



## Article

# Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators

Karna Vishnu Vardhana Reddy <sup>1</sup>, Irraivan Elamvazuthi <sup>1,\*</sup>, Azrina Abd Aziz <sup>1</sup>, Sivajothi Paramasivam <sup>2</sup>, Hui Na Chua <sup>3</sup> and S. Pranavanand <sup>4</sup>

<sup>1</sup> Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia; vishnu\_17009417@utp.edu.my (K.V.V.R.); azrina\_aaziz@utp.edu.my (A.A.A.)

<sup>2</sup> School of Engineering, UOWM KDU University College, Shah Alam 40150, Malaysia; siva@kdu.edu.my

<sup>3</sup> Department of Computing and Information Systems, School of Engineering, and Technology, Sunway University, Petaling Jaya 47500, Malaysia; huinac@sunway.edu.my

<sup>4</sup> Department of E.I.E., VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad 500090, India; pranavanand\_s@vnrvijet.in

\* Correspondence: irraivan\_elamvazuthi@utp.edu.my

**Abstract:** Cardiovascular diseases (CVDs) kill about 20.5 million people every year. Early prediction can help people to change their lifestyles and to ensure proper medical treatment if necessary. In this research, ten machine learning (ML) classifiers from different categories, such as Bayes, functions, lazy, meta, rules, and trees, were trained for efficient heart disease risk prediction using the full set of attributes of the Cleveland heart dataset and the optimal attribute sets obtained from three attribute evaluators. The performance of the algorithms was appraised using a 10-fold cross-validation testing option. Finally, we performed tuning of the hyperparameter number of nearest neighbors, namely, ‘k’ in the instance-based (IBk) classifier. The sequential minimal optimization (SMO) achieved an accuracy of 85.148% using the full set of attributes and 86.468% was the highest accuracy value using the optimal attribute set obtained from the chi-squared attribute evaluator. Meanwhile, the meta classifier bagging with logistic regression (LR) provided the highest ROC area of 0.91 using both the full and optimal attribute sets obtained from the ReliefF attribute evaluator. Overall, the SMO classifier stood as the best prediction method compared to other techniques, and IBk achieved an 8.25% accuracy improvement by tuning the hyperparameter ‘k’ to 9 with the chi-squared attribute set.

**Keywords:** heart disease; data pre-processing; attribute evaluation; machine learning classifiers; hyperparameter tuning



**Citation:** Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Appl. Sci.* **2021**, *11*, 8352. <https://doi.org/10.3390/app11188352>

Academic Editor: Giancarlo Mauri

Received: 27 July 2021

Accepted: 1 September 2021

Published: 9 September 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cardiovascular disease (CVD) is the biggest concern in the medical sector at present. It is one of the most lethal and chronic diseases, leading to the highest number of deaths worldwide. From the recent statistics reported by World Health Organization (WHO), about 20.5 million people die every year due to cardiovascular disease, which is approximately 31.5% of all deaths globally. It is also estimated that the number of annual deaths will rise to 24.2 million by 2030. About 85% of cardiovascular disease deaths are due to heart attack and strokes [1]. A heart attack is mainly caused when the blood flow to the heart is blocked due to the build-up of plaque in the arteries. Stroke is caused by a blood clot in an artery within the brain, which cuts off blood circulation to the brain [2]. Heart disease is triggered mostly when the heart is unable to provide enough blood supply to parts of the body [3,4]. It results in early symptoms, such as an irregular heartbeat, shortness of breath, chest discomfort, sudden dizziness, nausea, swollen feet, and a cold sweat. The accurate prediction and proper diagnosis of heart disease in time are indispensable for improving the survival rate of patients. The risk factors that cause CVD include high BP, cholesterol,

alcohol intake, and tobacco consumption, as well as obesity, physical inactivity, and genetic mutations. The early detection of signs and changes in lifestyle, such as physical activity, avoiding smoking, and appropriate medical examination by clinicians, can help to reduce mortality [5].

The techniques that are currently used to predict and diagnose heart disease are primarily based on the analysis of a patient's medical history, symptoms, and physical examination reports by doctors. Most of the time, it is difficult for medical experts to accurately predict a patient's heart disease, where they can predict with up to 67% accuracy [6] because, currently, the diagnosis of any disease is done concerning the similar symptoms observed from previously diagnosed patients [7]. Hence, the medical field requires an automated intelligent system for the accurate prediction of heart disease. This can be achieved by utilizing the huge amount of patient data that is available in the medical sector, along with machine learning algorithms [8]. In recent times, data science research groups have paid much attention to disease prediction. This is owing to the rapid development of advanced computer technologies in the healthcare sector, as well as the availability of massive health databases [9]. The combination of new deep-learning and intelligent decision-making systems has great potential to improve healthcare assistance in our society [10]. Data is the most valuable resource for obtaining new or additional knowledge and collecting important information. There is an enormous amount of data (big data) in various sectors, such as science, technology, agriculture, business, education, and health. This is completely unprocessed data, either in a structured or unstructured form [11]. It is necessary to extract valuable information from big data to store, process, analyze, manage, and visualize this data via performing data analysis [12].

Currently, in the healthcare sector, the information that is related to patients with medical reports is readily available in databases and is growing rapidly day by day. This raw data is highly redundant and unbalanced. It requires pre-processing to extract important features, reduce the execution time of training algorithms, and improve the classification efficiency [13]. The latest advancements in computing capacities and reprogramming capabilities of machine learning improve these processes and open doors for research opportunities in the healthcare sector [14], especially regarding the early prediction of the diseases, such as CVD and cancer, to improve the survival rate. Machine learning is used in a wide range of applications, from identifying risk factors for disease to designing advanced safety systems for automobiles. Machine learning offers predominant prediction modeling tools to address the current limitations [15]. It has good potential for transforming big data for prediction algorithm development. It relies on a computer to learn complex and non-linear interactions between attributes by minimizing the error between the predicted and observed outcomes [16]. The machine learns patterns from the features that are available in the existing dataset and applies them to the unknown dataset to predict the outcome. One of the powerful machine learning techniques for prediction is classification. Classification is a supervised machine learning method that is effective at identifying the disease when trained using appropriate data [17].

The main contribution of this research work was to implement an intuitive medical prediction system for the diagnosis of heart disease using contemporary machine learning techniques. In this work, different kinds of machine learning classifier algorithms, such as naïve Bayes (NB), logistic regression (LR), sequential minimal optimization (SMO), instance-based classifier (IBk), AdaBoostM1 with decision stump (DS), AdaBoostM1 with LR, bagging with REPTree, bagging with LR, JRip, and random forest (RF) were trained to select the best predictive model for the accurate heart disease detection at an initial stage. Three attribute selection techniques, such as correlation-based feature subset evaluator, chi-squared attribute evaluator, and ReliefF attribute evaluator, were utilized to obtain the optimal set of attributes that greatly influenced the performance of the classifiers when predicting the target class. Finally, tuning the hyperparameter “number of nearest neighbors” in the IBk classifier was performed on both the full attribute set and optimal sets obtained from attribute evaluators.

## 2. Related Works

This section discusses the state-of-the-art methods for heart disease diagnosis using machine learning techniques that were accomplished by various effective research works.

R. Perumal et al. [18] developed a heart disease prediction model using the Cleveland dataset of 303 data instances through feature standardization and feature reduction using PCA, where they identified and utilized seven principal components to train the ML classifiers. They concluded that LR and SVM provided almost similar accuracy values (87% and 85%, respectively) compared to that of k-NN with 69%. C. B. C. Latha et al. [19] performed a comparative analysis to improve the predictive accuracy of heart disease risk using ensemble techniques on the Cleveland dataset of 303 observations. They applied the brute force method to obtain all possible attribute set combinations and trained the classifiers. They achieved a maximum increase in the accuracy of a weak classifier of 7.26% based on ensemble algorithm, and produced an accuracy of 85.48% using majority vote with NB, BN, RF, and MLP classifiers using an attribute set of nine attributes. D. Ananey-Obiri et al. [20] developed three classification models, namely, LR, DT, and Gaussian naïve Bayes (GNB), for heart disease prediction based on the Cleveland dataset. Feature reduction was performed using single value decomposition, which reduced the features from 13 to 4. They concluded that both LR and GNB had predictive scores of 82.75% and AUC of 0.87. It was suggested that other models, such as SVM, k-NN, and random forest, be included.

N. K. Kumar et al. [21] trained five machine learning classifiers, namely, LR, SVM, DT, RF, and KNN, using a UCI dataset with 303 records and 10 attributes to predict cardiovascular disease. The RF classifier achieved the highest accuracy of 85.71% with an ROC AUC of 0.8675 compared to the other classifiers. A. Gupta et al. [22] replaced the missing values based on the majority label and derived 28 features using the Pearson correlation coefficient from the Cleveland dataset and trained LR, KNN, SVM, DT, and RF classifiers using the factor analysis of mixed data (FAMD) method; the results based on a weight matrix RF achieved the best accuracy of 93.44%. M. Sultana et al. [23] explored KStar, J48, sequential minimal optimization (SMO), BN, and MLP classifiers using Weka on a standard heart disease dataset from the UCA repository with 270 records and 13 attributes; they achieved the highest accuracy of 84.07% with SMO.

S. Mohan et al. [24] developed an effective hybrid random forest with a linear model (HRFLM) to enhance the accuracy of heart disease prediction using the Cleveland dataset with 297 records and 13 features. They concluded that the RF and LM methods provided the best error rates. S. Kodati et al. [25] developed a heart disease prediction system (HDPS) with the Cleveland dataset of 297 instances and 13 attributes using Orange and Weka data mining tools, where they evaluated the precision and recall metrics for the naïve Bayes, SMO, RF, and KNN classifiers. A. Ed-daoudy et al. [26] researched the Cleveland dataset of 303 records and 14 attributes from UCI. They evaluated the performance of the four main classifiers, namely, SVM, DT, RF, and LR, using Apache Spark with its machine learning library MLlib.

