest). This study found that random forest is the most effective technique, achieving an accuracy metric of 99%.

3. MARTIALS AND METHODS

This section will be divided into two parts. The first part will cover the dataset used and its significance. The second part will briefly describe the methods used to classify this dataset.

3.1 DATASET DESCRIPTION

In this article, a dataset is collected from the UC Irvine machine learning repository, which is open-source and available to all [39]. The dataset includes 76 attributes, and the number of instances is 303. The published literature indicates the use of a subset of only 14 attributes. The target field indicates the presence of heart disease. The attribute No. 14 will be considered to detect the disease's presence. A value of zero is assigned in the absence of a disease, while a value of 1 is given in the presence of a disease. Table 1 illustrates the details of the dataset used in this test. Moreover,

the dataset includes 160 patients with no heart disease results and 140 patients with heart disease results. Figure 3 illustrates the distribution of data for 300 patients.

Table 1. – The fourteen attributes with their significances.

S. No.	Attributes	Representative icon	Significances (Range)	Type
1	Age	Age	Patient Age in Years (29-71)	Numeric
2	Sex	Sex	Male=1 Female=0	Binary
3	Chest pain	Cp	Angina pain=0 Typical angina pectoris=1 non-angular pain=2 Asymptomatic=3	Categorical
4	Rest blood pressure	Trestbps	94-200	Numeric
5	Serum cholesterol	Chol	126-564	Numeric
6	Fasting blood sugar	Fbs	True=1 & False=0	Binary
7	Results electrocardiographic in repose	Restecg	Normal=0 Abnormality=1 Hypertrophy=2	Categorical
8	Maximum heart rate reached	Thalach	71-202	Numeric
9	Induced angina exercise	Exang	Yes=1 & No=0	Binary
10	Previous peak	Oldpeak	0-6.2	Numeric
11	ST segment slope of maximum exercise	Slope	Upsloping=0 Flat=1 Downsloping=2	Categorical
12	Frequency of fluoroscopy	Ca	0-3	Categorical
13	Infarction classification cardiac	Thal	Normal=0 Fixed default=1 Default=2	Categorical
14	Presence or not of heart disease	Num (Main Factor)	**Prediction Variable** **0** : No Heart Disease , **1** : Heart Disease Available	Categorical

Notice: All details in this table have been taken from the same repository.

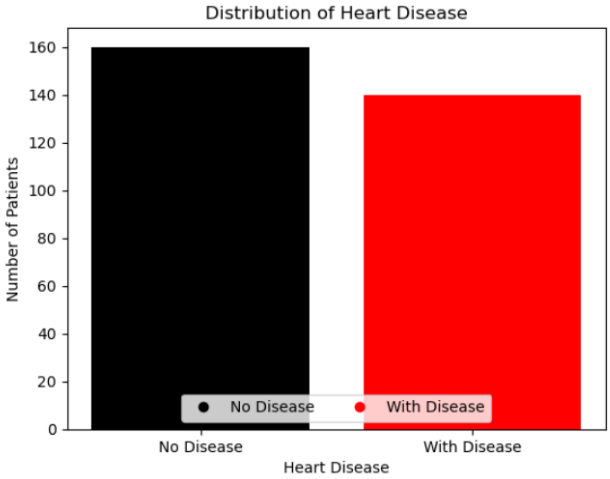


FIGURE 3. - Patients without disease and with disease.

3.2 MACHINE LEARNING TECHNIQUES

Machine learning is the most significant part of artificial intelligence that enables systems and applications to learn and enhance the experience without being explicitly programmed [40][41]. It relies on utilizing a set of algorithms and statistical measures to analyze and understand patterns in data of any size and achieve a set of processes, including predictions and diagnosis. These algorithms are influential in support experts in making the right decisions and give satisfactory effects. In this work, seven algorithms are employed for the early detection of heart disease, which are as follows:

- **K-nearest neighbor**

K-nearest neighbor (KNN) is a prevalent machine learning algorithm for classification and regression problems. This algorithm works by giving a class to an item based on the most similar items it has been trained with. The similarity is determined by the closeness of each item's training labels to the current item's label. For instance, the procedure was able to obtain the name and address of a new customer, along with a list of customer addresses labeled "happy" or

"unhappy". This algorithm can be utilized to find the closest customers from a customer list and assign a customer an 'unhappy' rating based on their similarity to unsatisfied customers. Hence, this algorithm has a vital role through a mathematical method to identify people who are likely to become unhappy by training it on a set of data, identifying patterns (unhappy people) and storing the results in a unique repository for these patterns. This approach can also be utilized for fraud detection [42], disease prediction [43][44], inventory management [45], and other classification problems. Recently, the K-NN algorithm has gained popularity for predicting heart disease. A study performed by Tak et al. used this algorithm in combination with statistical methods to predict heart failure disease for 299 patients with only five attributes (gender, smoking, diabetes mellitus, hypertension, and anemia). Their study reached a prediction accuracy of over 89% while an AUC of 93%. [46].

- **Random forest**

The random forest algorithm is one of the machine learning algorithms that seek to combine many weak learners based on decision trees to enhance generalization ability and is trained through bagging or bootstrap aggregating. This algorithm is utilized for classification (discrimination) and regression (prediction). Also, it makes many weak learners (in parallel) (decision trees) and computes the mean (or majority vote) of the output scores for each weak learner to make predictions. It is distinguished by its low computational cost because it can create several trees simultaneously and is characterized by efficiency and more accuracy than the decision tree algorithm. It has practical strategies for dealing with null values. In a study conducted by Mijwil et al. [47], they improved the performance of the random forest algorithm for imputation null values from different datasets from the UC Irvine machine learning repository. This study concluded that the ability of this algorithm to deal with null values, whatever the dataset. Moreover, it can make a reasonable prediction without hyperparameter tuning. The ability to solve the problem of overfitting in decision trees and eliminate unnecessary data with the ability to select a subset of features when splitting a node randomly.

