Journal of Computing and Communication Vol.2 , No.1 , PP. 50-65 , 2023

MLHeartDisPrediction: Heart Disease Prediction using Machine Learning

Diaa Salama AbdElminaam*[a,b], Nada Mohamed [a], Hady Wael [a], Abdelrahman Khaled [a],Adham Moataz [a]

[a] Department of Computer Science, Faculty of Computer Science, Misr International University, Cairo, Egypt

[d]epartment of Information Systems, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

*Corresponding Author: Diaa Salama [*diaa.salama@miuegypt.edu.eg*]

ARTICLE DATA

Article history:
Received 21 Nov 2022
Revised 24 Jan 2023
Accepted 25 Jan 2023
Available online

Keywords:
Heart Disease prediction
Machine Learning
Classification
Naïve Bayes
Gradient Boosting
Linear Regression
K-Nearest Neighbor.

ABSTRACT

Predicting critical health conditions in their early stages can make the difference between life and death, and one such health condition is heart disease. Over the last decade, the main reason for death has been heart disease. Heart Disease is an ailment that affects many lives, is severely life-threatening, and can impair a person's ability to live a conventional life. The delay in treating Heart Disease increases the endangerment of the afflicted person. Consequently, early diagnosis of it can help save countless lives. However, the reasons for Heart Disease are varied, making its prediction very complex. Our objective is to use Machine Learning to enhance the dependability and simplicity of the prediction of Heart Disease. It was concluded that three datasets should be used; two have an immense size, alongside many Machine Learning algorithms. The proposed algorithms were tested: k-Nearest Neighbor, Gradient Boosting, Random Forest, Naïve Bayes, Decision Tree, and Logistic Regression. After rigorous testing, the only algorithm, Logistic Regression, stayed dominant in most of the testing achieving accuracies of 91.6% and 90.8%. Still, on the last dataset, the best algorithm was a random forest which scored the highest accuracy in all the testing, 98.6%. As shown in this paper, Machine Learning is a superb approach to predicting Heart Disease, and results can be further improved with the help of medical professionals and more research.

1. Introduction

Heart Disease is an appalling illness that puts myriad lives in jeopardy, and the best way to fight it is to identify it in its early stages before it rots in the victim's body, as it is much easier to cure. [1]. Millions of individuals are affected by heart disease, which is still the reason behind mortality around the globe.

According to statistics by the World Health Organization (WHO), 17.9 million annual deaths occur because of heart disease. [1]. Moreover, one person dies from heart disease roughly every 34 seconds worldwide. Heart Disease can be very low profile until the victim experiences something as severe as a heart attack. Additionally, Heart Disease is hard to identify, as it has numerous risk factors that contribute to it. It can occur for multitudinous reasons including but not limited to: smoking, drinking alcohol [2], obesity, lack of physical activity, poor mental health, age, sex, number of hours slept, and many others. [3][4]. For example, it can be from cholesterol. [5].

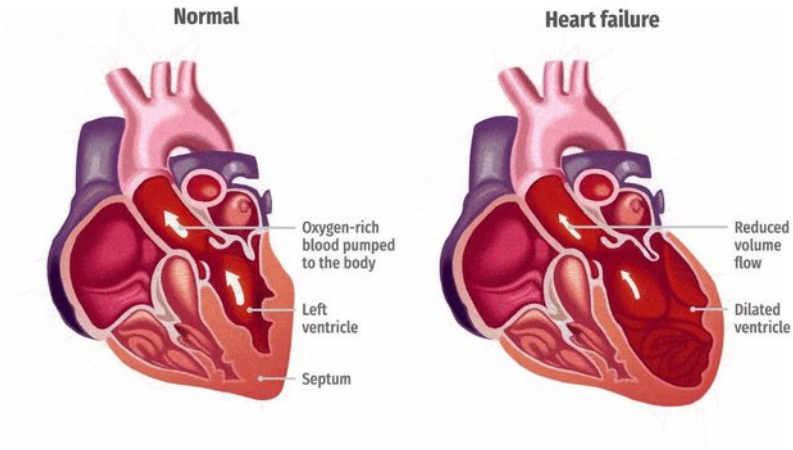

FIGURE 1. Illustration of a healthy heart and one with heart failure

Machine learning is a subfield of artificial intelligence still in its infancy. Its principal vision is to develop systems that can learn and form hypotheses based on their experiences. It builds a model by training machine learning algorithms on a training dataset. [6]. The model predicts the likelihood of heart disease based on the new input data. It constructs models by detecting obscure patterns in the input dataset using machine learning. [7]. For novel datasets, it makes accurate predictions. After the dataset has been processed and any null values have been filled. Using the new input data, the model is assessed for accuracy, then predicts the probability of heart disease. Machine learning techniques are categorized into supervised, unsupervised, and reinforcement learning.

In supervised learning, the model is trained on a labeled dataset. It has input data and output data. The data is categorized and partitioned into training and test datasets. [8]. The training dataset is used to train the model, while the testing dataset is used to evaluate the model's accuracy. The dataset contains models and their output. [9]. Its application is exemplified through classification and regression.

The clustering approach illustrates unsupervised learning. The data is used to train in unsupervised learning is not labeled or classed in the dataset. The objective is to find hidden patterns in the data. The model is being taught to recognize patterns. It may effortlessly anticipate hidden patterns for each new input dataset; after reviewing data, it draws inferences from datasets to define hidden patterns. There are no results in the dataset while using this method.

In reinforcement learning, it will not utilize labeled datasets, nor are the outputs related to data; instead, The model has been honed via experience. The model improves its presentation depending on its relationship with the environment and determines how to resolve its shortcomings and accomplish the intended result by assessing and evaluating various options. [10]. Classification algorithms are prominent supervised learning techniques for determining the likelihood of heart disease circumstances.

This is where Machine Learning takes part. Machine Learning is a technology that can apply to many fields and produce unexcelled results [11]; just some statistics need to be collected, and usually, they already are, and Machine Learning can start working and picking up on patterns that are too vague for traditional statistics to detect. Medical professionals already collect enormous amounts of data from their patients [12], so much data can be passed into the model to produce more accurate models and better results. Therefore, this makes predicting Heart Disease a prime field for Machine Learning [13].

