Innovations in Systems and Software Engineering

Performance Analysis of Supervised Classification Models on Heart Disease Prediction

--Manuscript Draft--

Manuscript Number:	ISSE-D-22-00068
Full Title:	Performance Analysis of Supervised Classification Models on Heart Disease Prediction
Article Type:	Original Research
Keywords:	Classifiers; Model Selection; Feature Selection; Exploratory Data Analysis; Evaluation metrics
Corresponding Author:	Ezekiel Adebayo Ogundepo, MSc University of Ilorin Faculty of Science Ilorin, Kwara NIGERIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Ilorin Faculty of Science
Corresponding Author's Secondary Institution:	
First Author:	Ezekiel Adebayo Ogundepo, MSc
First Author Secondary Information:	
Order of Authors:	Ezekiel Adebayo Ogundepo, MSc
	Waheed Babatunde Yahya, Professor
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	This paper presents a predictive analysis of data on heart disease patients to determine the possible risk factors associated with their heart disease status. Two independent (but similar) published heart disease datasets, the Cleveland data (used to build classification models) and the Statlog data (used for results' validation) were considered for analysis. Detail exploratory analysis using the Chi-square test of independence was performed on the Cleveland data after which ten standard classification models were constructed for class prediction. The classification models were constructed by partitioning the Cleveland data randomly into 208 (70%) training samples and 89 (30%) test samples over 200 replications. Preliminary results showed that some of the bio-clinical categorical variables are strongly associated with the heart disease conditions of the patients ($p < 0.001$). The classification results from the test samples indicated that the Support Vector Machine yielded the best predictive performances with 85% Accuracy, 82% Sensitivity, 88% Specificity, 87% Precision, 91% Area under the ROC curve (AUC), and 38% Log Loss value. These results were validated on the Statlog data in 10-fold cross-validation which were all consistent with those obtained from the Cleveland dataset.

Performance Analysis of Supervised Classification Models on Heart Disease Prediction

Ezekiel Adebayo Ogundepo[1*]; Waheed Babatunde Yahya[2]

[1,2] University of Ilorin, Department of Statistics, Ilorin, Nigeria

e-mail: ogundepoezekiel@gmail.com[1]; wbyahya@unilorin.edu.ng[2]

* Corresponding author: ogundepoezekiel@gmail.com

[1] https://orcid.org/0000-0003-3974-2733
[2] https://orcid.org/0000-0001-9193-5458

Abstract

This paper presents a predictive analysis of data on heart disease patients to determine the possible risk factors associated with their heart disease status. Two independent (but similar) published heart disease datasets, the Cleveland data (used to build classification models) and the Statlog data (used for results' validation) were considered for analysis. Detail exploratory analysis using the Chi-square test of independence was performed on the Cleveland data after which ten standard classification models were constructed for class prediction. The classification models were constructed by partitioning the Cleveland data randomly into 208 (70%) training samples and 89 (30%) test samples over 200 replications. Preliminary results showed that some of the bio-clinical categorical variables are strongly associated with the heart disease conditions of the patients ($p < 0.001$). The classification results from the test samples indicated that the Support Vector Machine yielded the best predictive performances with 85% Accuracy, 82% Sensitivity, 88% Specificity, 87% Precision, 91% Area under the ROC curve (AUC), and 38% Log Loss value. These results were validated on the Statlog data in 10-fold cross-validation which were all consistent with those obtained from the Cleveland dataset.

Keywords

Classifiers, Model Selection, Feature Selection, Exploratory Data Analysis, Evaluation metrics.

1.0 Introduction

Heart disease refers to several types of heart conditions, and it is one of the dominant and severe illnesses that affect people worldwide and often results in death [1, 2]. The World Health Organization (WHO) estimated that about 17 million lives are lost yearly due to heart-related diseases, which are relatively more prevalent in Asia, India, and the United States of America [3]. Heart conditions are usually dominant in males than females and primarily affects middle-aged and older people [4]. Medical research has identified that lifestyles, obesity, eating habits, and physical inactivity are significant factors leading to heart-related diseases [5]. Smoking, high blood pressure, ECG rate, cholesterol level, height, and hypertension can increase heart disease chances.

Heart disease prevention is essential and a sound data-driven system for predicting heart disease will enable us to learn how to reliably detect it and improve the entire research and prevention process, making sure that more people can live healthy lives. Many studies have shown that machine learning techniques can effectively predict heart disease given a set of prognostic and bio-clinical variables. Patel et al. (2013) [6] used Naive Bayes, classification by clustering, and Decision Tree with fewer features to predict heart disease risk. Yahya et al. (2014) [21] developed a sequential-based feature selection and classification method to efficiently select the core gene biomarkers to predict the histopathological responses of 43 patients with locally advanced rectal carcinomas. Pouriyeh et al. (2017) [8] investigated and compared the accuracy of seven classifiers for predicting heart disease in the Cleveland data. The experiments' results indicate that the SVM method using the boosting technique outperformed other methods. Latha and Jeeva (2019) [9] used an ensemble technique to predict the risk of heart disease, and the result showed ensemble methods could increase accuracy by 7%.