I. Tougui et al. [27] compared the performances of LR, SVM, KNN, ANN, NB, and RF models to classify heart disease with the Cleveland dataset with 297 observations and 13 features using six data mining tools: Orange, Weka, RapidMiner, Knime, MATLAB, and Scikit-Learn. V. Pavithra et al. [28] proposed a new hybrid feature selection technique with the combination of random forest, AdaBoost, and linear correlation (HRFLC) using the UCI dataset of 280 instances to predict heart disease. Eleven (11) features were selected using filter, wrapper, and embedded methods; an improvement of 2% was found for the accuracy of the hybrid model. C. Gazeloglu et al. [29] projected 18 machine learning models and 3 feature selection techniques (correlation-based FS, chi-square, and fuzzy rough set) to find the best prediction combination for heart disease diagnosis using the Cleveland dataset of 303 instances and 13 variables.

N. Louridi et al. [30] proposed a solution to identify the presence/absence of heart disease by replacing missing values with the mean values during pre-processing. They trained three machine learning algorithms, namely, NB, SVM (linear and radial basis func-

tion), and KNN, by splitting the Cleveland dataset of 303 instances and 13 attributes into 50:50, 70:30, 75:25, and 80:20 training and testing ratios. M. Kavitha et al. [31] implemented a novel hybrid model on the Cleveland heart dataset of 303 instances and 14 features with a 70:30 ratio for training and testing by applying DT, RF, and hybrid (DT + RF) algorithms. B. A. Tama et al. [32] designed a stacked architecture to predict heart disease using RF, gradient boosting machine, and extreme gradient boosting with particle swarm optimization (PSO) feature selection using various heart disease datasets, including the Cleveland with 303 instances and 13 attributes.

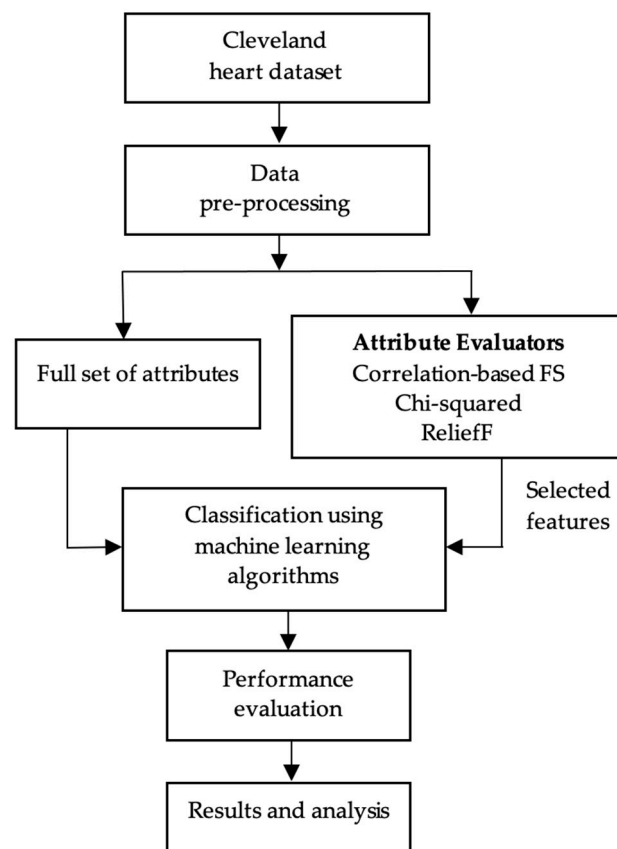
From the experimental works, it is understood that data pre-processing and feature selection can substantially enhance the classification accuracy of machine learning algorithms. During pre-processing, most researchers [18,19,21,22,26,29–32] replaced the missing values, either by using the mean value or the majority mark of that attribute, to make sure the dataset was comprehensive. In some works [20,24,25,27], the missing valued instances were removed. Feature selection is a challenging task due to the large exploration space. It grows exponentially according to the number of features available in the dataset. To solve this issue, an effective comprehensive search technique is required during feature selection. Furthermore, some studies have employed ensemble models, which combine multiple basic learning algorithms to obtain a better prediction accuracy. However, the performance of these techniques can further be improved regarding accurately predicting disease.

### 3. Materials and Methods

This section discusses the proposed methodology, which comprises the dataset description, data pre-processing, machine learning classifiers, attribute evaluators, and performance metrics.

#### 3.1. Proposed Research Methodology

The experimental workflow of the proposed methodology is shown in Figure 1. As a first step, we collected the Cleveland heart disease dataset in .csv format from the UCI machine learning repository. Then, we imported the dataset into the software tool and explored the attributes, types, value ranges, and other statistical information. The next step was pre-processing the data, which included tasks such as looking for the missing values in the dataset and replacing missing values, either with the user constant or mean value depending on the type of attribute, to make sure the machine learning classifiers provide better performance. Thereafter, classification was performed with cross-validation using several machine learning algorithms, such as NB, LR, SMO, IBk, AdaBoostM1 + DS, AdaBoostM1 + LR, bagging + REPTree, bagging + LR, JRip, and RF using the full set of attributes. Cross-validation is a resampling method that is used to assess the efficacy of the machine learning model by partitioning the original dataset into a training set to train the model and a test set to evaluate it. The observations in a dataset can be randomly split into  $k$  equal-sized groups. We then trained the model using  $k-1$  folds and validated the models using the remaining  $k$ th fold. We repeated this step until all  $k$  folds served as a test set and took the average of the recorded values as the performance metric of the model. This work considered  $k = 10$ , i.e., a 10-fold cross-validation. Further, we applied attribute evaluators, such as correlation-based feature selection with the BestFirst search method, chi-squared attribute evaluation with Ranker, and ReliefF attribute evaluation with Ranker using a full training set to obtain the optimal set of attributes for predicting heart disease risk and trained the classifiers again using cross-validation. Finally, we tuned the hyperparameter 'k' in the IBk classifier for enhanced performance and analyzed the results.



**Figure 1.** Experimental workflow of the proposed methodology.

### 3.2. Dataset Description and Statistics

The Cleveland heart dataset consists of 303 instances with 76 attributes, but only 14 attributes are considered more suitable for research experimental purposes. The attribute descriptions for the Cleveland heart dataset are given in Table 1.

**Table 1.** Attribute descriptions for the Cleveland heart dataset from the UCI machine learning repository [33].

Attribute	Description	Type of Attribute	Attribute Value Range
age	Age in years	Numeric	29 to 77
sex	Gender	Nominal	0 = female, 1 = male
cp	Chest pain type	Nominal	1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptomatic
trestbps	Resting blood pressure in mm Hg on admission to the hospital	Numeric	94 to 200
chol	Serum cholesterol in mg/dL	Numeric	126 to 564
fbs	Fasting blood sugar > 120 mg/dL	Nominal	0 = false, 1 = true

Table 1. Cont.

Attribute	Description	Type of Attribute	Attribute Value Range
restecg	Resting electrocardiographic results	Nominal	0 = normal, 1 = ST-T wave abnormality, 2 = definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	Numeric	71 to 202
exang	Exercise induces angina	Nominal	0 = no 1 = yes
oldpeak	ST depression induced by exercise relative to rest	Numeric	0 to 6.2
slope	The slope of the peak exercise ST segment	Nominal	1 = upsloping, 2 = flat, 3 = downsloping
ca	Number of major vessels colored by fluoroscopy	Nominal	0–3
thal	The heart status	Nominal	3 = normal, 6 = fixed defect, 7 = reversible defect
target	Prediction attribute	Nominal	0 = no risk of heart disease, 1 to 4 = risk of heart disease

The attributes with less than 10 classes are considered nominal or categorical types. The attribute 'sex' consists of two classes based on gender: 1 = male and 0 = female. The attribute 'cp' contains four classes of chest pain types: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, and 4 = asymptomatic. The attribute 'fbs' includes two classes regarding whether the fasting blood sugar >120 mg/dL: 1 = true and 0 = false. The attribute 'restecg' comprises three classes of resting electrocardiographic results: 0 = normal, 1 = abnormality in the ST-T wave, 2 = definite hypertrophy in the left ventricular. The attribute 'exang' consists of two classes based on exercise-induced angina: 1 = yes and 0 = no. The attribute 'slope' includes three classes of peak exercise ST segment slope: 1 = upslope, 2 = flat, and 3 = downslope. The attribute 'ca' comprises four classes based on the number of major vessels (0–3) that are colored using fluoroscopy. The attribute 'thal' contains three classes of heart status: 3 = normal, 6 = fixed, and 7 = reversible. The attribute 'target' consists of five classes of prediction: 0 = no risk of heart disease, and 1 to 4 = the risk of heart disease in various stages. Since the main purpose of this research work was to predict whether a patient was at risk of developing heart disease, the values in the range 1 to 4 were converted to 1. Therefore, the 'target' attribute consisted of only two classes: 0 and 1. The attributes 'age,' 'trestbps,' 'chol,' 'thalach,' and 'oldpeak' are considered as numeric/integer type attributes.

The statistical characteristics of the numeric attributes, such as the minimum, maximum, mean, standard deviation, missing, distinct, and unique values, are provided in Table 2(a). There are no missing values found in the numeric attributes of the Cleveland dataset.



**Table 2.** (a) The statistical outline of the numeric attributes. (b) The statistical outline of the nominal attributes.