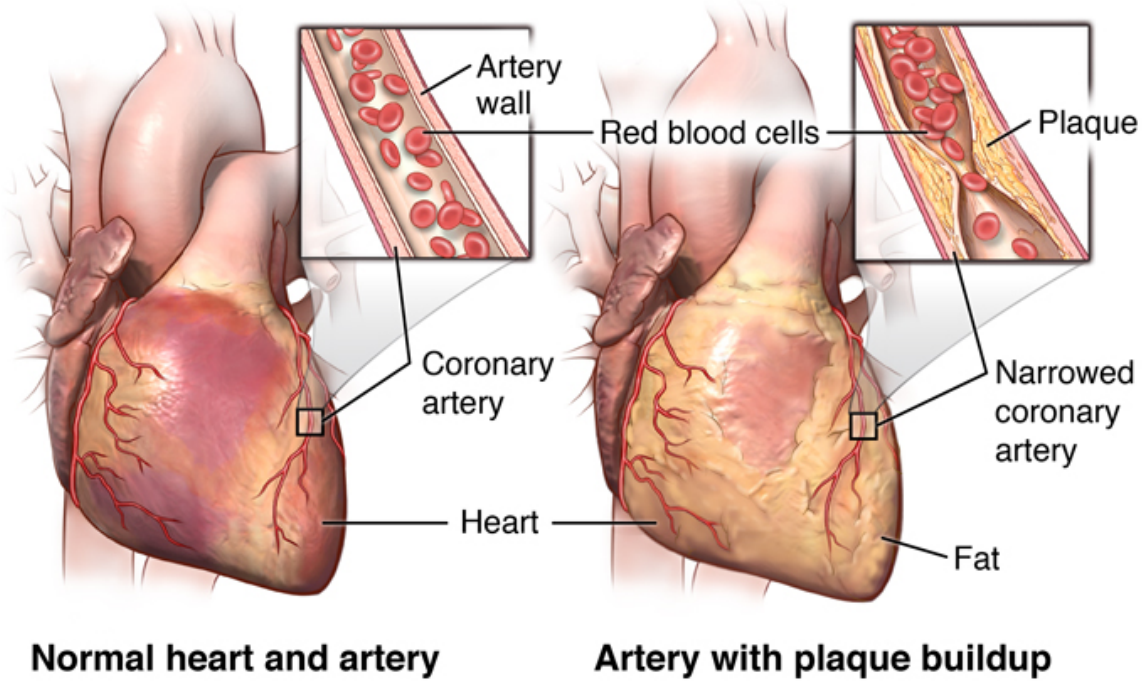

FIGURE 2. The heart and arteries show plaque buildup in the arterial wall. Normal and heart failure

This research paper utilizes classification techniques to forecast the risk of heart disease. This section depicts background information that illustrates topics such as heart disease and its symptoms and hazards, machine learning and its types with brief descriptions, and machine learning association with heart disease prediction and early prevention techniques. In this paper, the efforts will be highlighted to find the overall best model, the highest-ranking models will be compared, and explanations will be provided for the work done. As mentioned before, Heart Disease is a fatal illness that threatens countless lives around the globe. The earlier Heart Disease is diagnosed, the easier it is to cure. [14]. Machine Learning excels in predicting Heart Disease early since there are enormous amounts of data collected by medical professionals, which assists the Machine Learning algorithms in efficiently constructing the models, leading to more accurate predictions.

The contributions made to this topic are:

- The heart disease Prediction with machine learning.
- The testing of 6 machine learning algorithms.
- The use of 3 datasets, for a total of 574,440 entries with different sets of features spanning between 12 and 21 features.

The remaining sections in this paper are ordered as the following; related work is discussed in the second section. Moreover, The third section clarifies the proposed methodology of the research; it consists of a dataset description and used algorithms. The results of the proposed algorithms can be found in the fourth section and their analysis. The conclusion is located in the fifth section. An acknowledgment of all the supporting figures of this research is presented in the sixth section.

2. Related Work

The field of heart disease prediction is not unexplored; a lot of people investigated the field and come up with satisfactory results. We will mention some of the papers we read and analyzed to assist us in our research, and all the papers mentioned will be referenced in the references section.

Senthilkumar Mohan et al. [15] suggested a Hybrid Random Forest with Linear Model (HRFLM), a compound between a Random Forest algorithm and a Linear algorithm; HRFLM turned out to be very accurate with its predictions, with fewer errors than all the other tested algorithms. They used the UCI machine learning database for their data. The dataset had 303 records and used 13 of the 76 features. HRFLM scored an accuracy of 88.7%.

V.V. Ramalingam et al. [16] suggested Alternating Decision Trees with PCA, and it performed competently in contrast to Decision Trees which performed very poorly. They used Principal Component Analysis (PCA) for feature extraction from the dataset and selected the features using Correlation-based Feature Selection(CFS). Also, Support Vector Machines excelled in their testing. There were more methods mentioned in this paper, like Ant Colony Optimization.

Apurb Rajdhan et al. [17], after a lot of testing using a diverse array of algorithms, concluded that the Random Forest algorithm was the most accurate, with an accuracy rate of 90.16 % at predicting heart diseases. They used the UCI Cleveland heart disease dataset, which includes 76 features, from which they used 14. Some of these features are age, sex, the severity of chest pain, and the max heartbeat of the patient. In addition to Random forest, they tried Logistic Regression, Naive Bayes, and Decision trees, resulting in an accuracy of 85.25%, 85.25%, and 81.97%, respectively.

Jaymin Patel et al. [18] suggested the J48 technique, which produced satisfactory results and took minimal time to build. They used WEKA and the UCI Cleveland dataset, which comprised 76 features and 303 entries. They used features such as Diagnosis classes, sex, age, the severity of chest pain, and others. They used 10-fold Cross Validation with the J48 technique. Furthermore, they used w types of Reduced Error Pruning: Post pruning and Online pruning. The J48 technique had a test error of 0.1666667.

Devanch Shah et al. [19] considered 14 features and applied four machine learning algorithms: Decision Tree, Random Forest, K-Nearest Neighbor, and Naïve Bayes. They concluded that the best model was the K-Nearest Neighbor model with k=7, and it reached an accuracy of 83.16%. Moreover, their dataset needed preprocessing since it had large numbers, was noisy, and had missing data.

Youness Khourdifi et al. [20] concluded that each algorithm worked better in certain situations. Random Forest, K-Nearest Neighbor, and Neural Networks were the models that worked best with the dataset they used. Their results also showed that the optimization hybrid approach significantly increased prediction in medical datasets. They also suggested 2 dataset optimization methods: Particle Swarm optimization (PSO) and Ant Colony Optimization (ACO). They made a hybrid of both methods and used it with K-Nearest Neighbor, which resulted in an accuracy of 99.65%, and 99.6% with Random Forest. They got their dataset from the UCI machine learning repository.