- **Linear regression**

Linear regression is a popular statistical method for estimating the relationship between two unknown variables. The algorithm begins by creating a linear regression equation in which the unknown variable y is predicted by the known variables $x1$ and $x2$. Then, the algorithm calculates a predicted value for y based on the value of $x1$ and $x2$. The line of best fit is represented by an equation of the form $Y = a + bX$, where y is the dependent variable, x is the independent variable, a is the y-intercept, and b is the slope of the line. Also, it compares this predicted value with the observed value of y. If a difference between the predicted value and the observed value is large, then the algorithm concludes that there is a positive relationship between the two variables. If not, then there is a negatively related relationship. Regression models are commonly classified by the type of function used to represent the relationship between the response and the predictor variables. This algorithm can be operated for both simple and multiple regression analysis. In this work, linear regression was preferred over the logistic regression algorithm, because this algorithm has the ability to find the find the best-fitted line for separating two or more variables, as well as predicting the continuous dependent variable, unlike logistic regression, which has the ability to predict the categorical and separating discreet values.

- **Naïve bayes**

Naïve bayes is a family of probabilistic algorithms based on the Bayes Theorem, which remarks that the probability of an event happening is equal to the likelihood of the event multiplied by the prior probability of the event happening. In machine learning, these algorithms are often utilized for classification tasks. It is a superficial learning algorithm that provides mostly competitive classification accuracy. It operates on probability models that are derived based on the nature of the probability model, which includes strong independence assumptions. This algorithm has computational efficiency with many benefits and can be utilized in many tasks. Moreover, this algorithm is trained in a supervised learning setting, and it assumes that the attributes are conditional and class-based, i.e., each pair of attributes is classified as independent of the other. Therefore, this algorithm is considered one of the algorithms that has an important and vital role in analyzing the datasets by using Bayesian theory to calculate the posterior probability of event (A) given the prior probability of event (B) as shown in the mathematical formula (1)

$$P(y) = P(C_i|x_1, x_2, x_3, \ldots\ldots x_n) = \frac{P(x_1, x_2, x_3, \ldots\ldots x_n|C_i).P(C_i)}{P(x_1, x_2, x_3, \ldots\ldots x_n)} for\ 1 \leq i \leq k. \qquad (1)$$

- **Decision tree**

The decision tree algorithm is based on tree structure construction and is utilized for classification and regression assignments in analyzing a dataset and assisting in decision-making. It is a supervised algorithm with a predefined variable that is applied to the input and output variables, whether they are categorical or continuous. It has a flowchart-like structure where each internal (non-leaf) node has an attribute or feature, each branch represents a decision, each leaf node is the test result, and each leaf node (terminal) describes the class label. For the top node, which is the root node (uppermost node), the decision is represented by a path from the root node to the leaf node. This algorithm makes

the tree by repeatedly selecting the most useful feature to divide the data into two or more classes at each node. This algorithm makes the tree by recursively selecting the most reasonable feature to split the data at each node. The feature that results in the most homogeneous subsets of the data is determined as the splitting feature. This procedure is repeated on each subset until the stopping benchmark is achieved. The main objective is to create a predictive model that can see all observations about the sample (the branches) and make judgments about the target value of the sample (the leaves).

- **Multilayer perceptron**

The multilayer perceptron algorithm is supervised learning as it consists of an input layer, one or more hidden layers, and an output layer. This algorithm is used to produce the final prediction or decision from the input data, and each is classified with its own probability value. A meaningful query is how this algorithm is executed. To demonstrate the answer, this algorithm includes a hidden layer in which the gathered data is processed and crucial and relevant features are extracted. After that, it determines the probability of appropriate effects for each data category, employing a neural network consisting of a certain number of input and output nodes to track the patterns in the data. Besides, this algorithm has an activation function to execute calculations, and then each node is linked to its most immediate neighbors within the same network while feeding training data. The more training is accomplished, the more robust the network becomes in assigning probabilities to the input data so that it can ultimately make predictions with high accuracy and efficiency. The following points illustrate the features of this algorithm:

- MLPs can compare any nonlinear function employing mathematical strategies.
- This algorithm includes multiple layers of neurons, enabling it to learn and describe the features of the data set
- This algorithm uses backpropagation to train the neural network and improve implementation.
- MLP uses gradient descent as an optimization algorithm to reduce errors in the output stage.
- In this algorithm there are activation functions such as sigmoid, tanh and ReLU
- Also, this algorithm has many parameters that must be trained to work perfectly.
- This algorithm can approximate any process if the number of neurons and layers is large enough.

- **Support vector machines**

This algorithm is one of the most widely applied machine learning algorithms in many fields, as it contributes to the decision-making process by mathematical practices [48-50]. This algorithm classifies the data set and is subject to supervision. In addition, it combines a large number of features. It is optimized for non-linear data utilizing kernel trick, which locates the data in a higher dimensional space where linear bounds can be used in performance. Its high speed and adequate performance characterise this algorithm compared to neural networks with limited samples. Thus, this algorithm is the most suitable for text classification. Moreover, it is known for its high accuracy and ability to manage extensive data through specific work steps. This algorithm is widely employed in many applications, such as natural language processing, bioinformatics, and image recognition. Equation (2) denotes the classification model function through which the discriminant feature is employed to determine the classification model.

$$f(X_i) = [WX_{i+}b] \qquad (2)$$

Where:

$X_i \rightarrow$ training /test pattern

$W \rightarrow$ weight vector value

$b \rightarrow$ bias correction factor

The total of the product line of vector elements can be specified by operating the input space and vector combination $[W_i * X_i]$. Equation 3 illustrates the regression model for two functionalities in the standard:

$$f(X_i) = [W_1 * X_1 + W_2 * X_2 + b] \qquad (3)$$

After the training stage, the SVM supplies estimates for W_1, W_2, and b.

