Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 3152618, 12 pages
https://doi.org/10.1155/2021/3152618

*Research Article*

# Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine

**Ebenezer Owusu, Prince Boakye-Sekyerehene, Justice Kwame Appati, and Julius Yaw Ludu**

*Department of Computer Science, University of Ghana, Legon, Accra, Ghana*

Correspondence should be addressed to Justice Kwame Appati; jkappati@ug.edu.gh

Received 9 August 2021; Revised 4 December 2021; Accepted 10 December 2021; Published 23 December 2021

Academic Editor: Hubert Cecotti



Heart diseases are a leading cause of death worldwide, and they have sparked a lot of interest in the scientific community. Because of the high number of impulsive deaths associated with it, early detection is critical. This study proposes a boosting Support Vector Machine (SVM) technique as the backbone of computer-aided diagnostic tools for more accurately forecasting heart disease risk levels. The datasets which contain 13 attributes such as gender, age, blood pressure, and chest pain are taken from the Cleveland clinic. In total, there were 303 records with 6 tuples having missing values. To clean the data, we deleted the 6 missing records through the listwise technique. The size of data, and the fact that it is a purely random subset, made this approach have no significant effect for the experiment because there were no biases. Salient features are selected using the boosting technique to speed up and improve accuracies. Using the train/test split approach, the data is then partitioned into training and testing. SVM is then used to train and test the data. The C parameter is set at 0.05 and the linear kernel function is used. Logistic regression, Nave Bayes, decision trees, Multilayer Perceptron, and random forest were used to compare the results. The proposed boosting SVM performed exceptionally well, making it a better tool than the existing techniques.

## 1. Introduction

Heart disease refers to a variety of conditions that affect the heart from contamination to genetic deficiencies and blood-vessel diseases. These defects are among the topmost causes of deaths globally for all races. In 2016, about 28.2 million adults in the United State were diagnosed with this condition [1] and in 2015 nearly 634000 people died [2] making it the foremost cause of deaths. According to the American Heart Association, a nonprofit organization that funds cardiovascular medical research, one American has a heart attack every 40 seconds [3]. Per the data, there are 720,000 new cases of heart attacks and 335,000 chronic attacks in the United States each year. The form of heart or cardiovascular disease- (CVD-) related morbidity and mortality has been rather fascinating in Sub-Saharan Africa, an area thought to have the world's youngest population. Sub-Saharan Africa remained the only region in the globe where heart disease-related fatalities increased between 1990 and 2013 [4]. The World Health Organization (WHO), for example, has listed heart disease as one of the top two causes of death in Ghana, after diarrheal infections [5]. In 2008, heart disease was the leading cause of death in Ghana among all non-communicable diseases (NCDs) and the major cause of institutional deaths, accounting for 14.5 percent of all deaths reported [6].

Traditionally, a patient's need to know the status of his heart condition was based on the doctor's view. Before doing any test, the doctor will likely perform a few physical checks and interrogate the patient to examine his medical history, regardless of the severity of the cardiac problem. With the exception of blood tests and chest X-rays, any heart disease diagnosis may include the involvement of an electrocardiogram (ECG), which records electrical signals that aid in the discovery of anomalies in the heart's rhythm and structure. Holter monitoring echocardiogram, stress test,

Cardiac Catheterizations, Cardiac Computerized Tomography (CT) Scan, and Cardiac Magnetic Resonance Imaging (MRI) are some of the other therapies. A Holter monitor is a small, wearable device that captures an ECG during a 24- to 72-hour period. Holter monitoring detects heart rhythm abnormalities that are not at all noticeable on a standard ECG. The echocardiogram consists of an ultrasound image of the chest and detailed images of the heart's construction and function. A stress test, often known as a treadmill test or an exercise test, is used by doctors to determine how well the patient's heart can endure workload. The patient will engage in some physical activity or take drugs to raise their heart rate for this test. After that, the actual examination and various photographs of the heart are taken to analyze the underlying reality. In case you ask your doctor if you have heart disease, the standard procedure is for him to assess the likelihood based on risk factors. Age, diabetes, smoking, high blood pressure, being male, and cholesterol are all significant risk factors. According to previous studies, nearly half of those who had coronary attacks had two risk factors: being male and being over 60[7]. As a result, it is incredibly exciting that technology has enabled early diagnosis and risk assessment straightforward before people develop the disease.

Owing to the increased risk of heart disease and the fact that current research forecasts computer-assisted treatments, this study aims to suggest two novel approaches to the problem. To begin, we offer a better algorithm that enhances diagnosis, and then we explain how the proposed method is unquestionably superior to earlier proposed techniques by demonstrating the technique's real implementation. Tables 1, 2, 3, and 4 and Figure 1 demonstrate unequivocally that the suggested method is superior to earlier proposed methods. The remaining part of the study is structured as follows: previous related studies and their challenges are presented in Section 2. The proposed technique and how data is preprocessed as well as previous algorithms employed to solve the problem are discussed in Section 3. The result of the study is then discussed in Section 4. The conclusions are finally drawn in Section 5.

## 2. Related Studies

Several methods have been used to predict the risks of getting heart disease. Genetic algorithms, for example, have been used in a variety of applications. According to [8], the neurofuzzy system combines the capabilities of neuro-adaptive capability and fuzzy logic reasoning for the prediction of the heart disease risk level. The algorithms are generally used for weight optimization when training the model, but there is a serious drawback. Genetic algorithms do not guarantee an optimal solution; hence, the weight optimization may not be completely accurate. In comparison to SVM, Naive Bayes, decision tree, and random forest and genetic algorithms are more complicated to implement and require a large number of parameters to be set in order to achieve a result that is close to optimal. As a result, for small datasets like the Cleveland utilized in this investigation, the genetic algorithm is not appropriate.

The Iterative Dichotomiser 3 (ID3) algorithm, a type of decision tree building algorithm [9], is a relatively simple algorithm that has proven to be effective in other areas but has the drawback of only handling categorical data, so it cannot be used in Cleveland, which is plagued by missing values. If the sample data tested is tiny, this approach is prone to overfitting. As a result, it cannot be used for this research.

Deep neural networks [10], which have shown greater performance in prediction, were also excluded from this study because what is learned with deep neural nets is difficult to comprehend. Furthermore, because learning is progressive, deep neural nets require a large amount of data to train the learning algorithms [11]. When compared to random forest, logistic regression, Nave Bayes, neural networks, and decision trees, the proposed boosting SVM algorithm utilized in this study performed well. On small datasets, these solution approaches are among the best-performing algorithms, and they are also a lot easier to grasp.

