Hindawi
BioMed Research International
Volume 2023, Article ID 6864343, 15 pages
https://doi.org/10.1155/2023/6864343

Research Article

Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques

Niloy Biswas,[1] **Md Mamun Ali**,[1] **Md Abdur Rahaman**,[1] **Minhajul Islam**,[1] **Md. Rajib Mia**,[1] **Sami Azam**,[2] **Kawsar Ahmed**,[3,4] **Francis M. Bui**,[4] **Fahad Ahmed Al-Zahrani**,[5] **and Mohammad Ali Moni**[6]

[1]*Department of Software Engineering (SWE), Daffodil International University (DIU), Sukrabad, Dhaka 1207, Bangladesh*
[2]*College of Engineering, IT, and Environment, Charles Darwin University, Casuarina, NT 0909, Australia*
[3]*Group of Biophotomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh*
[4]*Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK, S7N 5A9, Canada*
[5]*Department of Computer Engineering, Umm Al-Qura University, Mecca 24381, Saudi Arabia*
[6]*Artificial Intelligence & Digital Health, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St. Lucia, QLD 4072, Australia*

Correspondence should be addressed to Kawsar Ahmed; kawsar.ict@mbstu.ac.bd and Mohammad Ali Moni; m.moni@uq.edu.au

Received 21 October 2022; Revised 8 December 2022; Accepted 1 April 2023; Published 2 May 2023

Academic Editor: Alejandro L. Borja

Almost 17.9 million people are losing their lives due to cardiovascular disease, which is 32% of total death throughout the world. It is a global concern nowadays. However, it is a matter of joy that the mortality rate due to heart disease can be reduced by early treatment, for which early-stage detection is a crucial issue. This study is aimed at building a potential machine learning model to predict heart disease in early stage employing several feature selection techniques to identify significant features. Three different approaches were applied for feature selection such as chi-square, ANOVA, and mutual information, and the selected feature subsets were denoted as SF1, SF2, and SF3, respectively. Then, six different machine learning models such as logistic regression (C1), support vector machine (C2), K-nearest neighbor (C3), random forest (C4), Naive Bayes (C5), and decision tree (C6) were applied to find the most optimistic model along with the best-fit feature subset. Finally, we found that random forest provided the most optimistic performance for SF3 feature subsets with 94.51% accuracy, 94.87% sensitivity, 94.23% specificity, 94.95 area under ROC curve (AURC), and 0.31 log loss. The performance of the applied model along with selected features indicates that the proposed model is highly potential for clinical use to predict heart disease in the early stages with low cost and less time.

1. Introduction

Nowadays, machine learning algorithms are vastly used all over the world. In the healthcare industry, machine learning is widely used for predicting disease at an early stage. It saves a lot of people's lives worldwide by predicting their disease at an early stage. Even then, every year, thousands of people are affected and died from heart disease. If machines can predict the early stage of the disease, then, this prediction should reduce the death risk of heart disease. The heart is a significant limb of the human body, and heart disease is the major reason for death in the present world. When it is unable to perform properly, various limbs are obstructed, and then, the brain and several limbs do not work, and a person will die within a few seconds. It is one of the foremost diseases that most commonly affects middle or old-aged people and creates severe complications in the human body [1]. It is difficult to diagnose heart disease because of the number of risk factors. The main symptoms of heart disease are body physical weakness, chest pain, shortness of breath, and rapid or irregular heartbeat [2]. The incidence of heart disease is much higher in the United States (US), and every 34 seconds, one person died due to heart disease [3]. Approximately, almost 26 million people all over the world are affected by heart disease [4]. Every year, 17.9 million people are affected by heart disease, and the worldwide death rate of heart disease is 32% [5]. From 2005 to 2015, India lost up to $237 billion, due to heart-related diseases, estimates made by the World Health Organization (WHO) [5]. Both males and females suffer from heart disease (HD) [6]. Heart diseases are also revealed in older age and middle life, because of exposure to unhealthy lifestyles for many years. After finishing this research, we can predict heart disease at an early stage. This prediction will help millions of heart disease patients worldwide, and millions of lives will be saved. We see heart disease causes a huge loss in the global economy, and predicting it in the early stage will save billions of dollars. For prediction, six machine learning algorithms are used to find the best accuracy. Then, come to the latest conclusion as to which algorithm is better among them.

2. Related Work

In this section, previous heart disease-related study using machine learning methods is discussed, which motivated this work. In this paper, according to Ramalingam et al. [7], a machine learning approach has been employed on some medical datasets and experiments of numerous data. This paper contributes to various model-based algorithms and techniques. Using some supervised algorithms such as Naive Bayes, random forest (RF), decision trees (DT), support vector machine (SVM), and K-nearest neighbor (KNN) are found in these researchers. Based on the accuracy, the implementation of various techniques used in the research was compared. The results accuracy of NB was 84.1584% with SVM-RFE (recursive feature elimination) selected in the 10 most significant features. According to Pouriyeh et al. [8] using 13 attributes, in this research, the NB algorithm has performed an accuracy of 83.49%. In 1951, Fix and Hodges [9] proposed a nonparametric method for pattern classification which is popularly known as the KNN rule. Accuracy of DT and KNN was 82.17% and 83.16%, respectively. Palaniappan and Awang [10] predict the intelligent heart disease prediction in ML algorithms. The algorithms are collectively proposed to achieve accuracy. Using DT, NB, and NN technique to perdition HD, the accuracy of the DT, NB, and NN was 80.4%, 86.12%, and 85.68%. Rabbi et al. [11] used Cleveland standard heart disease dataset and classified the three-technique to prove the accuracy. Predicting the accuracy of the computer-based prediction algorithm, SVM, KNN, and artificial neural network (ANN) are used. In the accuracy, KNN (82.963%) and ANN (73.3333%) are used. They proposed SVM as the best classification algorithm with the highest accuracy to predict heart disease. In the paper, Haq et al. [12] used the UCI dataset to develop using popular algorithms, the cross-validation method, three feature selection (FS) algorithms, and seven classifier performance evaluation metrics such as classification accuracy, specificity, Matthews' correlation, sensitivity, and execution time. Impact on classifier's performance terms to accuracy and execution time is featured. Three feature selection algorithms, mRMR, relief, and LASSO, were used to select the important features, to develop performance, specificity, sensitivity, and accuracy.

Above all those previous studies [7], Ramalingam et al. did a survey which is heart disease prediction using machine learning techniques. The best data will give the best performance of each algorithm [8]. This author worked on the UCI data set with a comprehensive investigation on the comparison of machine learning techniques on heart disease domain. However, the performance of those techniques depends on feature selection algorithms [9]. Palaniappan and Awang use data mining techniques to predict heart disease; this work was done on 909 patients' data. However, data mining is much more effective with big amounts of data [10]. According to Rabbi et al., this paper is done by the same techniques using several algorithms which are given less than 90% accuracy, and those algorithms are applied on MATLAB, and using Python for feature selection techniques, it could be performed better [11]. Haq et al. use much better techniques. But it is not given more than 90% accuracy [12]. If it can handle data more carefully, it may give the best accuracy. Finally, it can be said that they tried to find the best accuracy for predicting heart disease from the UCI dataset's clinical information of patients and correctly predicted below the average of 80% of heart disease patients. They tried to find the best accuracy using all of the features or use some specific feature selection algorithm for a specific machine learning algorithm, and they do not visualize any correlation between features. Also, every other study only shows the prediction score of any algorithm, and they do not describe other performance evaluation matrices like sensitivity, specificity, log loss, and others.