4. PERFORMANCE COMPARISON

This section will discuss the algorithms' performance and ability to classify the dataset. The focus will be on the performance of the algorithms to determine the most acceptable mathematical structure that has the ability to separate the dataset. The MATLAB 2019b platform is employed because it includes a set of machine learning and deep learning tools (these tools are statistics, classification, regression, clustering, and creating deep neural networks) with acceptable ability to train algorithms on data and analyze it. The primary goal of this platform is to help programmers design and improve machine learning models using modern methods and getting rid of traditional programming methods. In this work, we relied on computer specifications that include the following:

- Windows 10 pro.
- 8-GB RAM
- Intel(R) Core i5-1135G7 (2.40GHz)

- NVIDIA GeForce (MX350) and 2-GB RAM.

The performance of the involved algorithms (accuracy, sensitivity, specificity, precision, F1-score, AUC-ROC) mainly relies on the five necessary metrics, using a set of mathematical equations (4 to 8) mentioned below:

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \quad (4)$$

$$Sensitivity = \frac{TP}{TN+TP} \quad (5)$$

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$F1-score \ \ is\ estimated\ as = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (8)$$

The sample is divided into four natural categories mentioned in the above equations to compare the pros and cons of the algorithms in the performance, which are: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). In this work, the ratio of the training set to the prediction set is 80:20, where the data from the training set are employed for training as well as the trained model and the prediction set are employed as inputs for the prediction component to determine the algorithms path in the classification process. In addition, the AUC-ROC curve is employed to find classification problems in various configurations based on the threshold value. AUC is the area under the ROC curve, and it means the measurement of a two-dimensional area under the ROC curve, while ROC is an analytical method, and it represents the probability curve. This indicator is the ability of the algorithm to distinguish between categories. Therefore, the more increased the AUC, the more useful the model distinguishes between patients with and without heart disease. Figure 4 illustrates the steps for carrying out this work. In the ROC curve, the x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR). They can be calculated from the following equations:

$$True\ Positive\ Ratio = Sensitivity = \frac{TP}{TP+FP} \quad (9)$$

$$False\ Positive\ Ratio = 1 - Specificity = \frac{TN}{TN+FP} \quad (10)$$

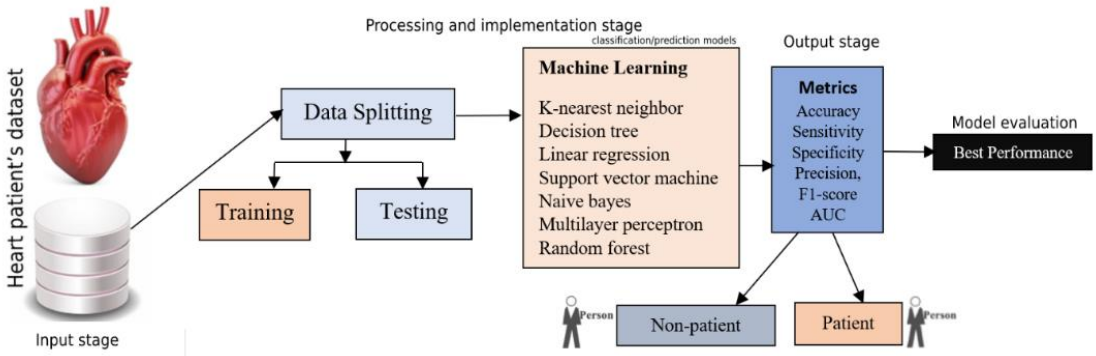


FIGURE 4. - Workflow stages.

To explain the steps, starting from the information generated based on the symptoms in the dataset to make a diagnosis of heart patients, which would serve as inputs to the applied algorithms. Then, a confusion matrix is introduced to check the level of accuracy in data classification for each algorithm, and its performance is measured employing an AUC-ROC curve. A Cross-validation value ($k = 5$) is utilized in modeling all algorithms with training datasets. Now, the results obtained by applying the algorithms to the testing data will be reviewed (see Figures 5-11). Table 3 fully illustrates all the effects obtained. Figure 12 illustrates a comparison of the performance of the algorithms through accuracy and AUC only. Through all the below practices, it is clear that the most suitable performance is the multilayer perceptron algorithm. The algorithm performed well in terms of accuracy, specificity, precision, and F1-score, and adequately in terms of sensitivity. Additionally, the algorithm yielded an excellent result with regards to the AUC metric, although it was not the highest among other algorithms (LR, NB, and KNN), which achieved a better data separation ratio. Algorithm 1 illustrates the multilayer perceptron algorithm's steps in training the dataset. The prediction speed of each algorithm is demonstrated in Table 2 and KNN has the quickest training time.

Algorithm 1**:** The steps for training a multilayer perceptron (MLP) algorithm

Start

*− **Import** model Libraries/**Step 1***

*−**Input**: train_data, test_data*

*−**Initialize** the weights of the network randomly.*

*−**Feed** the input data into the input layer of the network.*

−Perform a dot product of the input layer with the first set of weights, and then pass the result through an activation function.

−Perform the same dot product and activation function operation on the result from the previous step and the next set of weights in the next layer.

*−**Repeat** step 6 for each subsequent layer until the output layer is reached.*

−Compare the output of the network to the desired output and calculate the error.

−Use backpropagation to propagate the error back through the network and adjust the weights accordingly.