(a)							
Attribute	Min.	Max.	Mean	StdDev	Missing	Distinct	Unique
age	29	77	54.439	9.039	0	41	4 (1%)
trestbps	94	200	131.69	17.6	0	50	17 (6%)
chol	126	564	246.693	51.777	0	152	61 (20%)
thalach	71	202	149.607	22.875	0	91	28 (9%)
oldpeak	0	6.2	1.04	1.161	0	40	10 (3%)
(b)							
Attribute	Label	Count	Proportion	Missing	Distinct		
sex	0	97	32%	0	2		
	1	206	68%				
cp	1	23	7.6%	0	4		
	2	50	16.5%				
	3	86	28.4%				
	4	144	47.5%				
fbs	0	258	85.15%	0	2		
	1	45	14.85%				
restecg	0	151	49.83%	0	3		
	1	4	1.32%				
	2	148	48.84%				
exang	0	204	67.33%	0	2		
	1	99	32.67%				
slope	1	142	46.86%	0	3		
	2	140	46.20%				
	3	21	6.93%				
ca	0	176	58.08%	4 (1.32%)	4		
	1	65	21.45%				
	2	38	12.54%				
	3	20	6.6%				
thal	3	166	54.79%	2 (0.66%)	3		
	6	18	5.95%				
	7	117	38.6%				
target	0	164	54%	0	2		
	1	139	46%				

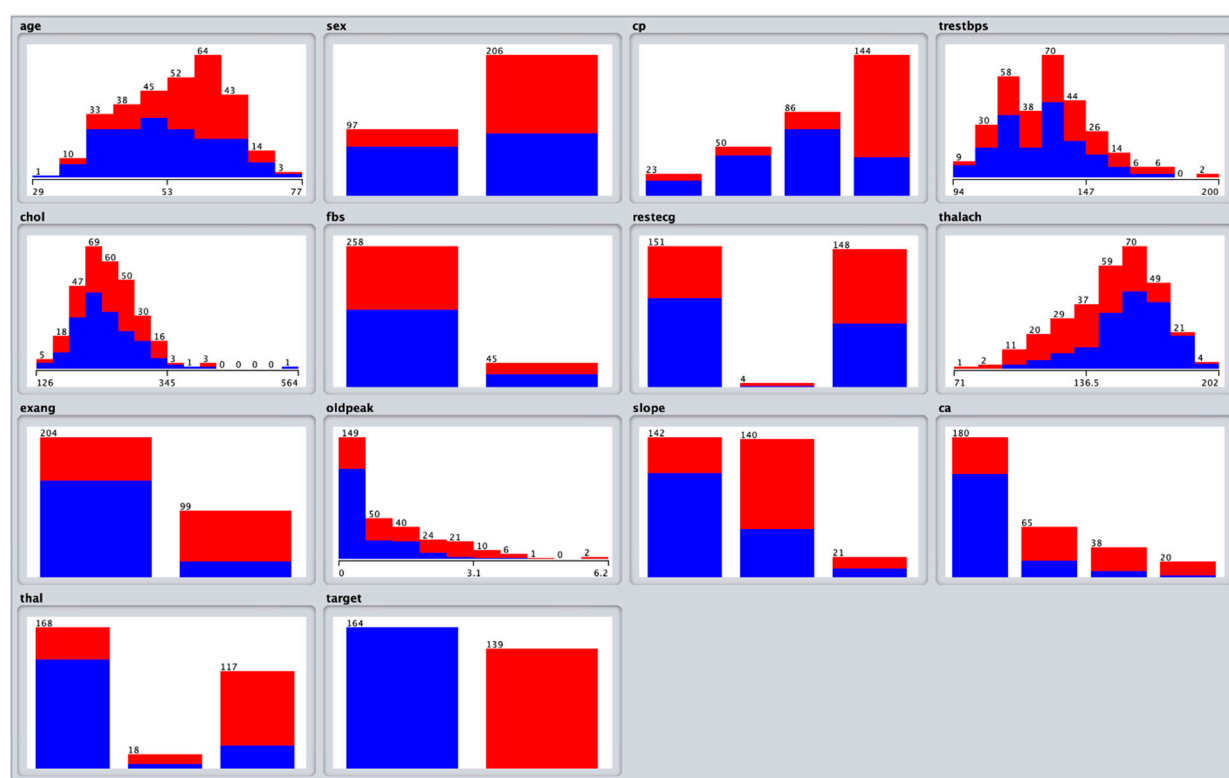
Min.—minimum, Max.—maximum, StdDev—standard deviation.

The statistical characteristics of the nominal attributes, such as label, count, missing, and distinct values, are provided in Table 2(b). There are six (6) instances in total out of 303 that were found to have missing values, which accounted for 2% of the whole dataset: four (4) from the ‘ca’ attribute, and two (2) from the ‘thal’ attribute. The target class labels 0 (no risk of heart disease) consisted of 164 instances and label 1 (risk of heart disease) consisted of 139 instances, which accounted for 54% and 46% of the dataset, respectively.

### 3.3. Pre-Processing of Dataset

Having missing data means that the dataset is incomplete. In statistics, missing values or missing data occur when no data value is stored for the variable in an observation. These missing values are represented by blank/dashes. The main reason for having missing values is that respondents forget/refuse/fail to answer certain questions. Other reasons include sensor failure, loss of data while transferring, internet connection disruption, and wrong mathematical calculations, such as dividing by zero. It is always hard to predict

when missing values are present in the dataset because sometimes, they affect results and sometimes not. In a dataset, each variable may only have a small number of missing responses, but in combination, the missing data could be numerous. The analysis might run but the results may not be statistically significant because of the missing data. For research purposes, replacing missing values either by a user constant or the mean value will be more effective than removing those observations from the dataset. There are some missing values in the Cleveland heart dataset, namely, from the nominal attributes ‘ca’ and ‘thal’, which were replaced with the user constant based on the majority mark. The attribute ‘ca’ has four missing values and has the value 0 as the majority mark in 176 observations out of 299. Meanwhile, the attribute ‘thal’ has two missing values and has the value 3 as the majority mark in 166 observations out of 301. Therefore, to make sure the dataset is complete, the missing values in ‘ca’ and ‘thal’ were replaced by the corresponding majority marks 0 and 3, respectively. A visualization of all 14 attributes of the Cleveland heart dataset is presented in Figure 2.



**Figure 2.** Visualization of attributes of the Cleveland heart dataset.

### 3.4. Machine Learning Models for Classification

Researchers have applied multiple supervised machine learning algorithms on a single dataset to identify the best classifier for disease prediction. This section discusses the various classifiers that were used in this work to predict heart disease risk.

Naïve Bayes (NB) is based on the Bayes theorem, which assumes that the training observations are samples from a set of statistical distributions. Each response class has its distribution. Each distribution in the model provides a probability that a new data point would be found at its location. For the normal distribution, the parameters are the mean and standard deviation [34].



Logistic regression (LR) is an equation where each predictor is multiplied by a coefficient and summed together. This sum becomes the argument for the logistic function to predict the class [35]. For a single observation  $x$  with  $n$  features the response  $y$  is given by

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} \quad (1)$$

Sequential minimal optimization (SMO) is an algorithm that is used to solve very large quadratic programming (QP) optimization problems quickly without any extra storage requirement while training a support vector machine (SVM) [23]. SMO selects two Lagrange multipliers and analytically finds the optimal values for these multipliers to solve the SVM QP problem [36].

The instance-based classifier IBk, also known as k-nearest neighbors, determines the class of observation by comparing it to nearby observations from the training data set. The distance measure that is used to determine the neighbors can be selected from a range of options. The model uses majority voting from the  $K$  nearest data points and assigns a class to the unknown observation [37].

Bagging, also known as the bootstrap aggregation technique, is a simple and powerful ensemble technique that is used to decrease the variance of the decision tree classifier [37]. It provides the learning algorithm with a training set consisting of a random sample of  $m$  training examples that are selected from the initial training set of  $m$  items on each run.

Boosting is an example of an ensemble technique that creates a robust classifier from several weak classifiers. Adaptive boosting (AdaBoostM1) is a successful boosting algorithm that was developed for binary classification and is used to boost the performance of any machine learning algorithm. The decision tree with one level or one decision for classification, called the decision stump, is the most suitable algorithm to work with AdaBoost [38].

JRip is a rule-based classifier that utilizes repeated incremental pruning to produce error reduction (RIPPER). It is a bottom-up approach for learning rules that treats specific judgments of examples in the training data as a class and finds a set of rules that covers all members of the class [39].

Random forest constructs a forest of random trees by creating a set of decision trees from a random sample of the training set to minimize the variance at the expense of a small increase in bias (controlling over-fitting) and results in a final prediction model that should be more accurate and reliable. While growing the trees, the random forest adds more randomness to the algorithms by using random thresholds for each attribute [40].

### 3.5. Attribute Evaluators

Three attribute evaluators correlation-based feature selection with the BestFirst search method, chi-squared attribute evaluation with Ranker, and ReliefF attribute evaluation with Ranker are used in this work.

The correlation-based feature selection technique considers the individual predictive capacity of each attribute, as well as the degree of redundancy between them when determining the value of a subset of attributes. The subsets of attributes with a low inter-correlation but high correlation with the class are preferred [29]. Table 3 shows the attribute set that was obtained from the correlation-based feature selection method.

**Table 3.** Attribute sets that were obtained from the correlation-based feature selection.

Attribute No.	Attribute Name
3	cp
7	restecg
8	thalach
9	exang
10	oldpeak
12	ca
13	thal

The chi-squared attribute evaluation technique is an attribute ranking filter that computes the value of the chi-squared statistic with respect to the class to determine the rank of an attribute using the Ranker search method [40,41]. The rank values of the Cleveland attributes using chi-squared techniques are shown in Table 4. We created an attribute space with the 10 best predictor attributes by removing the three least ranked ones, namely, fbs, trestbps, and chol, from the dataset and trained the machine learning algorithms.

