Hindawi
Health & Social Care in the Community
Volume 2023, Article ID 1406060, 10 pages
https://doi.org/10.1155/2023/1406060

*Research Article*

# A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction

**Arsalan Khan,[1] Moiz Qureshi ,[2] Muhammad Daniyal,[3] and Kassim Tawiah [4,5]**

[1]*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Department of Statistics, Shaheed Benazir Bhutto University, Shaheed Benazirabad, Nawabshah, Pakistan*
[3]*Department of Statistics, The Islamia University of Bahawalpur, Bahawalpur, Pakistan*
[4]*Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana*
[5]*Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

Correspondence should be addressed to Kassim Tawiah; kassim.tawiah@uenr.edu.gh

Received 29 September 2022; Revised 19 October 2022; Accepted 27 October 2022; Published 20 February 2023

Academic Editor: Andrea Maugeri



Cardiovascular disease (CVD) is a life-threatening disease rising considerably in the world. Early detection and prediction of CVD as well as other heart diseases might protect many lives. This requires tact clinical data analysis. The potential of predictive machine learning algorithms to develop the doctor's perception is essential to all stakeholders in the health sector since it can augment the efforts of doctors to have a healthier climate for patient diagnosis and treatment. We used the machine learning (ML) algorithm to carry out a significant explanation for accurate prediction and decision making for CVD patients. Simple random sampling was used to select heart disease patients from the Khyber Teaching Hospital and Lady Reading Hospital, Pakistan. ML methods such as decision tree (DT), random forest (RF), logistic regression (LR), Naïve Bayes (NB), and support vector machine (SVM) were implemented for classification and prediction purposes for CVD patients in Pakistan. We performed exploratory analysis and experimental output analysis for all algorithms. We also estimated the confusion matrix and recursive operating characteristic curve for all algorithms. The performance of the proposed ML algorithm was estimated using numerous conditions to recognize the best suitable machine learning algorithm in the class of models. The RF algorithm had the highest accuracy of prediction, sensitivity, and recursive operative characteristic curve of 85.01%, 92.11%, and 87.73%, respectively, for CVD. It also had the least specificity and misclassification errors of 43.48% and 8.70%, respectively, for CVD. These results indicated that the RF algorithm is the most appropriate algorithm for CVD classification and prediction. Our proposed model can be implemented in all settings worldwide in the health sector for disease classification and prediction.

## 1. Introduction

The heart is a major part of the human or animal body that plays an essential role in the life of mammals. The heart pumps blood throughout the body parts, thereby supplying oxygen to all parts of the body and controlling the pressure of the blood. The heart performs its function together with the nervous system and the endocrine system. The nervous system helps to control the heart rate while the endocrine system sends hormones as well as blood pressure by causing the human blood vessels to either spasm or relax. However, when the human brain is at rest or under stress, it transmits signals telling your heart to beat more quickly. In stressful situations, our heart beats faster than usual leading to serious heart problems. Aside from stress, heart problems escalate with excessive drinking of liquor, smoking, and heavy fat intake [1, 2]. The rate of health hazards in humans rises as a function of unhealthy dietary habits, excessive stress, lack of good sleep, and lifestyle changes [2].

Cardiovascular disease (CVD) is one of the most noticeable heart diseases which has affected people of all ages. CVD is caused by excessive intake of alcohol, smoking, high blood pressure, high cholesterol level, poor diet, and family history [3]. Del Paoli et al. [4] showed that high blood pressure, unhealthy arguments, and alcohol are highly correlated with CVD. It has been proven that men are at a higher risk of

CVD compared to women [5]. Age is one of the most significant factors for heart disease [6].

In addition to CVD, coronary disease, myocarditis, congenital heart disease, arrhythmias, cardiomyopathy, congestive heart failure, angina pectoris, and myocardial infarction have been classified as acute heart diseases. Each type of heart disease has its symptoms. However, it is very abstruse to identify these heart diseases sharing common high-risk factors like cholesterol level and blood pressure, diabetes, abnormal pulse rate (PR), and many more [7].

The lack of physical fitness due to lifestyle changes may also lead to heart disease for all age groups. A survey reported that seventeen million people in recent years lost their lives due to heart failure [8]. The early detection of heart disease may save a lot of lives provided the patients take their treatments together with their medication seriously and on time [8]. The predicted global number of casualties from CVD in 2015 was 17.7 million, of which 7.4 million were as a result of coronary heart disease and 6.7 million by stroke. According to the World Health Organization (WHO), approximately 54% of deaths from non-communicable diseases in Pakistan are due to cardiovascular problems [9]. Although 17.3 million deaths were caused due to heart disease in 2008, studies by the WHO in 2018 estimated deaths due to heart disease to be around 56.9 million globally [10].

Deep learning models like the backpropagation neural network (BNN) are highly effective for predicting diseases [11]. Likewise, feature selection approaches like decision tree (DT), logistic regression (LR), random forest (RF), Naïve Bayes (NB), and support vector machine (SVM) have been observed to be equally effective in disease prediction [12, 13]. Soni et al. [14] used predictive data mining techniques for the prediction of cardiovascular disease by evaluating the highest accuracy in the DT among a class of predictive machine learning models such as K-nearest neighbour algorithms, neural network classification, and Bayesian classification algorithms [15, 16].

Data mining techniques are very essential in effective healthcare delivery as they can assist in determining whether a patient has a disease or not in healthcare centres (hospitals or clinics). Additionally, it can be employed to rapidly and automatically diagnose people with diseases with great satisfaction [17]. The prediction approach of these techniques may enable all participants in making rational decisions, especially professionals who must make decisions about how to treat patients [18].

Hybrid machine learning models have been applied to predict heart diseases as well as perform optimum classification methods for prediction. Hybrid models give a better optimum output depending on the machine learning method implemented for the execution [8]. Similarly, random forest, decision trees, and hybrid algorithms have been used to predict diseases with high accuracy. The hybrid algorithms were found to have a high accuracy in the neighbourhood of 88.7% for the prediction of disease compared to other models [8].