*−**Repeat** steps 4 → 9 for a number of iterations or until the error is below a certain threshold.*

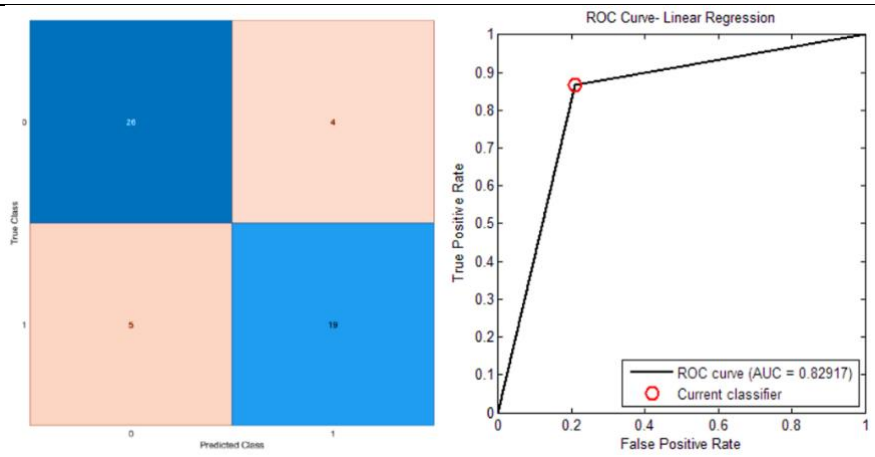
−Use the trained network for predictions on new, unseen data.

***Note**: the step 5 → 9 is the forward pass and step 10 is backpropagation.*

−Output: predicate output = Model.predicate

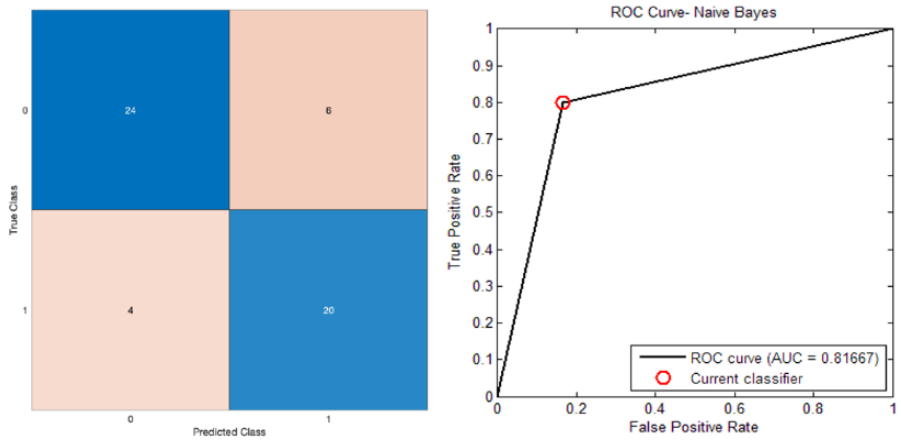
Stop

The weight update rule is used to adjust the weights of the network during training. The most common weight update rule used in this algorithm is backpropagation algorithm. The activation function is used to introduce non-linearity in the network.



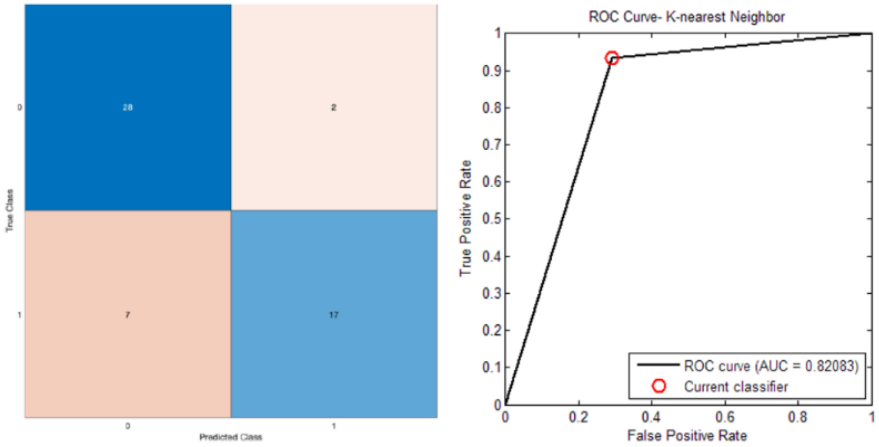
(a) (b)

FIGURE 5. Linear regression performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 26 people without heart disease and 19 people with heart disease. Also, incorrect effects appeared for the four non-patients and five patients.



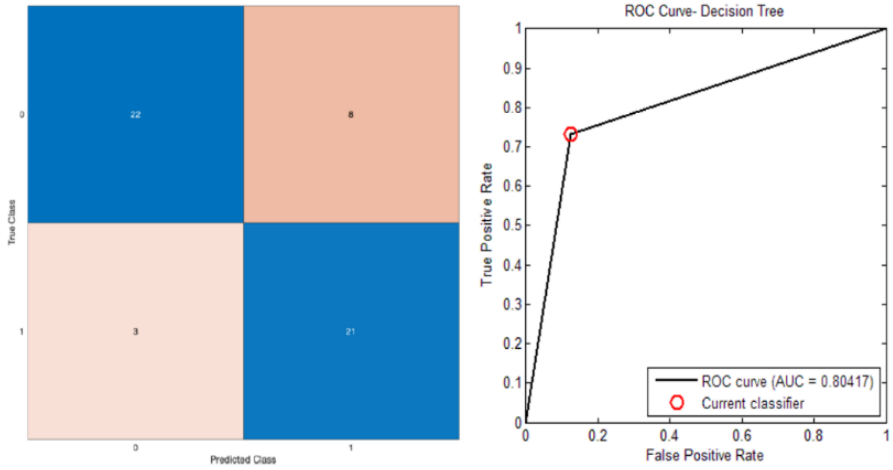
(a) (b)

FIGURE 6. Naive bayes performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 24 people without heart disease and 20 people with heart disease. Also, incorrect effects appeared for the six non-patients and four patients.