This work aims to evaluate the performances of ten classifiers such as Logistic regression, Random Forest, Decision Tree, Naive Bayes, k-Nearest Neighbor, Extreme Gradient Tree, Conditional Random Forests, Linear Discriminant Analysis, Artificial Neural Network and Support Vector Machine in predicting heart disease. The results from this work would serve as a helpful guide in a future study in choosing the appropriate machine learning technique(s) to analyze the kind of data presented in this work as there is no free lunch of statistical machine learning models [10].

2.0 Materials and methods

2.1 Materials

This study employed two heart disease datasets open to the public at the UCI (University of California, Irvine C.A) Machine Learning Repository. The first data, hereafter called the Cleveland data, were obtained from a heart disease study in the Cleveland database [11] and are available at *https://bit.ly/cleveland-heart-disease-database.* The second data, hereafter called the Statlog, were obtained from the Statlog heart disease database [12] and is available at *https://bit.ly/statlog-heart-disease-database*. The two datasets have thirteen (13) features, representing the bio-clinical variables obtained on heart disease patients and their cardiovascular statistics. These prognostic variables were used in the various models in this study to predict whether a patient has heart disease or not.

The Cleveland dataset has 303 patients, out of which 139 (45.9%) patients had heart disease, and 164 (54.1%) patients did not have the condition. However, during the preprocessing of the Cleveland dataset, we removed six (6) patients with incomplete information, which reduced the total samples to 297 patients, out of which 137 (46%) patients had heart disease, and 160 (54%) did not have the condition. The Statlog data contained 270 patients, out of which 120 (44.4%) patients had heart disease, and 150 (55.6%) patients did not have the condition.

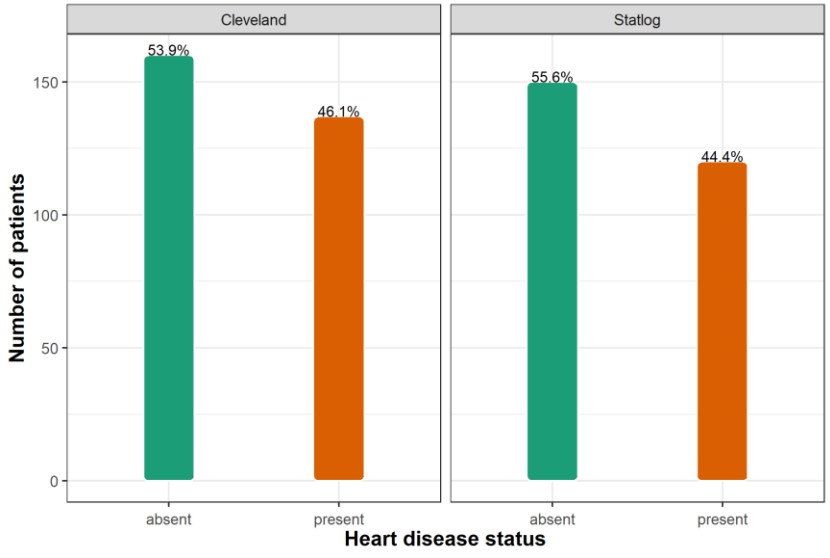

Fig. 1 A bar chart showing the distribution of heart disease status between Cleveland and Statlog heart disease datasets

2.1.1 The Training and Validation Data

This study used the Cleveland dataset as the training data to build all the ten classification models. The Statlog dataset, although entirely independent of the Cleveland data, is similar in structure to the Cleveland data was used as the validation data for efficiency and assessment of the extent of reproducibility of the results.

Table 1 shows the descriptive statistics of the continuous variables in Cleveland heart disease data. The patients' mean age was 54.54 years with a standard deviation of 9.05 years, while the minimum and maximum ages were 29 and 77 years, respectively. Tables 2 and 4 show a detailed exploratory data analysis using the Chi-square test of independence for Cleveland and Statlog heart disease datasets; to get more insight into the features that are useful in determining whether a patient has heart disease before applying standard machine learning models for the class prediction.

Table 1: Summary statistics of the continuous features in the Cleveland heart disease dataset

Determinants	Minimum	Maximum	Mean	Median	Standard deviation
Age (in years)	29	77	54.5421	56	9.0497
Resting blood pressure in mm Hg	94	200	131.6936	130	17.7628
Serum cholestoral in mg/dl	126	564	247.3502	243	51.9976
Maximum heart rate achieved	71	202	149.5993	153	22.9416
ST depression	0	6.2	1.0556	0.8	1.1661

Table 2: Frequency (percentage in parenthesis) distribution of determinants of heart disease across the 297 patients in the Cleveland data