**Table 4.** Attribute sets that were obtained from the chi-squared attribute evaluation.

Attribute No.	Attribute Name	Rank
13	thal	82.6845
3	cp	81.8158
12	ca	72.6169
10	oldpeak	61.5234
9	exang	56.5193
8	thalach	51.5870
11	slope	45.7846
1	age	24.8856
2	sex	23.2181
7	restecg	10.0515
6	fbs	0.1934
4	trestbps	0
5	chol	0

The ReliefF attribute evaluation technique is also an attribute ranking filter that evaluates the value of an attribute by sampling an instance many times and comparing the value of the supplied attribute for the closest instances of the same and different classes [37]. It can work with data from both discrete and continuous classes. This method utilizes all the instances while sampling, the number of nearest neighbors  $k = 10$ , and the Ranker search method to provide the rank values [42] of the Cleveland attributes, which are recorded in Table 5. The classifiers were trained with the top nine attributes by discarding the four lowest-ranked attributes, namely, age, trestbps, fbs, and chol, from the dataset.

**Table 5.** Attribute sets that were obtained from the ReliefF attribute evaluation.

Attribute No.	Attribute Name	Rank
12	ca	0.18812
3	cp	0.17789
13	thal	0.11452
2	sex	0.09307
11	slope	0.06898
9	exang	0.06667
7	restecg	0.05842
10	oldpeak	0.02350
8	thalach	0.02118
1	age	0.01786
4	trestbps	0.01577
6	fbs	0.01386
5	chol	0.00181

### 3.6. Performance Metrics

The performance metrics used in this research work, namely, accuracy, mean absolute error (MAE), sensitivity (recall), fallout, precision, F-measure, specificity, and ROC area, are discussed here. The confusion matrix shown in Table 6 depicts various performance metrics for evaluating a classifier. True positives are the responses equal to the positive class that are correctly predicted as positive. True negatives are the responses equal to the negative class that are correctly predicted as negative. False positives are the responses equal to the negative class but are predicted as positive. False negatives are the responses equal to the positive class but are predicted as negative.

**Table 6.** Confusion matrix.

		Predicted Class	
		High Risk (1)	Low Risk (0)
Actual class	High risk (1)	True Positive (TP)	False Negative (FN)
	Low risk (0)	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (2)$$

$$\text{MAE} = \frac{\sum |\text{Predicted value} - \text{Actual value}|}{\text{Number of predictions}} \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (5)$$

$$\text{Fallout} = \frac{FP}{TN + FP} \times 100\% \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

ROC Area: The area under the ROC curve measures the quality of a model's predictions regardless of what classification threshold is chosen. The ROC curve represents the

true positive rate (sensitivity or recall) vs. the false positive rate (fallout) at every  $0 \rightarrow 1$  threshold.

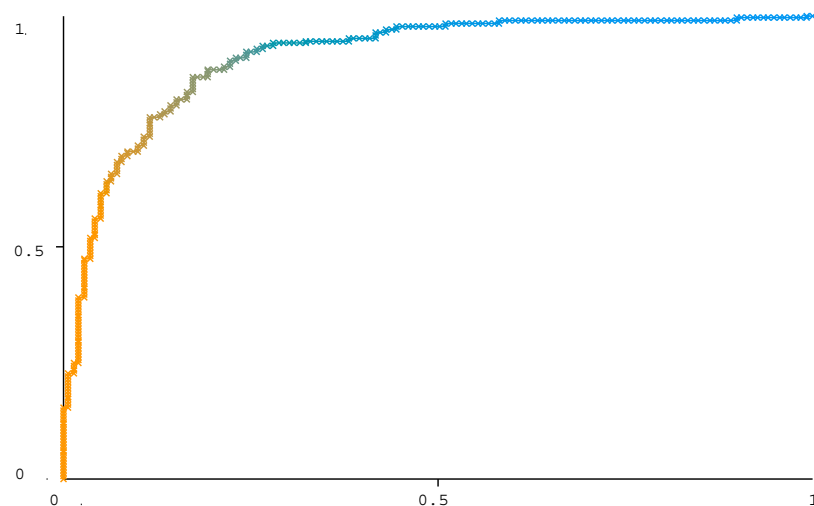
#### 4. Results

The results of the machine learning classifiers using the full set of attributes and optimal set that was obtained from attribute evaluators, tuning the parameter 'k' in the IBk method, and comparison with related works are discussed in the following.

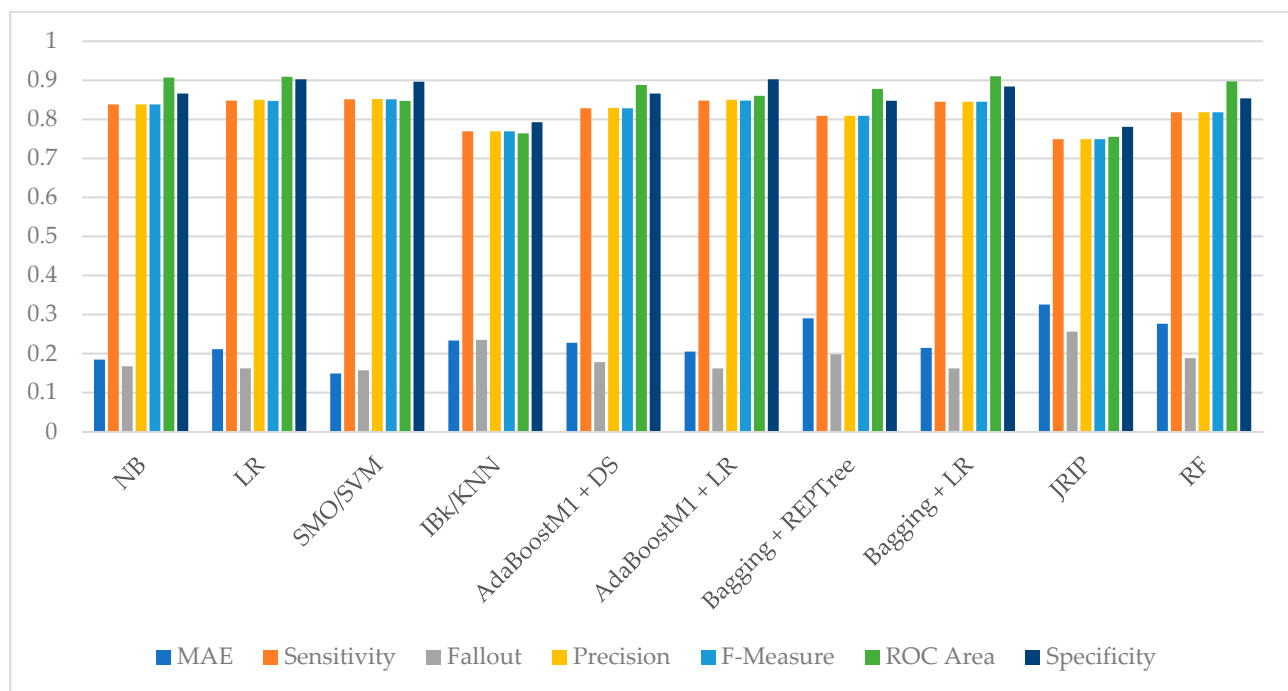
As shown in Table 7, the highest accuracy of 85.148% was attained with the sequential minimal optimization (SMO) algorithm, followed by logistic regression (LR) with an accuracy of 84.818% using the full set of attributes from the Cleveland dataset using the 10-fold cross-validation test option. The SMO algorithm also provided the best MAE of 0.148, sensitivity of 0.851, fallout of 0.157, precision of 0.852, F-measure of 0.851, and specificity of 0.90 compared to other machine learning algorithms. The meta classifier bagging with LR achieved a high ROC area value of 0.91, followed by the single classifier LR with a ROC area of 0.909. The LR and AdaBoostM1 with LR classifiers also reached a specificity of 0.90. NB provided the second-best MAE of 0.184. The visualization of the threshold curve for the target class provided an ROC area of 0.91 using the bagging with LR meta classifier using the full set of attributes, as shown in Figure 3. The bar plot of the performance metrics using the full set of attributes of the Cleveland heart dataset is shown in Figure 4.

**Table 7.** Performance of machine learning classifiers based on the full set of attributes using 10-fold cross-validation.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	83.828	0.184	0.838	0.167	0.838	0.838	0.907	0.870
LR	84.818	0.210	0.848	0.162	0.85	0.847	0.909	0.900
SMO	85.148	0.148	0.851	0.157	0.852	0.851	0.847	0.900
IBk/KNN	76.897	0.233	0.769	0.235	0.769	0.769	0.764	0.790
AdaBoostM1 + DS	82.838	0.227	0.828	0.178	0.829	0.828	0.888	0.870
AdaBoostM1 + LR	84.818	0.204	0.848	0.162	0.850	0.848	0.860	0.900
Bagging + REPTree	80.858	0.290	0.809	0.198	0.809	0.809	0.878	0.850
Bagging + LR	84.488	0.214	0.845	0.162	0.845	0.845	0.910	0.880
JRip	74.917	0.325	0.749	0.256	0.749	0.749	0.755	0.780
RF	81.848	0.276	0.818	0.188	0.818	0.818	0.897	0.850