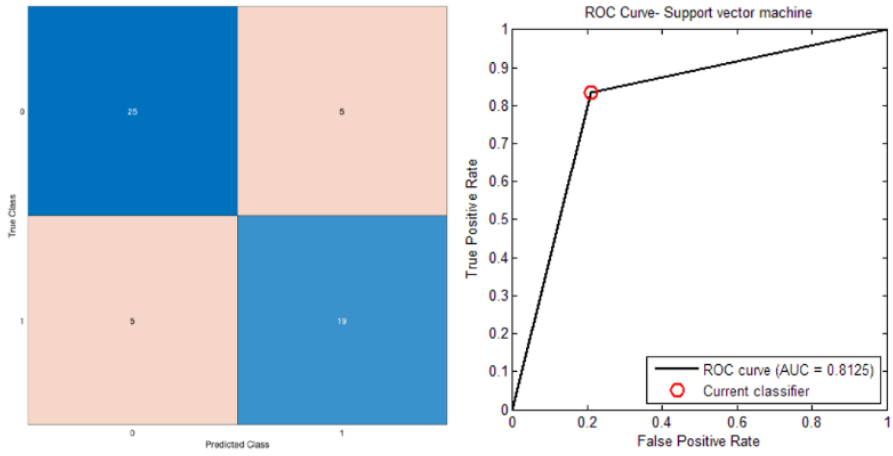
(a) (b)

FIGURE 7. K-nearest neighbor performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 28 people without heart disease and 17 people with heart disease. Also, incorrect effects appeared for the two non-patients and seven patients.



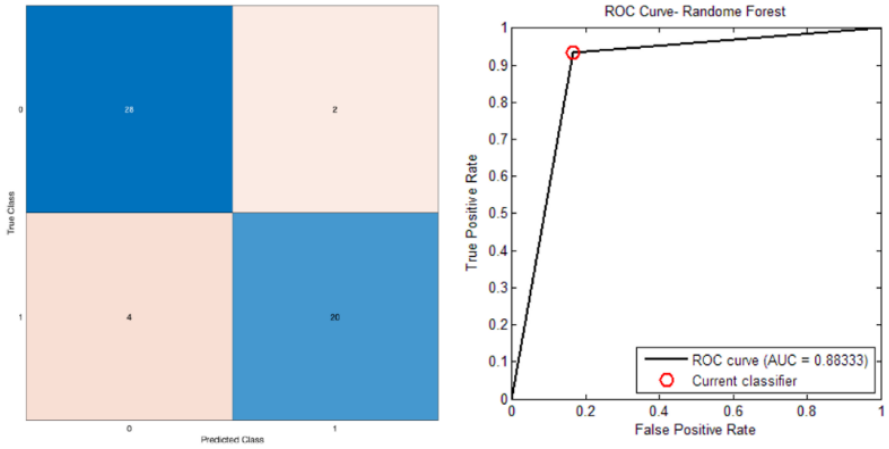
(a) (b)

FIGURE 8. Decision tree performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 22 people without heart disease and 21 people with heart disease. Also, incorrect effects appeared for the eight non-patients and three patients.



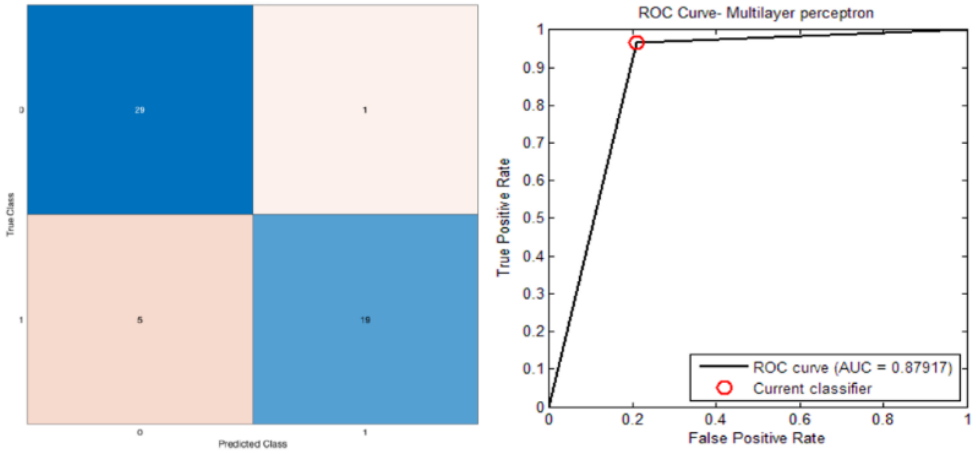
(a) (b)

FIGURE 9. Support vector machines performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 25 people without heart disease and 19 people with heart disease. Also, incorrect effects appeared for the five non-patients and five patients.



(a) (b)

FIGURE 10. Random forest performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 28 people without heart disease and 20 people with heart disease. Also, incorrect effects appeared for the two non-patients and four patients.



(a) (b)

FIGURE 11. Multilayer perceptron performance (a) Confusion matrix (b) ROC curve. The correct prediction was obtained for 29 people without heart disease and 19 people with heart disease. Also, incorrect effects appeared for the one non-patients and five patients.

Table 2. –Effects of all algorithms applied on the Cleveland Heart Disease Dataset.