Determinants (factors)	Factor levels	Heart disease status		Total	P-value
		Absent (0)	Present (1)		
		160 (54%)	137 (46%)	297 (100%)	
Sex of the patients	Male (ref)	89 (44.3%)	112 (55.7%)	201 (67.7%)	<0.001
	Female	71 (74.0%)	25 (26.0%)	96 (32.3%)	
Chest pain type	Typical angina (ref)	16 (69.6%)	7 (30.4%)	23 (7.7%)	
	Asymptomatic	39 (27.5%)	103 (72.5%)	142 (47.8%)	<0.001
	Non-anginal pain	65 (78.3%)	18 (21.7%)	83 (27.9%)	
	Atypical angina	40 (81.6%)	9 (18.4%)	49 (16.5%)	
FBS > 120 mg/dl	No (ref)	137 (53.9%)	117 (46.1%)	254 (85.5%)	0.999 *
	Yes	23 (53.5%)	20 (46.5%)	43 (14.5%)	
REC	2 (ref)	67 (45.9%)	79 (54.1%)	146 (49.2%)	
	0	92 (62.6%)	55 (37.4%)	147 (49.5%)	0.008
	1	1 (25.0%)	3 (75.0%)	4 (1.3%)	
Exercise induced angina	No (ref)	137 (68.5%)	63 (31.5%)	200 (67.3%)	<0.001
	Yes	23 (23.7%)	74 (76.3%)	97 (32.7%)	

SST	Downsloping (ref)	9 (42.9%)	12 (57.1%)	21 (7.1%)	
	Flat	48 (35.0%)	89 (65.0%)	137 (46.1%)	<0.001
	Upsloping	103 (74.1%)	36 (25.9%)	139 (46.8%)	
Major vessel	0 (ref)	129 (74.1%)	45 (25.9%)	174 (58.6%)	
	3	3 (15.0%)	17 (85.0%)	20 (6.7%)	<0.001
	2	7 (18.4%)	31 (81.6%)	38 (12.8%)	
	1	21 (32.3%)	44 (67.7%)	65 (21.9%)	
Thallium stress test	Fixed defect (ref)	6 (33.3%)	12 (66.7%)	18 (6.1%)	
	Normal	127 (77.4%)	37 (22.6%)	164 (55.2%)	<0.001
	Reversable defect	27 (23.5%)	88 (76.5%)	115 (38.7%)	

N.B.: The *p*-value is from the Pearson chi-square test of independence. The symbol (∗) indicates that the chi-square test of independence is not significant at the 5% level. (ref) indicates the reference category of a factor as used in fitting our various models. FBS = Fasting blood sugar; REC = Resting electrocardiographic results; SST = Slope of the peak exercise ST segment, Major vessel = Number of major vessels colored by fluoroscopy.

For the Statlog heart disease data, the mean age of the patients in the sample was 54.53 years with a standard deviation of 9.11 years, while the minimum and maximum ages were 29 and 77 years, respectively, are shown in Table 3.

Table 3: Summary statistics of the continuous features in the Statlog heart disease dataset

Determinants	**Minimum**	**Maximum**	**Mean**	**Median**	**Standard deviation**
Age (in years)	29	77	54.433	55	9.109
Resting blood pressure in mm Hg	94	200	131.344	130	17.862
Serum cholesterol in mg/dl	126	564	249.659	245	51.686
Maximum heart rate achieved	71	202	149.678	153.5	23.166
ST depression	0	6.2	1.05	0.8	1.145

Table 4: Frequency (percentage in parenthesis) distribution of determinants of heart disease across the 270 patients in the Statlog data

Determinants (factors)	**Factor levels**	**Heart disease status**		**Total**	**P-value**
		Absent (0)	**Present (1)**		
		150 (55.6%)	**120 (44.4%)**	**270 (100%)**	
Sex of the patients	Male (ref)	83 (45.4%)	100 (54.6%)	183 (67.8%)	<0.001
	Female	67 (77.0%)	20 (23.0%)	87 (32.2%)	
Chest pain type	Typical angina (ref)	15 (75.0%)	5 (25.0%)	20 (7.4%)	
	Asymptomatic	38 (29.5%)	91 (70.5%)	129 (47.8%)	<0.001
	Non-anginal pain	62 (78.5%)	17 (21.5%)	79 (29.3%)	
	Atypical angina	35 (83.3%)	7 (16.7%)	42 (15.6%)	
FBS > 120 mg/dl	No (ref)	127 (55.2%)	103 (44.8%)	230 (85.2%)	0.999 *
	Yes	23 (57.5%)	17 (42.5%)	40 (14.8%)	
REC	2 (ref)	64 (46 .7%)	73 (53.3%)	137 (50.7%)	
	0	85 (64.9%)	46 (35.1%)	131 (48.5%)	0.008
	1	1 (50.0%)	1 (50.0%)	2 (0.7%)	
Exercise induced angina	No (ref)	127 (70.2%)	54 (29.8%)	181 (67%)	<0.001
	Yes	23 (25.8%)	66 (74.2%)	89 (33%)	
SST	Downsloping (ref)	8 (44.4%)	10 (55.6%)	18 (6.7%)	<0.001

	Flat	44 (36.1%)	78 (63.9%)	122 (45.2%)	
	Upsloping	98 (75.4%)	32 (24.6%)	130 (48.1%)	
Major vessel	0 (ref)	120 (75.0%)	40 (25.0%)	160 (59.3%)	
	3	3 (15.8%)	16 (84.2%)	19 (7%)	<0.001
	2	7 (21.2%)	26 (78.8%)	33 (12.2%)	
	1	20 (34.5%)	38 (65.5%)	58 (21.5%)	
Thallium stress test	Fixed defect (ref)	6 (42.9%)	8 (57.1%)	14 (5.2%)	
	Normal	119 (78.3%)	33 (21.7%)	152 (56.3%)	<0.001
	Reversable defect	25 (24.0%)	79 (76.0%)	104 (38.5%)	

N.B.: The *p*-value is from the Pearson chi-square test of independence. The symbol (∗) indicates that the chi-square test of independence is not significant at the 5% level. (ref) indicates the reference category of a factor as used in fitting our various models. FBS = Fasting blood sugar; REC = Resting electrocardiographic results; SST = Slope of the peak exercise ST segment, Major vessel = Number of major vessels colored by fluoroscopy.