**Figure 3.** ROC curve of bagging with LR meta classifier using the full set of features.

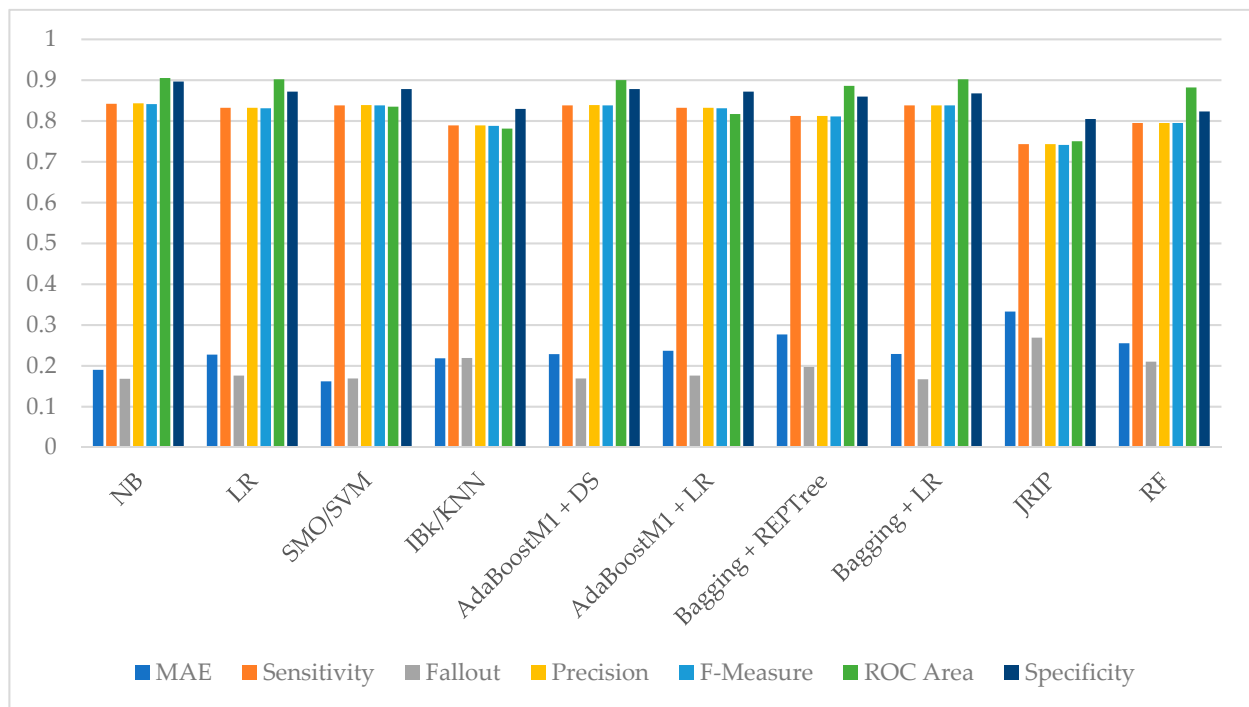


**Figure 4.** Performance metrics using the full set of attributes from the Cleveland heart dataset.

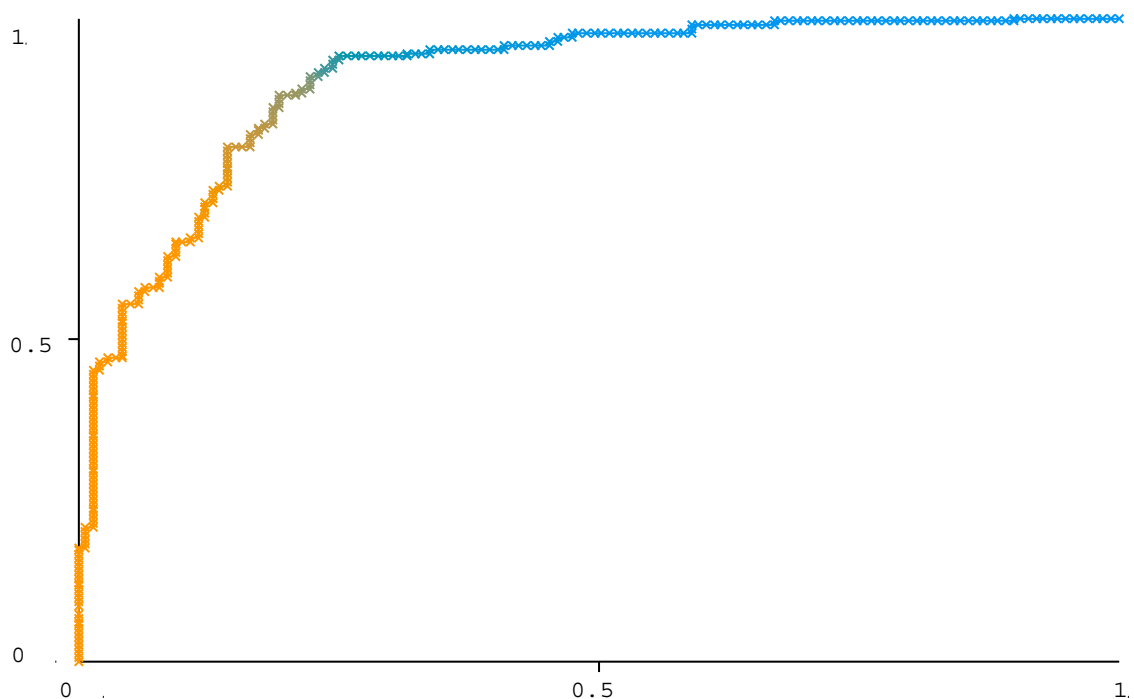
Table 8 shows that the highest accuracy of 84.158% was achieved with the naïve Bayes (NB) algorithm, followed by the SMO, AdaBoostM1 with decision stump (DS), and bagging + LR algorithms, with an accuracy of 83.828% when using the optimal attribute set that was obtained from the correlation-based feature selection method. The SMO algorithm provided the best MAE of 0.161, the NB classifier produced a high sensitivity of 0.842, precision of 0.843, F-measure of 0.841, ROC area of 0.905, and specificity of 0.90 compared to other algorithms. The meta classifier bagging with LR achieved the best fallout value of 0.167. The performance metrics of the ML classifiers with the optimal set obtained using correlation-based feature selection are graphically presented in Figure 5. The ROC curve of the naïve Bayes classifier using the correlation-based feature selection set provided with an area of 0.905 is shown in Figure 6.

**Table 8.** Performance of the machine learning classifiers using the optimal attribute set found based on the correlation-based feature selection technique.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	84.158	0.190	0.842	0.168	0.843	0.841	0.905	0.900
LR	83.168	0.227	0.832	0.176	0.832	0.831	0.902	0.870
SMO	83.828	0.161	0.838	0.169	0.839	0.838	0.835	0.880
IBk/KNN	78.877	0.218	0.789	0.219	0.789	0.788	0.781	0.830
AdaBoostM1 + DS	83.828	0.228	0.838	0.169	0.839	0.838	0.900	0.880
AdaBoostM1 + LR	83.168	0.236	0.832	0.176	0.832	0.831	0.817	0.870
Bagging + REPTree	81.188	0.276	0.812	0.197	0.812	0.811	0.886	0.860
Bagging + LR	83.828	0.229	0.838	0.167	0.838	0.838	0.902	0.870
JRip	74.257	0.332	0.743	0.269	0.743	0.741	0.750	0.800
RF	79.538	0.255	0.795	0.210	0.795	0.795	0.882	0.820



**Figure 5.** Performance metrics using the optimal set obtained from the correlation-based feature selection technique.



**Figure 6.** ROC curve of an NB classifier with a correlation-based feature selection set.

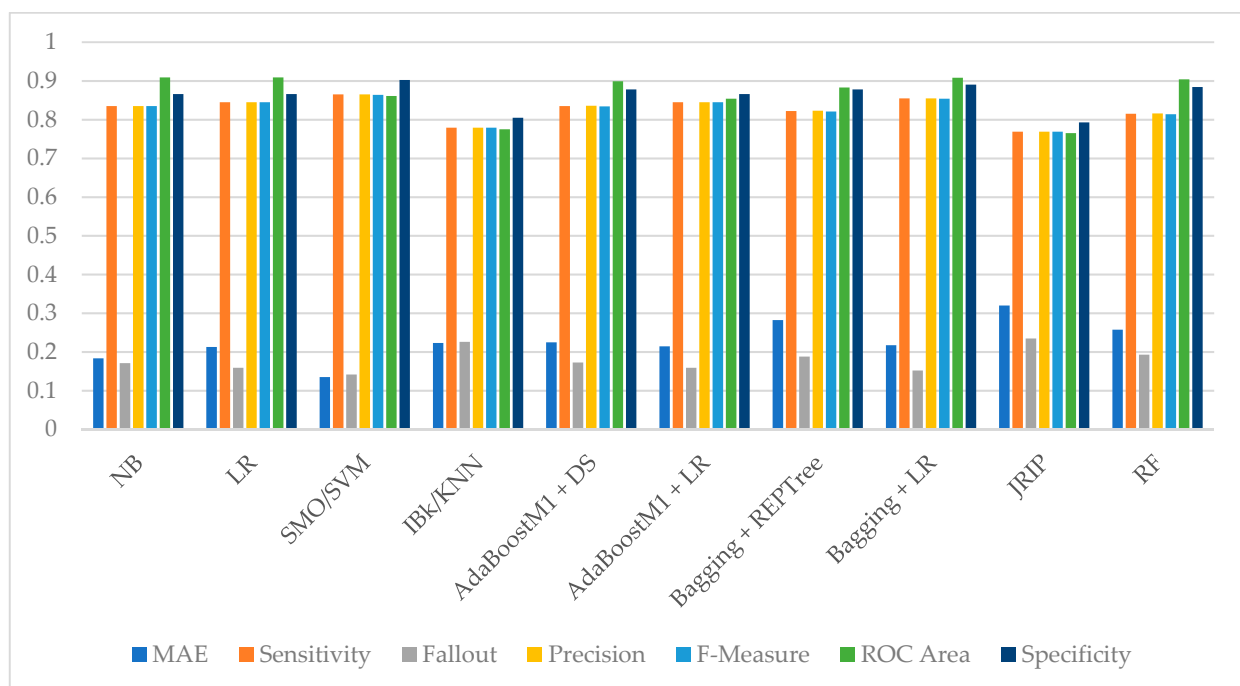
As shown in Table 9, the maximum accuracy of 86.468% was attained with the sequential minimal optimization (SMO) algorithm, followed by bagging with an LR classifier, with a maximum accuracy of 85.478% using the optimal attribute set obtained from chi-squared attribute evaluation technique. The SMO algorithm also offered the best MAE of 0.135, sensitivity of 0.865, fallout of 0.142, precision of 0.865, F-measure of 0.864, and specificity of 0.90 relative to other classifiers. Both the naïve Bayes and logistic regression classifiers