	Performance metrics							
Techniques	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC	Training Time*	Prediction speed (obs/sec)
LR	83.3%	86.6%	79.1%	83.8%	85.2%	82.9%	72.94	7400
NB	81.4%	80%	83.3%	85.7%	82.7%	81.6%	81.22	14000
KNN	83.3%	93.3%	70.8%	80%	86.1%	82%	35.67	5500
DT	79.6%	73.3%	87.5%	88%	80.1%	80.4%	77.12	12000
SVM	81.5%	83.3%	79.1%	83.3%	83.3%	81.2%	50.67	26000
RF	88.8%	93.3%	83.3%	87.5%	90.3%	88.3%	100.11	9400
MLP	88.9%	96.6%	79.1%	85.2%	90.6%	87.9%	113.91	2000

*The time in seconds (sec), obs= observation

From the table above, it is clear that the MLP algorithm has outperformed due to its ability to model complex and non-linear relationships. Also, this algorithm has the ability to learn relevant features and show strong generalization automatically. Moreover, it is reported that the DT algorithm is the least performing among the applied algorithms because it is considered computationally efficient on a small dataset, and it works robustly with outliers.

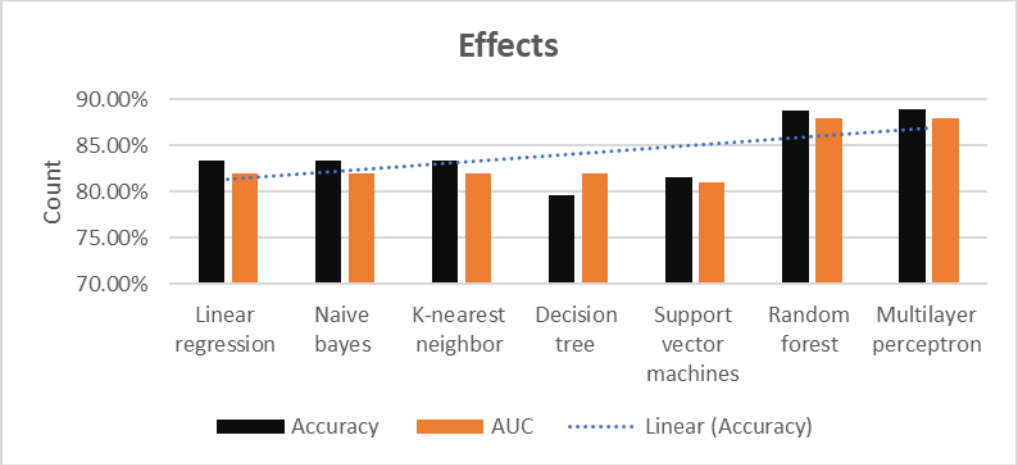


FIGURE 12. Comparison of the performance of algorithms using the metrics of accuracy and AUC.

5. CONCLUSIONS

This work achieved a set of conclusions regarding the performance of the seven algorithms, as it showed that the area under the curve (AUC) was perfect at random forest and multilayer perceptron, which provides us with the idea that these algorithms work sufficiently in separating data. The multilayer perceptron algorithm gave an exceptional performance through the accuracy metric, as it achieved an accuracy of more than 88%, which is the highest accuracy obtained in this study. Accordingly, the results of the experiments indicate that the multilayer perceptron algorithm achieved the most acceptable performance. The random forest algorithm is also perfect in performance, achieving an accuracy of more than 88%. As for the worst performance, it can be stated that the practice of the decision tree algorithm is the lowest among all the algorithms, as it achieved an accuracy of over 79%, with an area under the ROC curve of 80%, which is also the lowest. As for the rest of the algorithms, they were close to each other in performance. Through the prediction speed, the most suitable algorithm is multilayer perceptron, which recorded a lower prediction speed of 2000, although its training time is the largest among all algorithms.

ABBREVIATIONS AND ACRONYMS

Acronym	Description
AI	Artificial intelligence
AUC	Area under the ROC Curve
CART	Classification and Regression Tree
DL	Deep Learning
DT	Decision Tree
ECG	Electrocardiogram
K-NN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LR	Linear Regression
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
MNB	Multinomial Naive Bayes
NB	Naïve Bayes
RF	Random Forest
ROC Curve	Receiver Operating Characteristic Curve
SVFM	Support Vector Feature Measure
SVM	Support Vector Machine
UCI	University of California Irvine

Funding

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol.2, pp:100060, November 2022. https://doi.org/10.1016/j.health.2022.100060

[2] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "C-CADZ: computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset," *Applied Intelligence*, vol. 52, pp: 2436–2464, June 2021. https://doi.org/10.1007/s10489-021-02467-3

[3] S. H. Hyun, K. D. Bhilare, G. In, and J. Kim, "Effects of Panax ginseng and ginsenosides on oxidative stress and cardiovascular diseases: pharmacological and therapeutic roles," *Journal of Ginseng Research*, vol.46, no.1, pp:33-38, January 2022. https://doi.org/10.1016/j.jgr.2021.07.007

[4] L. Xu, M. Zimmermann, H. Forkey, J. Griffin, C. Wilds, et al., "How to Mitigate Risk of Premature Cardiovascular Disease Among Children and Adolescents with Mental Health Conditions," *Current Atherosclerosis Reports*, vol. 24, pp:253–264, March 2022. https://doi.org/10.1007/s11883-022-00998-9

[5] S. Wichansawakun, K. Chupisanyarote, W. Wongpipathpong, G. Kaur, and H. S. Buttar, "Antioxidant diets and functional foods attenuate dementia and cognition in elderly subjects," *Functional Foods and Nutraceuticals in Metabolic and Non-Communicable Diseases*, chapter 37, pp:533-549, 2022. https://doi.org/10.1016/B978-0-12-819815-5.00028-8

[6] C. Albus, C. Herrmann-Lingen, V. Köllner, R. Känel, and G. Titscher, "Psychosomatic Problem Areas and Comorbidities Using the Example of Coronary Heart Disease," *Psychocardiology*, pp:63–128, December 2022. https://doi.org/10.1007/978-3-662-65322-7_5