In this study, heart disease (HD) datasets from UCI Machine Learning repository [13] are used. This work is related to the supervised problem of machine learning. Although there has been a lot of research on heart disease, they have tried to solve it using different algorithms. However, it is a complex problem that cannot be solved with a simple machine learning algorithm. This project will be solved by some algorithms such as linear regression (LR) and decision tree (DT). For these analyses, some feature selection methods were applied to the datasets. Several classifiers show the best accuracy in heart disease. In addition, machine learning algorithms play vital roles to predict various health-related diseases in the early stages. The visual representation of the sequential steps for predicting heart disease analysis workflow used in this study is shown in Figure 1.

3. Methodology

In this study, Python 3.8 was used to perform the experiment because it is more accessible to everyone, and it makes it easier to perform rapid testing of algorithms. The workflow of the study is mentioned in Figure 1. The following subsections briefly describe the research methods used in this study.

3.1. Dataset. In this study, the UCI Cleveland dataset [13] is used. This dataset was used in so much research and analysis. We use it for predicting heart disease. The UCI heart disease dataset contains 303 patient records, and each record

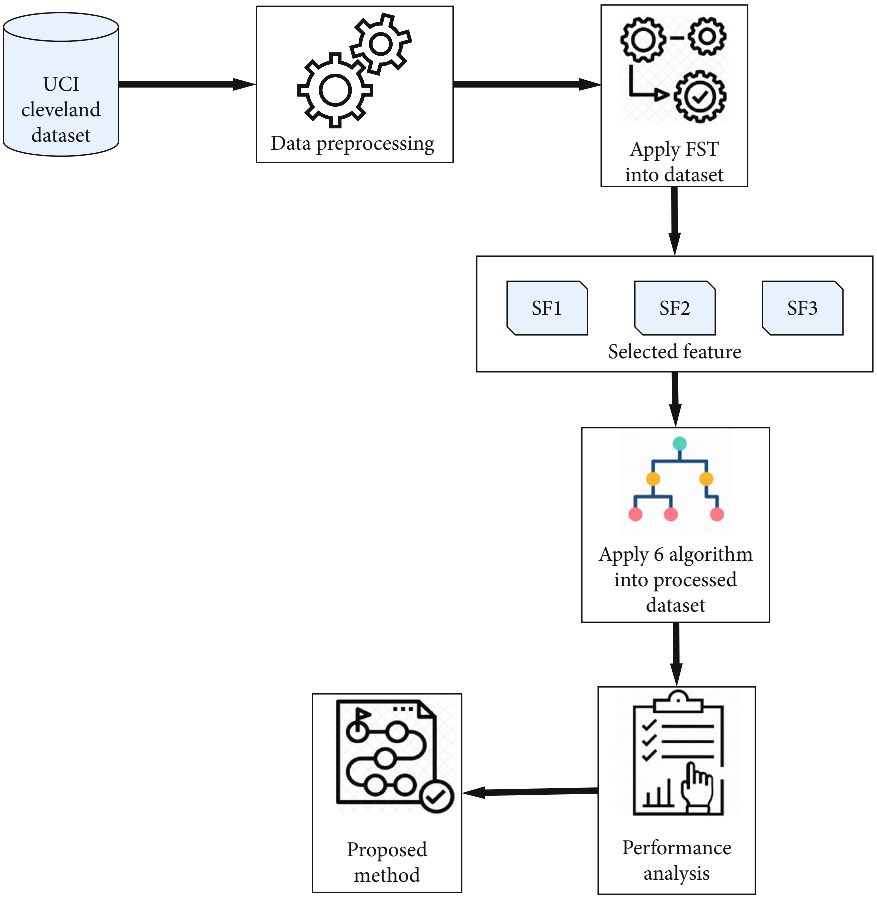

FIGURE 1: Workflow of predicting heart disease.

TABLE 1: Heart disease dataset description.

Serial no.	Feature name	Code	Description
1	Age	AGE	The patient's age in years.
2	Sex	SEX	The patient's sex: male = 1, female = 0
3	cp	CPT	Chest pain type: 0 = typical angina,1 = atypical angina, 2 = nonanginal pain, 3 = asymptomatic
4	trestbps	RBP	Resting blood pressure (in mm)
5	chol	CM	The patient's cholesterol measurement in mg/dl
6	fbs	FBS	The patient's fasting blood sugar > 120 mg/dl. 1 = true, 0 = false
7	restecg	REC	Resting electrocardiographic results: 0 = nothing to note, 1 = having ST-T wave abnormality, 2 = possible or definite left ventricular hypertrophy
8	Thalach	MHR	Maximum heart rate achieved
9	exang	EIA	Exercise-induced angina: 1 = yes, 0 = no
10	Oldpeak	OP	ST depression induced by exercise relative to rest checks the stress of the heart during exercise. The weak heart will stress more.
11	Slope	PES	The slope of the peak exercise ST segment: 0 = up sloping, 1= flat sloping, 2 = down sloping
12	ca	NMV	Number of primary vessels (0-3) colored by fluoroscopy.
13	thal	TS	Thallium stress result: 1, 3 = normal, 6 = fixed defect, 7 = reversible defect

Table 2: Brief description of different feature selection techniques.

FST	Description	Code
ANOVA *F* value	Calculate analysis of variance (ANOVA) between features for classification algorithms.	FST1
Chi-square	Calculate the chi-squared score, which is used to select the highest valued feature between each nonnegative feature.	FST2
Mutual information (MI)	Calculate mutual information between the attributes, which measures the relation between the features.	FST3

Table 3: Feature score using FST1.

Order	Feature	Feature name	Code	Scores
1	9	exang	EIA	70.95
2	3	cp	CPT	69.77
3	10	Oldpeak	OP	68.55
4	8	Thalach	MHR	65.12
5	12	ca	NMV	64.05
6	11	Slope	PES	40.90
7	13	thal	TS	31.80
8	2	Sex	SEX	25.79
9	1	Age	AGE	16.12
10	4	trestbps	RBP	6.46
11	7	restecg	REC	5.78
12	5	chol	CM	2.20
13	6	fbs	FBS	0.24

Table 4: Feature score using FST2.

Order	Feature	Feature name	Code	Scores
1	8	Thalach	MHR	188.32
2	10	Oldpeak	OP	72.64
3	12	ca	NMV	70.89
4	3	cp	CPT	62.60
5	9	exang	EIA	38.91
6	5	chol	CM	23.94
7	1	Age	AGE	23.29
8	4	trestbps	RBP	14.82
9	11	Slope	PES	9.80
10	2	Sex	SEX	7.58
11	13	thal	TS	5.90
12	7	restecg	REC	2.98
13	6	fbs	FBS	0.20

has 13 features. Two classes represent heart patients or normal cases in our target label. The dataset matrix information is given in Table 1.

3.2. Data Preprocessing. In this study, data were preprocessed after collection. There are 4 records on NMV and 2 records on TS that are incorrect in the Cleveland dataset. All those records with incorrect values are replaced with optimal values. Next, StandardScaler is used for ensuring that every feature has mean 0 and variance 1 and bringing all the features to the corresponding coefficient.

3.3. Feature Selection. Feature selection plays an important role in the machine learning process because sometimes, the dataset contains many irrelevant features that are affecting the accuracy of the algorithms. Feature selection helps to reduce those unconnected features and improve the performance of the algorithms [14]. It used different feature ranking techniques [15] to rank the most important feature based on their relevance. In this study, three well-known feature selection algorithms are used to identify important features based on their score.