2.2 Methods

Machine learning (ML) techniques can be implemented in different forms. Notable among these include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning methods. The supervised learning algorithms consist of a label and a set of features. The task is to learn a function that maps an input to an output based on example, input-output pairs. A supervised learning algorithm analyses the training data and produces an inferred function for mapping test data or new input (*X*) to predict the output or label (*Y)*. The two examples of supervised learning techniques are classification and regression. The main difference between these two methods is that regression has the label of the data continuous while classification has the categorical label (or discrete). This paper applied ten classifiers or classification learning methods on two real-life data to predict whether a patient has heart disease or not.

2.2.1 Classification machine learning methods

A classifier is a supervised ML method that learns from the training data and uses that knowledge to predict unseen or future (test) data. This section presents a brief definition of the most widely used classification learning methods employed in this work.

Decision Tree (DT)

A decision tree (DT) is a tree-like structure that consists of a root node, branches, and leaf nodes [8]. It is a non-parametric model that can efficiently deal with large and complex datasets without imposing complicated distributional assumptions. We can implement DT in both classification and regression tasks. It is easy to interpret, robust to outliers, and can also work in the presence of

missing values without needing to resort to imputation. The main disadvantage of the decision tree model is that it can be subject to overfitting and underfitting when using a small data set [13].

Extreme Gradient Tree (XGBTree)

Extreme Gradient Tree (XGBTree), which is also known as Extreme Gradient Boosting (XGBoost) method, is an efficient and scalable implementation of gradient boosted decision trees that are designed for execution speed and model performance [14]. XGBTree offers state-of-the-art results on many challenges and can automatically handle missing data and support tree construction's parallelization. Other features that make XGBTree efficient are cache access patterns, data compression capabilities, and sharding to build a scalable tree boosting system [15].

Conditional Random Forests (Cforest)

Conditional Random Forests (Cforest) is a bagging tree ensemble technique similar to the random forest. The main difference between Cforest and the classical random forest is how the trees are aggregated during the training phase. Cforest uses conditional inference trees as base learners, which puts more weight on terminal nodes with a higher cost. The Cforest approach is better than the classical tree algorithms because the trees are unbiased and do not artificially favour splits in variables with many categories or continuous variables [16].

Logit

Logit, also known as logistic regression, is a statistical model that uses a logistic curve to model the probability of a particular class or event existing. Logistic regression studies the association between a categorical dependent variable and a set of independent (explanatory) variables [17]. Logistic regression can be binomial or multinomial. Binomial or binary logistic regression refers to the instance in which the observed outcome can have only two possible outcomes. Multinomial logistic regression refers to cases where the outcome can have three or more possible types.

Random Forest (Rforest)

Random forest (Rforest) is a popular machine learning model used for classification or regression tasks. Random forest falls under a class of algorithms called bagging, which from numerous experiments have shown to outperform single tree models. It constructs multiple decision trees trained on a bootstrap dataset and aggregates the result. The idea behind this is the belief that multiple decision trees will optimally converge to the perfect decision. One significant advantage

of the random forest is that it drastically decreases the model's variance without increasing the bias [18].

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that uses a robust margin to separate instances into different classes. Given a set of training samples, an SVM model represents the samples as points in space so that there is a clear separation margin between them. New test samples are mapped into that same space and are classified based on the margin's side where they fall.

Artificial Neural Networks (ANNs)

Artificial Neural networks (ANNs) are computational networks mainly inspired by biological neurons [19]. It represents a collection of connected computation units where each unit provides the input to the next unit in the chain. ANNs mainly used an unstructured supervised learning task and have shown great success compared to traditional tree algorithms. One of the reasons for its success is that artificial neural networks can accurately extract features from unstructured data without humans' intervention and accurately learn and simulate high dimensional, non-linear data without prior knowledge. Generally, ANNs are arranged in layers: an input layer, one or more hidden layers, and the output layer [20].

The k-Nearest Neighbour (kNN)

The k-Nearest Neighbour (kNN) is perhaps the simplest, most popular, highly efficient, and intuitive algorithm for pattern recognition [21]. The strategy for predicting an observation class is to identify the k closest neighbours from the training dataset and then assign the class with the most prevalent class among its nearest neighbours. The kNN works well with a small number of input variables (p) but struggles when the number of inputs (features) is large [22]. The kNN does not have a training phase; hence it is referred to as a lazy learner. We can implement kNN for both the classification and regression problems.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a statistical predictive classification method that assigns an unknown class label to one of the classes based on a multivariate observation [23]. LDA finds a linear combination of features that separates two or more classes of observations and makes some

simplifying assumptions about the data—normally distributed data, statistically independent features, and identical covariance matrices for every class. LDA makes predictions by estimating the probability that a new set of inputs belongs to each class, and the class that gets the highest probability is the output class. The model uses Bayes Theorem to estimate the probabilities [24].