Abhijeet Jagtap et al. [21] started by clearing the first hurdle in their path, which was their dataset. The raw data range was significantly variable, and the dataset had many missing values. They split the data into two sections: 75% for training and 25% for testing. Afterward, they applied three algorithms to the dataset: Support Vector Machines, Logistic Regression, and Naïve Bayes. They concluded that Support Vector Machines were the most accurate of the three algorithms, with an efficiency of 64.4%.

Amin Ul Haq et al. [22] applied seven different algorithms to their dataset: the Cleveland heart disease dataset and the algorithms were KNN, SVM, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and Neural Networks. They used Lasso to select the most important attributes from the dataset. The most notable statistics were the accuracy of the Support Vector Machine at g = 0.0001 and c = 100, Logistic Regression, and Neural Networks, which were 88%, 87%, and 86%, respectively. It is worth noting that Logistic Regression performed better with a fold cross-validation of 10; it achieved 89% accuracy.

Harshit Jindal et al. [23] used a combination of three Machine Learning algorithms: KNN, Logistic Regression, and Random Forest. Their combined model achieved an accuracy of 87.5%. They concluded that their high accuracy is mostly the result of the increased medical attributes they used. They used 13 attributes which include blood pressure, age, cholesterol, fasting sugar, chest pain, sex, and others. The dataset contained 304 entries in total.

Rahul katarya et al. [24] tested nine algorithms: Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Support Vector Machines, Random Forest, Artificial Neural Networks, Deep Neural Networks, and Multilayer Perceptron. They acquired their dataset from the UCI repository then, they chose 14 out of the 76 features and normalized them, and replaced any missing data with the python library NAN. Some of the chosen features were age,sex, cholesterol,type of chest pain sugar, resting blood pressure, fasting blood, and others. The top 3 highest accuracies were 95.6%, 93.4%, and 92.3% with Random Forest, Logistic Regression, and Artificial Neural Networks and Support vector machines, which shared the same accuracy.

3. Proposed Methodology

Numerous algorithms were used, and research was done on each algorithm before training the model using them on the datasets. Figure 3 represents the steps the datasets went through to get the results.

4.1. Datasets Descriptions

The first dataset consists of 17 features, and it has 319,785 records. The dataset was split into two partitions: 70% for training and 30% for testing. A detailed description of the features can be found below.

BMI is an abbreviation for Body mass index, computed by multiplying a person's weight in kg by the square of their height. [24]. Stroke represents if someone had a stroke before. Physical health represents how good the person's physical health is. Mental Health represents how good a person's mental health is. DiffWalking represents if the person has difficulty walking or not. Age Category represents which age range the person belongs to. Race represents the ethnicity of the person. Diabetic represents whether the person has diabetes or not. The physical activity represents whether the person partakes in physical activity or not. General Health represents how better the person's general health is. Sleep Time is the number of hours the person sleeps. Asthma represents whether the person has asthma or not. Kidney Disease represents whether a person suffers from any kidney disease. Skin Cancer represents whether a person has skin cancer; finally, the target is Heart Disease.

The second dataset consists of 21 features, and it has 253,680 records. The dataset was normalized and split into two partitions: 70% for training and 30% for testing. A detailed description of the features can be found below.

HeartDiseaseorAttack is the target, representing whether the person has or will have a heart disease or a heart attack. High BP represents whether the person suffers from high blood pressure or not. HighChol represents whether the person suffers from high cholesterol or not. CholCheck represents whether the person treats their high cholesterol or not. BMI is computed by dividing a person's weight in kilograms by the square of their height. [24]. A stroke represents if someone had a stroke before. Diabetes represents whether the person has diabetes or not. PhysActivity represents whether the person partakes in physical activity or not. Fruits represent whether the person consumes fruits or not. Veggies represent whether the person consumes vegetables or not. HvyAlcoholConsump represents if the person consumes large amounts of alcohol. AnyHealthCare represents whether the person receives any healthcare. NoDocBcCost represents if the person doesn't receive healthcare due to financial reasons. GenHlth represents how better the person's general health is. MentHlth is a number representing how better a person's mental health is. PhysHlth is a number representing how good the person's physical health is. Diffwalk represents if the person has difficulty walking or not. Education is a number that represents the quality of a person's education. Income is a number that represents the amount of money the person makes.