Nyaga et al. [19], by summarizing available information on aetiology, rates, treatment, covariates, and mortality prevalence arising from heart failure in sub-Saharan Africa, created CVD models. Prasad et al. [20] implemented procedures geared towards predicting heart problems by recapitulating recent studies that utilized artificial intelligence procedures. Wu et al. [21] initiated new CVD forecasting structures by incorporating several procedures in a single hybridized phoned protocol. Their result validated accuracy in diagnosing by implementing a mixture of styles emanating from all methods.

In recent medical fields, a lot of information on diseases is generated through numerous sources. These available data need to be purified as fast as possible with different preprocessing techniques for the required information to fast-track the diagnosis of diseases. This study seeks to develop and propose new methodologies by the utilization of machine learning algorithms to increase the accuracy of the detection of CVD. We investigated and predicted CVD based on hybrid machine learning methods. We used hybrid machine learning models to predict CVD and perform optimum classification methods for the predictions. Our models and approach can be applied in all hospital settings across the world for effective prediction and diagnosis of CVD and other heart diseases. We are hopeful that our suggested technique will be utilized for the detection and prediction of other diseases in general.

We have discussed the materials and methods applied in the proceeding section followed by the results and discussion. The paper ends with the conclusions of the study.

## 2. Materials and Methods

*2.1. Data.* The data were collected from the two largest teaching hospitals, the Lady Reading Hospital (LRM) and the Khyber Teaching Hospital (KTH), in Khyber Pakhtunkhwa (KPK), one of the four provinces of Pakistan. Ethical approval for the inclusion of heart disease patients was sought from the Human Ethical Committees of the two teaching hospitals. The ethics approval certificate number for the Lady Reading Hospital is B371/12/07/2022, while that of the Khyber Teaching Hospital is A418/12/07/2022. A simple random sampling technique was employed in the collection of sample units included in the survey. The sample data consisted of a total of 518 randomly selected heart disease patients.

*2.1.1. Variables in the Study.* The CVD data included the individual output with corresponding factors. The all-inclusive dataset contained the following attributes: age, gender, height, weight, systolic, diastolic, cholesterol, glucose, smoke, alcohol intake, physical activity, cardiovascular disease, and body mass index (BMI). The response variable, CVD, was classified into two categories "presence" and "absence." Furthermore, the data were cleaned of noise, inconsistencies, or any missing observations. We found a few missing observations in the data because some of the patients were discharged from the ward without any proper residential address or mobile/telephone numbers to trace them. As a result, it was very difficult to contact them. Since

our analysis is based on complete data, we replaced the missing data by implementing the usual statistical method such as using median/mode for the categorical data to replace the missing values with the corresponding value. Thus, the data cleaning was completed using the corresponding statistical tools for the preprocessing stage.

Different data mining techniques were utilized in association, classification, clustering, pattern evaluation, and prediction. In the methods section below, we have discussed the techniques extensively.

### 2.2. Methods

*2.2.1. Classification.* Classification is the process of categorizing a given set of data into classes. Classification can be performed for both structured and unstructured data. Predicting the class of the provided data points is the first step in the procedure [22]. Common names for the classes include target, label, and categories. Different statistical and mathematical procedures such as linear programming, decision trees, and neural networks involve classification [23]. That notwithstanding, CVD detection can be recognized through classification procedures because it has two categories, that is, one has CVD or not [24].

*2.2.2. Decision Tree (DT) Algorithm.* The decision tree (DT) is one of the most important predictive modelling and classification methods in learning algorithms that are widely used in practical approaches in supervised learning techniques [25, 26]. It utilizes algorithms that can detect different ways of splitting datasets based on numerous situations. In the classification tree, the response variable is considered a discrete set of values for tree models [26]. DT is a useful contemporary approach to solving decision-making challenges by building models that can be used for prediction through systematic analysis. Internal nodes of a DT indicate a test of the features, branches represent the result, and leaves reflect the decisions that are produced after further computation [27, 28]. We performed our DT as follows:

(I) Divide the dataset into two subdata, that is, training and testing datasets.

(II) In the initial stage, the entire training data are considered the root.

(III) Continuous values are discretized before the model building, whereas categorical values are preferable for feature values.

(IV) Establish subsets such that each subset includes data with the aforementioned feature attributes.

(V) Finally, steps I–IV are repeated for each subset until we get the tree leaves.

In the DT, the prediction for a record class label begins at the root. The values are compared with the root features in the succeeding record characteristics. In this contrast, the equivalent value of the next node to go is displayed [29–31].

*2.2.3. Random Forest (RF) Algorithm.* A random forest (RF) is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}; €k); k = 1, 2, \ldots\}$ where $€k$ are independent and identically distributed random vectors where each tree casts a unit vote for the most popular class at the input of the predictor, $\mathbf{x}$ [32–35].

The RF is an ensemble learning approach for regression or classification used to develop a large number of decision trees at training time. The average prediction of the separated tree is returned for regression purposes, while in the classification, the RF output is the class predicted by the maximum trees. The RF algorithm developed by Ho [36] used a stochastic subspace approach and was reintroduced as a technique for the implementation of a collection of tree predictors by Breiman [37]. RF implements bootstrapping to randomly select training and testing datasets from the original data. After selecting the training dataset, the remaining dataset called out of bag (OOB) is used to estimate the goodness of fit [37].

In the growing phase of the RF, classification and regression tree techniques are developed for tree growth by splitting the local training set at each node with value 1 to a randomly selected subset of the response variable. The growth of the tree continues to the largest extent possible since it does not consider pruning. The phases of bootstrapping and growing of the tree require independent random input quantities. We assumed that these inputs are independent and identically distributed among trees. In that manner, each tree can be viewed as independently sampled for a given training data [37, 38].

For prediction purposes, each tree as well as their terminal nodes are assigned to a class in the forest. Predictions by the trees are performed through voting processes in such a way that the forest returns a class with the maximum number of votes by random selection [39].