Miranda et al. [12] used the Naive Bayes algorithm to forecast this health concern and looked at the related risk levels for adults in their study. In this study, blood and urine test results from the clinical laboratory were used as training datasets. The difficulty with this study is that the authors failed to explore ECG and echocardiography analysis, both of which are crucial in detecting cardiovascular diseases, and the accuracy of 80% obtained is comparably poor. Again, since all the properties in Naive Bayes are expected to be mutually independent, using this predictor to predict heart disease is challenging because finding a collection of predictors that are totally independent of one another is extremely difficult in real life.

In addition, neural networks are widely employed [13, 16]. To predict cardiovascular heart disease, Nandy et al. [14] employed a swarm-artificial neural network. The goal of the research was to increase accuracy. While the study's findings were promising, the accuracy of 95.78% needed to be improved, especially when compared to the study we recommended. Sayad and Halkarnikar [17] proposed a data mining and artificial neural network-based detection approach for cardiac disease. A multilayer perceptron neural network (MLPNN) and a back-propagation algorithm were used in this investigation. The residual dataset was separated into two parts after preprocessing. The MLPNN with backpropagation approach had a 92% accuracy, which is below average. Kim and Kang [18] developed a neural network-based technique for predicting the risk of heart disease using the Korea National Health and Nutritional Examination Survey (KNHANES-VI) dataset [19]. This method consists of two steps. A feature sensitivity-based feature selection is the first phase, followed by a neural network-based prediction model. 3031 people were judged to be at low risk out of 4146, whereas 1115 were found to be at high risk. Dutta et al. [20] suggested a convolutional neural network for predicting heart disease by classifying clinical data that was highly class-imbalanced. The study's findings, on the other hand, were not encouraging.

Table 1: Comparative performance of the training and testing accuracies of methods.

| Method | Accuracy | | |
|---|---|---|---|
| | Training (%) | Testing (%) | Testing time (s) |
| Random forest | 100 | 83.33 | 3.0 |
| Multilayer Perceptron | 75.36 | 80.0 | 5.8 |
| Decision tree | 92.15 | 83.33 | 4.0 |
| Naïve Bayes | 82.13 | 85.5 | 3.2 |
| Logistic regression | 84.06 | 84.44 | 4.5 |
| Boosting SVM | 99.92 | 99.75 | 2.1 |

Table 2: Comparative confusion matrices of different methods.

| Method | Confusion matrix | |
|---|---|---|
| Random forest | 47 | 8 |
| | 7 | 28 |
| Multilayer Perceptron | 46 | 9 |
| | 9 | 26 |
| Decision tree | 48 | 7 |
| | 8 | 27 |
| Naïve Bayes | 47 | 8 |
| | 5 | 30 |
| Logistic regression | 45 | 10 |
| | 4 | 31 |
| Boosting SVM | 51 | 4 |
| | 2 | 33 |

Table 3: Comparing classification report on test data.

| Method | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Random forest | 0.87 | 0.85 | 0.86 | 55 |
| | 0.78 | 0.80 | 0.79 | 35 |
| Multilayer Perceptron | 0.84 | 0.84 | 0.84 | 55 |
| | 0.74 | 0.74 | 0.74 | 35 |
| Decision tree | 0.86 | 0.87 | 0.86 | 55 |
| | 0.79 | 0.77 | 0.78 | 35 |
| Naïve Bayes | 0.90 | 0.85 | 0.88 | 55 |
| | 0.79 | 0.82 | 0.82 | 35 |
| Logistic regression | 0.92 | 0.82 | 0.87 | 55 |
| | 0.76 | 0.89 | 0.82 | 35 |
| Boosting SVM | 0.94 | 0.87 | 0.90 | 55 |
| | 0.82 | 0.89 | 0.85 | 35 |

Table 4: Performances of different methods on Cleveland datasets.

| Author | Method | Accuracy (%) |
|---|---|---|
| Mirza et al. [31] | RBFSVM | 87.114 |
| Amen et al. [32] | Logistics regression | 82 |
| Sajja et al. [33] | SVM | 92–94 |
| Waris & Koteeswaran [34] | Novel KNN | 93 |
| Gupta et al. [35] | Naive Bayes | 88.16 |
| Saini et al. [36] | Hybrid classifier with weighted voting (HCWV) | 82.54 |
| Abdeldjouad et al. [37] | GFS-logicboost-C | 94.17 |
| Motarwar et al. [38] | AdaBoost | 80.32 |
| Alotaibi [39] | Decision tree | 93.19 |
| Gupta et al. [40] | Ensemble of Naïve Bayes, AdaBoost, and boosted tree | 87.97 |
| Proposed method | Boosting SVM | 99.92 |