Naive Bayes (NB)

Naive Bayes is a probabilistic machine learning classifier that is based on the Bayes Theorem and can be used for a wide variety of classification tasks [21]. The naive Bayes assumption is that the features that go into the model are independent of each other. It is a simple yet powerful algorithm among popular classification methods because it can quickly make predictions in real-time.

2.2.2 Evaluation of Classification Learning Methods

This section briefly discussed the different metrics for assessing the quality and performance of various classification models used in this study. The performance measures employed here include the Accuracy, Sensitivity, Specificity, Precision, Log Loss, and the Area under Receiver Operating Characteristics curve (AUC).

Accuracy or Correct Classification Rate: The accuracy of any given classifier is the ratio of response class labels that the classifier predicted correctly over the total number of predictions.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For a binary classifier, the classification rate is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

where TP = True Positive; TN = True Negative; FP = False Positive and FN = False Negative.

In a class-imbalanced dataset, accuracy may not be an excellent method to evaluate classifiers; instead, we may use recall, precision, or the F1 score.

***True Positive Rate* (TPR)**: The True Positive Rate (TPR), often called the Sensitivity or Recall, is the ratio of the total number of samples that are correctly classified as having the response of

interest (disease present) divided by the total number of samples that have the response of interest in the test data. Sensitivity is obtained as:

$$\text{Sensitivity (TPR)} = \frac{TP}{TP + FN} = \frac{\sum_i^n \mathrm{I}_{[y_i = \hat{y}_i \text{ and } y_i = 1]}}{\sum_i^n \mathrm{I}_{[y_i = 1]}}$$

True Negative Rate **(TNR)**: The True Negative (TNR), often called the Specificity, is the ratio of the total number of samples that are correctly classified as not having the response of interest (disease absent) divided by the total number of samples that do not have the response of interest in the test data. Specificity is obtained as:

$$\text{Specificity (TNR)} = \frac{TN}{TN+FP} = \frac{\sum_i^n \mathrm{I}_{[y_i = \hat{y}_i \text{ and } y_i = 0]}}{\sum_i^n \mathrm{I}_{[y_i = 0]}}$$

Precision: The precision measures how often a classifier correctly predicts the response of interest (disease present). For instance, when the classifier predicts the heart disease to be present in a set of samples, how often is it correct? Precision can be expressed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\sum_i^n \mathrm{I}_{[y_i = \hat{y}_i \text{ and } y_i = 1]}}{\sum_i^n \mathrm{I}_{[\hat{y}_i = 1]}}$$

Log Loss: Logarithmic Loss, also known as Log Loss, is another metric to assess the goodness of a classifier. Mathematically, Log-Loss is expressed as:

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$$

where: n is the sample size; y_i is the true class label that assumes indicator zero (0) if sample i does not have the response of interest (disease absent) and assumes indicator one (1) if sample i has the response of interest (disease present); $\hat{p}_i$ is the model's predicted probability that sample i is of class c in y_i.

Log Loss has no upper bound, and it exists in the range $[0, \infty)$ [25]. In general, the least Log Loss gives greater accuracy for the classifier. Therefore, the goal is to minimize the Log Loss, and a perfect classifier would have a Log Loss near zero while less ideal classifiers would have larger values of Log Loss.

***The area under a ROC curve* (AUC)**: The area under the ROC curve (AUC) metric is used to calculate the area under the receiver operating characteristics (ROC) curve. The AUC score is always bounded between zero and one, and a very poor classifier has an AUC of around 0.5 [26]. The AUC of a classifier represents the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation. Thus, it is a useful metric even for datasets with highly unbalanced classes [26, 27]. Comparing the performance of different classifiers with the ROC curve is not easy [28, 29]. This is because no scalar value represents the expected performance [30].

3.0 Analysis and Results

The predictive performance of machine learning models depends on the dataset's structure and proper data preparation will ensure the models work optimally.

3.1 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) summarises the main characteristics of the heart disease data by using tables and visuals. We performed several EDA on the Cleveland dataset to get more insight into the features that are useful in separating the classes of whether or not a patient would have heart disease before applying machine learning models for prediction.