3.3.1. ANOVA F Value. ANOVA test is a prediction technique to measure similarity or pertinent feature and to reduce the high dimensional data and identify the important feature by feature space and improving the classification accuracy. Here, the formula [16] is used:

$$F = \frac{\sum_{j=1}^{i} N_j \left(\underline{x}_j - x\right)^2 / (J - 1)}{\left(\sum_{j=1}^{i} \left(N_j - 1\right)^{s2} j / (N - 1)\right)}. \tag{1}$$

3.3.2. Chi-Square. This test is a statistical hypothesis testing system, and also, it is written as x^2 test. It is calculated between the observed value and the expected value. This formula [17] is given below.

$$X^2 = \sum \frac{\left(o_j - e_j\right)}{e_i}. \tag{2}$$

3.3.3. Mutual Information (MI). A couple of decennial mutual information has acquired considerable attention for its application in both machine learning. MI is calculated between two variables and features [18], and this is the mathematical equation for calculating mutual information between the features.

$$I(X\,;Y) = H(Y) - H\left(\frac{Y}{X}\right). \tag{3}$$

As previously mentioned in this experiment, ML algorithms were used such as LR, SVM, KNN, RF, NB, and DT.

3.4. Classification and Modeling. The models used for predicting heart disease are described sequentially. Each algorithm is applied following that sequence. Various types of classification algorithms are available for data analysis. In

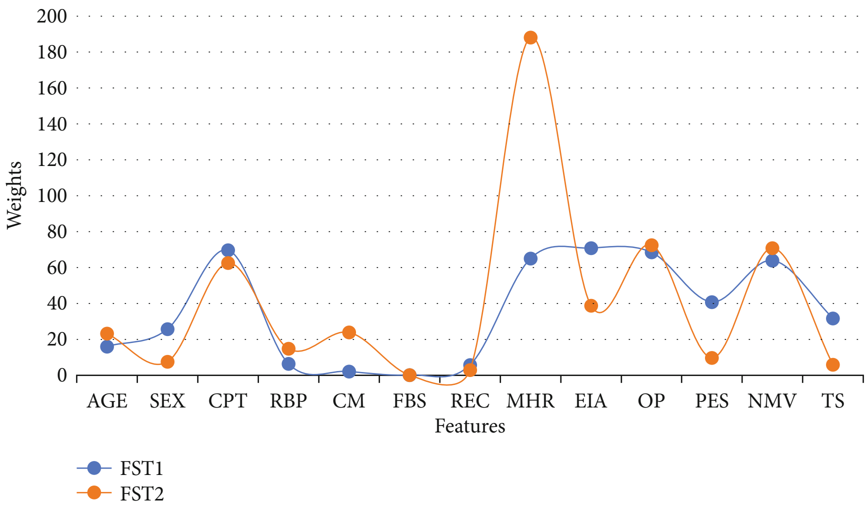

Figure 2: Feature score by FST1 and FST2.

Table 5: Feature score using FST3.

Order	Feature	Feature name	Code	Scores
1	3	cp	CPT	0.17
2	13	thal	TS	0.14
3	12	ca	NMV	0.11
4	9	exang	EIA	0.10
5	8	Thalach	MHR	0.10
6	10	Oldpeak	OP	0.09
7	5	chol	CM	0.08
8	11	Slope	PES	0.08
9	2	Sex	SEX	0.05
10	4	trestbps	RBP	0.03
11	1	Age	AGE	0.01
12	6	fbs	FBS	0.00
13	7	restecg	REC	0.00

this study, six types of classification algorithms are used. A brief discussion of each algorithm is given below.

3.4.1. Logistic Regression. Logistic regression model, the probabilities for classification problems with two possible outcomes, can be regarded as y when $y \in [0, 1], 0$ is a negative class and 1 is a positive class [12], and a hypothesis is designed based on it $h(\theta) = (\theta^n A)$. Consider that the hypothesis value is $h\theta(a) \geq 0.5$, then predict value $y = 1$. Consider that the hypothesis value is $h\theta(a) \leq 0.5$, then predict value $y = 0$. Here, the logistic regression sigmoid function is written:

$h\theta(a) = m(\theta^n \mathrm{A})$, where

$$f(y) = \frac{1}{1 + a^{-y}},$$
$$h(a) = \frac{1}{1 + a^{-y}}. \quad (4)$$

3.4.2. Support Vector Machine. SVM creates an effective decision boundary (hyperplane) between the two classes [19]. The main focus when drawing a decision boundary is centered on the maximum distance of the nearest data point of both classes. Although the radial base function is used as a kernel, SVM automatically determines centers, mass, and doorstep and reduces the upper limit of the expected test error. In the case of the study, we consider the support vector function as a radial base function. Here, p is the length of the vector. It clarifies as

$$R\left(p, p'\right) = \text{expexp}\left(-\frac{\left\|p - p'\right\|^2}{2\sigma\text{^}2}\right). \quad (5)$$

Here, $\left\|p - p'\right\|^2$ is identified as the squared Euclidean distance between vector and σ.

3.4.3. K-Nearest Neighbor. KNN uses a training set directly for classifying the test data. Which refers to the number of KNN. To test each data, it calculates all the training data and the distance between them. Then, test data will be assigned to be used by multiplicity voting and class label. The Euclidean distance measure equation is given below:

$$W_e = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}. \quad (6)$$

3.4.4. Random Forest. Random forest is the most powerful algorithm of supervisory machine learning algorithms. It is principally used for classification problems. As we see, a forest is made up of many trees, which means almighty forest. This algorithm similarly builds a decision tree based on data samples. Here, we use it for efficient heart disease results.

3.4.5. Naive Bayes. In potential, the Bayes theorem is used for calculating probability and conditional probabilities. A patient may have certain symptoms (side effects). The possibility of the proposed conclusion being true may be due to the use of the Bayes hypothesis. Here, M = target variable

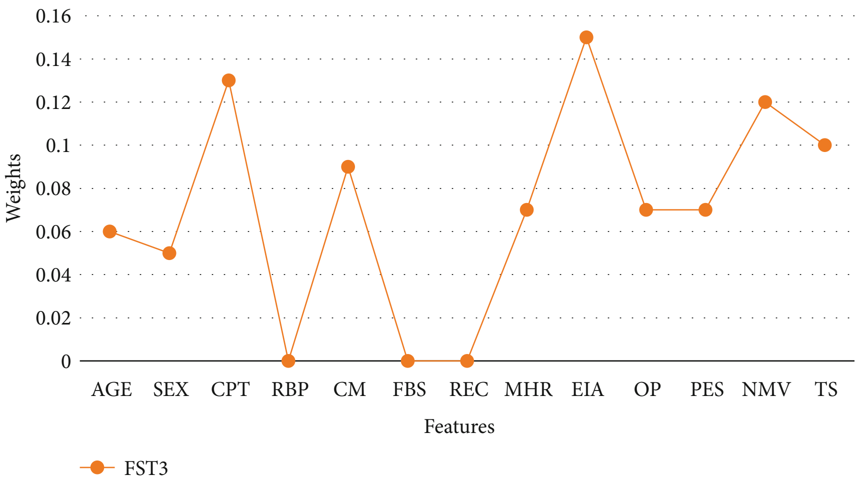

Figure 3: Feature score by FST3.

Table 6: Selected features.