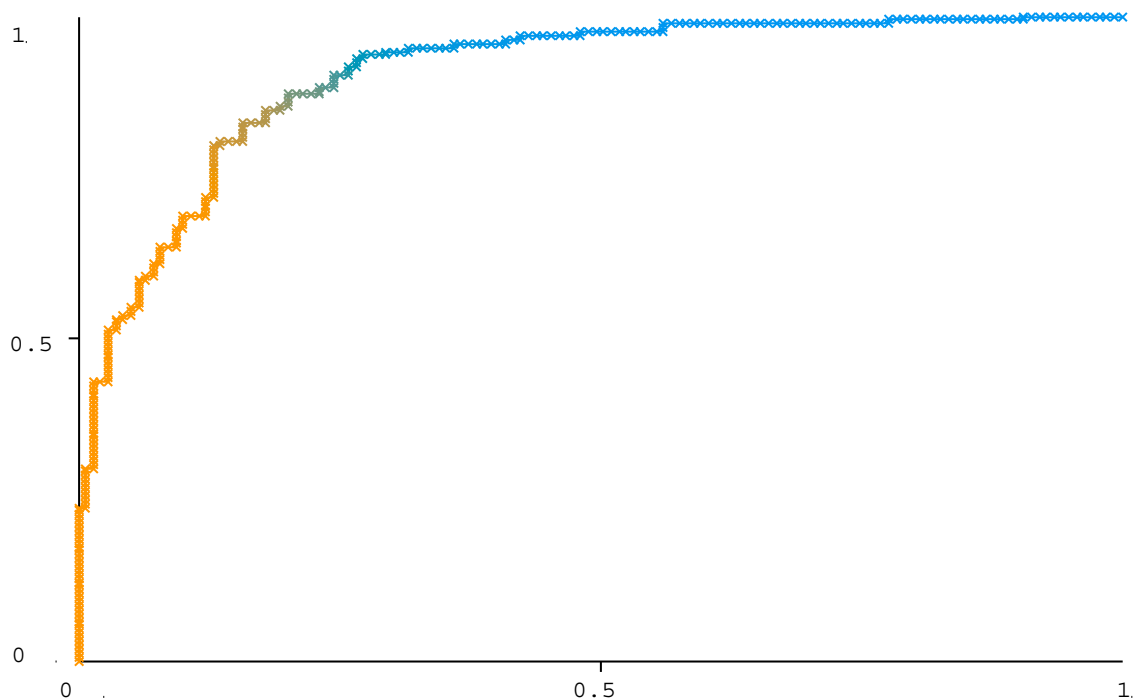
achieved a high ROC area value of 0.909. The graphical representation of performance metrics using the chi-squared attribute evaluation method is shown in Figure 7. The ROC curve with an area of 0.909, which was found using the naïve Bayes and logistic regression models using the chi-squared attribute evaluation set, are shown in Figure 8a,b respectively.

**Table 9.** Performance of the machine learning classifiers based on the optimum attribute set found using the chi-squared attribute evaluation technique.

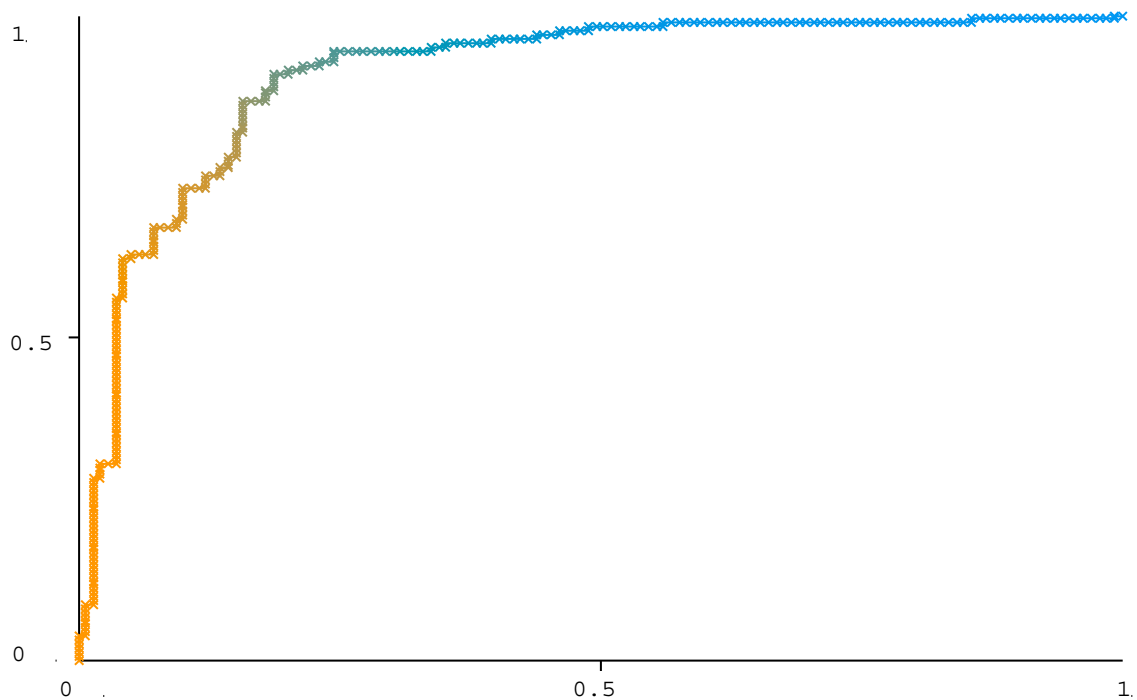
Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	83.498	0.183	0.835	0.171	0.835	0.835	0.909	0.870
LR	84.488	0.212	0.845	0.159	0.845	0.845	0.909	0.870
SMO	86.468	0.135	0.865	0.142	0.865	0.864	0.861	0.900
IBk/KNN	77.887	0.223	0.779	0.226	0.779	0.779	0.775	0.800
AdaBoostM1 + DS	83.498	0.224	0.835	0.173	0.836	0.834	0.899	0.880
AdaBoostM1 + LR	84.488	0.214	0.845	0.159	0.845	0.845	0.854	0.870
Bagging + REPTree	82.178	0.282	0.822	0.188	0.823	0.821	0.883	0.880
Bagging + LR	85.478	0.217	0.855	0.152	0.855	0.854	0.908	0.890
JRip	76.897	0.319	0.769	0.235	0.769	0.769	0.765	0.790
RF	83.168	0.257	0.815	0.193	0.816	0.814	0.904	0.880



**Figure 7.** Performance metrics based on the optimal set obtained using the chi-squared attribute evaluation technique.



(a)



(b)

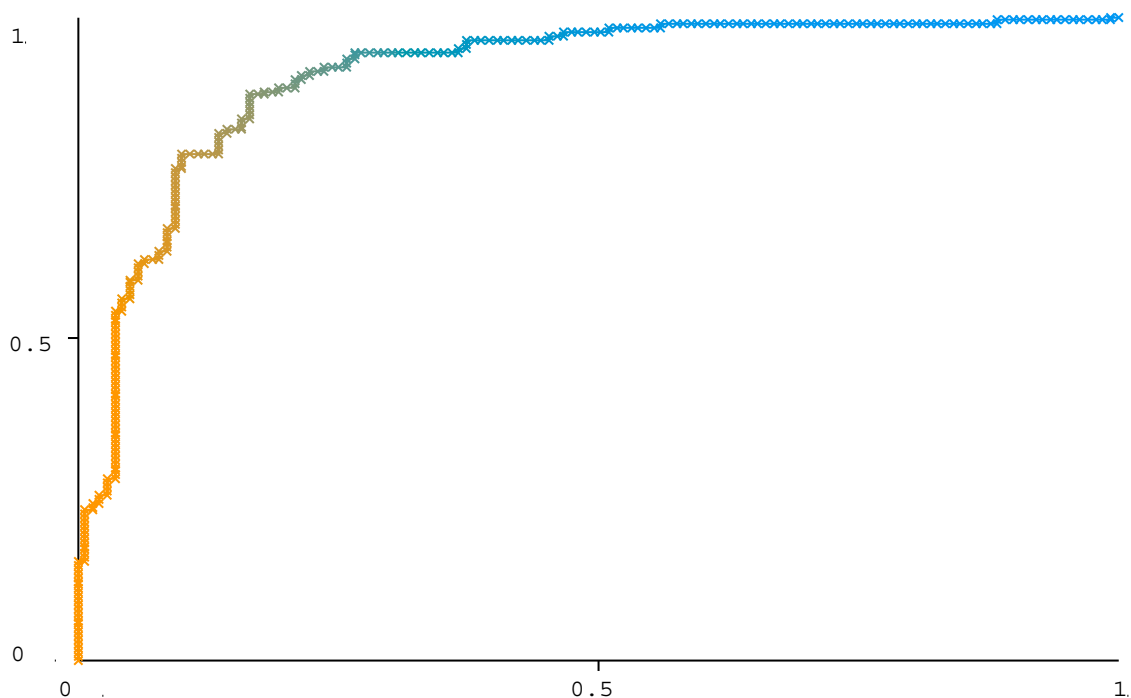
**Figure 8.** ROC curves of the optimal set obtained using chi-squared attribute evaluation: (a) naïve Bayes classifier and (b) logistic regression classifier.