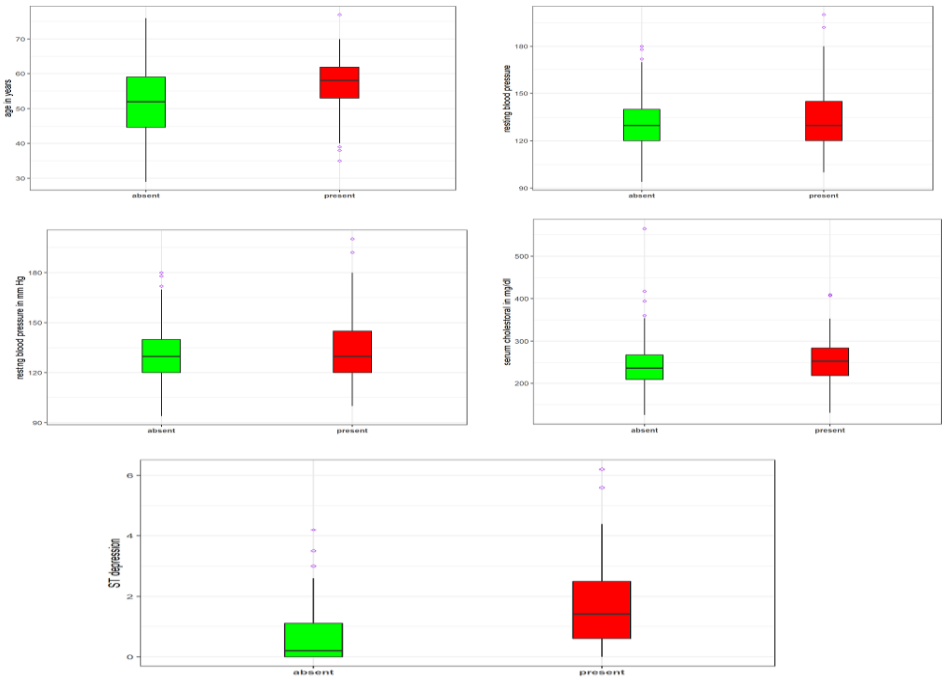

Fig. 2 Box plots of the metrical (continuous) features by class labels in the Cleveland data

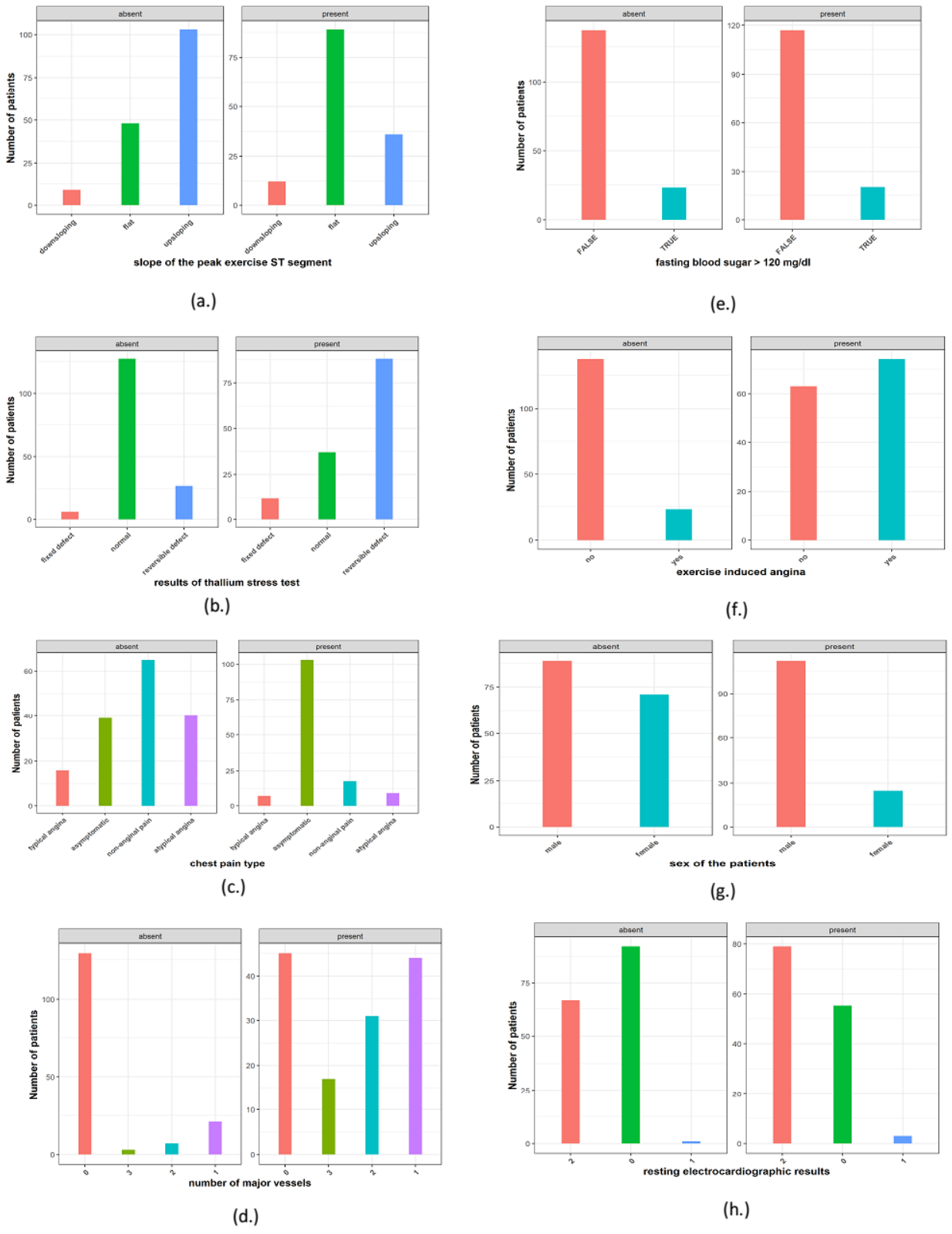

Fig. 3 ((a.) – (h.)**):** Bar plots of features faceted by the class label in the Cleveland data.