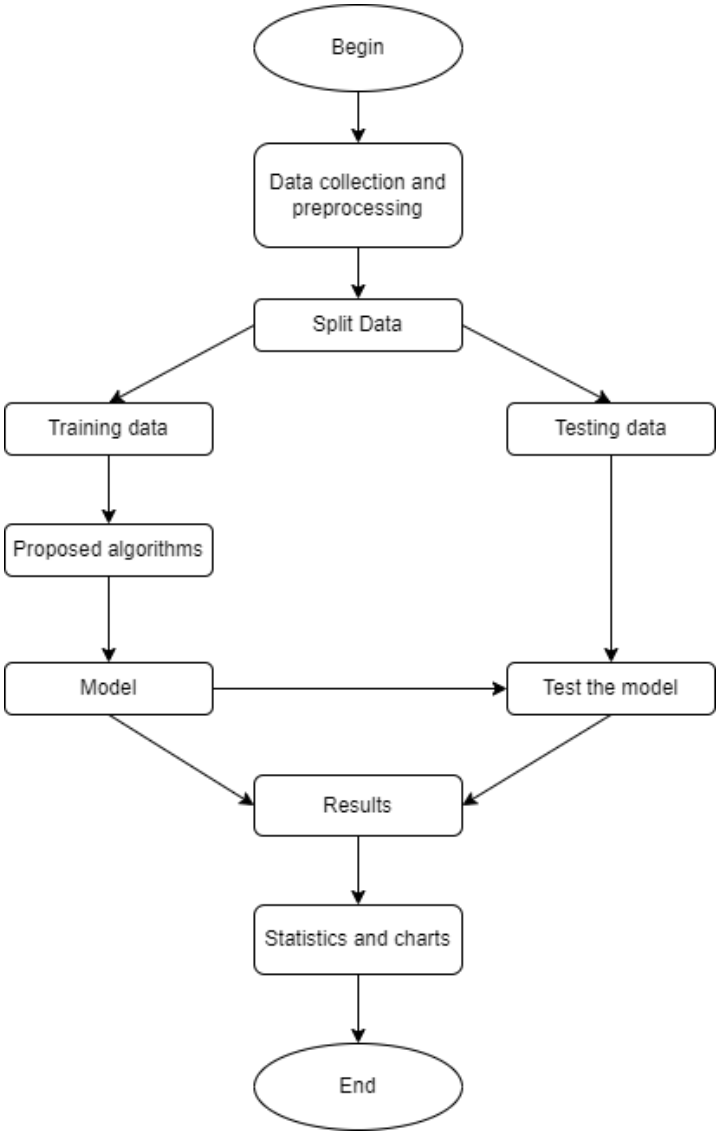

FIGURE 2. Heart disease prediction process

TABLE 1: Features of Dataset 1

Feature	Type	Values
BMI	Numerical	From 12.02 to 94.85
Smoking	Classification	Yes or No
Alcohol	Classification	Yes or No
Stroke	Classification	Yes or No
Physical Health	Numerical	From 0 to 30
Mental Health	Numerical	From 0 to 30
Diff Walking	Classification	Yes or No
Sex	Classification	Male or Female

TABLE 2: Features of Dataset 2

Feature	Type	Values
HeartDiseaseorAttack (Target)	Classification	Yes or No
HighBP	Classification	Yes or No
HighChol	Classification	Yes or No
CholCheck	Classification	Yes or No
BMI	Numerical	From 12 to 98
Smoker	Classification	Yes or No
Stroke	Classification	Yes or No
Diabetes	Classification	Yes or No
PhysActivity	Classification	Yes or No
Fruits	Classification	Yes or No
Veggies	Classification	Yes or No
HvyAlcoholConsump	Classification	Yes or No
AnyHealthCare	Classification	Yes or No
NoDocbcCost	Classification	Yes or No
GenHlth	Numerical	From 1 to 5
MentHlth	Numerical	From 0 to 30
PhysHlth	Numerical	From 0 to 30
DiffWalk	Classification	Yes or No
Sex	Classification	Male or Female
Age	Classification	Classes from 1 to 13
Education	Numerical	From 1 to 6
Income	Numerical	From 1 to 8

The second dataset consists of 21 features, and it has 253,680 records. The dataset was normalized and split into two partitions: 70% for training and 30% for testing. A detailed description of the features can be found below.

HeartDiseaseorAttack is the target, representing whether the person has or will have a heart disease or a heart attack. HighBP represents whether the person suffers from high blood pressure or not. HighChol represents whether the person suffers from high cholesterol or not. CholCheck represents whether the person treats their high cholesterol or not. BMI is computed by dividing a person's weight in kilograms by the square of their height. [24]. A stroke represents if someone had a stroke before. Diabetes represents whether the person has diabetes or not. PhysActivity represents whether the person partakes in physical activity or not. Fruits represent whether the person consumes fruits or not. Veggies represent whether the person consumes vegetables or not. HvyAlcoholConsump represents if the person consumes large amounts of alcohol. AnyHealthCare represents whether the person receives any healthcare. NoDocBcCost represents if the person doesn't receive healthcare due to financial reasons. GenHlth represents how better the person's general health is. MentHlth is a number representing how better a person's mental health is. PhysHlth is a number representing how good the person's physical health is. Diffwalk represents if the person has difficulty walking or not. Education is a number that represents the quality of a person's education. Income is a number that represents the amount of money the person makes.

The third and final dataset consists of 13 features, and it has 1025 records.The dataset was normalized and split into two partitions 70% for training and 30% for testing. A detailed description of the features can be found below.

Cp represents the type of chest pain the person suffers from. Trestbps represents the resting blood pressure of the person. Chol represents the cholesterol in mg/dl. Fbs represents if the fasting blood sugar is normal or abnormal. Restecg represents the results of the resting electrocardiograph. Thalach represents the peak heart rate reached by the person. Exang represents whether the person had angina that was induced by exercise. Oldpeak represents the ST depression influenced by exercise concerning rest. The slope represents the slope of the maximum exercise ST segment. Ca represents the number of vital vessels colored by fluoroscopy. A condition that is our target and represents whether the person has or will be having heart disease.

TABLE 3: Features of Dataset 3

Feature	Type	Values
BMI	Numerical	From 12.02 to 94.85
Age	Numerical	From 29 to 77
Sex	Classification	Male or Female
CP	Classification	1, 2, or 3
Trestbps	Numerical	From 94 to 200
Chol	Numerical	From 126 to 564
FBS	Classification	Yes or No
Restecg	Classification	0, 1, or 2
Thalach	Numerical	From 71 to 202
Exang	Classification	Yes or No
Oldpeak	Numerical	From 0 to 6.2
Slope	Classification	0, 1, or 2
Ca	Numerical	From 0 to 3
Thal	Classification	0, 1, or 2
Condition (Target)	Classification	Yes or No

4. Methods

The mentioned datasets were passed into six different Machine Learning algorithms: Logistic Regression, Gradient Boosting, K Nearest Neighbor (k-NN), Random Forest, Decision Tree, and Naive Bayes. For each of the algorithms, their statistics were generated. These statistics were: Accuracy, Recall, Precision, and Specificity. Afterward, the results were charted and compared. The results, charts, and discussion can be found later in the paper. [25].

4.1. Gradient Boosting:

Gradient boosting is a boosting utilized in machine learning. It is built on the presumption that the prediction error is lessened when the head potential future model is matched with former versions. To minimize error, the rudimentary concept defines the target results for the succeeding model. [26].

4.2. Decision Tree:

The supervised learning type includes the decision tree algorithm. Both regression and classification issues may be handled using them. Each node in the tree corresponds to a class label, with attributes expressed on the tree's inner node. Any Boolean function with discrete characteristics may be described

using the decision tree. The entropy varies when a node is employed in a decision tree, and it breaks down the training dataset into smaller groupings. The information denotes the increase in entropy. [27].

Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of S with $A = v$, and Values(A) is the set of all possible values of A, then

$$\text{Gain}\ (S, A) = \text{Entropy}\ (S) - \sum_{v \in \text{Values}\ (A)} \mid \frac{|S|}{S|} \cdot \text{Entropy}\ (S_v) \quad (2)$$

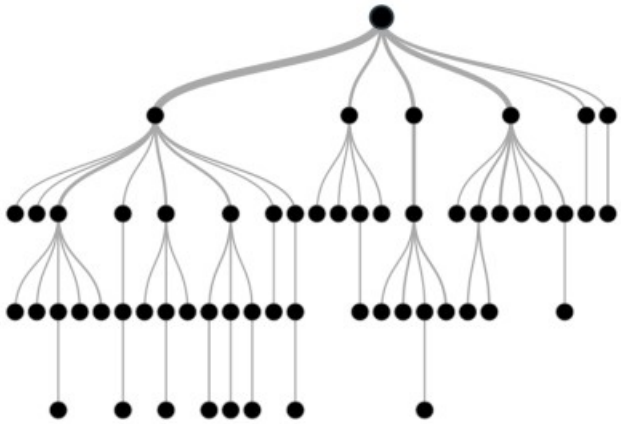

FIGURE 2. Illustration of decision tree

4.3. Naive Bayes (NB)

Naive Bayes is a simple yet capable categorization algorithm built on the Bayes Theorem. It presupposes predictor independence, which means that the traits or characteristics are unrelated or connected in any way. Even though there is a dependence, each of these qualities or attributes contributes to the probability independently, which is why it is termed Naive. [28].

$$P(c \mid x) = \frac{P(x|c)P(c)}{P(x)} \quad (3)$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c) \quad (4)$$

4.4. K – Nearest Neighbor

Hodges and fix established a classification of nonparametric pattern algorithm described as the (KNN) K-Nearest Neighbor rule in 1951. [29]. The KNN technique is one of the best basic and most powerful classification methods. It doesn't make any assumptions about data and is classification usable jobs where very little or no prior knowledge of the distribution of data is access able. This algorithm is used to find The value of the found data points in it is allocated to the nearest data points in the training set to the data point for which a target value is assigned. [30].

4.5. Random Forest:

Random Forest is one of the supervised machine learning algorithms that can be used for classification and regression tasks but works better in classification tasks. This algorithm considers multiple decision trees before giving an output. This technique is founded on the notion that a greater number of trees would ultimately guide the rectified selection. It employs a voting approach for classification and then

determines the class, whereas it uses the mean of all the decision tree outputs for regression. [31]. Random Forest Algorithm is extremely efficient with large datasets with high dimensionality. [32].

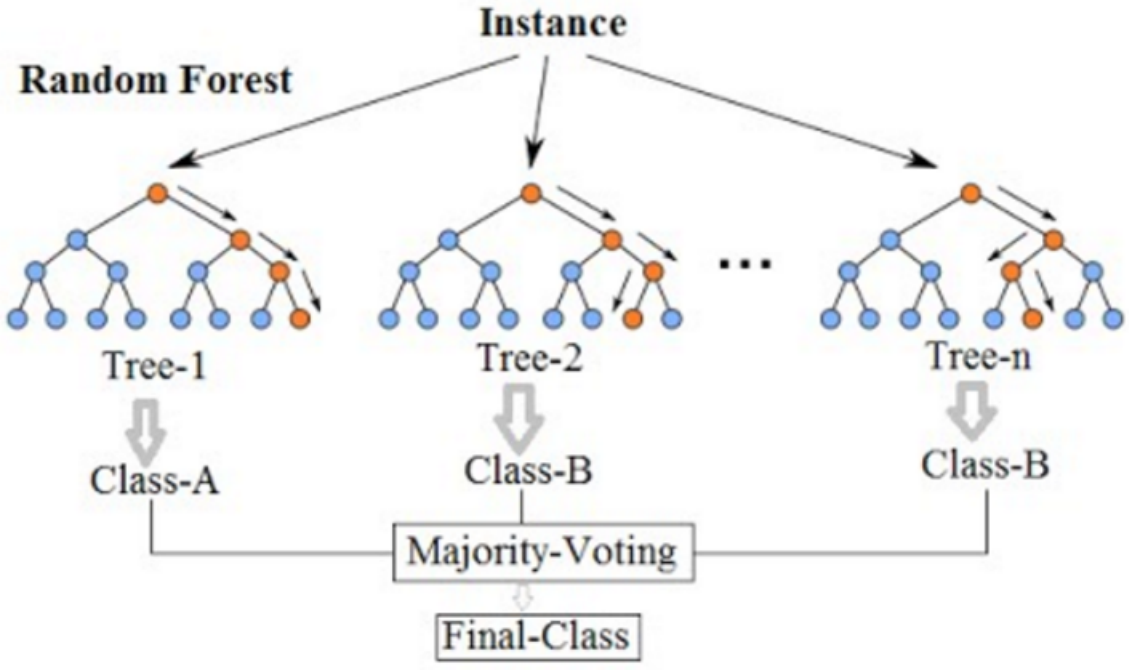

FIGURE 3. Random forest demonstration

4.6. Logistic Regression

Logistic regression is one of the supervised machine learning algorithms, and it is a classification method. The models of Logistic regression are classified as "statistical models that describe a relationship between an independent variable" in the Medical Subject Headings (MeSH) thesaurus used by the National Library of Medicine and qualitative dependent variable (that is, one that can only take particular discrete values, such as the presence or absence of a disease). The effects of feature variables on categorical outcomes are studied using logistic regression models. The label is habitually binary, such as the residence or non-existence of disease (e.g., non-lymphoma), Hodgkin's, which is called a binary logistic model. A multiple or multi variable logistic regression model is one of the most often used predictive methods when there are several features. (e.g.treatments and risk factors). [33] So, Logistic regression is reliable in predicting the likelihood of a person having heart disease or not.

5. Experimental Results

Accuracy is the count of legitimately anticipated data from all the data. The count of accurately anticipated positives taken from the anticipated positives is the Precision-Recall is the number of correctly anticipated positives from all the true positives. The number of accurately anticipated negatives out of all the expected negatives is known as specificity. Several evaluation metrics are used to evaluate the performance of the classification. The most common metrics include accuracy (ACC), precision (PREC), sensitivity (recall) (REC), specificity, and f-score (F1). They are calculated as follows:

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$prec = \frac{TP}{TP+FP} \quad (6)$$

$$Rec = \frac{TP}{TP+FN} \quad (7)$$

$$\text{Specificity} = (\text{TN}/\text{TN} + \text{FP}) \quad (8)$$

Where TP, TF, FP, and FN indicate the true positive, true negative, false positive, and false negative respective.

5.1. First dataset results

The results collected from Gradient Boosting, Naïve Bayes, Logistic Regression, Random Forest, k-Nearest Neighbor, and Decision Tree are shown below.
The following results are from the first dataset.