[7] J. L. Rodgers, J. Jones, S. I. Bolleddu, S. Vanthenapalli, L. E. Rodgers L., et al., "Cardiovascular Risks Associated with Gender and Aging," *Journal of Cardiovascular Development and Disease*, vol.6 no.2, pp:1-18, April 2019. https://doi.org/10.3390/jcdd6020019

[8] P. Gać, K. Czerwińska, P. Macek, A. Jaremków, G. Mazur, et al., "The importance of selenium and zinc deficiency in cardiovascular disorders," *Environmental Toxicology and Pharmacology*, vol.82, pp:103553, February 2021. https://doi.org/10.1016/j.etap.2020.103553

[9] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, et al, "Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no.10, pp:e56–e528, January 2019. https://doi.org/10.1161/CIR.0000000000000659

[10] P. Dewan, R. Rørth, P. S. Jhund, L. Shen, V. Raparelli, et al., "Differential Impact of Heart Failure With Reduced Ejection Fraction on Men and Women," *Journal of the American College of Cardiology*, vol.73, no.1, pp: 29–40, January 2019.

[11] Z. Gao, Z. Chen, A. Sun, and X. Deng, "Gender differences in cardiovascular disease," *Medicine in Novel Technology and Devices*, vol.4, pp:100025, December 2019. https://doi.org/10.1016/j.medntd.2019.100025

[12] T. Ohkuma, Y. Komorita, S. A. E. Peters, and M. Woodward, "Diabetes as a risk factor for heart failure in women and men: a systematic review and meta-analysis of 47 cohorts including 12 million individuals," *Diabetologia*, vol. 62, pp:1550–1560, July 2019. https://doi.org/10.1007/s00125-019-4926-x

[13] M. Walli-Attaei, P. Joseph, A. Rosengren, C. K. Chow, S. Rangarajan, et al., "Variations between women and men in risk factors, treatments, cardiovascular disease incidence, and death in 27 high-income, middle-income, and low-income countries (PURE): a prospective cohort study," *Lancet*, vol. 396, no.10244, pp: 97-109, July 2020. https://doi.org/10.1016/S0140-6736(20)30543-2

[14] H. Ritchie, F. Spooner, and M. Roser, "Causes of death," *Our World in Data*, December 2019, https://ourworldindata.org/causes-of-death

[15] A. Scirè, F. Tropeano, A. Anagnostopoulos, and I. Chatzigiannakis, "Fog-Computing-Based Heartbeat Detection and Arrhythmia Classification Using Machine Learning," *Algorithms*, vol.12, no.2, pp:1-21, January 2019. https://doi.org/10.3390/a12020032

[16] G. Pillar, M. Berall, R. B. Berry, T. Etzioni, Y. Henkin, et al., "Detection of Common Arrhythmias by the Watch-PAT: Expression of Electrical Arrhythmias by Pulse Recording," *Nature and Science of Sleep*, vol.14, pp:751-763, April 2022. https://doi.org/10.2147/NSS.S359468

[17] S. Sahay and A. K. Wadhwani, "Computer-Aided Interpretation of ECG Signal—A Challenge," *Medical Imaging and Health Informatics*, pp:335-358. May 2022. https://doi.org/10.1002/9781119819165.ch18

[18] E. Merdjanovska and A. Rashkovska, "Comprehensive survey of computational ECG analysis: Databases, methods and applications," *Expert Systems with Applications*, vol.203, pp:117206, October 2022. https://doi.org/10.1016/j.eswa.2022.117206

[19] Nuubo wearable ECG, https://www.nuubo.com/en-us/health-professionals/nuubo-system

[20] B. Saravi, F. Hassel, S. Ülkümen, A. Zink, S. Couillard-Despres, et al., "Artificial Intelligence-Driven Prediction Modeling and Decision Making in Spine Surgery Using Hybrid Machine Learning Models," *Journal of Personalized Medicine*, vol.12, no.4, pp:1-24, March 2022. https://doi.org/10.3390/jpm12040509

[21] J. Li, G. Yin, X. Wang, and W. Yan, "Automated decision making in highway pavement preventive maintenance based on deep learning," *Automation in Construction*, vol.135, pp:104111, March 2022. https://doi.org/10.1016/j.autcon.2021.104111

[22] K. Aggarwal, M. M. Mijwil, Sonia, AH. Al-Mistarehi, S. Alomari, M. Gök, A. M. Alaabdin, and S. H. Abdulrhman, "Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning," *Iraqi Journal for Computer Science and Mathematics*, vol.3, no.1, pp:115-123, January 2022. https://doi.org/10.52866/ijcsm.2022.01.01.013

[23] K. Moulaei, M. Shanbehzadeh, Z. M. Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," *BMC Medical Informatics and Decision Making*, vol. 22, no.2, pp:1-12, January 2022. https://doi.org/10.1186/s12911-021-01742-0

[24] S. Sanchez-Martinez, O. Camara, G. Piella, M. Cikes, M. Á. González-Ballester, et al., "Machine Learning for Clinical Decision-Making: Challenges and Opportunities in Cardiovascular Imaging," *Frontiers in Cardiovascular Medicine*, vol.8, no.765693, pp:1-11, January 2022. https://doi.org/10.3389/fcvm.2021.765693

[25] B. S. Shukur and M. M. Mijwil, "Involving Machine learning as Resolutions of Heart Diseases," *International Journal of Electrical and Computer Engineering*, vol.13, no.2, pp:2177-2185, April 2023. http://doi.org/10.11591/ijece.v13i2.pp2177-2185