As presented in Figure 2, the box plots are very useful, and by construction, we focused on the overlap of the quartiles of the distribution. In this case, we might ask whether there are sufficient differences in the quartiles for the feature to help separate the label classes? All the numerical features in the Cleveland data help separate between the present and absence of heart disease cases. As one might expect, older people tend to have heart disease compared to younger ones.

The bar plots in Figure 3 (a.) to (h.) contain a lot of information. The key to interpreting these plots is to compare the proportion of the categories for each factor level. If these proportions are distinctly different within each class label, such a feature will likely help separate the class labels.

There are several cases evident in the bar plots in Figure 3:

a) Some features such as the slope of the peak exercise ST segment, thallium stress test, chest pain type, and the number of major vessels coloured by fluoroscopy showed significant differences in the number of samples that belong to the various factor levels or groups. That is evidence that those factors might be discriminatory of the response class in the data.

b) Other features such as fasting blood sugar, exercise-induced angina, and sex show slight differences, which are unlikely to be significant.

c) A feature like resting electrocardiographic results has dominant categories with very few cases of other categories. This feature will likely have very little power to separate the cases.

It is important to note that only a few of these categorical features might help separate the response cases.

3.2 Training and evaluation of the classification models

As remarked earlier in section two, the Cleveland data were used as the training data to build the various classification models in this study. The models' efficiency and goodness were validated and evaluated on the second independent dataset, the Statlog data.

After applying a simple backward feature selection, also known as recursive feature elimination using random forest on the Cleveland dataset, fifteen (15) out of the nineteen (19) features were selected as the best crop of features for the final models. During the models' constructions, 80%

of the Cleveland heart disease data were randomly selected and employed for training each model, while the remaining 20% was used as the test data for models' evaluation over 200 replications. Each model's performance, averaged over the 200 replications, was determined for each model's assessment criteria discussed earlier. However, several tuning of the models' hyperparameters were performed with the **caret** package in R over 10-fold cross-validation to ensure models' stability and efficiency. Without loss of generality, the results of the ten classifiers' performances analyzed here on the test data over 200 replications are graphically presented in Figures 4 through 10 for each of the evaluation metrics used in this study.