*2.2.4. Logistic Regression (LR) Algorithm.* The logistic regression (LR) model is the most accurate in the case of the dichotomous categorical response variable [40]. In the machine learning (ML) algorithm, the LR model can be used for classification purposes [40, 41]. We used the LR model for the classification problem satisfying the cardiovascular-affected respondents. It is implemented on the idea of likelihood by assigning observations to a discrete class being performed using logistic regression [42]. The exponential logit function is utilized for output transformation. The cost function is often restricted by the LR hypothesis to a range between 0 and 1. Consequently, according to the regression hypothesis, linear functions cannot be implemented here because they can have values of either >1 or ≤0. We classified and predicted the CVD patients in the machine learning LR [43] using the function

$$f(x) = \begin{cases} 1, & \text{CVD Present,} \\ 0, & \text{CVD Absent.} \end{cases} \quad (1)$$

*2.2.5. Naïve Bayes (NB) Algorithm.* The Naïve Bayes (NB) method is a supervised learning approach that is based on the Bayes theorem. The NB machine learning method

applies probabilistic techniques in solving classification problems [44]. The main assumption of the NB is the independence (free from multicollinearity) of the predictors fitted in the probabilistic models [45]. A class of classification algorithms predicated on the Bayes theorem is referred to as Naïve Bayes classifiers. It is characterized as a collection of algorithms whereby each algorithm follows the same guiding principle that every combination of features classified is independent of each other pair [46]. In our case, we used the NB classifier to partition the response variable CVD patients into those who have CVD or not for all patients with heart disease [44, 47].

*2.2.6. Support Vector Machine (SVM) Algorithm.* Among the different classification techniques, the support vector machine (SVM) is well known for its discriminative power for classification. The SVM is widely considered in recent times due to its efficiency in most different pattern classification techniques [48]. It has numerous applications ranging from bioinformatics to involuntary language recognition as well as handwritten typescript recognition with sufficient accomplishment. Kim et al. [49] proved that the SVM displays exceptional performance in the classification for prognostic prediction of class III malocclusion. Based on [50], we discuss a brief mathematical theory of the SVM below.

By assuming the binary classification of our response variable, CVD with the convention of linear divisibility for training samples, we have

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, \tag{2}$$

where $x_i \in \mathbb{H}$, such that the design matrix $\mathbf{X}$ belongs to the d-dimensional response space, and the response variable, CVD, is represented by $y_i$, which has a binary class in the vector $\mathbf{Y}$ with $y_i \in (0, 1)$ in our study. The appropriate discriminating equation is given by

$$f(x) = \operatorname{sgn}\{(z, x) + \beta\}. \tag{3}$$

Similarly, $\mathbf{Z}$ represents the vector that determines the coordination of the hyperplane (discriminating plane), and so $\mathbf{Z}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ are offsets [48, 51, 52]. We have infinite possible hyperplanes that are efficiently classified by the training data which can be applied to the validation dataset. The optimal classifier identifies the similar optimal generalized hyperplanes that are nearer or even away from each cluster of objects [53]. The input set of coordinates is considered optimally separated by the hyperplane if there is accuracy in the separation with a maximum distance existing between the nearest components and the support vectors leading to the identification of a specific hyperplane [53, 54].

We used R version 4.1.2 for all our analyses.

## 3. Results and Discussion

The descriptive analysis of the attributes at the aggregate and age levels of the responses of all randomly selected patients with heart disease in the study is represented in Table 1. The table illustrates the numerical output of the cardiovascular disease-associated risk factors. Table 1 indicates the variability in the age proportion of the CVD-affected patients. The exploratory analysis revealed that almost 52.1% of the respondents had CVD at an aggregate level. Furthermore, there was a noticeable variation in the proportion of heart disease concerning different factors such as gender, physical activity, smoking, and so on that correlated with CVD. For instance, a maximum of 4.25% of 60-year-old patients were estimated to have CVD, whereas a maximum of 0.19% of 45-year-old patients had it.

Figure 1 shows the gender, cholesterol level, and glucose levels for all randomly selected CVD patients in the study. The figure shows that a greater proportion of the patients had CVD. Figure 2 presents a line graph for the proportion of gender with respect to the age of patients. The figure shows that CVD is predominant in males compared to females since a greater proportion of the males had the disease. Moreover, the proportion of CVD patients increases from forty years to sixty-one years, which confirms the result of Gulfam Ahmad and Jasim Shah [6].

To achieve our goal, we employed the binary classifier based on a supervised machine learning algorithm for classification to predict the association for the appropriate class of patients [55–57] as proposed by Ramesh et al. [58] and Boukhatem [42]. Table 2 indicates the output of the predictive models that were used for the prediction of CVD.

All five ML algorithms (i.e., DT, SVM, NB, LR, and RF) were used to build the CVD prediction model in two different stages. In the initial stage, the data were split into two separate 70% and 30% groups for training and validation, respectively. In the second stage, however, the data were split into 75% and 25% for training and validation, respectively. The RF model had the highest accuracy of 85.01% with a 95% confidence interval of (0.6608, 0.8043), followed by DT with 83.72% accuracy with a 95% confidence interval of (0.654, 0.7986). The SVM and LR algorithms had the same accuracy of 83.08%, respectively, with a 95% confidence interval of (0.654 and 0.7986), respectively. The NB had the least accuracy of 74.74% with a 95% confidence interval of (0.567, 0.7221). This shows that the RF algorithm is the best predictor of CVD patients. Our outcome confirms the results obtained by the authors in [6, 55–58].