Selected feature	Selected features
SF1	Age, sex, CPT, RBP, CM, FBS, REC, MHR, EIA, OP, PES, NMV, TS
SF2	Age, sex, CPT, CM, MHR, EIA, OP, PES, NMV, TS
SF3	Age, sex, CPT, MHR, EIA, OP, PES, NMV, TS

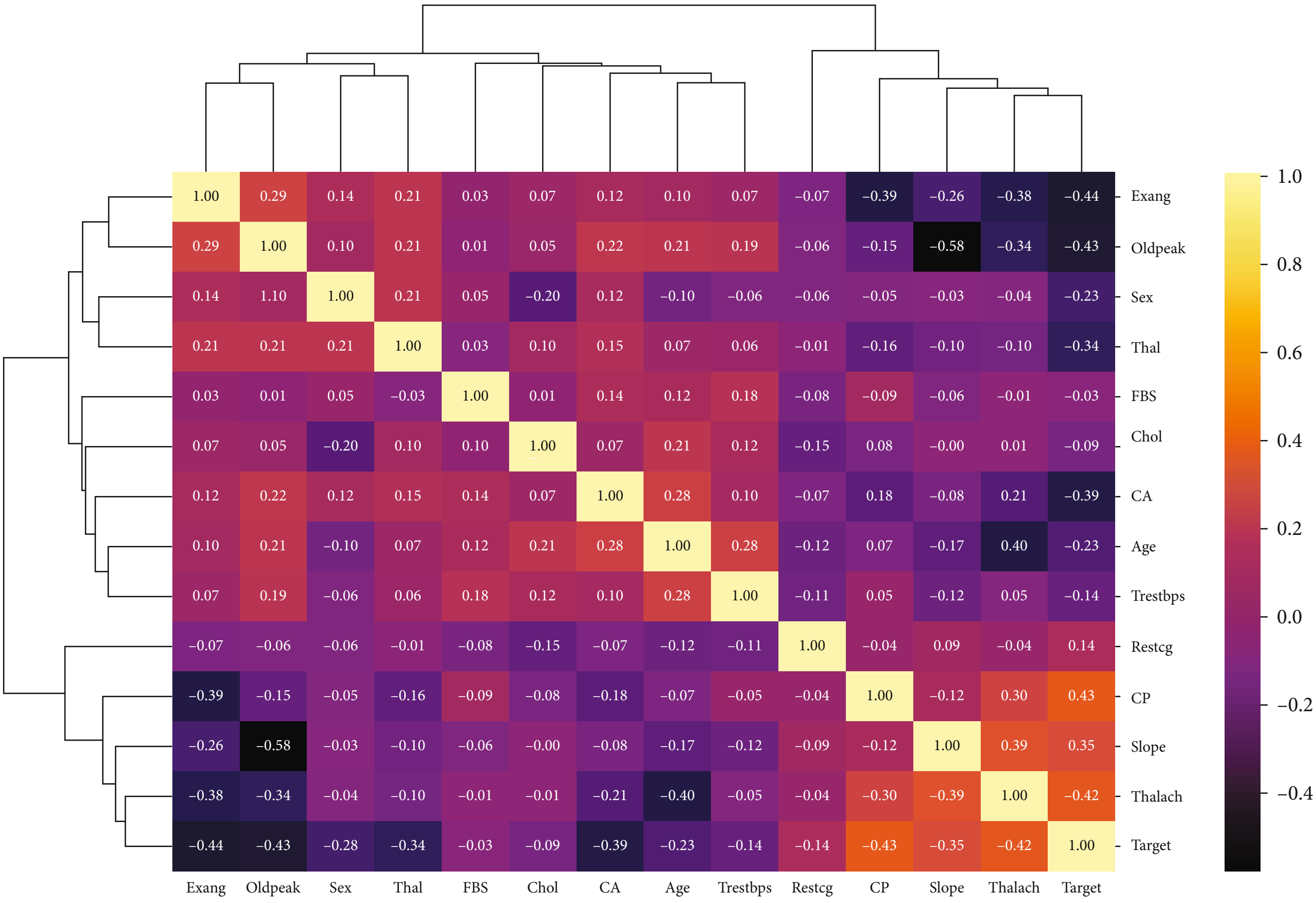

Figure 4: Correlation matrix heat map.

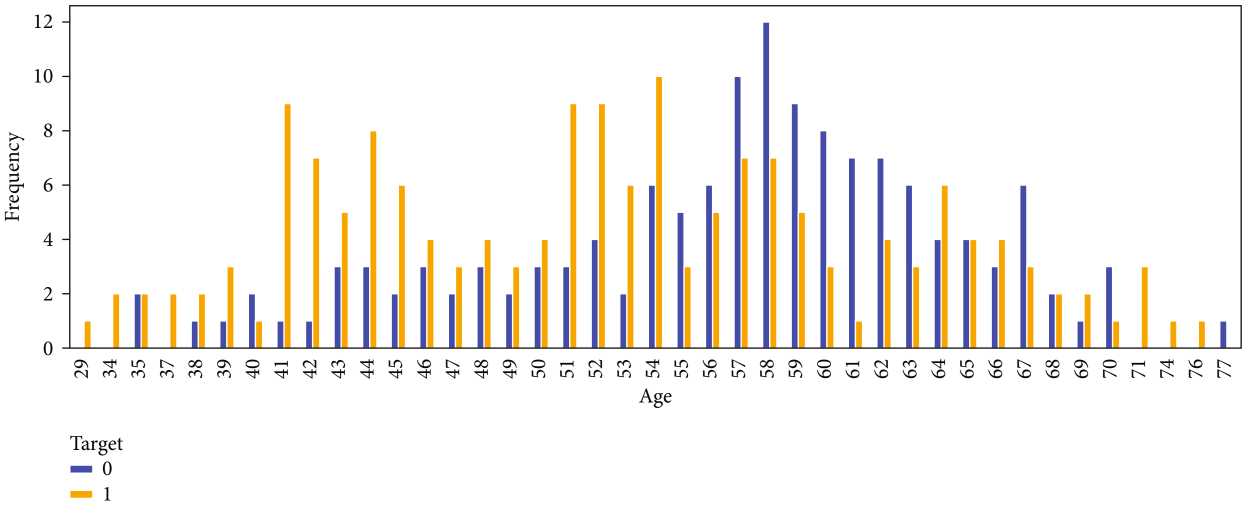

Figure 5: Correlation between patients' age with the disease.

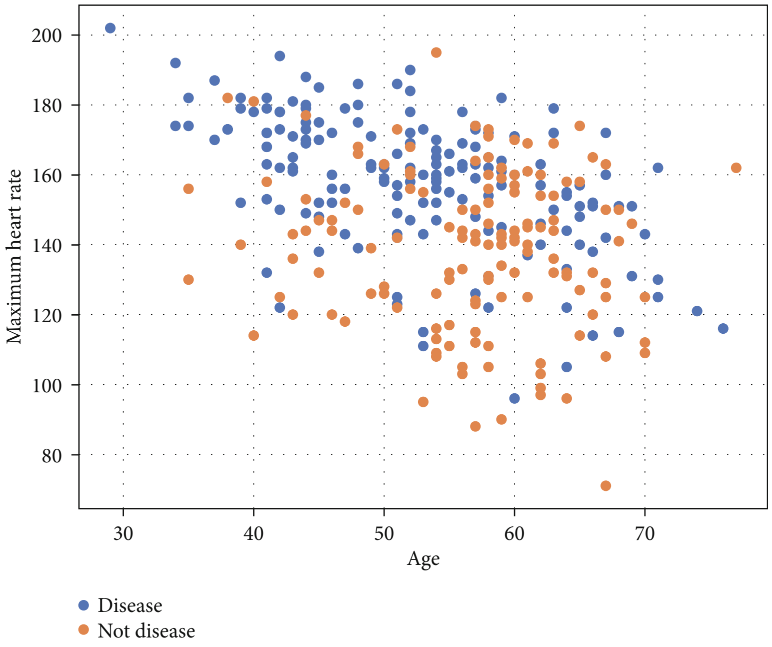

Figure 6: Correlation between patients' maximum heart rate with the disease.