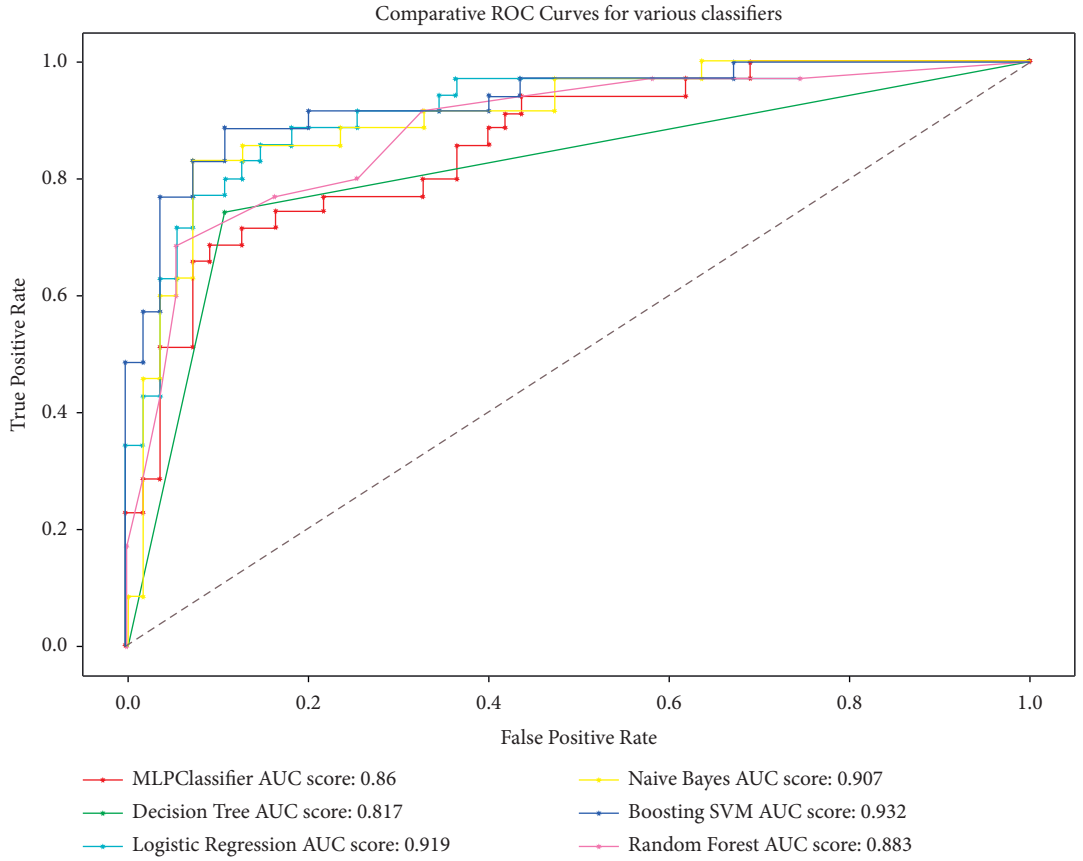


Figure 1: Comparative ROC of various classifiers.

While neural networks are gaining popularity and appear to be realistic, they suffer from data overfitting and temporal complexity. When dimensionality is low, neural networks also fail to converge.

For the same reason, the random forest has been employed in various investigations [21]. Javeed et al. [22] used the Cleveland datasets to construct a random search algorithm (RSA) for feature selection and a random forest model for heart failure prediction. To improve the suggested diagnostic system, the grid search method was applied. Two types of testing were conducted to determine the accuracy of the proposed approach. The first trial only builds a random forest model, whereas the second trial builds the specified RSA-based random forest model. The proposed method has a classification accuracy of 93.33%, and that is not really impressive. Jabbar et al. [23] also proposed a random forest-based classification and feature selection by chi-square and genetic algorithm to predict the risk of heart disease on the Cleveland dataset. The proposed technique outperformed other methods such as Naïve Bayes, decision tree, and neural networks. However, the study's accuracy was only 84%, making it worthless for actual deployment. Decision tree prediction for heart disease has also been proposed [24, 25]. Decision trees, on the other hand, do not work well with missing attributes in the Cleveland datasets if they are not treated with considerable attention, making the outcome inaccurate. The use of logistic regression techniques in the prediction of cardiac disorders is very common. For example, Soleimani and Neshati [26] utilized three logistic regression models with 28 features to predict heart disease risk using 711 data from patients with factors such as severe chest pain, back pain, cold chills, shortness of breath, nausea, and vomiting. However, the study's accuracy of 94.9% was not particularly noteworthy.

A Support Vector Machine (SVM) has also become highly popular. The SVM with sequential minimal optimization strategies was investigated in 2015, with prediction accuracies ranging from 82% to 90%, which was not promising. However, new research into SVM algorithms is yielding better results. Harimoorthy and Thangavelu [27], for example, recently used R studio's SVM-radial bias kernel approach to predict heart disease with 98.7% accuracy.

Based on the favorable results with SVM, we were encouraged to do further examination to improve the technique in the proposed study.

## 3. Materials and Methods

*3.1. Datasets Description.* The Cleveland dataset was used in this study. It is a Cleveland Clinic Foundation dataset containing 14 variables related to patients' vital signs in relation to heart disease. The remaining property is used as the target or projected class, and thirteen of the fourteen qualities are used as predictor variables. Sex, age, type of chest pain, serum cholesterol, resting blood pressure, fasting blood sugar, resting maximum heart rate, electrocardiography, and ST

segment elevation are among the study's 13 predictor variables. The expected characteristics include exercise-induced angina, depression, slope, thallium test result, number of vessels damaged by fluoroscopy, and diagnosis. There were 303 data sets in total, with 6 missing values. The 303 records were reduced to 297 by deleting the 6 tuples that have missing records through the listwise method. Looking at the large size of the data, and the fact that it is a purely random subset, this method had no significant effect on the rest of the data used for the experiment because there were no biases. Table 5 contains descriptions of the datasets.

*3.2. The Proposed Framework.* The proposed framework for the study is shown in Figure 2.

The framework demonstrates the whole methodology of the proposed technique. The explanations are as follows.

*3.3. Feature Importance Estimation.* The feature importance score assigns a numerical value to each data feature; the higher the score, the more significant the feature to the output variable. We extracted the top features for the dataset using the Extra Tree Classifier. The amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for, is used to evaluate the relevance of a single decision tree. The purity (Gini index) was used to choose the separation points. The relevance of each attribute is then summed across all decision trees in the model. The Gini index in Algorithm 1 is presented as follows:

The entire method is developed with the goal of maximizing purity in each split. Purity is defined in (1) as the degree to which the groupings are homogeneous:

$$\text{Gini} = 1 - \sum_{j} p_j^2, \tag{1}$$

where $p_j$ is the probability of an object being classified to a particular class with label $j$ number of times. Figure 3 shows the degree of importance of each feature.

*3.4. Feature Correlation Matrix.* A correlation is a term that describes how features are related to one another. The heatmap makes it simple to see which features are most closely associated with the target variable. Using the seaborn library, we created a heatmap of connected features. Pearson's correlation coefficient was used in this study. This correlation evaluates how closely two numerical sequences are positively connected. We plotted Pearson's heatmap to see the correlation of independent variables. By using AdaBoost as feature selection algorithm, only selected features which have correlation above 0.5, taking into consideration absolute values, were selected. The Seaborn functions automatically perform the statistical estimation required to complete operation. The factors in deep blue in Figure 4 show the highest correlation, namely, max. heart rate and age and ST depression and max. heart rate, indicating that both "age" and "max. heart rate" will play a significant role in predicting heart disease.