Sensitivity, mathematically defined as the ratio of the total number of true-positive patients to the sum of the number of true-positive and false-negative patients, was used to find the proportion of true patients suffering from CVD [59, 60]. Similarly, the specificity is described according to respondents that are not affected by cardiovascular disease. Specificity, mathematically defined as the ratio of the total number of true negatives to the sum of the number of true negatives and false-positive patients [61], was also used to determine the true proportion of true patients who are not suffering from CVD [62]. The RF algorithm estimated sensitivity and specificity as 86.11% and 65.48%, respectively. That is, our algorithm correctly classified 86.11% of the patients to have CVD but failed to identify 13.89% as having CVD. Similarly, the test correctly classified 65.48% of patients as not having CVD while 34.52% of them were misclassified. Although the DT was not the best in terms of accuracy of prediction, it had the highest sensitivity

Table 1: Descriptive analysis of both response and predictive variables at aggregate and age levels of CVD patients.

| Age | CVD patient | Gender | Height | Weight | Systolic | Diastolic | Cholesterol | Glucose | Smoke | Alcohol | Exercise | BMI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.521 | 0.639 | 164.11 | 74.33 | 128.05 | 91.803 | Extremely high | Extremely high | 904 | 0.959 | 0.786 | 27.685 |
| 40 | 0.0097 | 0.0135 | 168 | 69.1 | 120 | 79.9 | Normal | Extremely high | 0.0212 | 0.0232 | 0.0251 | 24.4 |
| 41 | 0.0039 | 0.0116 | 169 | 74.8 | 122 | 80 | High | High | 0.0212 | 0.0232 | 0.0154 | 26.1 |
| 42 | 0.0077 | 0.0174 | 166 | 78.5 | 128 | 156 | Normal | Extremely high | 0.0232 | 0.0232 | 0.0174 | 28.8 |
| 43 | 0.0039 | 0.0000 | 172 | 61.5 | 125 | 80 | Normal | Normal | 0.0039 | 0.0039 | 0.0019 | 20.8 |
| 44 | 0.0058 | 0.0135 | 168 | 74.5 | 120 | 78.5 | High | Normal | 0.0212 | 0.0212 | 0.0212 | 26.4 |
| 45 | 0.0019 | 0.0039 | 164 | 68.9 | 112 | 71.2 | Extremely high | Extremely high | 0.0135 | 0.0135 | 0.0135 | 25.7 |
| 46 | 0.0116 | 0.0290 | 167 | 85.7 | 122 | 73.3 | High | High | 0.0405 | 0.0425 | 0.0347 | 30.1 |
| 47 | 0.0058 | 0.0135 | 163 | 69.6 | 122 | 75 | Normal | Normal | 0.0193 | 0.0232 | 0.0193 | 26.2 |
| 48 | 0.0232 | 0.0290 | 164 | 76 | 127 | 130 | High | High | 0.0347 | 0.0386 | 0.0270 | 28.3 |
| 49 | 0.0058 | 0.0097 | 160 | 78 | 126 | 81.2 | Normal | Normal | 0.0135 | 0.0154 | 0.0116 | 30.5 |
| 50 | 0.0251 | 0.0444 | 163 | 74.7 | 126 | 109 | High | Extremely high | 0.0502 | 0.0502 | 0.0463 | 28.3 |
| 51 | 0.0270 | 0.0347 | 165 | 74.6 | 131 | 84.4 | High | Normal | 0.0444 | 0.0463 | 0.0347 | 27.5 |
| 52 | 0.0328 | 0.0328 | 164 | 75.7 | 132 | 121 | High | High | 0.0444 | 0.0405 | 0.0367 | 28.1 |
| 53 | 0.0116 | 0.0328 | 164 | 72.3 | 124 | 81.3 | Normal | Extremely high | 0.0425 | 0.0502 | 0.0386 | 26.8 |
| 54 | 0.0290 | 0.0386 | 161 | 73.2 | 129 | 82.5 | Extremely high | Normal | 0.0483 | 0.0541 | 0.0483 | 28.3 |
| 55 | 0.0232 | 0.0425 | 161 | 73.6 | 128 | 80 | High | High | 0.0521 | 0.0521 | 0.0386 | 28.2 |
| 56 | 0.0309 | 0.0309 | 165 | 73 | 133 | 81.9 | High | Extremely high | 0.0444 | 0.0463 | 0.0386 | 26.7 |
| 57 | 0.0290 | 0.0309 | 164 | 69.4 | 122 | 80.2 | Extremely high | High | 0.0425 | 0.0483 | 0.0347 | 25.9 |
| 58 | 0.0270 | 0.0193 | 164 | 72.9 | 131 | 81.7 | High | High | 0.0386 | 0.0463 | 0.0309 | 27.1 |
| 59 | 0.0367 | 0.0367 | 163 | 74.9 | 134 | 82.9 | Normal | Normal | 0.0483 | 0.0502 | 0.0405 | 28 |
| 60 | 0.0425 | 0.0367 | 160 | 71 | 132 | 82 | Extremely high | High | 0.0637 | 0.0656 | 0.0560 | 29 |
| 61 | 0.0425 | 0.0367 | 165 | 75.1 | 133 | 117 | Normal | Normal | 0.0483 | 0.0521 | 0.0405 | 27.8 |
| 62 | 0.0232 | 0.0232 | 167 | 79.5 | 128 | 81.3 | High | High | 0.0386 | 0.0425 | 0.0367 | 28.4 |
| 63 | 0.0135 | 0.0212 | 165 | 80.3 | 129 | 82.1 | Extremely high | High | 0.0251 | 0.0270 | 0.0193 | 29.6 |
| 64 | 0.0251 | 0.0193 | 163 | 75.1 | 135 | 85.6 | High | Extremely high | 0.0328 | 0.0328 | 0.0309 | 28.5 |
| 65 | 0.0232 | 0.0174 | 165 | 70 | 131 | 139 | Extremely high | Extremely high | 0.0270 | 0.0270 | 0.0270 | 25.7 |

CVD patient: proportion of affected CVD patients. Gender: proportion of male patients. Height: mean of height predictor. Weight: mean of weight predictor. Systolic: mean of systolic predictor. Diastolic: mean of diastolic predictor. Cholesterol level: median value of cholesterol. Smoke: proportion of smoker patients. Alcohol: proportion of alcohol patients. Physical activity: proportion of physical activity.