[26] C. Liu, X. Wang, C. Liu, Q. Sun, and W. Peng, "Differentiating novel coronavirus pneumonia from general pneumonia based on machine learning," *BioMedical Engineering OnLine*, vol.19, no.66, pp:1-14, August 2020.https://doi.org/10.1186/s12938-020-00809-9

[27] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol.13, pp:8-17, 2015. https://doi.org/10.1016/j.csbj.2014.11.005

[28] A. K. Arshadi, J. Webb, M. Salem, E. Cruz, S. Calad-Thomson, et al., "Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development," *Frontiers in Artificial Intelligence*, vol.3, no.65, pp:1-13, August 2020. https://doi.org/10.3389/frai.2020.00065

[29] S. A. Bagabir, N. K. Ibrahim, H. A. Bagabir, and R. H. Ateeq, "Covid-19 and Artificial Intelligence: Genome sequencing, drug development and vaccine discovery," *Journal of Infection and Public Health*, vol.15, no.2, pp:289-296, February 2022. https://doi.org/10.1016/j.jiph.2022.01.011

[30] M. Supriya and A. J. Deepa, "Machine learning approach on healthcare big data: a review," *Big data and information analytics*, vol.5, pp:58-75, October 2020. https://doi.org/10.3934/bdia.2020005

[31] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol.22, no.5, pp:1589 - 1604, September 2018. https://doi.org/10.1109/JBHI.2017.2767063

[32] H. Omotunde and M. R. Mouhamed, "The Modern Impact of Artificial Intelligence Systems in Healthcare: A Concise Analysis," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol.2023, pp.66-70, November 2023. https://doi.org/10.58496/MJAIH/2023/013

[33] R. Hagan, C. J. Gillan, and F. Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease," *Informatics in Medicine Unlocked*, vol.24, pp:100606, 2021. https://doi.org/10.1016/j.imu.2021.100606

[34] S. K. Arunachalam and R. Rekha, "A novel approach for cardiovascular disease prediction using machine learning algorithms," *Concurrency Computation Practice and Experience*, vol.34, no.19, pp:e7027, April 2022. https://doi.org/10.1002/cpe.7027

[35] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol.2022, no.1410169, pp:1-9, March 2022. https://doi.org/10.1155/2022/1410169

[36] S. Sajeev, A. Maeder, S. Champion, A. Beleigoli, C. Ton, X. Kong, and M. Shu, "Deep Learning to Improve Heart Disease Risk Prediction," In Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting, pp:96–103, October 2019. https://doi.org/10.1007/978-3-030-33327-0_12

[37] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution," *Future Science OA*, vol. 7, No. 6, pp:1-9, July 2021. https://doi.org/10.2144/fsoa-2020-0206

[38] S. Shehzadi, M. A. Hassan, M. Rizwan, N. Kryvinska, and K. Vincent, "Diagnosis of Chronic Ischemic Heart Disease Using Machine Learning Techniques," *Computational Intelligence and Neuroscience*, vol.2022, no.3823350, pp:1-9, June 2022. https://doi.org/10.1155/2022/3823350

[39] Heart disease dataset, UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/heart+disease

[40] M. Supriya and V. K. Chattu, "A Review of Artificial Intelligence, Big Data, and Blockchain Technology Applications in Medicine and Global Health," *Big Data and Cognitive Computing*, vol.5, no.3, pp:1-20, September 2021. https://doi.org/10.3390/bdcc5030041

[41] A. Haleem A., M. Javaid, M. A. Qadri, R. P. Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literature-based study," *International Journal of Intelligent Networks*, vol.3, pp:119-132, 2022. https://doi.org/10.1016/j.ijin.2022.08.005

[42] S. Nami and M. Shajari, "Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors," *Expert Systems with Applications*, vol.110, pp:381-392, November 2018. https://doi.org/10.1016/j.eswa.2018.06.011

[43] S. F. Khorshid and A. M. Abdulazeez, "Breast Cancer Diagnosis Based on K-Nearest Neighbors: A Review," *PalArch's Journal of Archaeology of Egypt / Egyptology*, vol.18, no.4, pp:1927-1951, 2021.

[44] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol.12, no.6256, pp.1-11, April 2022. https://doi.org/10.1038/s41598-022-10358-x

[45] M. Kück and M. Freitag, "Forecasting of customer demands for production planning by local K-nearest neighbor models," *International Journal of Production Economics*, vol.231, pp:107837, January 2021. https://doi.org/10.1016/j.ijpe.2020.107837

[46] A. Tak, P. M. Parihar, S. Mathur, B. Das, and D. Sawal, "Survival prediction in heart failure using machine learning algorithms," *Research Square*, pp:1-11, August 2022. https://doi.org/10.21203/rs.3.rs-1960150/v1

[47] M. M. Mijwil, A. W. Abdulqader, S. M. Ali, and A. T. Sadiq, "Null-Values Imputation Using Different Modification Random Forest Algorithm," *IAES International Journal of Artificial Intelligence*, vol.12, no.01, pp:374-383, March 2023. http://doi.org/10.11591/ijai.v12.i1.pp374-383.

[48] S. M. Khazaal and H. Maarouf, "Predicting Coronary Artery Disease Utilizing Support Vector Machines: Optimizing Predictive Model," *Mesopotamian Journal of Artificial Intelligence in Healthcare*, vol.2023, pp.21-26, March 2023. https://doi.org/10.58496/MJAIH/2023/004

[49] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vo.3, pp.100071, June 2022. https://doi.org/10.1016/j.dajour.2022.100071

[50] J. Hu, T. Zhou, S. Ma, D. Yang, M. Guo, and P. Huang, "Rock mass classification prediction model using heuristic algorithms and support vector machines: a case study of Chambishi copper mine," *Scientific Reports*, vol.12, no.928, pp.1-20, January 2022. https://doi.org/10.1038/s41598-022-05027-y