As shown in Table 10, the highest accuracy of 86.138% was achieved with the sequential minimal optimization (SMO) algorithm, followed by the bagging with LR algorithm, with the highest accuracy of 85.148% based on the optimum attribute set, which was obtained using the ReliefF attribute evaluation method. Furthermore, the SMO algorithm produced the best MAE of 0.138, sensitivity of 0.861, fallout of 0.145, precision of 0.862,

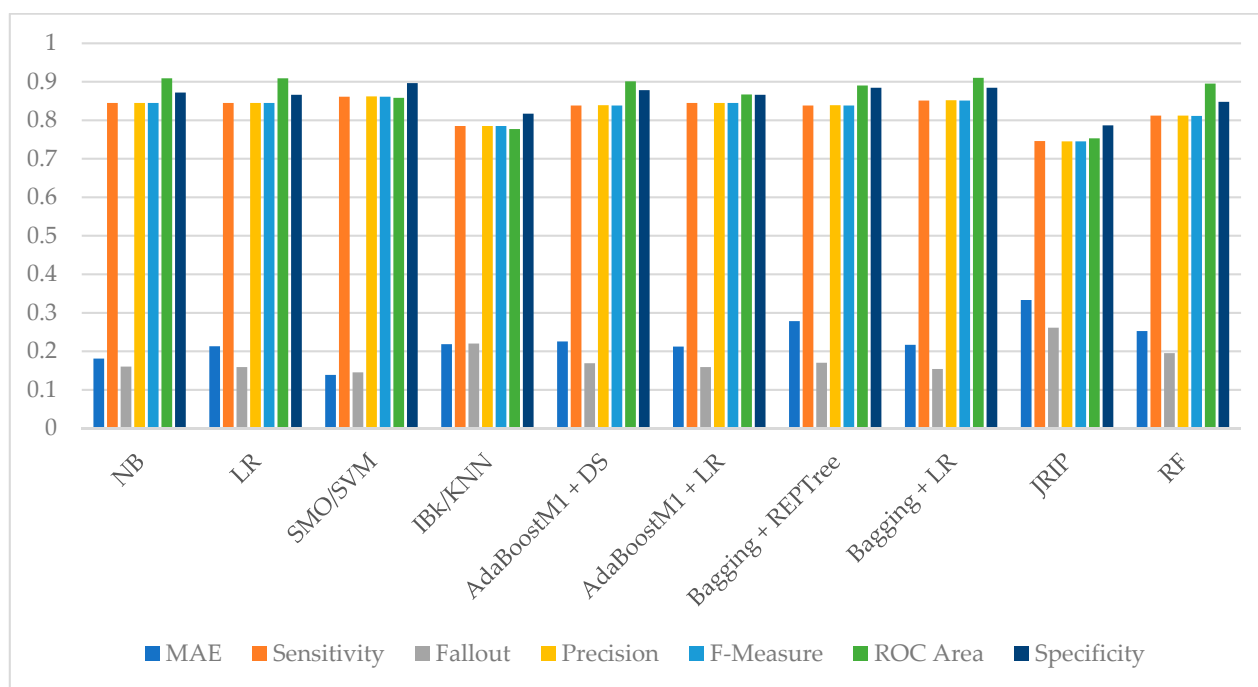
F-measure of 0.861, and specificity of 0.90 compared with the other machine learning classifiers. The meta classifier bagging + LR achieved a high ROC area value of 0.91. The visualization of the threshold curve for the target class produced an ROC area of 0.91 using bagging with LR meta classifier, which is shown in Figure 9. Figure 10 shows a graphical representation of performance metrics using the ReliefF attribute evaluation approach.

**Table 10.** Performance of machine learning classifiers based on the optimal attribute set using the ReliefF attribute evaluation technique.

Classifier	Accuracy	MAE	Sensitivity	Fallout	Precision	F-Measure	ROC Area	Specificity
NB	84.488	0.180	0.845	0.160	0.845	0.845	0.909	0.870
LR	84.488	0.212	0.845	0.159	0.845	0.845	0.909	0.870
SMO	86.138	0.138	0.861	0.145	0.862	0.861	0.858	0.900
IBk/KNN	78.547	0.218	0.785	0.220	0.785	0.785	0.777	0.820
AdaBoostM1 + DS	83.828	0.225	0.838	0.169	0.839	0.838	0.901	0.880
AdaBoostM1 + LR	84.488	0.211	0.845	0.159	0.845	0.845	0.867	0.870
Bagging + REPTree	83.828	0.278	0.838	0.170	0.839	0.838	0.890	0.880
Bagging + LR	85.148	0.216	0.851	0.154	0.852	0.851	0.910	0.880
JRip	74.587	0.333	0.746	0.261	0.745	0.745	0.753	0.790
RF	81.188	0.252	0.812	0.195	0.812	0.811	0.895	0.850



**Figure 9.** ROC curve of the bagging with LR meta classifier using the ReliefF attribute evaluation set.



**Figure 10.** Performance metrics based on the optimal set obtained using the ReliefF attribute evaluation.

A comparison of the accuracy values using the full set of attributes of the Cleveland dataset and optimal attribute sets obtained using various attribute selection techniques performed in this work is shown in Table 11. The correlation-based feature selection method was not able to provide an accuracy value greater than that of the full attribute space. However, there was an improvement in accuracy of about 2% and 1% from the IBk and AdaBoostM1 + DS classifiers, respectively. Besides providing the highest accuracy of 86.468% from SMO, the chi-squared attribute evaluation technique improved most of the classifiers' performances, except for the NB, LR, and AdaBoostM1 + LR classifiers. An increase in the accuracy of about 2% was attained using the JRip algorithm and 1% using the IBk and Bagging + LR algorithms. The ReliefF attribute evaluation method offered the highest improvement in accuracy of about 3% when using the bagging + REPTree classifier, followed by 1.65% from the IBk classifier and about 1% from the SMO and AdaBoostM1 + DS classifiers.

**Table 11.** Accuracy comparison of the attribute selection techniques.

Classifier	Full Attributes	CfsSubset	Diff.	Chi-Squared	Diff.	ReliefF	Dif.
NB	83.828	84.158	0.330	83.498	−0.330	84.488	0.660
LR	84.818	83.168	−1.650	84.488	−0.330	84.488	−0.330
SMO	85.148	83.828	−1.320	86.468	1.320	86.138	0.990
IBk/KNN	76.897	78.877	1.980	77.887	0.990	78.547	1.650
AdaBoostM1 + DS	82.838	83.828	0.990	83.498	0.660	83.828	0.990
AdaBoostM1 + LR	84.818	83.168	−1.650	84.488	−0.330	84.488	−0.330
Bagging + REPTree	80.858	81.188	0.330	82.178	1.320	83.828	2.970
Bagging + LR	84.488	83.828	−0.660	85.478	0.990	85.148	0.660
JRip	74.917	74.257	−0.660	76.897	1.980	74.587	−0.330
RF	81.848	79.538	−2.310	83.168	1.320	81.188	−0.660

Diff.—difference.

Besides the training of the ML classifier on the full and optimal attribute sets obtained from the attribute evaluators, the hyperparameter ‘number of nearest neighbors  $k$ ’ tuning was performed for various values of  $k = 3, 5, 7, 9, 11, 13, 15, 17, 19$ , and  $21$  in the IBk classifier. The best accuracy, accuracy improvement, and other performance metrics for specific ‘ $k$ ’ values that were attained from the parameter tuning are presented in Table 12. Though the accuracy delivered by the IBk classifier was slightly less than that of the SMO classifier that was obtained from the chi-squared attribute set, i.e., 86.468%, there was a significant improvement in accuracy by tuning the hyperparameter ‘ $k$ ’ value in all the cases. We observed that the greatest accuracy improvement of about 8.25% came from the chi-squared attribute evaluation with  $k = 9$  compared to that of default parameter  $k = 1$ . The performance comparison of this research work with the related works is presented in Table 13.

**Table 12.** Performance comparison of the KNN algorithm by tuning the parameter ‘ $k$ .’

Attribute set	Acc. ( $k = 1$ )	Acc. ( $k$ )	Acc. Impr. (%)	MAE	Sen.	Fallout	Pre.	F-Mea.	ROC Area	Spe.
Full attributes	76.897	-	-	0.184	0.769	0.235	0.769	0.769	0.764	0.790
Full attributes	76.897	84.158 ( $k = 5$ )	7.260	0.228	0.842	0.166	0.842	0.841	0.893	0.880
CfsSubset	78.877	83.498 ( $k = 11$ )	4.620	0.237	0.835	0.178	0.839	0.833	0.889	0.910
Chi-Squared	77.887	86.138 ( $k = 9$ )	8.250	0.224	0.861	0.146	0.862	0.861	0.905	0.900
ReliefF	76.897	84.488 ( $k = 9$ )	7.590	0.224	0.845	0.165	0.847	0.844	0.904	0.900

Acc.—accuracy, Impr.—improvement, Sen.—sensitivity, Pre.—precision, F-Mea.—F-measure, Spe.—specificity.

**Table 13.** Performance comparison of related works.

Research Author	Method	# Attr.	Acc. (%)	Pre.	Sen.	AUC
R. Perumal et al. [18]	LR with PCA	7	87.0	-	0.85	-
C.B.C Latha et al. [19]	Majority vote with NB, BN, RF, and MP	9	85.48	-	-	-
D. Ananey-Obiri et al. [20]	LR and GNB with Single value decomposition	4	82.75	-	-	0.87
N. K. Kumar et al. [21]	Random Forest	10	85.71	-	-	0.8675
A. Gupta et al. [22]	FAMD + RF	28	93.44	-	0.8928	-
M. Sultana et al. [23]	SMO	14	84.0741	-	-	0.8392
S. Kodati et al. [25]	SMO	14	-	0.84	0.8365	-
I. Tougui et al. [27]	ANN	14	85.86	-	0.8394	-
V. Pavithra et al. [28]	HRFLC (RF + AdaBoost + Pearson Coefficient)	11	79.0	0.78	0.79	-
C. Gazeloglu et al. [29]	Correlation-based feature selection with NB	6	84.818	-	-	0.905
C. Gazeloglu et al. [29]	Fuzzy Rough Set and Chi-square FS with Radial bias function (RBF) Network	7	81.188	-	-	0.261
B. A. Tama et al. [32]	Two-tier ensemble PSO	7	85.6	-	-	0.8586
S. M. Saqlain et al. [43]	Forward feature selection with Radial Basis Function SVM	7	81.19	-	72.92	-
<b>Proposed method</b>	<b>Chi-Squared + SMO</b>	<b>11</b>	<b>86.468</b>	<b>0.865</b>	<b>0.865</b>	<b>0.861</b>

Attr.—attributes, Acc.—accuracy, Pre.—precision, Sen.—sensitivity, AUC—area under the ROC curve.