Table 4
Statistics of Algorithms with 70/30 Data Split for the first data set

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.916	0.890	0.916	0.182
Gradient Boosting	0.916	0.890	0.916	0.172
k-NN	0.908	0.870	0.908	0.159
Random Forest	0.906	0.875	0.906	0.205
Decision Tree	0.895	0.926	0.895	0.254
Naïve Bayes	0.876	0.890	0.877	0.453

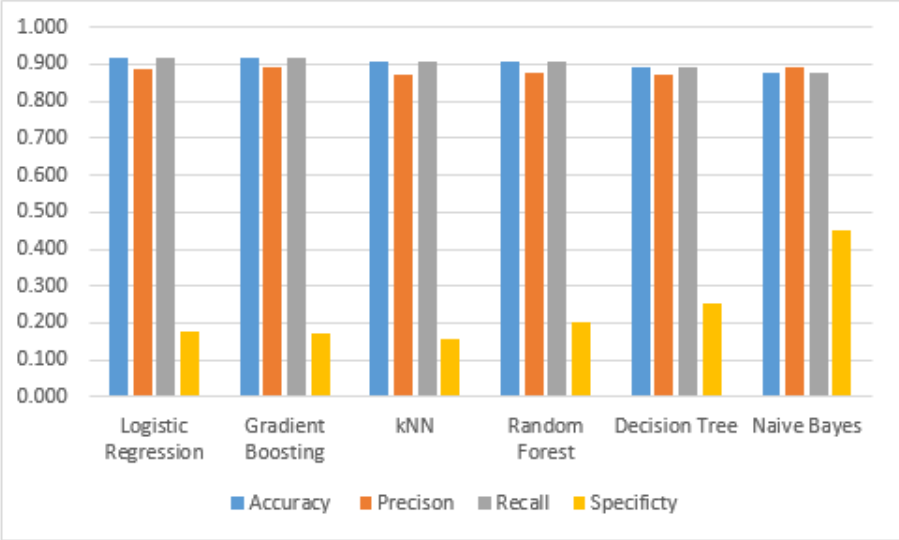

FIGURE 6. First dataset performance results with data split

Logistic Regression and Gradient Boosting are the dominant algorithms, sharing the same accuracy of 0.916, with Logistic Regression being slightly better in terms of precision and specificity. K-NN was a close second with an accuracy of 0.908, and it is worth noting that k-NN took an extremely long time to create the model. Naive Bayes was the worst with an accuracy of 0.875 but was the fastest while creating the model.

Table 5
Statistics of Algorithms with 10 K-fold for the first dataset

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.916	0.889	0.916	0.180
Gradient Boosting	0.916	0.890	0.916	0.171
k-NN	0.908	0.870	0.908	0.154

Random Forest	0.908	0.875	0.906	0.202
Decision Tree	0.872	0.926	0.893	0.252
Naïve Bayes	0.916	0.889	0.916	0.180

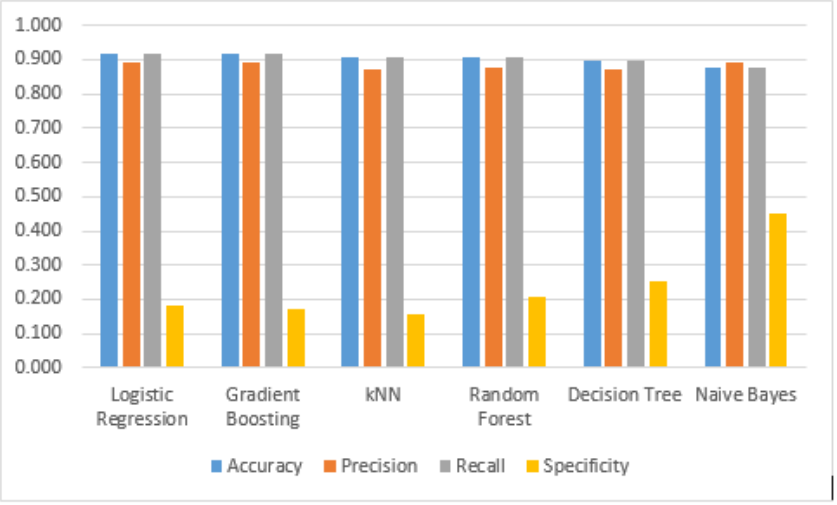

FIGURE 7. First dataset performance results with ten k-fold

Using k-fold on the dataset had slight improvements with most algorithms. Logistic regression had 0.009 of increased precision. Gradient Boosting suffered a very slight loss of 0.001 in terms of specificity. Random Forest, Decision Tree, and Naive Bayes all suffered losses in terms of accuracy. K-NN had a slightly decreased specificity compared to the 70/30 data split.

5.2. Second dataset results

The following results are from the second dataset.

Table 6
Statistics of Algorithms with 70/30 Data Split for the first data set

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.908	0.881	0.908	0.207
Gradient Boosting	0.908	0.881	0.908	0.192
k-NN	0.898	0.867	0.898	0.221
Random Forest	0.903	0.871	0.903	0.212
Decision Tree	0.865	0.926	0.887	0.269
Naïve Bayes	0.847	0.885	0.847	0.537

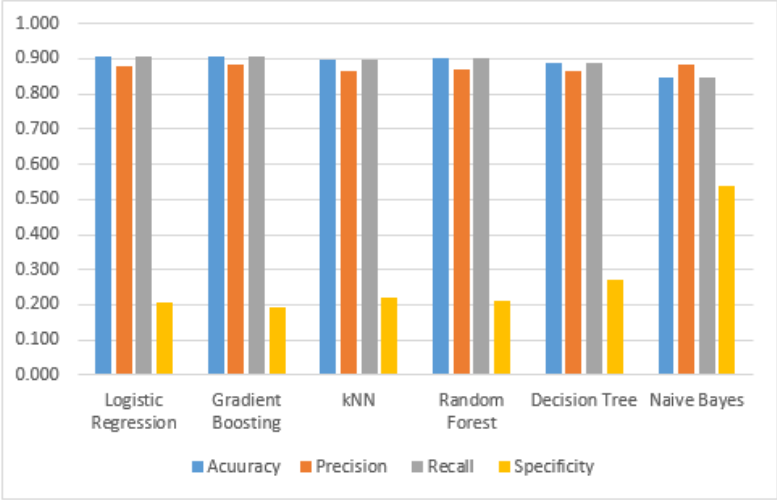

FIGURE 8. Second dataset performance results with data split

The second dataset had generally less accuracy than the first dataset in all its conditions. Logistic Regression and Gradient Descent share the same accuracy again, which is 0.908 in this case, with Logistic Regression having a slightly higher specificity. K-NN had a much lower accuracy than usual, scoring 0.898, while Random Forest had a lower accuracy of 0.903 but ranking higher among the algorithms.