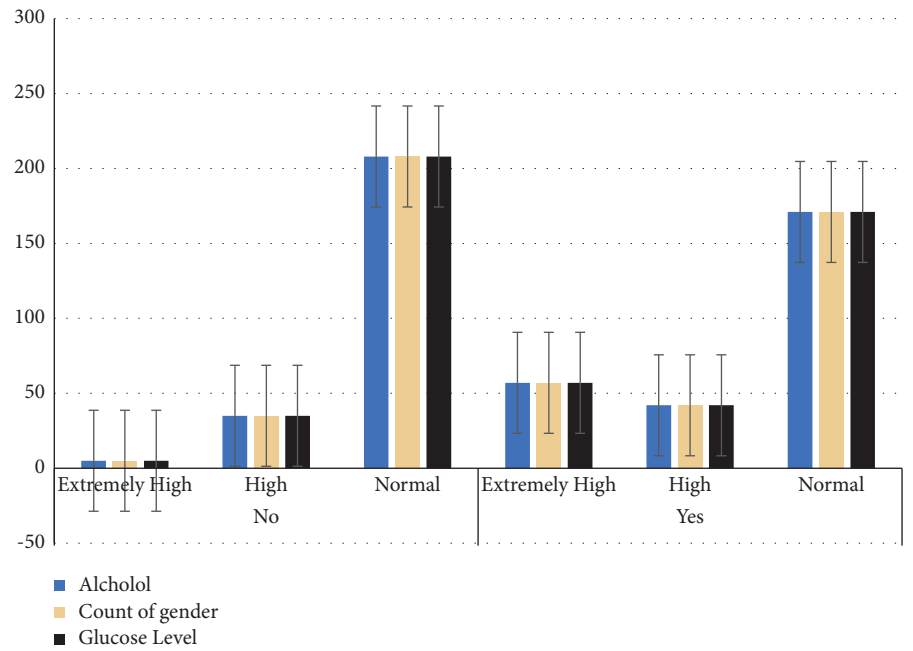


Figure 1: Bar graph with error bars for patient CVD status with gender, cholesterol level, and glucose level.

(90.28%). Our results confirm those of Boukhatem et al. [63]. Figure 3 shows the visualization of all ML algorithm outputs, thereby confirming the superiority of the RF.

Table 3 represents the confusion matrix of the predictive model for 25% of our validation data. The confusion matrix is used to evaluate the performance of the

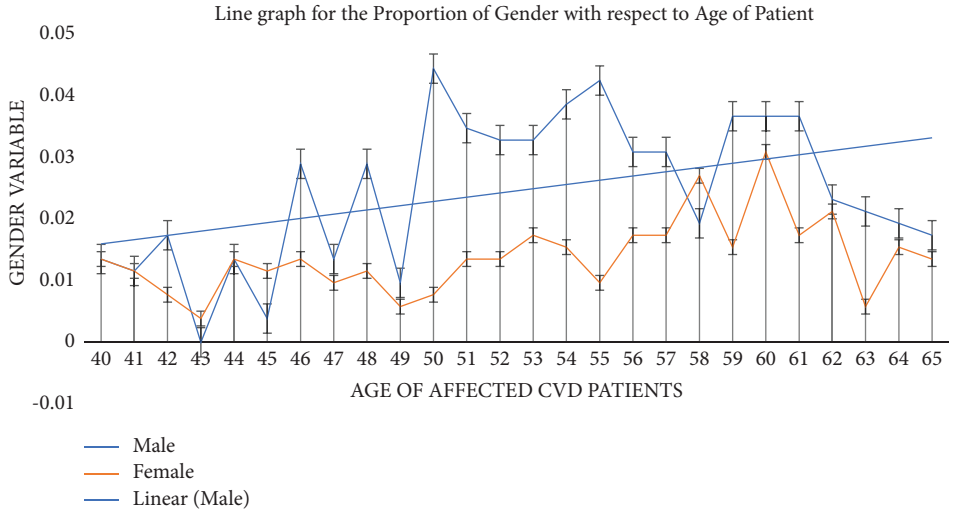


Figure 2: Line chart with error bars for the proportion of gender with respect to the age of patients.

Table 2: An experimental output of the predictive models for CVD patients.

| Output | DT | SVM | NB | LR | RF |
|---|---|---|---|---|---|
| Accuracy | 0.8372 | 0.8308 | 0.7474 | 0.8308 | 0.8501 |
| 95% confidence interval | (0.6608, 0.8043) | (0.654, 0.7986) | (0.567, 0.7221) | (0.654, 0.7986) | (0.6745, 0.8158) |
| Sensitivity | 0.9028 | 0.8472 | 0.8889 | 0.8333 | 0.8611 |
| Specificity | 0.5952 | 0.631 | 0.4405 | 0.6429 | 0.6548 |
| +Predicted value | 0.6566 | 0.663 | 0.5766 | 0.6667 | 0.6813 |
| −Predicted value | 0.8772 | 0.8281 | 0.8222 | 0.8182 | 0.8862 |
| Prevalence | 0.4615 | 0.4615 | 0.4615 | 0.4615 | 0.4615 |
| Detection rate | 0.4167 | 0.391 | 0.4103 | 0.3846 | 0.3974 |
| Detection prevalence | 0.6346 | 0.5897 | 0.7115 | 0.5769 | 0.5833 |
| Balanced accuracy | 0.849 | 0.8391 | 0.7647 | 0.8381 | 0.8579 |