Table 7: Accuracy of different algorithms.

	Selected features	C1	C2	C3	C4	C5	C6
	SF1	93.41	78.02	87.91	90.11	89.01	83.52
Dataset	SF2	93.41	76.92	86.81	89.01	90.11	92.31
	SF3	93.41	75.82	84.61	94.51	90.11	91.21

and N = attributes. The formula is given below:

$$P\left(\frac{M}{N}\right) = \frac{P(N/M)P(M)}{P(N)}. \tag{7}$$

3.4.6. Decision Tree. Decision trees are the most powerful way to classify problems. In this method, the entropy for each property is calculated in two or more similar sets based on more predictive values, and then, the data set is divided on the basis of minimum entropy or maximum data gain.

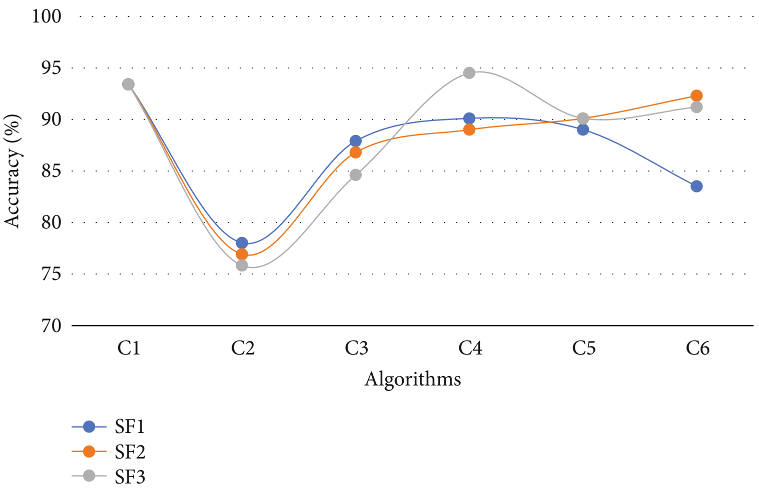

Figure 7: Accuracy of different algorithms.

Table 8: Sensitivity of different algorithms.

	Selected features	C1	C2	C3	C4	C5	C6
Dataset	SF1	94.74	70.83	87.18	94.28	87.5	80.49
	SF2	94.74	69.38	83.33	91.66	87.8	94.6
	SF3	94.74	71.42	80.95	94.87	87.8	92.1

The entropy and information gain formula are given as follows:

$$\text{Entropy}(E) = \sum_{i=1}^{c} - q\, iq_i,$$
$$\text{Info} - \text{gain}(E, G) = \text{Entropy}(E) - \sum_{v \in \text{Values}(G)} \frac{|Gv|}{|E|}(Sv). \tag{8}$$

Multiplex evaluation metrics such as accuracy, sensitivity, specificity, AUROC, and log loss were evaluated to present the results of different algorithms and comparison performance based on these metrics. These matrices were represented by calculating the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values. The below section describes more about these metrics. After completing the analysis, the best algorithm is represented which achieves the highest outcomes.

3.4.7. Performance Evaluation Matrices

(1) Accuracy. The accuracy is determined by the matrices called confusion matrices. The confusion matrices are $N \times N$ matrices, which are used for assessing the performance of the classification model. The formula used to calculate the accuracy is

$$A_{cc} = \frac{(TP + TN)}{(TP + TN + FP + TN)}. \tag{9}$$

(2) Sensitivity. It is the measurement of the proportion of true positive cases and predicts that all values are positive. For calculating sensitivity, the used formula is

$$S_{en} = \frac{(TP)}{(TP + FN)}. \tag{10}$$

(3) Specificity. It calculates the proportion of true negative cases and predicts that all values are negative. The formula used to calculate the specificity is.

$$Sp_{ec} = 1 - \left(\frac{FP}{FP + TN}\right). \tag{11}$$

(4) AUROC. This evaluation matrix is used for checking classification model performance. For calculating AUROC, the used formula is

$$TPR = \left(\frac{TP}{TP + FN}\right),$$
$$FPR = 1 - \left(\frac{FP}{FP + FN}\right). \tag{12}$$

(5) Log loss. This is a classification loss function used to evaluate the performance of machine learning algorithms. The closer to zero will be the value of the log loss model and will become more accurate. For calculating log loss, the used formula is

$$Lg = \frac{-\sum_{y=1}^{j}\sum_{x=1}^{n} f(x, y) \log\left(p()x, y\right)}{n}. \tag{13}$$

4. Experimental Setting

In this analysis, Jupyter notebook is used to perform heart disease prediction of the dataset. It helps to create documents with live codes and easy to visualize various data relation diagrams of the dataset. In this analysis, firstly, the UCI

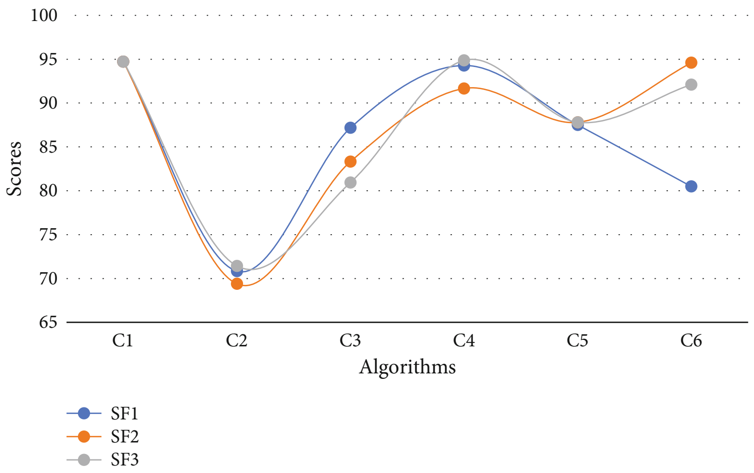

Figure 8: Sensitivity of different algorithms.

Table 9: Specificity of different algorithms.

	Selected features	C1	C2	C3	C4	C5	C6
Dataset	SF1	92.45	86.05	88.46	87.5	90.2	86.0
	SF2	92.45	85.71	89.79	87.27	92.0	90.70
	SF3	92.45	79.59	87.75	94.23	92.0	90.57

HD dataset is cleaned using Pandas 1.1 and NumPy 1.19.0 libraries of Python and then preprocessed it using the StandardScaler algorithm from Scikit-learn [20] library of Python. Secondly, some feature selection algorithm is applied to find the feature importance, then made three different selected feature (SF) sets. Thirdly, the dataset was split into train and test sets, 70% of the data is used as a train set, and the rest is used as a test set. In the last, this 70% test data was used to train six different machine learning algorithms. The algorithm with the highest performance was used for predicting heart disease. The used PC for performing all the computations is Intel(R) Core(™) i5-7200U @ 2.50GHz.

4.1. Experimental Results. In this study, the Scikit-learn package of Python [20] is used for feature selection and classification tasks. First, different algorithms, logistic regression, decision tree, random forests, support vector machine, Gaussian NB, and K-nearest neighbor (denoted as C1, C2, C3, C4, C5, and C6, respectively), were applied to the processed dataset using all the feature and have checked the performance. In the second, Matplotlib and seaborn library of Python are used to visualize correlation matrix heat map and other correlations between different features. Third, different feature selection methods of univariate selection algorithm ANOVA F value, chi-square, and mutual information (MI) that are given in Table 2 (denoted as FST1, FST2, and FST3, respectively) were applied. Fourth, different algorithm performances were evaluated for the selected features. Accuracy, sensitivity, specificity, AUROC, and log loss were used to prove the results of those analyses. All features were standardized using StandardScaler before applying them to the algorithms.