*3.5. Boosting SVM Classification.* Boosting is an ensemble meta-algorithm that, in essence, removes dataset biases for machine learning algorithms and upgrades weak learners to strong learners. The goal of the boosting strategy is to enhance prediction accuracy. The following is a description of the adaptive boosting algorithm that was used:

Let $p$ be denoted by positive and $g$ negative samples and let each sample be $(S_i, y_i)$ where $y \in \{\pm 1\}$ represents the corresponding class label. The feature selection algorithm is formulated as follows:

Step 1: initialize the sample distribution by weighting every training sample equally such that the initial weights become $w_{1,i} = 1/2p$ and $w_{1,i} = 1/2g$ for $y = 1$ and -1, respectively. For the iteration $t = 1, 2, \ldots, T$, where $T$ is the final iteration, execute the following.

Step 2: normalize $w_{t,i} \leftarrow w_{t,i} / \sum_{i=1}^{N} w_{t,i}$, where $w_t$ is a probability distribution and $N$ is total number of features.

Step 3: train a weak classifier $h_t$ for feature $j$, which uses a single feature. The training error $\xi_t$ is estimated with respect to $w_t$ as stated in the following equation:

$$\xi_t = \sum_{r} w_{t,i} \left| h_t(x_i) - y_i \right|^2. \tag{2}$$

Step 4: select the hypothesis $h_t^1$ with the most discriminating information, that is to say, the hypothesis with the least classification error $\xi_t^1$, on the weighted samples.

Step 5: compute the weight $\omega_t$ that weights $h_t^1$ by its classification performance as in the following equation:

$$\omega_t = \frac{1}{2} \ln\left[\frac{1}{\xi_t^1} - 1\right]. \tag{3}$$

Step 6: the weight distribution is then updated and normalized with the following equation:

$$w_{t+1,i} \approx w_{t,i}.e^{-\omega_t y_i h_i^1\left(S_i^1\right)}. \tag{4}$$

Step 7: the final feature selection hypothesis $H(S)$ which is a function of the selected features is denoted by the following equation:

$$H(S) = \text{sgn}\left[\sum_{t=1}^{T} \omega_t h_t^1\left(S_t^1\right)\right]. \tag{5}$$

Input the Cleveland training datasets sets, represented by $\{(y_1, x_1), \ldots, (y_N, x_N)\}$. $N = a + b$; where $a$ datasets have $y_i = +1$ and $b$ datasets have $y_i = -1$. The $b$ datasets represent the 0 attributes of the datasets. The scale parameters $x$ and $y$ are the feature vectors selected by the AdaBoost algorithm. The maximal margin separating the hyperplane becomes an optimization problem shown in the following equations:

Table 5: Description of the attributes.

| No | Attribute | Description | Ranges |
|---|---|---|---|
| 1 | Age | Ages of patients taken in years. | 29 to 27 |
| 2 | Sex | 0 for female, 1 for male. | 0, 1 |
| 3 | Chest pain type | There are four types—1 for angina, 2 for atypical angina, 3 for nonangina pain, and 4 for asymptomatic angina. | 1, 2, 3, 4 |
| 4 | Resting blood pressure | Blood pressure of the patient when at rest in mm Hg. | 94 to 200 |
| 5 | Serum cholesterol | The amount of cholesterol in the blood in mg/dL. | 126 to 564 |
| 6 | Fasting blood sugar | Amount of sugar present at fasting. 0 for false—fasting blood sugar is not above 120 mg/dL; 1 for true—fasting blood sugar is above 120 mg/dL. | 0, 1 |
| 7 | Resting electrocardiograph | Values produced by electrocardiography at rest. 0 is normal; 1 is having ST-T wave abnormality; 2 for showing probable or definite left ventricular hypertrophy. | 0, 1, 2 |
| 8 | Maximum heart rate | Maximum heart rate of patient. | 71 to 202 |
| 9 | Exercise-induced angina | Whether or not the patient gets angina when exercise is performed. They are 0 for no and 1 for yes. | 0, 1 |
| 10 | ST depression | Finding on an electrocardiogram wherein the trace of the ST segment is abnormally low below the baseline. Values contain ST depression induced by exercise relative to rest. The abbreviation ST in medical terms means sinus tachycardia. | 1 to 3 |
| 11 | Slope | The slope of the ST segment for peak exercise by the patient. 1 for upsloping, 2 for flat, and 3 for downsloping. | 1, 2, 3 |
| 12 | Number of vessels | Number of vessels colored by fluoroscopy. | 0 to 3 |
| 13 | Thallium stress test result | How well blood flows to the heart while at rest or during exercise. 3 is normal, 6 is a fixed defect, and 7 is a reversible defect. | 3, 6, 7 |
| 14 | Diagnosis | Predicted attribute that contains values showing no presence or presence of heart disease to varying degrees. 0 for no presence, 1 for least likelihood, 2 for moderate likelihood, 3 for a high likelihood, and 4 for very high likelihood. Values 1 through 4 are compressed to a single value, 1, representing the presence of heart disease. | 0 or 1 |