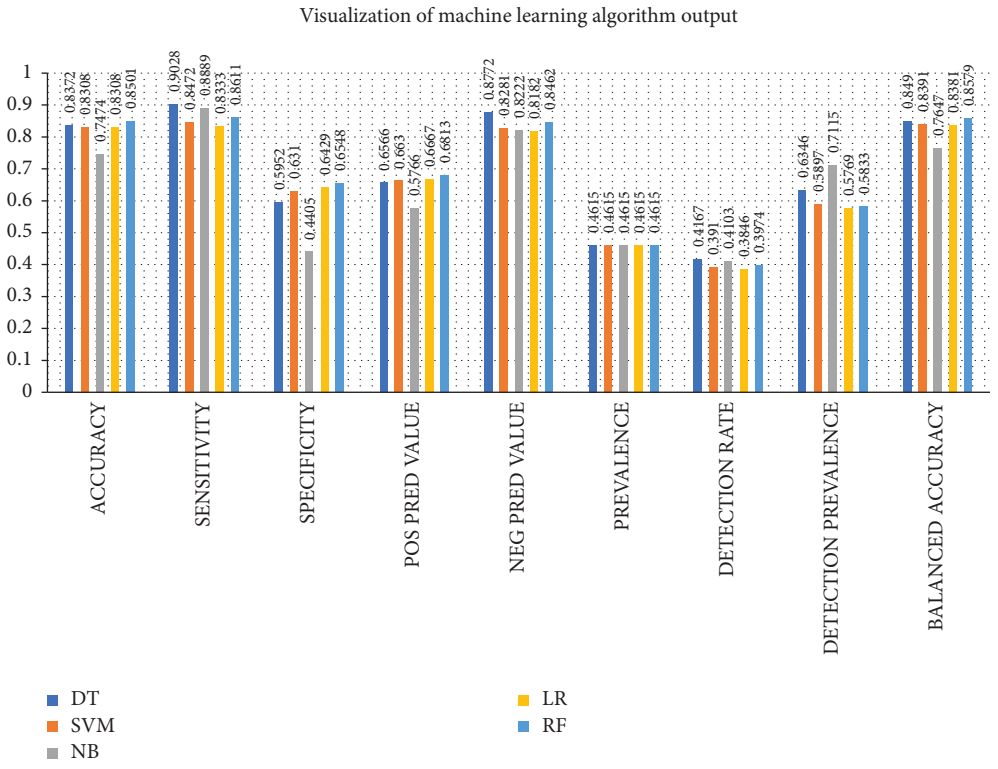


Figure 3: Visualization of the ML algorithm output.

Table 3: Confusion matrix for predictive models.

| Prediction | 1 | 2 | Misclassification error |
|---|---|---|---|
| Decision tree | | | |
| 1 | 65 | 34 | 0.3434 |
| 2 | 7 | 50 | 0.1228 |
| Support vector machine | | | |
| 1 | 61 | 31 | 0.3370 |
| 2 | 11 | 53 | 0.1719 |
| Naïve Bayes model | | | |
| 1 | 64 | 47 | 0.4234 |
| 2 | 8 | 37 | 0.1778 |
| Logistic regression model | | | |
| 1 | 60 | 30 | 0.3333 |
| 2 | 12 | 54 | 0.1818 |
| Random forest model | | | |
| 1 | 61 | 26 | 0.2989 |
| 2 | 6 | 63 | 0.0870 |

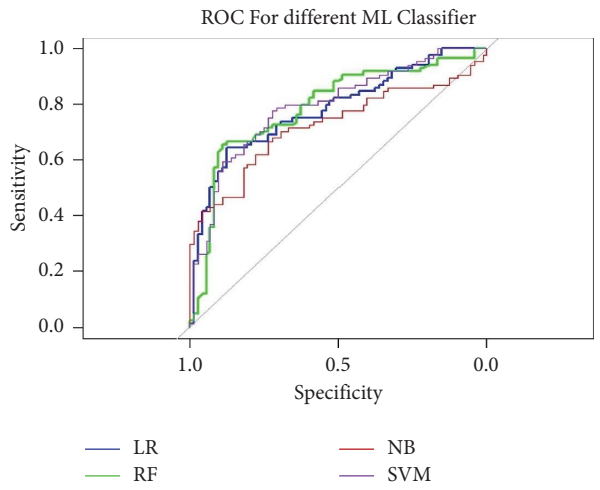


Figure 4: Recursive operating characteristic curve (ROC) for different ML classifiers.

classification algorithm by associating the actual target values for the response variable, CVD patients, with a predicted output of the response by the machine learning model. Just as expected, the RF had the best performance for all evaluation metrics for the confusion matrix. The confusion matrix essentially provides the misclassification error rates for all our ML algorithms. The misclassification error rates for the respondents who are affected were 0.087, 0.1228, 0.1719, 01778, and 0.1818, for the RF, DT, SVM, NB, and LR, respectively, in decreasing order of performance. Thus, the RF performed the best among all competing algorithms, while the LR had the poorest performance among them. Our results are similar to those obtained by O'Kelly et al. [64].

Furthermore, the recursive operating characteristic curve (ROC) was used for the visualization of the accuracy. The ROC uses a matrix to execute the performance of classification algorithms by visualizing the true-positive rate with a corresponding false-positive rate, thereby measuring and highlighting the specificity and sensitivity of the classifiers. Figure 4 shows the ROC for the different classifiers.

The ROC also indicates that the RF algorithm's performance is the best among all classes of ML algorithms. The ROC ranges from 0 to 1, where the nearest to 0 value means it is inept for a given classifier, whereas a value nearest to 1 signifies a more capable algorithm for the classifier. The ROC value is 0.8737 for the RF algorithm which precisely signifies good prediction and classification. The highest ROC for the RF algorithm implies a better ability to discriminate the classes, while the highest accuracy signifies the well-performing ability of the algorithm and the sense of prediction just as in [15, 42, 56].