This research work utilized the Cleveland heart dataset to achieve the highest accuracy of 85.148% with the SMO model based on the full set of attributes and an accuracy of 84.158% with the NB model based on an optimal set of seven attributes obtained from correlation-based feature selection. The SMO classifier further achieved the best prediction accuracies of 86.468% and 86.138% from the optimal sets obtained from chi-squared (11 attributes) and ReliefF (10 attributes) techniques, respectively. The best values of other performance metrics, namely, MAE (0.135), sensitivity (0.865), specificity (0.90), fallout (0.142), precision (0.865), and F-measure (0.864), was obtained from SMO with the chi-squared method. The bagging + LR classifier provided an ROC area of 0.91 on both the full attributes and optimal sets obtained from the ReliefF method. Nevertheless, the ensemble classifiers AdaBoost and bagging fell short in their predictions compared to the SMO, while the bagging + REPTree classifier achieved the highest improvement in accuracy of about 3% with the ReliefF method. Tuning of the hyperparameter 'k' in IBk reached an improvement in accuracy of 8.25% with the chi-squared evaluator for k = 9. Overall, the SMO classifier showed better performance on the full attributes and optimal sets obtained from the chi-squared and ReliefF attribute evaluators, whereas the NB classifier showed a better performance with the correlation-based feature selection technique.

## 5. Conclusions

In this study, three attribute evaluator techniques were utilized to select significant attributes from the Cleveland heart dataset to improve the performance of machine learning classifiers when predicting heart disease risk. A remarkable performance was achieved by the SMO classifier using the chi-squared attribute evaluation method. Eventually, we noticed that there was a significant improvement in the prediction performance with appropriate attribute selection and tuning the hyperparameters of the classifiers. Although the performance of the classifiers looks satisfactory, a smaller dataset of 303 instances, 10 machine learning classifiers, and 3 feature selection methods were used in this research. There is a huge scope to explore various machine learning algorithms and feature selection techniques. In the future, we intend to combine multiple datasets to obtain a higher number of observations and conduct more experiments by selecting appropriate attributes to improve the classifier's predictive performance.

**Author Contributions:** Conceptualization, K.V.V.R. and I.E.; methodology, K.V.V.R., I.E., A.A.A. and H.N.C.; software, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); validation, K.V.V.R., I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); formal analysis, K.V.V.R. and I.E.; investigation, K.V.V.R.; resources, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); data curation, K.V.V.R.; writing—original draft preparation, K.V.V.R.; writing—review and editing, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); visualization, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); supervision, I.E., A.A.A., S.P. (Sivajothi Paramasivam), H.N.C. and S.P. (S. Pranavanand); project administration, I.E., A.A.A. and H.N.C.; funding acquisition, I.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Universiti Teknologi PETRONAS, grant number 0153AB-M66 and the APC was funded by 0153AB-M66.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: [<https://archive.ics.uci.edu/ml/datasets/heart+disease>] accessed on 15 August 2021.

**Conflicts of Interest:** The authors declare no conflict of interest.



## References

1. WHO. Available online: [https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab\\_1](https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1) (accessed on 9 February 2021).
2. Healthline. Available online: <https://www.healthline.com/health/stroke-vs-heart-attack#treatment> (accessed on 20 February 2021).
3. Chicco, D.; Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–16. [CrossRef] [PubMed]
4. Karthick, D.; Priyadarshini, B. Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. In Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018, Coimbatore, India, 19–20 January 2018; pp. 1182–1186. [CrossRef]
5. Obasi, T.; Shafiq, M.O. Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. In Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, USA, 9–12 December 2019; pp. 2393–2402. [CrossRef]
6. Sharma, H.; Rizvi, M.A. Prediction of Heart Disease using Machine Learning Algorithms: A Survey. *Int. J. Recent Innov. Trends Comput. Commun.* **2017**, *5*, 99–104.
7. Ramalingam, V.V.; Dandapath, A.; Raja, M.K. Heart disease prediction using machine learning techniques: A survey. *Int. J. Eng. Technol.* **2018**, *7*, 684–687. [CrossRef]
8. Alaa, A.M.; Bolton, T.; Di Angelantonio, E.; Rudd, J.H.F.; Van Der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS ONE* **2019**, *14*, e0213653. [CrossRef]
9. Uddin, S.; Khan, A.; Hossain, E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–16. [CrossRef]
10. Song, Q.; Zheng, Y.-J.; Yang, J. Effects of Food Contamination on Gastrointestinal Morbidity: Comparison of Different Machine-Learning Methods. *Int. J. Environ. Res. Public Heal.* **2019**, *16*, 838. [CrossRef]
11. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities. *IEEE Access* **2017**, *5*, 8869–8879. [CrossRef]
12. Aljanabi, M.; Qutqut, M.; Hijawi, M. Machine Learning Classification Techniques for Heart Disease Prediction: A Review. *Int. J. Eng. Technol.* **2018**, *7*, 5373–5379. [CrossRef]
13. Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. *IEEE Access* **2020**, *8*, 184087–184108. [CrossRef]
14. Swain, D.; Pani, S.K.; Swain, D. A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. In Proceedings of the 2018 International Conference on Advanced Computation and Telecommunication, ICACAT, Bhopal, India, 28–29 December 2018; pp. 1–6. [CrossRef]
15. Weng, S.F.; Reys, J.M.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **2017**, *12*, e0174944. [CrossRef]
16. Khan, Y.; Qamar, U.; Yousaf, N.; Khan, A. Machine Learning Techniques for Heart Disease Datasets: A Survey. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22–24 February 2019; pp. 27–35. [CrossRef]
17. Goel, S.; Deep, A.; Srivastava, S.; Tripathi, A. Comparative Analysis of various Techniques for Heart Disease Prediction. In Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, Mathura, India, 21–22 November 2019; pp. 88–94. [CrossRef]
18. Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 4225–4234.
19. Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked* **2019**, *16*, 100203. [CrossRef]
20. Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. *Int. J. Comput. Appl.* **2020**, *176*, 17–21. [CrossRef]
21. Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21. [CrossRef]
22. Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. *IEEE Access* **2019**, *8*, 14659–14674. [CrossRef]
23. Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka, Bangladesh, 22–24 September 2016; pp. 1–5. [CrossRef]
24. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554. [CrossRef]
25. Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai University Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Glob. J. Comput. Sci. Technol.* **2018**, *18*.

26. Ed-Daoudy, A.; Maalmi, K. Performance evaluation of machine learning based big data processing framework for prediction of heart disease. In Proceedings of the International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco, 26–27 December 2019; pp. 1–5. [\[CrossRef\]](#)
27. Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. *Health Technol.* **2020**, *10*, 1137–1144. [\[CrossRef\]](#)
28. Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. *Mater. Today Proc.* **2021**, *22*, 660–670. [\[CrossRef\]](#)
29. Gazeloğlu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. *Prog. Nutr.* **2020**, *22*, 660–670. [\[CrossRef\]](#)
30. Louridi, N.; Amar, M.; El Ouahidi, B. Identification of Cardiovascular Diseases Using Machine Learning. In Proceedings of the 7th Mediterranean Congress of Telecommunications 2019, CMT 2019, Fez, Morocco, 24–25 October 2019; pp. 1–6. [\[CrossRef\]](#)
31. Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021; pp. 1329–1333. [\[CrossRef\]](#)
32. Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. *BioMed Res. Int.* **2020**, *2020*. [\[CrossRef\]](#)
33. Heart Disease Dataset. Available online: <https://archive.ics.uci.edu/ml/datasets/heart+disease> (accessed on 24 May 2021).
34. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* **2018**, *2018*. [\[CrossRef\]](#)
35. Maini, E.; Venkateswarlu, B.; Maini, B.; Marwaha, D. Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India. *Med. J. Armed Forces India* **2021**, *77*, 302–311. [\[CrossRef\]](#)
36. Zeng, Z.-Q.; Yu, H.-B.; Xu, H.-R.; Xie, Y.-Q.; Gao, J. Fast training Support Vector Machines using parallel sequential minimal optimization. In Proceedings of the 2008 3rd International Conference on Intelligent System and Knowledge Engineering, ISKE 2008, Xiamen, China, 17–19 November 2008; Volume 1, pp. 997–1001. [\[CrossRef\]](#)
37. Ghosh, P.; Azam, S.; Jonkman, M.; Karim, A.; Shamrat, F.M.J.M.; Ignatious, E.; Shultana, S.; Beeravolu, A.R.; De Boer, F. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. *IEEE Access* **2021**, *9*, 19304–19326. [\[CrossRef\]](#)
38. Kang, K.; Michalak, J. Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. [1802.03522] Enhanced Version of AdaBoostM1 with J48 Tree Learning Method. Available online: [arxiv.org](https://arxiv.org/abs/1802.03522) (accessed on 27 June 2021).
39. Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinform.* **2020**, *21*, 1–18. [\[CrossRef\]](#)
40. Muhammad, Y.; Tahir, M.; Hayat, M.; Chong, K.T. Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci. Rep.* **2020**, *10*, 1–17. [\[CrossRef\]](#)
41. Al Janabi, K.B.; Kadhim, R. ('Weka' Feature Selection-bad results) Data Reduction Techniques: A Comparative Study for Attribute Selection Methods. *Int. J. Adv. Comput. Sci. Technol.* **2018**, *8*, 1–13.
42. Spencer, R.; Thabtah, F.; Abdelhamid, N.; Thompson, M. Exploring feature selection and classification methods for predicting heart disease. *Digit. Health* **2020**, *6*, 1–10. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Saqlain, S.M.; Sher, M.; Shah, F.A.; Khan, I.; Ashraf, M.U.; Awais, M.; Ghani, A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowl. Inf. Syst.* **2018**, *58*, 139–167. [\[CrossRef\]](#)