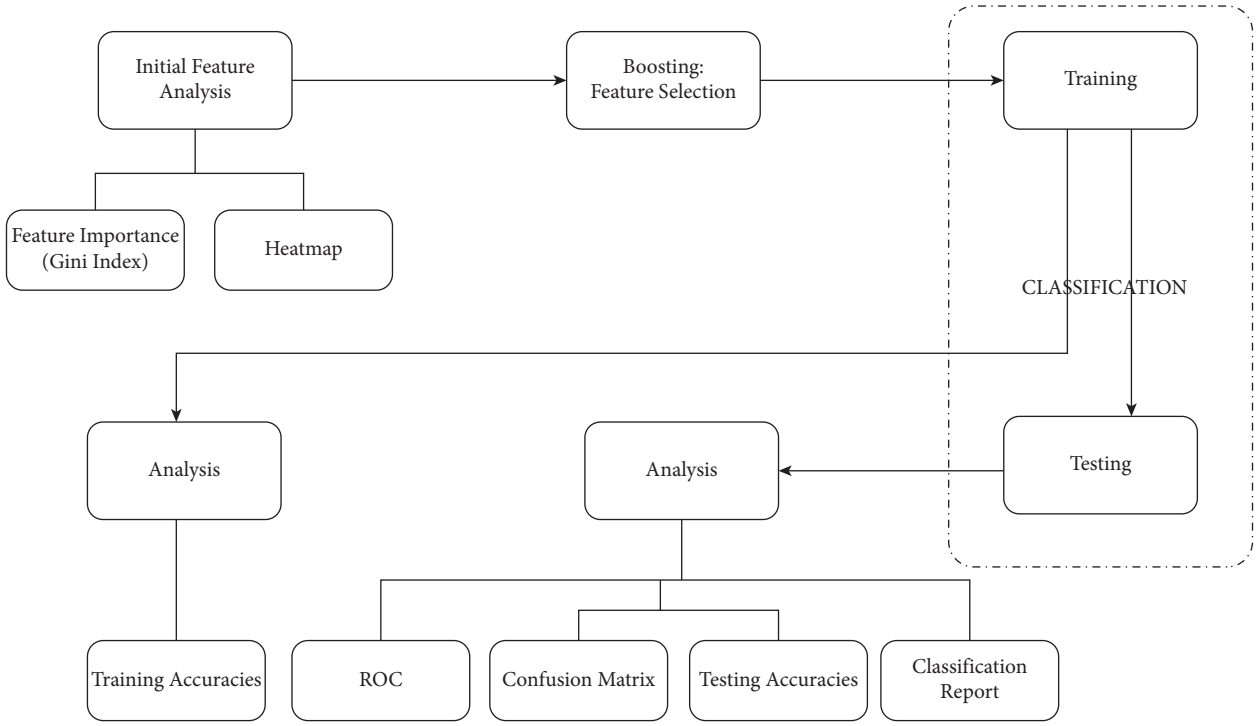


Figure 2: The proposed framework.

```
Gini Index:
  for each branch in split:
    Calculate percent branch represents #Used for weighting
    for each class in branch:
      Calculate probability of class in the given branch.
      Square the class probability.
    Sum the squared class probabilities.
    Subtract the sum from 1.
  Weight each branch based on the baseline probability.
  Sum the weighted gini index for each split.
```

Algorithm 1: Gini index computation.

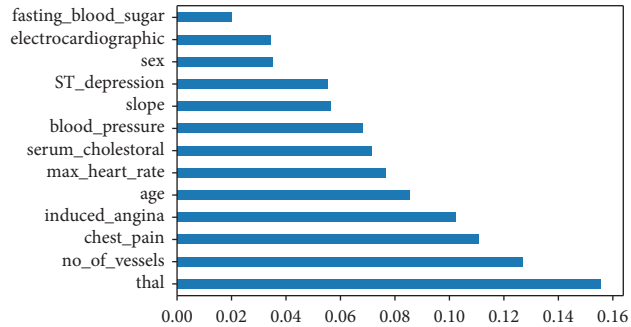


Figure 3: Importance of each feature.

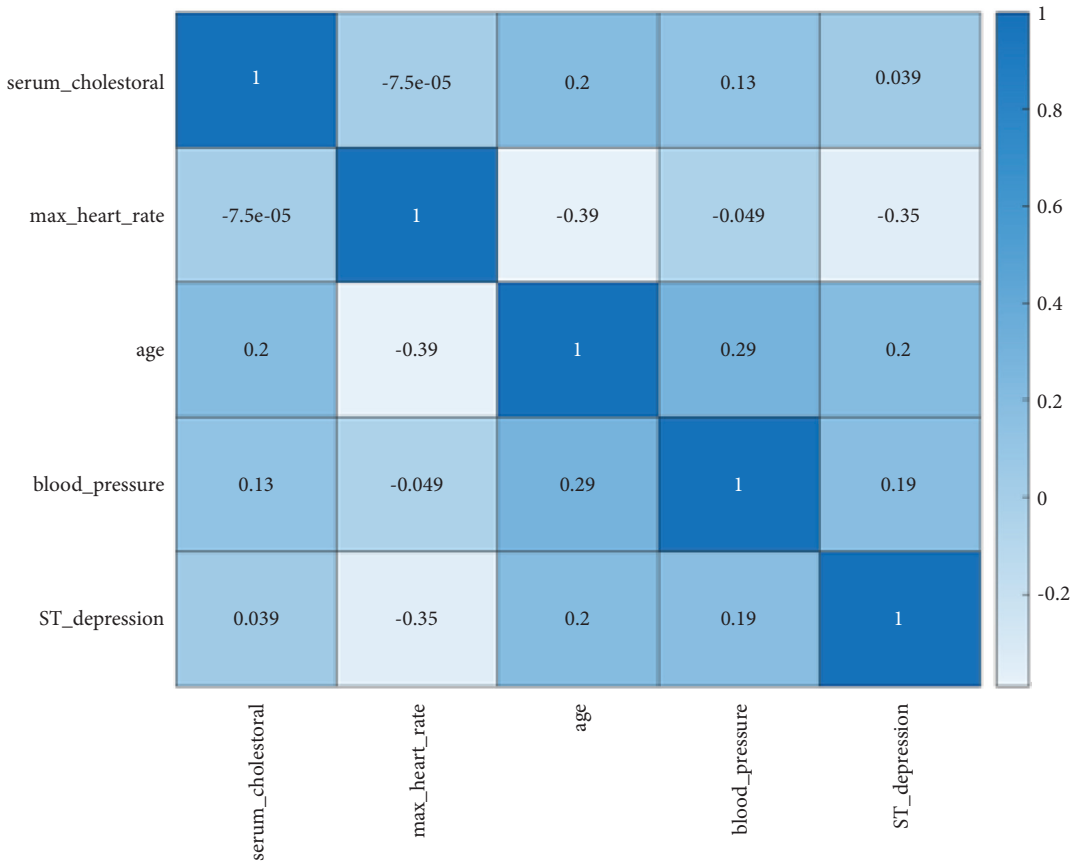


Figure 4: A correlation matrix with heatmap.

$$w^T x + k = 0, \tag{6}$$

$$\min_w \frac{1}{2}\left(w^T w\right), \tag{7}$$

subject to the constraints in the following equation:

$$y_i\left(\left(w^T x_i\right) + k\right) \geq 1. \tag{8}$$

Since $w^T x + k = 0$ and $c(w^T x + k) = 0$ define the same plane, $w$, $c$ is the regularization parameter. $w^T(x_+) + k = 0$ and $w(x_-) + k = 0$, where $(x_+)$ and $(x_-)$ are the respective positive and negative support vectors. The margin is then denoted by the following equation:

$$\frac{w}{\|w\|}\left((x_+) - (x_-)\right) = \frac{w^T\left((x_+) - (x_-)\right)}{\|w\|} = \frac{2}{\|w\|}. \tag{9}$$