## 4. Conclusion

Heart diseases are considered a significant apprehension in medical data analysis. The potential of predictive machine learning algorithms to develop the doctor's perception is essential to all stakeholders in the health sector since it can augment the efforts of doctors to have a healthier climate for patient diagnosis and treatment. This study investigated the performance of predictive ML algorithms for CVD patients. CVD is one of the leading causes of mortality worldwide. We used data from the Lady Reading Hospital and the Khyber Teaching Hospital in Khyber Pakhtunkhwa Province, Pakistan. Ethical approval for the inclusion of heart disease patients was sought from the Human Ethical Committees of the two teaching hospitals. Five machine learning algorithms (i.e., DT, RF, LR, NB, and SVM) were implemented for the classification and prediction of CVD. We performed exploratory analysis and experimental output analysis for all algorithms. We also estimated the confusion matrix and recursive operating characteristic curve for all algorithms. The performance of the proposed ML algorithm was estimated using numerous conditions to recognize the best suitable machine learning algorithm in the class of models. The RF algorithm had the highest accuracy of prediction, sensitivity, and recursive operative characteristic curve of 85.01%, 92.11%, and 87.73%, respectively, for CVD. It also had the least specificity and misclassification errors of 43.48% and 8.70%, respectively, for CVD. These results indicated that the RF algorithm is the most appropriate for

CVD classification and prediction. Our proposed model can be implemented in all settings worldwide in the health sector for disease classification and prediction. It can also be implemented in other sectors with a similar function. The main limitation of the study is that detailed patient data and clinical datasets across the globe may be required if we need to have more powerful and considerable prediction models. For improving the accuracy of the ML models and algorithm, high-dimensional data would be more suitable. The ML algorithms used are limited to heart disease prediction studies. Future studies should look into exploring other ML techniques in selecting significant characteristics.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Authors' Contributions

Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, and Kassim Tawiah were responsible for conceptualization, methodology, validation, and visualization. Arsalan Khan, Moiz Qureshi, and Muhammad Daniyal were responsible for data curation, formal analysis, and original draft preparation. Kassim Tawiah and Muhammad Daniyal were responsible for review and editing.

## Acknowledgments

We are grateful to the authorities of the Lady Reading Hospital and the Khyber Teaching Hospital in Khyber Pakhtunkhwa (KPK) Province, Pakistan, for the opportunity to conduct the study and providing us with the ethical approval certificate and waiving the consent. We appreciate all participants for taking time to contribute to this study.

## References

[1] M. G. Tektonidou, "Cardiovascular disease risk in antiphospholipid syndrome: thrombo-inflammation and atherothrombosis," *Journal of Autoimmunity*, vol. 128, Article ID 102813, 2022.

[2] World Health Organization, *The World Health Report: 2000: Health Systems: Improving Performance*, World Health Organization, Geneva, Switzerland, 2000.

[3] M. A. Said, Y. J. van de Vegte, M. M. Zafar et al., "Contributions of interactions between lifestyle and genetics on coronary artery disease risk," *Current Cardiology Reports*, vol. 21, no. 9, pp. 1–8, 2019.

[4] M. De Paoli, D. W. Wood, M. K. Bohn et al., "Investigating the protective effects of estrogen on $\beta$-cell health and the progression of hyperglycemia-induced atherosclerosis," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 323, no. 3, pp. E254–E266, 2022.

[5] S. Jóźwik, A. Wrzeciono, B. Cieślik, P. Kiper, J. Szczepańska-Gieracha, and R. Gajda, "The use of virtual therapy in cardiac rehabilitation of male patients with coronary heart disease: a randomized pilot study," *Healthcare*, vol. 10, no. 4, p. 745, 2022.

[6] H. Gulfam Ahmad and M. Jasim Shah, "Prediction of cardiovascular diseases (cvds) using machine learning techniques in health care centers," *Azerbaijan Journal of High Performance Computing*, vol. 4, no. 2, pp. 267–279, 2021.

[7] C. D. Patnode, N. Redmond, M. O. Iacocca, and M. Henninger, "Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults without known cardiovascular disease risk factors: updated evidence report and systematic review for the us preventive services task force," *JAMA*, vol. 328, no. 4, pp. 375–388, 2022.

[8] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 1329–1333, Coimbatore, India, January 2021.

[9] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: a case study in Pakistan," *PLoS One*, vol. 14, no. 2, Article ID e0210602, 2019.

[10] K. Hill, "Review of the world health report 2000: health systems: improving performance, by World Health Organization," *Population and Development Review*, vol. 27, no. 2, pp. 373–376, 2001.

[11] N. Al-Milli, "Backpropogation neural network for prediction of heart disease," *Journal of Theoretical and Applied Information Technology*, vol. 56, no. 1, pp. 131–135, 2013.

[12] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 619–623, Islamabad, Pakistan, January 2019.

[13] A. Aleem, G. Prateek, and N. Kumar, "Improving heart disease prediction using feature selection through genetic algorithm," in *Advanced Network Technologies and Intelligent Computing ANTIC*, I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma, and P. Verma, Eds., Springer, Berlin, Germany, pp. 765–776, 2021.

[14] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: an overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[15] W. M. Jinjri, P. Keikhosrokiani, and N. L. Abdullah, "Machine learning algorithms for the classification of cardiovascular disease-a comparative study," in *Proceedings of the 2021 International Conference on Information Technology (ICIT)*, pp. 132–138, Amman, Jordan, July 2021.

[16] M. N. Uddin and R. K. Halder, "An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach," *Informatics in Medicine Unlocked*, vol. 24, Article ID 100584, 2021.

[17] M. Kumar, S. Shambhu, and A. Sharma, "Classification of heart diseases patients using data mining techniques," *International Journal of Research in Electronics and Computer Engineering*, vol. 6, no. 3, pp. 1495–1499, 2018.

[18] K. Sudhakar and D. M. Manimekalai, "Study of heart disease prediction using data mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 1, pp. 1157–1160, 2014.

[19] U. F. Nyaga, J. J. Bigna, V. N. Agbor, M. Essouma, N. A. Ntusi, and J. J. Noubiap, "Data on the epidemiology of heart failure in Sub-Saharan Africa," *Data in Brief*, vol. 17, pp. 1218–1239, 2018.

[20] R. Prasad, P. Anjali, S. Adil, and N. Deepa, "Heart disease prediction using logistic regression algorithm using machine learning," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 3S, pp. 659–662, 2019.

[21] C. M. Wu, M. Badshah, and V. Bhagwat, "Heart disease prediction using data mining techniques," in *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology (DSIT 2019)*, pp. 7–11, New York, NY, USA, July 2019.

[22] A. D. Gordon, *Classification*, Chapman and Hall/CRC, London, UK, 2nd edition, 1999.