4.2. Result of Different Feature Selection Techniques. ANOVA F value method calculates the F value between features based on the weights of the features. The score of ANOVA F value is given in Table 3. In this score, the three most important features are EIA, CPT, and OP, and the less important features are RES, CM, and FBS, respectively. Another method is chi-square, which calculates the chi-square score between every feature and the target. The scores of chi-square are given in Table 4. In this method, the three most important features are MHR, OP, and NMV, and the less important features are TS, REC, and FBS, respectively. The rank of features in the FST1 and FST2 methods are shown in Figure 2. The third method used in FST3 is mutual information (MI), which calculates the mutual information between each feature, which measures dependency between the features. If the score is zero, then, two features are independent, and the more score will increase, the more the features will be dependent. The scores of mutual information are given in Table 5. Here, the three most dependent features are CPT, TS, and NMV, and the independent features are fbs and restecg. The rank of the feature in FST3 method is shown in Figure 3. Those three tables present significant features for the prediction of heart disease. Besides, FBS, REC, RBP, and CM have an overall lower score for all three FSTs, and in this study, those features are not used in the different algorithms. From all those features, three different sets of features are selected based on their score. Each of the three sets of features was denoted by SF1, SF2, and SF3, respectively. Those selected feature sets are shown in Table 6.

4.3. Visualizing Correlation between Features. Firstly, a clustered heat map is visualized that is shown in Figure 4. This heat map shows the correlation amongst the different features of the dataset. The correlation values show that almost all of this dataset's features are significantly less correlated with each other. This implies that only a few features can be eliminated. In this heat map, CPT, MHR, and PES

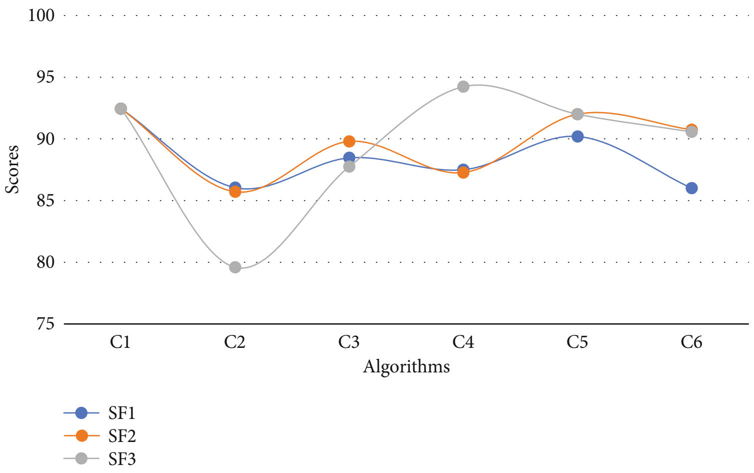

Figure 9: Specificity of different algorithms.

Table 10: AUROC of different algorithms.

	Selected features	C1	C2	C3	C4	C5	C6
	SF1	94.56	76.54	94.09	93.77	95.05	91.89
Dataset	SF2	96.03	76.27	93.43	94.41	95.54	91.81
	SF3	96.08	79.48	93.87	94.95	95.49	93.8

show the highest positive correlation between the target, and EIA, OP, and NMV show the highest negative correlation between the target attribute. However, FBS, CM, RBP, and REC show the lowest correlation score between the target. This is similar to the other feature selection technique feature score, and these features are eliminated in different SF.

Secondly, a relation is shown between age and the target attribute that is shown in Figure 5. It shows that around nine patients aged 41, 51, and 52 and 11 patients, aged 54 suffered from heart disease. It suggests that between the ages of 41 to 54 and mostly the mid-aged people suffered from heart disease.

Thirdly, a relation between MHR and target is shown in Figure 6. It shows that older people have a lower heart rate than young aged. And higher heart rate slightly increases the possibility of heart disease.

4.4. Experimental Analysis of Accuracy. The processed dataset was analyzed using different algorithms, and Table 7 shows the accuracy of each algorithm. Relevant to the accuracy of each algorithm, the highest accuracy (94.51%) was calculated by C4 for SF3; C4 also gave (90.11% and 89.01%) accuracy for SF1 and SF2. The second highest accuracy (93.41%) was calculated by C1 for all three SFs. On the other hand, the poor accuracy (75.82%) was calculated by C2 for SF3. C4 also gave low accuracy (78.02% and 76.92%) for SF1 and SF2. The other algorithm's accuracy was between 84.61 and 92.31%. In addition, the result shows that the best algorithm for the dataset is C4 for SF3. All the accuracies of different algorithms for different SFs are shown in Figure 7.

4.5. Experimental Analysis of Sensitivity. In this analysis, the sensitivity was analyzed for all those algorithms. The score of the sensitivity for all those algorithms was shown in Table 8. The poorest sensitivity (69.38) was given by C2 for SF2. C2 also gave (70.83 and 71.42) scores for SF1 and SF2. And the highest sensitivity was 94.87 given by C4 for SF3 also; the second-highest sensitivity was 94.74 given by C1 for all the SFs. The other algorithm's sensitivity was between 80.49 and 94.6. In addition, the result shows that C4 gave the best score for SF3. All the sensitivity scores of different algorithms for different SFs are shown in Figure 8.

4.6. Experimental Analysis of Specificity. The specificity was explored for all of those algorithms, and the scores of specificity for different algorithms are shown in Table 9. During analysis, C2 gave the most inferior score (79.69) for SF3, and C4 gave the highest score (94.23) for SF3. C4 also gave sensitivity scores (87.50 and 87.27) for SF1 and SF2. C1 gave the second highest score (92.45) for all those SFs. The other algorithms gave scores between 87.27 and 92.0. In addition, the result shows that C4 gave the best score for SF3. All the specificity scores of different algorithms for different SFs are shown in Figure 9.

4.7. Experimental Analysis of AUROC. AUROC were analyzed to evaluate the predictions made for the heart disease dataset. The scores of AUROC for different algorithms were shown in Table 10. In this analysis, the poorest AUROC score (76.27) was given by C2 for SF2. C2 also gave scores (76.54) and (79.48) for SF1 and SF3. C1 gave the highest score (96.08) for SF3. C1 also gave AUROC scores (94.56 and 96.03 for SF1 and SF2. C5 gave the second highest score (95.54) for SF2. The other algorithms gave AUROC scores between 91.81 and 95.49. In addition, the result shows that C1 gave the best score for SF3. All the AUROC scores of different algorithms for different SFs are shown in Figures 10–12.