The optimal plane is solved by using the convex quadratic programming problem in the following equation:

$$\begin{cases} \min\limits_{w\in R^p, \xi\in R^+} \frac{1}{2}\left(w^T w\right) + c\sum\limits_{i=1}^{N}\xi_t, \\ \text{s.t. } y_i\left(\left(w^T x_i\right) + k\right) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{cases} \tag{10}$$

for $i = 1, \ldots, N, c = 0.05$. The decision boundary of the classifier is expressed as the sum over the support vectors in the following equation:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i Q(x_i, x) + b\right), \tag{11}$$

where $x_i$ is the support vector data, $\alpha_i$ is the Lagrange multiplier, and $y_i$ is the label of membership class $(+1, -1)$ with $n = 1, 2, 3, \ldots, N$. The product $Q(x_i, x)$ represents a linear kernel function, given by the following equation:

$$Q(x_i, x) = \varphi(x_i)\varphi(x). \tag{12}$$

The linear kernel function $Q(x_i, x)$ transforms the original data space into a new space with a higher dimension; this includes the transformation function with dot product, $\varphi(x)$. The reason is to make transformed data easily separable.

*3.6. Model Evaluation Metrics.* An important component of the study is to assess the performance of the proposed method. This is accomplished by comparing the performance of the proposed technique to that of some standard techniques using some acceptable measures. The confusion matrix, classification report, Receiver Operating Characteristic (ROC) curve, and Area under the Curve (AUC) data were used to evaluate the model's performance. The model's test and training accuracies must also be assessed.

*3.6.1. Receiver Operating Characteristic Curve.* A Receiver Operating Characteristic curve is a graph that depicts a classification model's performance over all categorization levels. The curve represents a comparison of the True Positive Rate (TPR) and the False Positive Rate (FPR) in the following equations:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{13}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{14}$$

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively.

*3.6.2. Area under the Curve.* The Area under the Curve (AUC) is the most well-known quantitative index to describe accuracy.

The AUC is computed as follows:

$$\text{AUC} = \frac{1 + \text{TPR} - \text{FPR}}{2}. \tag{15}$$

Generally, an area of 1 means a perfect test and area of 0.5 represents a worthless test. The general acceptable interpretation of AUC values is displayed in Table 6.

*3.7. Comparative Algorithms*

*3.7.1. Comparing SVM with Boosted SVM.* Preliminary experiment was conducted using Support Vector Machine (SVM) and the boosted SVM with the same linear kernel function to determine whether the proposed boosted SVM has significant advantages over the traditional SVM. The results show that the accuracies for SVM and the boosting SVM in terms of training and testing accuracies are 86.83% and 83.41% against 99.92% and 99.75%, respectively. This result is statistically significant ($p < 0.5$). Thus, we follow up to compare the results of the proposed method against Logistic regression, Naïve Bayes, decision tree, Multilayer Perceptron, and random forest which are extensively used in this domain.

*3.7.2. Logistic Regression.* Logistic regression is the best regression analysis to use when the dependent variable or response variable is binary [28]. It works by combining the input variable ($X$) in a linear form and using coefficients to predict an output variable ($Y$) which is a binary value of 0 or 1. The logistic regression technique models the chance of an outcome based on the individual characteristics or input variables ($X$). It is represented mathematically as follows:

$$\log_{10}\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n, \tag{16}$$

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

Table 6: Interpretation of AUC values.

| AUC value | Connotation |
|---|---|
| 0.9 < **AUC** < 1.0 | Excellent |
| 0.8 < **AUC** < 0.9 | Good |
| 0.7 < **AUC** < 0.8 | Fair |
| 0.6 < **AUC** < 0.7 | Poor |
| 0.5 < **AUC** < 0.6 | Insignificant |

*3.7.3. Naïve Bayes.* A Naive Bayes classifier is a simple probabilistic classifier modelled on the application of Bayes' theorem, with strong (Naive) independence assumptions [29]. Naïve Bayes classifier can be trained very efficiently in the context of supervised learning. The Bayesian rule is given in the following equation:

$$P(H|X) = \frac{P(X|A)P(H)}{P(H)}. \tag{17}$$

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

*3.7.4. Decision Tree.* The Gini index, impurity (information gain) approach, which evaluates the degree or chance of a given variable being incorrectly classified when it is randomly chosen, was utilized to compare with the proposed method. The term "information gain" refers to the process of determining which characteristic or attribute provides the most information about a class. The Gini impurity is calculated by summing the probabilities $p_i$, of a class with label $i$, times the probability $\sum_{k\neq i} 1 - p_i$ of a mistake in categorizing that item. The computation is given in the following equation:

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2, \tag{18}$$

where $p_i$ is the probability of an object being classified to a particular class.

*3.7.5. Multilayer Perceptron.* The Multilayer Perceptron (MLP) network is trained using the backpropagation [30], which uses data to adjust the network's weights and thresholds to minimize the error in its predictions on the training set. First, it computes the total weighted input $x_j$, using the following equation:

$$X_j = \sum y_i w_{ij}, \tag{19}$$

where $y_i$ is the activity level of the $j$-th unit in the previous layer and $w_{ij}$ is the weight of the connection between the $i$-th and the $j$-th unit. Next, the unit calculates the activity $y_j$ using the sigmoid function.

*3.7.6. Random Forest.* The training algorithm used is the bagging or the bootstrapping aggregating trees. This creates an ensemble of trees where multiple training sets are generated with replacement, meaning data instance can be repeated. The algorithm is represented as follows.

Given a training set $X = x_1, \ldots, x_n$ with a response, $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample of the training set and fits trees to these samples:

For $b = 1, \ldots, B$

(i) Sample, with replacement, n training examples from $X$, $Y$; call $X_b$, $Y_b$.

(ii) Train a classification tree $f_b$ on $X_b$, $Y_b$.

When training is done, predictions for unseen samples $x'$ are done by determining the average of the predictions from all the individual regression trees on $x'$ as stated in the following equation:

$$\widehat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x'). \tag{20}$$

The process above depicts the original tree bagging algorithm. Random forest, on the other hand, differs in only one way: its algorithm chooses a random subset of features at each candidate split in the learning process (ensemble learning method that tries to reduce the correlation between estimators in an ensemble by training them on random samples of features rather than the entire feature set), also known as feature bagging. The Gini impurity was employed as the criterion because the random forest is based on decision tree and the study is based on classification.