[23] E. M. De Villiers, C. Fauquet, T. R. Broker, H. U. Bernard, and H. Zur Hausen, "Classification of papillomaviruses," *Virology*, vol. 324, no. 1, pp. 17–27, 2004.

[24] X. Liu, X. Wang, Q. Su et al., "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8272091, 11 pages, 2017.

[25] T. G. Dietterich, "Machine learning," *Annual Review of Computer Science*, vol. 4, no. 1, pp. 255–306, 1990.

[26] P. H. Swain and H. Hauska, "The decision tree classifier: design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[27] M. T. Huyut and H. Üstündağ, "Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study," *Medical Gas Research*, vol. 12, no. 2, pp. 60–66, 2022.

[28] S. Christa, V. Suma, and U. Mohan, "Regression and decision tree approaches in predicting the effort in resolving incidents," *International Journal of Business Information Systems*, vol. 39, no. 3, pp. 379–399, 2022.

[29] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015.

[30] A. Akavia, M. Leibovich, Y. S. Resheff, R. Ron, M. Shahar, and M. Vald, "Privacy-preserving decision trees training and prediction," *ACM Transactions on Privacy and Security*, vol. 25, no. 3, pp. 1–30, 2022.

[31] A. Hamoud, "Selection of best decision tree algorithm for prediction and classification of students' action," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, vol. 16, no. 1, pp. 26–32, 2016.

[32] T. A. Pham and V. Q. Tran, "Developing random forest hybridization models for estimating the axial bearing capacity of pile," *PLoS One*, vol. 17, no. 3, Article ID e0265747, 2022.

[33] X. Wu, C. Peng, P. T. Nelson, and Q. Cheng, "Random forest-integrated analysis in AD and LATE brain transcriptome-wide data to identify disease-specific gene expression," *PLoS One*, vol. 16, no. 9, Article ID e0256648, 2021.

[34] F. Santos, V. Graw, and S. Bonilla, "A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon," *PLoS One*, vol. 14, no. 12, Article ID e0226224, 2019.

[35] A. W. Oehm, A. Springer, D. Jordan et al., "A machine learning approach using partitioning around medoids clustering and random forest classification to model groups of farms in regard to production parameters and bulk tank milk antibody status of two major internal parasites in dairy cows," *PLoS One*, vol. 17, no. 7, Article ID e0271413, 2022.

[36] T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282, Lausanne, Switzerland, August 1995.

[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[38] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2022.

[39] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: random forest," in *Information Computing and Applications, Lecture Notes in Computer Science*, B. Liu, M. Ma, and J. Chang, Eds., pp. 246–252, Springer, Berlin, Germany, 2012.

[40] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[41] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic Regression*, Springer-Verlag, Berlin, Germany, 2002.

[42] C. Y. Bakhoum, S. Madala, C. K. Long et al., "Retinal vein occlusion is associated with stroke independent of underlying cardiovascular disease," *Eye*, 2022.

[43] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar, and T. M. Nithya, "A cardiovascular disease prediction using machine learning algorithms," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 2, pp. 904–912, 2021, https://www.annalsofrscb.ro/index.php/journal/article/view/104048.

[44] R. G. Nadakinamani, A. Reyana, S. Kautish et al., "Clinical data analysis for prediction of cardiovascular disease using machine learning techniques," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2973324, 13 pages, 2022.

[45] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, Article ID 105361, 2020.

[46] K. Rrmoku, B. Selimi, and L. Ahmedi, "Application of trust in recommender systems—utilizing naive Bayes classifier," *Computation*, vol. 10, no. 1, p. 6, 2022.

[47] S. K. Sameer and P. Sriramya, "Improving the accuracy for prediction of heart disease by novel feature selection scheme using decision tree comparing with naive-bayes classifier algorithms," in *Proceedings of the 2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, pp. 1–8, Dubai, UAE, February 2022.

[48] M. Tanveer, T. Rajani, R. Rastogi, Y. H. Shao, and M. A. Ganaie, "Comprehensive review on twin support vector machines," *Annals of Operations Research*, pp. 1–46, 2022.

[49] B. M. Kim, B. Y. Kang, H. G. Kim, and S. H. Baek, "Prognosis prediction for Class III malocclusion treatment by feature wrapping method," *The Angle Orthodontist*, vol. 79, no. 4, pp. 683–691, 2009.

[50] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.

[51] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science & Business Media, Berlin, Germany, 2008.

[52] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[53] J. Fan, J. Zheng, L. Wu, and F. Zhang, "Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models," *Agricultural Water Management*, vol. 245, Article ID 106547, 2021.

[54] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A comprehensive comparative study of artificial neural network

(ANN) and support vector machines (SVM) on stock forecasting," *Annals of Data Science*, pp. 1–26, 2021.

[55] S. Dhar, K. Roy, T. Dey, P. Datta, and A. Biswas, "A hybrid machine learning approach for prediction of heart diseases," in *Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–6, Greater Noida, India, December 2018.

[56] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data," *Medical Engineering & Physics*, vol. 105, Article ID 103825, 2022.

[57] R. Gulhane and S. Gupta, *Machine Learning Approach for Predicting the Heart Disease*, Elsevier, Amsterdam, Netherlands, 2022.

[58] R. Tr, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian Journal of Computer Science*, vol. 2022, pp. 132–148, 2022.

[59] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational & Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[60] E. Diday and J. C. Simon, "Clustering analysis," in *Digital Pattern Recognition*, pp. 47–94, Springer, Berlin, Germany, 1976.

[61] T. S. Madhulatha, "An overview on clustering methods," 2012, https://arxiv.org/abs/1409.23291205.1117.

[62] E. H. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, no. 1, pp. 22–32, 1969.

[63] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart disease prediction using machine learning," in *Proceedings of the 2022 Advances in Science and Engineering Technology International Conferences (ASET)*, pp. 1–6, Dubai, UAE, February 2022.

[64] A. C. O'Kelly, E. D. Michos, C. L. Shufelt et al., "Pregnancy and reproductive risk factors for cardiovascular disease in women," *Circulation Research*, vol. 130, no. 4, pp. 652–672, 2022.