4.8. Experimental Analysis of Log Loss. In this analysis, log loss was explored. The results given by different algorithms are shown in Table 11. In this experiment, C2 gave the highest

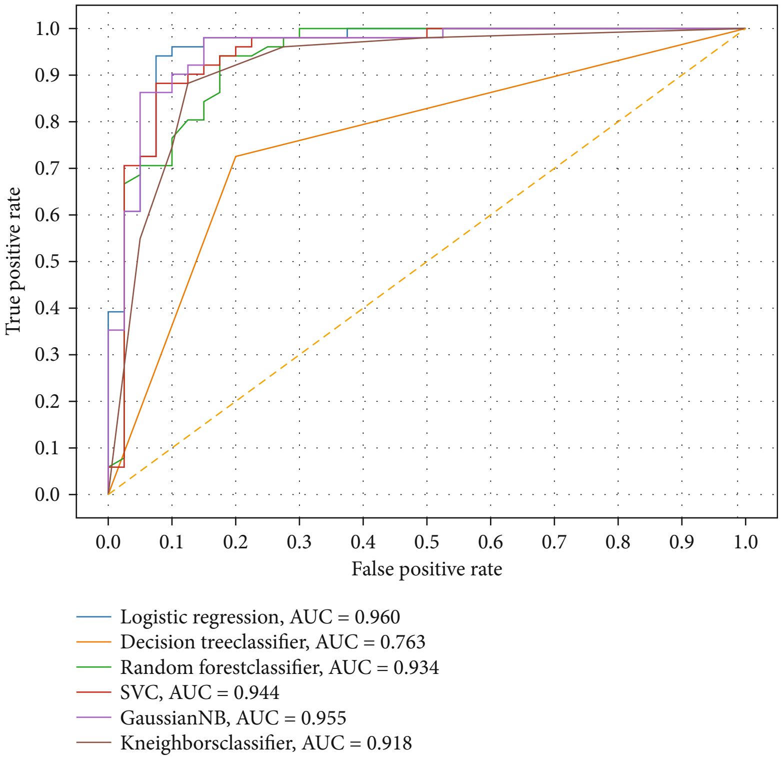

Figure 10: AUROC for SF1.

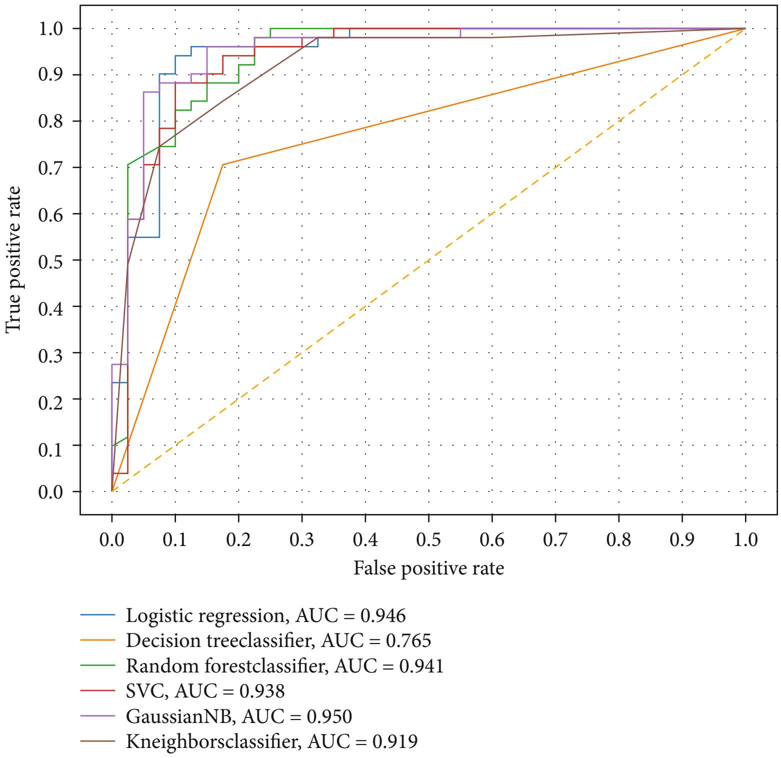

Figure 11: AUROC for SF2.

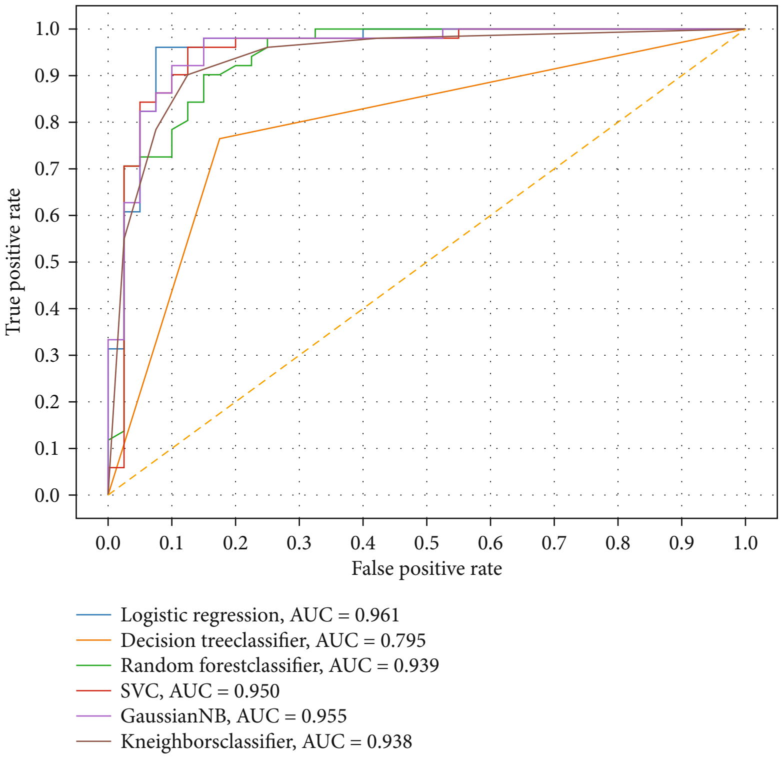

Figure 12: AUROC for SF3.

Table 11: Log loss of different algorithms.

	Selected features	C1	C2	C3	C4	C5	C6
Dataset	SF1	0.29	7.59	0.35	0.33	0.31	1.02
	SF2	0.27	7.97	0.36	0.32	0.29	0.67
	SF3	0.27	8.35	0.34	0.31	0.29	0.62

score (8.35) for SF3. C2 also gave scores (7.59 and 7.97) for SF1 and SF2. Therefore, the lowest log loss value (0.27) was given by C1 for SF2 and SF3 both. The other algorithms gave log loss scores between 0.29 and 1.02. All the log loss scores of different algorithms for different SFs are shown in Figure 13.

5. Discussion

In this research, various machine learning algorithms were used for the early detection of heart disease, and the UCI Cleveland dataset was used for training and testing purposes. Specifically, six well-known algorithms such as LR, DT, RF, SVM, Gaussian NB, and KNN were used with different selected features. And univariate selection algorithms, ANOVA *F* value, chi-square, and mutual information (MI) are used to classify significant features which are more important for predicting heart disease. To check the performance of the different algorithms, different evaluation metrics which are accuracy, sensitivity, specificity, AUROC, and log loss were used. The experimental result shows that the algorithm C4 achieves the highest accuracy (94.51%) for SF3, and C1 achieved the second highest accuracy (93.41%) for all three SFs shown in Table 7. In terms of sensitivity and specificity, C4 also achieved the highest sensitivity (94.87) and specificity score (94.23) for SF3 shown in Tables 8 and 9. Then, for AUROC, C1 gave the highest AUROC score (96.08) for SF3 as shown in Table 10. Then, for log loss, C1 gives the lowest log loss value (0.27) for SF2 and SF3 both, as shown in Table 11. Because of the highest performance of C4 with SF3, it is the best predictive model in terms of accuracy, sensitivity, and specificity. And for AUROC and log loss, C1 is the better predictive model for SF2 and SF3, which is the second-best predictive model overall. In this analysis, we find that SVM has given the best performance for accuracy, sensitivity, and specificity, and LR is given the best performance for AUROC and log loss. Consequently, it is authorized to judge that the support vector machine is an efficient algorithm for heart disease prediction. If compressing between several machine learning algorithms, it was performing above 90 percent accuracy most of the time.