## 4. Results and Discussion

The results of the study are presented as follows: Table 1 shows the different models' training and testing accuracies and its processing time when run on 4 CPUs), ~2.2 GHz processor of 8192 MB RAM. Table 2 shows the confusion matrices and Table 3 shows the classification report.

For each method, the value at the upper left corner is the true positive and the one at the upper right corner is the false positive. The lower right corner is the true negative and the lower left corner is the false negative.

Precision refers to the accuracy with which a judgment is made. The upper row values represent the likelihood of heart illness, whereas the lower row values indicate the likelihood of a decision. The harmonic mean of precision and recall is represented by the F1 score. This is a performance-based statistical measure. The capacity to determine the number of samples that test positive for a specific attribute is known as recall. Figure 1 compares the

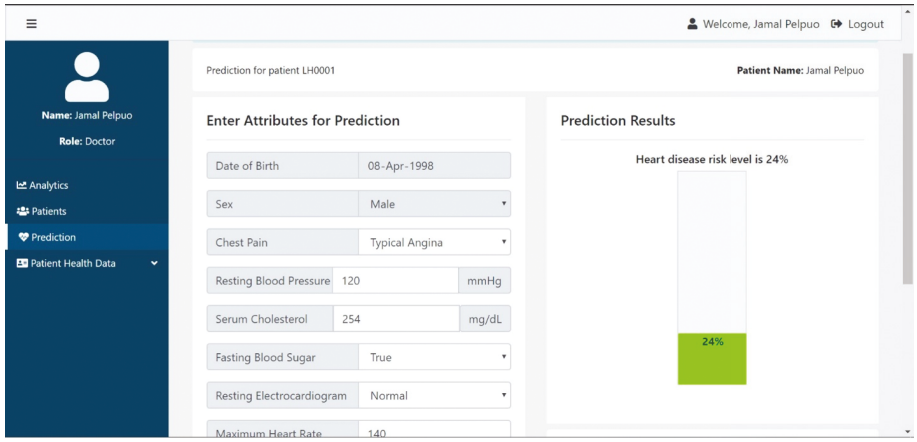


Figure 5: Prediction page showing prediction result for patient with low risk level.

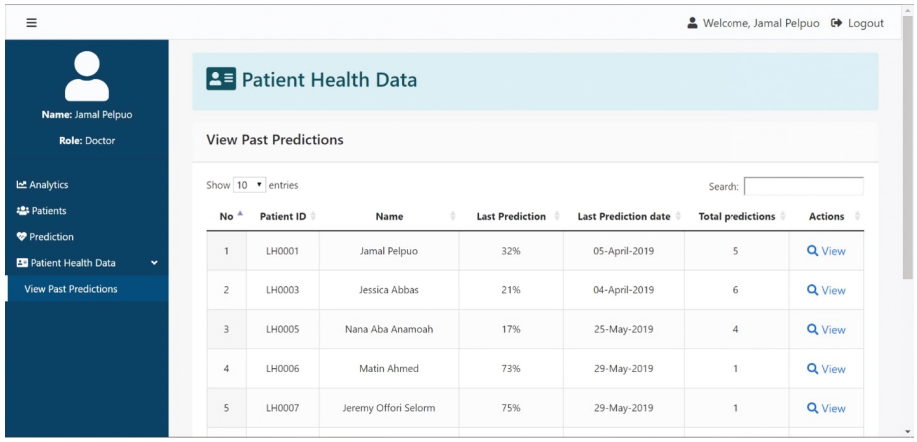


Figure 6: Screen showing prediction history of all patients.

performance of all of the solution models and Table 4 shows the performances of different methods on the Cleveland dataset. We conducted a one-way ANOVA for the results to find if there is a statistically significant difference between the outcome of the proposed technique result and the others in terms of boosting SVM versus random forest, boosting SVM versus Multilayer Perceptron, boosting SVM versus decision tree, boosting SVM versus Naïve Bayes, and finally boosting SVM versus logistic regression. The analysis of the variances, followed by Tukey simultaneous plot at 95% CI, shows that the corresponding means are significantly different ($p < 0.5$) which demonstrates that boosting SVM is the best. Also, tests for the training speed were conducted and the results again show that there was statistically significant difference between groups ($p = 0.029$). A further Tukey post hoc analysis shows that the processing time for the boosting SVM was significantly smaller than all the other techniques after pairing boosting SVM and random forest ($p = 0.041$), boosting SVM and Multilayer Perceptron ($p = 0.027$), boosting SVM and decision tree ($p = 0.038$), boosting SVM and Naïve Bayes ($p = 0.04$), and boosting SVM and logistic regression ($p = 0.035$). All comparatives show that the boosting SVM methodology is extremely promising.

Figures 5 and 6 demonstrate the test application as a proof of concept using the boosting SVM algorithm.

## 5. Conclusion

The study emphasizes the seriousness of cardiac disease and the need of detecting early warning signs. Many machine learning algorithms based on random forest, logistic regression, Multilayer Perceptron, Naive Bayes, and decision trees are being investigated in light of recent studies that call for the automatic detection of dangers. This study proposed a boosting SVM technique to further investigate how to improve prediction accuracy. The technique is based on the Cleveland datasets, which have been utilized successfully and extensively in earlier studies. To reduce misclassification, we preprocessed the data by normalizing it and removing the redundant ones. The feature importance is also computed, which assigns a score to each characteristic in the data; the greater the score, the more relevant the feature to the output variable. Also a heatmap of linked features is produced. The heatmap demonstrates that the most

important factors in predicting heart disease are age and maximum heart rates. Finally, classification is performed using the proposed boosting SVM. For the analysis, confusion matrices, classification reports, ROC, and AUC are all used, and the findings reveal that the provided methodologies performed the best. The proposed method has a recognition accuracy of 99.75%, which is much higher than previous studies. The algorithm has now been enacted and has shown to be pretty useful. In the future, we plan to develop a new ensemble model that combines SVM and AdaBoost to improve accuracy and speed, as well as releasing the app on both Android and iOS.