Table 7

Statistics of Algorithms with 10 K-fold for the first dataset

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.908	0.881	0.908	0.208
Gradient Boosting	0.908	0.881	0.908	0.196
k-NN	0.897	0.865	0.897	0.220
Random Forest	0.902	0.871	0.902	0.216
Decision Tree	0.887	0.864	0.887	0.263
Naïve Bayes	0.847	0.884	0.847	0.534

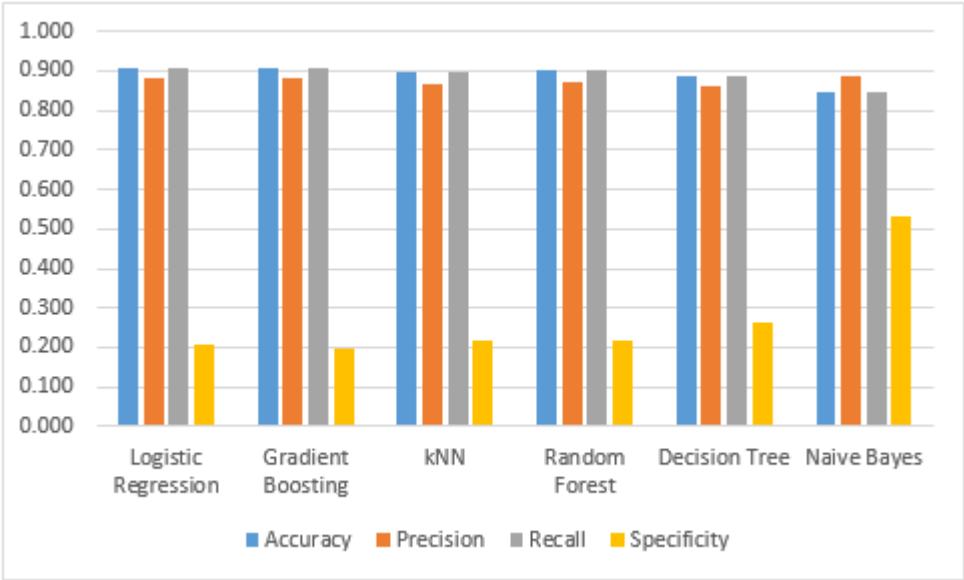

FIGURE 9. The second dataset performance results with 10 k-fold

The results using k-fold were very similar and sometimes worse than the 70/30 data split, and it also maintained the same hierarchy.

5.3. Third dataset results

The following results are from the third dataset.

Table 8
Statistics of Algorithms with 70/30 Data Split for the third data set

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.843	0.846	0.843	0.841
Gradient Boosting	0.962	0.962	0.962	0.962
k-NN	0.848	0.848	0.848	0.848
Random Forest	0.961	0.961	0.961	0.961
Decision Tree	0.921	0.922	0.921	0.922
Naïve Bayes	0.843	0.843	0.843	0.842

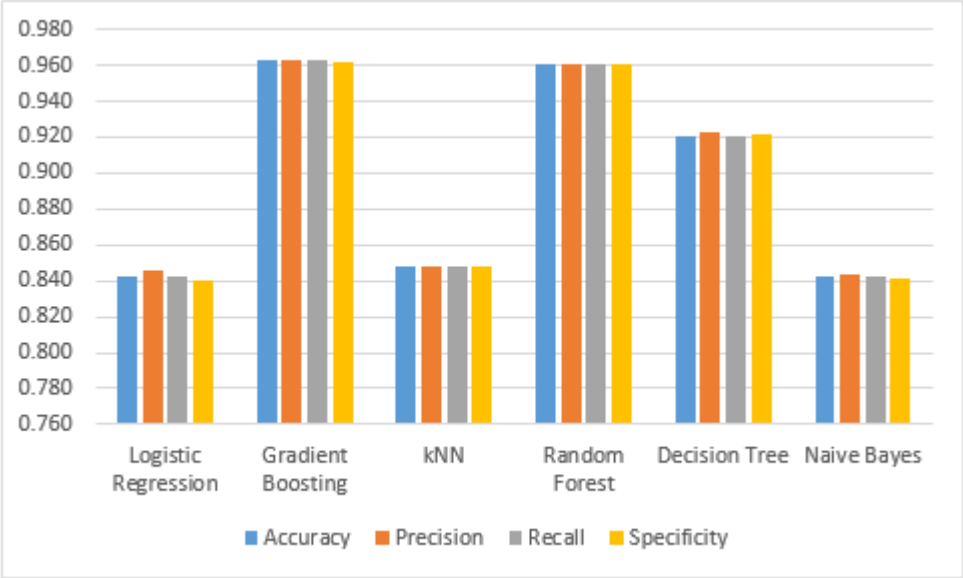

FIGURE 10. Third dataset performance results with data split

T Surprisingly, Gradient Boosting had the best accuracy in this dataset, scoring 0.962. Results were better for almost all the algorithms, but that could be attributed to the significantly smaller dataset. All the algorithms had a much higher specificity. Logistic regression lost first place for the first time in our testing, scoring a mediocre 0.843 in terms of accuracy.

Table 9
Statistics of Algorithms with 10 K-fold for the first dataset

Model/Measures	Accuracy	Precision	Recall	Specificity
Logistic Regression	0.849	0.853	0.849	0.844
Gradient Boosting	0.965	0.965	0.965	0.965
k-NN	0.847	0.848	0.847	0.844
Random Forest	0.986	0.986	0.986	0.986
Decision Tree	0.951	0.952	0.951	0.952
Naïve Bayes	0.849	0.849	0.849	0.848