5.1. Comparisons with Other Work. Comparing our analysis with previous studies we found, Mohan et al. [21] developed a heart disease prediction model by using the HRFLM method. Their model predicted (88.47%) accuracy, (92.8%) sensitivity, and (82.6%) specificity for the UCI heart disease dataset, and they used all thirteen features. Amin et al. [22] predicted heart disease 87.41% accurately using Naive Bayes and logistic regression algorithm. A previous study [23] has 56.76% accuracy using J48 with reduced error pruning algorithm. There are more previous studies shown in Table 12, where their overall accuracy is between 87.41 and 83.70%. Besides, no study has evaluated the heart disease prediction

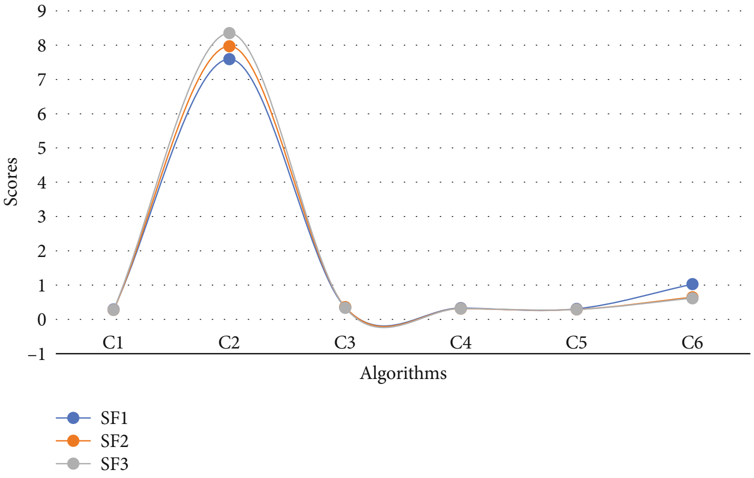

Figure 13: Log loss of different algorithms.

Table 12: Compare our predictive results with the previous results.

Authors	Methods	Acc.(%)	Sens. (%)	Spec. (%)	AUROC (%)	Log loss
Our study	SVM and LR	94.51	94.87	94.23	96.08	0.27
Mohan et al. [21]	HRFLM	88.47	92.8	82.6	-	-
Amin et al. [22]	Naïve Bayes and Logistic Regression	87.41	-	-	-	-
Latha & Jeeva [24]	NB, BN, RF, and MP	85.48	-	-	-	-
Patel et al. [23]	J48 with ReducedErrorpruning Algorithm	56.76	-	-	-	-
Tomar & Agarwal [25]	Feature selection-based LSTSVM	85.59	0.8571	0.8913	-	-
Buscema et al. [26]	TWIST algorithm	84.14	-	-	-	-
Subbulakshmi et al. [27]	ELM	87.5	-	-	-	-
Srinivas et al. [28]	Na¨ıve Bayes	83.70	-	-	-	-
Polat & Gunes [29]	Combining of RBF kernel F-score feature selection and LS-SVM classifier	83.70	83.92	83.54	0.831	-
Kahramanli & Allahverdi [30]	Hybrid neural network method	86.8	-	-	-	-

in detail; while in our study, a range of metrics (accuracy, sensitivity, specificity, AUROC, and log loss) is evaluated, and different feature selection algorithms are used for selected important features that also improve the performance of algorithms.

6. Conclusion

In summary, we implemented different feature selection techniques and found the most significant features which are highly valuable for heart disease prediction, then applied six different machine learning algorithms for those selected features. Every algorithm performed a separate score using different selected features. SVM and LR performance were more significant among all other algorithms. However, the amount of heart disease data available was not large enough for a better predictive model. This experiment will be more accurate if the same analysis is performed in a large real-world patient's data. In future, more experiments will be performed to find more efficient algorithms like deep learning algorithms, for this prediction to achieve better performance of the algorithms using more effective feature selection techniques.

Data Availability

The data are available by contacting the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

K.A. and M.A.M. provided the idea and designed the experiments; N.B., M.M.A., M.A.R., M.R.M., M.I., and K.A. analyzed the data and wrote the manuscript. N.B., M.M.A., M.A.R. M.I., F.M.B., S.A., F.A.A., and M.R.M. helped perform the experimental analysis with constructive discussions. F.M.B. and F.A.A. supported the funding. All authors discussed the results and contributed to the manuscript.

Acknowledgments

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number: IFP22UQU4170008DSR06 and also in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] M. Heron, "Deaths: leading causes for 2008," *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, vol. 60, no. 6, pp. 1–94, 2012.

[2] M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," *International Journal of Control Theory and Applications*, vol. 9, no. 27, pp. 255–260, 2016.

[3] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., "Forecasting the future of cardiovascular disease in the United States," *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[4] G. Savarese and L. H. Lund, "Global public health burden of heart failure," *Cardiac Failure Review*, vol. 3, no. 1, pp. 7–11, 2017.

[5] March 2022, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[6] C. Beyene and P. Kamat, "Survey on prediction and analysis the occurrence of heart disease using data mining techniques," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 165–174, 2018.

[7] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques : a survey," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.

[8] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, pp. 204–207, Heraklion, Greece, 2017.

[9] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108–115, Doha, Qatar, 2008.

[11] M. F. Rabbi, M. P. Uddin, M. A. Ali et al., "Performance evaluation of data mining classification techniques for heart disease prediction," *American Journal of Engineering Research*, vol. 7, no. 2, pp. 278–283, 2018.

[12] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018.

[13] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, *Heart Disease*, UCI Machine Learning Repository, 1988.

[14] S. Razia, J. C. Babu, K. H. Baradwaj, K. S. S. R. Abhinay, and M. Anusha, "Heart disease prediction using machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 10316–10320, 2019.

[15] H. Gadde, "Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning," *International Journal of Engineering Trends and Applications*, vol. 7, no. 4, 2020.

[16] I. D. Mienye and Y. Sun, "Effective feature selection for improved prediction of heart disease," in *Pan-African Artificial Intelligence and Smart Systems Conference*, pp. 94–107, Springer, Cham, 2021.

[17] V. Vakharia, V. K. Gupta, and P. K. Kankar, "A comparison of feature ranking techniques for fault diagnosis of ball bearing," *Soft Computing*, vol. 20, no. 4, pp. 1601–1619, 2016.

[18] N. Carrara and J. Ernst, "On the estimation of mutual information," in *39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 33, Garching, Germany, 2020no. 1.

[19] T. Akter, M. S. Satu, M. I. Khan et al., "Machine learning-based models for early stage detection of autism spectrum disorders," *IEEE Access*, vol. 7, pp. 166509–166527, 2019.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[22] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[23] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," *Heart Disease*, vol. 7, no. 1, pp. 129–137, 2015.

[24] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, article 100203, 2019.

[25] D. Tomar and S. Agarwal, "Feature selection based least square twin support vector machine for diagnosis of heart disease," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 2, pp. 69–82, 2014.

[26] M. Buscema, M. Breda, and W. Lodwick, "Training with Input Selection and Testing (TWIST) Algorithm: A Significant Advance in Pattern Recognition Performance of Machine Learning," *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 1, article 27937, 2013.

[27] C. V. Subbulakshmi, S. N. Deepa, and N. Malathi, "Extreme learning machine for two category data classification," in *2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 458–461, Ramanathapuram, India, 2012.

[28] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering*, vol. 2, no. 2, pp. 250–255, 2010.

[29] K. Polat and S. Güneş, "A new feature selection method on classification of medical datasets: kernel F-score feature selection," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10367–10373, 2009.

[30] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 82–89, 2008.