## Data Availability

The data for this study are publicly available at https://archive.ics.uci.edu/ml/datasets/heart+disease.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## References


[1] National Center for Health Statistics (NCHS), *Health, United States, 2016, with Chartbook on Long-Term Trends in Health*, Government Printing Office, Hyattsville, MD, USA, 2017.

[2] S. L. Murphy, J. Xu, K. D. Kochanek, S. C. Curtin, and E. Arias, "Deaths: final data for 2015," *National Vital Statistics Reports*, vol. 66, no. 6, pp. 1–75, 2017.

[3] E. J. Benjamin, S. S. Virani, C. W. Callaway et al., "Heart disease and stroke statistics-2018 update: a report from the American heart association," *Circulation*, vol. 137, no. 12, p. e67, 2018.

[4] G. A. Roth, M. H. Forouzanfar, A. E. Moran et al., "Demographic and epidemiologic drivers of global cardiovascular mortality," *New England Journal of Medicine*, vol. 372, no. 14, pp. 1333–1341, 2015.

[5] WHO, *World Health Statistics 2010*, World Health Organization, Geneva, Switzerland, 2010.

[6] W. K. Bosu, "Accelerating the control and prevention of noncommunicable diseases in Ghana: the key issues," *Postgraduate Medical Journal of Ghana*, vol. 2, no. 1, pp. 32-33, 2013.

[7] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[8] M. Shah, P. Shukla, and S. Nikam, "Cardio-vascular disease 9 prediction using genetic algorithm and neuro fuzzy system," *International Journal of Latest Trends in Engineering and Technology*, vol. 8, no. 2, pp. 104–110, 2017.

[9] Z. Zhang, S. Zhang, S. Geng, Y. Jiang, H. Li, and D. Zhang, "Application of decision trees to the determination of the year-end level of a carryover storage reservoir based on the iterative dichotomizer 3," *International Journal of Electrical Power & Energy Systems*, vol. 64, pp. 375–383, 2015.

[10] H. C. Shin, H. R. Roth, M. Gao et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[11] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, Article ID e5909, 2020.

[12] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detection of cardiovascular disease risk's level for adults using naive bayes classifier," *Healthcare Informatics Research*, vol. 22, no. 3, pp. 196–205, 2016.

[13] T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network," in *Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 719–723, Antalya, Turkey, October 2017.

[14] S. Nandy, M. Adhikari, V. Balasubramanian, V. G. Menon, X. Li, and M. Zakarya, "An intelligent heart disease prediction system based on swarm-artificial neural network," *Neural Computing and Applications*, pp. 1–15, 2021.

[15] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47, pp. 44–48, 2012.

[16] S. M. Awan, M. U. Riaz, and A. G. Khan, "Prediction of heart disease using artificial neural network," *VFAST Transactions on Software Engineering*, vol. 13, no. 3, pp. 102–112, 2018.

[17] A. T. Sayad and P. P. Halkarnikar, "Diagnosis of heart disease using neural network approach," *International Journal of Advances in Science Engineering and Technology*, vol. 2, no. 3, pp. 88–92, 2014.

[18] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *Journal of Healthcare Engineering*, vol. 2017, Article ID 2780501, 13 pages, 2017.

[19] S. Kweon, Y. Kim, M. J. Jang et al., "Data resource profile: the Korea national health and nutrition examination survey (KNHANES)," *International Journal of Epidemiology*, vol. 43, no. 1, pp. 69–77, 2014.

[20] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," *Expert Systems with Applications*, vol. 159, Article ID 113408, 2020.

[21] Y. K. Singh, N. Sinha, and S. K. Singh, "Heart disease prediction system using random forest," in *Proceedings of the International Conference on Advances in Computing and Data Sciences*, pp. 613–623, Springer, Berlin, Germany, November 2016.

[22] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019.

[23] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach," *Journal of Network and Innovative Computing*, vol. 4, pp. 175–184, 2016.

[24] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease," in *Information and Communication Technology for Competitive Strategies*, pp. 447–454, Springer, Berlin, Germany, 2019.

[25] K. Saxena and R. Sharma, "Efficient heart disease prediction system," *Procedia Computer Science*, vol. 85, pp. 962–969, 2016.

[26] P. Soleimani and A. Neshati, "Applying the regression technique for prediction of the acute heart attack," *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 9, no. 11, pp. 767–771, 2015.

[27] K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 3715–3723, 2021.

[28] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica: Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014.

[29] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.

[30] E. Owusu, J. D. Abdulai, and Y. Zhan, "Face detection based on multilayer feed-forward neural network and haar features," *Software: Practice and Experience*, vol. 49, no. 1, pp. 120–129, 2019.

[31] I. Mirza, A. Mahapatra, D. Rego, and K. Mascarenhas, "Human heart disease prediction using data mining techniques," in *Proceedings of the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pp. 1–5, IEEE, Mumbai, India, December 2019.

[32] K. Amen, M. Zohdy, and M. Mahmoud, "Machine learning for multiple stage heart disease prediction," in *Proceedings of the 7th International Conference on Computer Science, Engineering and Information Technology*, pp. 205–223, Copenhagen, Denmark, September 2020.

[33] G. S. Sajja, M. Mustafa, K. Phasinam, K. Kaliyaperumal, R. J. M. Ventayen, and T. Kassanuk, "Towards application of machine learning in classification and prediction of heart disease," in *Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1664–1669, IEEE, Coimbatore, India, August 2021.

[34] S. F. Waris and S. Koteeswaran, "Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python," *Materials Today: Proceedings*, 2021, In press.

[35] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, "Heart disease prediction using classification (naive bayes)," in *Proceedings of the First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, pp. 561–573, Springer, Berlin, Germany, April 2020.

[36] M. Saini, N. Baliyan, and V. Bassi, "Prediction of heart disease severity with hybrid data mining," in *Proceedings of the 2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pp. 1–6, IEEE, Noida, India, August 2017.

[37] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," in *Proceedings of the International Conference on Smart Homes and Health Telematics*, pp. 299–306, Springer, Berlin, Germany, June 2020.

[38] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive approach for heart disease prediction using machine learning," in *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–5, IEEE, Vellore, India, February 2020.

[39] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 261–268, 2019.

[40] N. Gupta, N. Ahuja, S. Malhotra, A. Bala, and G. Kaur, "Intelligent heart disease prediction in cloud environment through ensembling," *Expert Systems*, vol. 34, no. 3, Article ID e12207, 2017.