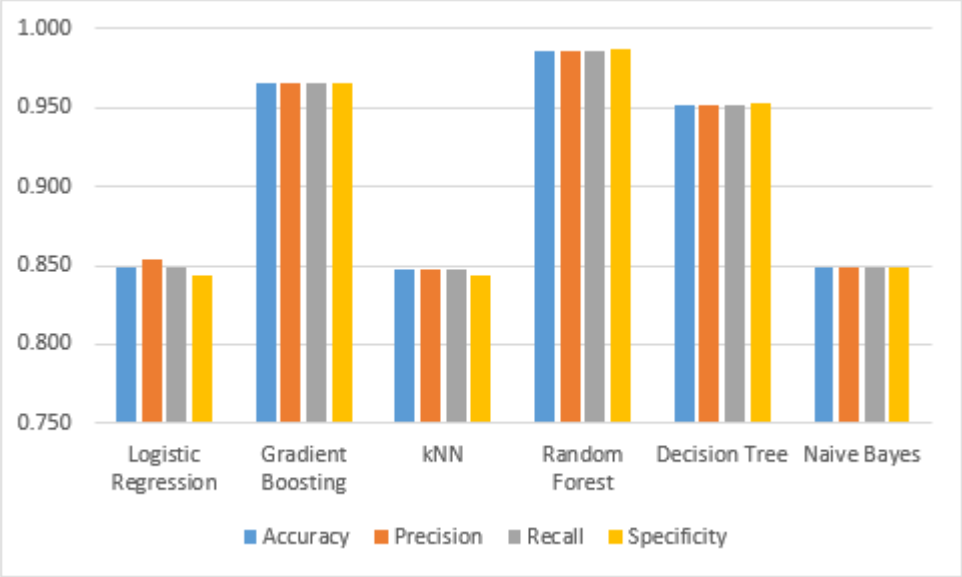

FIGURE 11. The second dataset performance results with 10 k-fold

Using k-fold, this testing had the highest accuracy. Random Forest scored an accuracy of 0.986. Gradient Boosting came second with an accuracy of 0.965. As seen in the previous testing, specificity was much higher in all the algorithms.

6. Conclusion

An early diagnosis is crucial to saving as many lives as possible, and Machine learning proved to be a great approach to detect this cunning disease prematurely. Logistic regression excelled in predicting heart disease in most datasets with accuracies of 91.6%, and 90.8%, but it was beaten in the last dataset only by Random Forest which had an accuracy of 98.6%. With more research and guidance from medical professionals, prediction accuracy can grow even more. Machine Learning can be applied to many fields, not just medicine, and it can be used to predict anything from stock prices to the results of sports matches, making it a very useful tool for humanity. And this tool will only keep improving and producing better results.

References

[1] F. Przerwa, A. Kukowka, K. Kotrych, and I. Uzar, "Probiotics in prevention and treatment of cardiovascular diseases," Herba Polonica, vol. 67, no. 4, pp. 77–85, 2021.

[2] J. Rehm, C. T. Sempos, and M. Trevisan, "Average volume of alcohol consumption, patterns of drinking and risk of coronary heart disease-a review," Journal of Cardiovascular Risk, vol. 10, no. 1, pp. 15–20, 2003.

[3] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives bayesian," in 2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019, pp. 292–297.

[4] R. Sun, M. Liu, L. Lu, Y. Zheng, and P. Zhang, "Congenital heart disease: causes, diagnosis, symptoms, and treatments," Cell biochemistry and biophysics, vol. 72, no. 3, pp. 857–860, 2015.

[5] W. P. Castelli, R. D. Abbott, and P. M. McNamara, "Summary estimates of cholesterol used to predict coronary heart disease." Circulation, vol. 67, no. 4, pp. 730–734, 1983.

[6] S. Anitha and N. Sridevi, "Heart disease prediction using data mining techniques," Journal of analysis and Computation, 2019.

[7] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," Cluster computing, vol. 17, no. 3, pp. 881–891, 2014.

[8] H. Sharma and M. Rizvi, "Prediction of heart disease using machine learning algorithms: A survey," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 8, pp. 99–104, 2017.

[9] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison," Computers in Biology and Medicine, vol. 136, p. 104672, 2021.

[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in 2008 IEEE/ACS international conference on computer systems and applications. IEEE, 2008, pp. 108–115.
[11] B. Mahesh, "Machine learning algorithms-a review," International Journal of Science and Research (IJSR).[Internet], vol. 9, pp. 381–386, 2020.
[12] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," IEEE journal of biomedical and health informatics, vol. 19, no. 4, pp. 1209–1215, 2015.
[13] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1. IOP Publishing, 2021, p. 012072.
[14] S. P. Shaji et al., "Predictionand diagnosis of heart disease patients using data mining technique," in 2019 international conference on communication and signal processing (ICCSP). IEEE, 2019, pp. 0848–0852.
[15] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE access, vol. 7, pp. 81 542–81 554, 2019.
[16] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," International Journal of Engineering Technology, vol. 7, no. 2.8, pp. 684–687, 2018.
[17] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Heart disease prediction using machine learning," International Journal of Research and Technology, vol. 9, no. 04, pp. 659–662, 2020.
[18] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," Heart Disease, vol. 7, no. 1, pp. 129–137, 2015.
[19] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," SN Computer Science, vol. 1, no. 6, pp. 1–6, 2020.
[20] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, pp. 242–252, 2019.
[21] A. Jagtap, P. Malewadkar, O. Baswat, and H. Rambade, "Heart disease prediction using machine learning," International Journal of Research in Engineering, Science and Management, vol. 2, no. 2, pp. 352–355, 2019.
[22] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, vol. 2018, 2018.
[23] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," Health and Technology, vol. 11, no. 1, pp. 87–97, 2021.
[24] E. B. Rimm, M. J. Stampfer, E. Giovannucci, A. Ascherio, D. Spiegelman, G. A. Colditz, and W. C. Willett, "Body size and fat distribution as predictors of coronary heart disease among middle-aged and older us men," American journal of epidemiology, vol. 141, no. 12, pp. 1117–1127, 1995.
[25] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452–457.
[26] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized xgboost based diagnostic system for effective prediction of heart disease," Journal of King Saud University-Computer and Information Sciences, 2020.
[27] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, 2011, pp. 23–30.
[28] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012.
[29] E. Fix and J. L. Hodges, "Discriminatory analysis.nonparametric discrimination: Consistency properties," International Statistical Review/Revue Internationale de Statistique, vol. 57, no. 3, pp. 238–247, 1989.
[30] B. Deekshatulu, P. Chandra et al., "Classification of heart disease using k-nearest neighbor and genetic algorithm," Procedia technology, vol. 10, pp. 85–94, 2013.
[31] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of heart disease using random forest and feature subset selection," in Innovations in bio-inspired computing and applications. Springer, 2016, pp. 187–196.
[32] S. Asadi, S. Roshan, and M. W. Kattan, "Random forestswarm optimization-based for heart diseases diagnosis," Journal of Biomedical Informatics, vol. 115, p. 103690, 2021.
[33] T. G. Nick and K. M. Campbell, "Logistic regression," Topics in biostatistics, pp. 273–301, 2007.