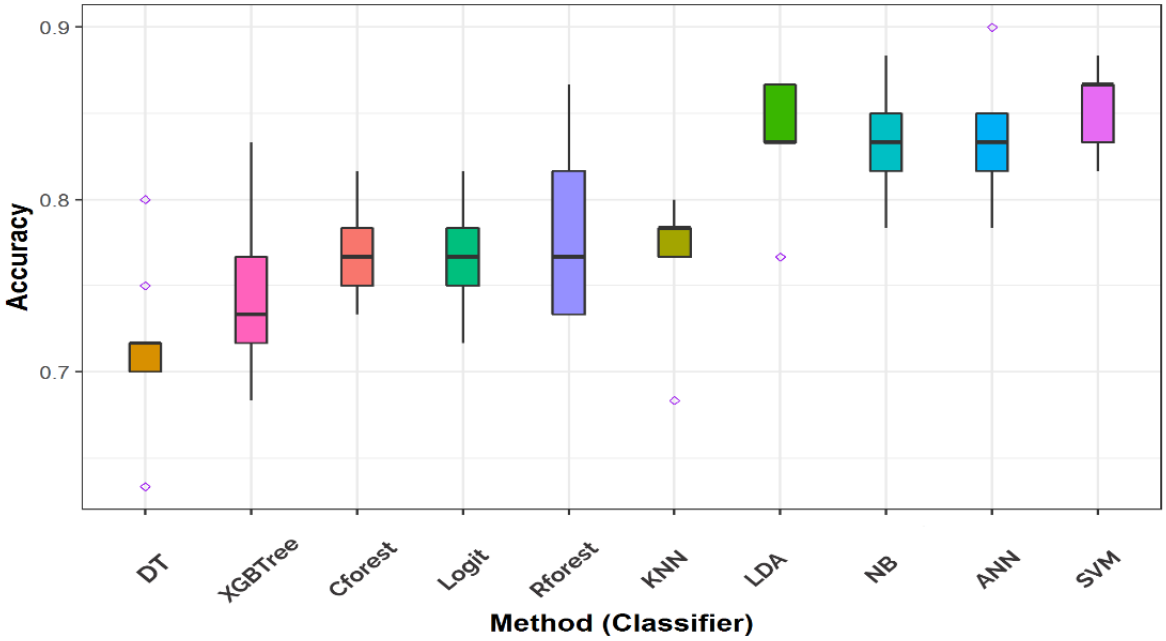

Fig. 4 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on Accuracy

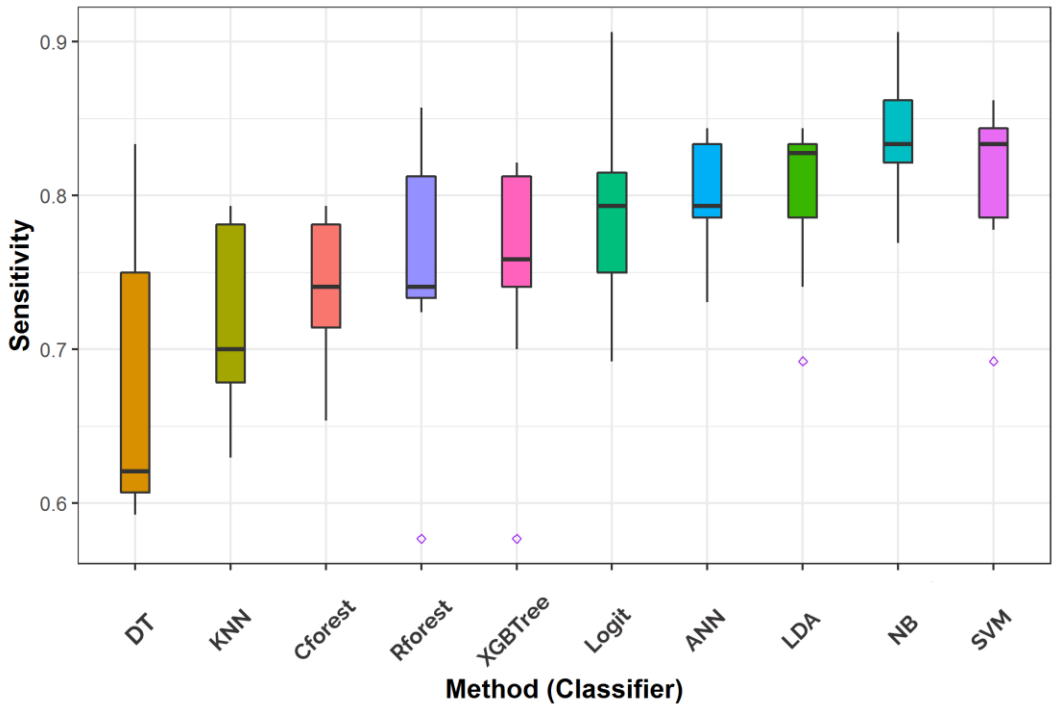

Fig. 5 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on Sensitivity

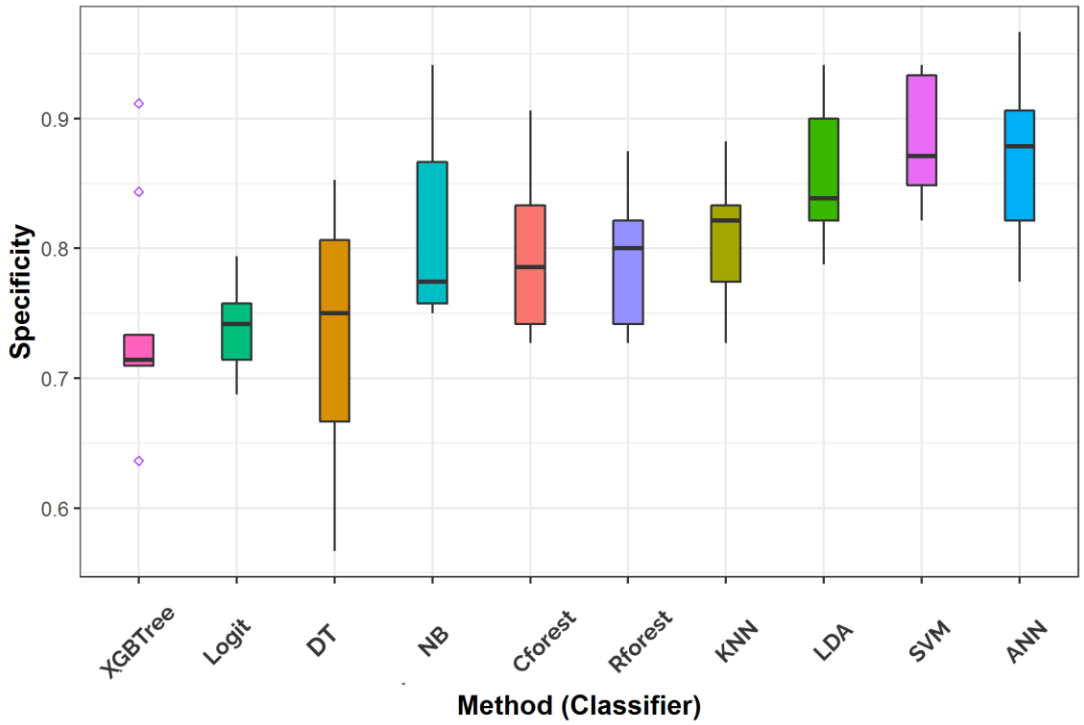

Fig. 6 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on Specificity

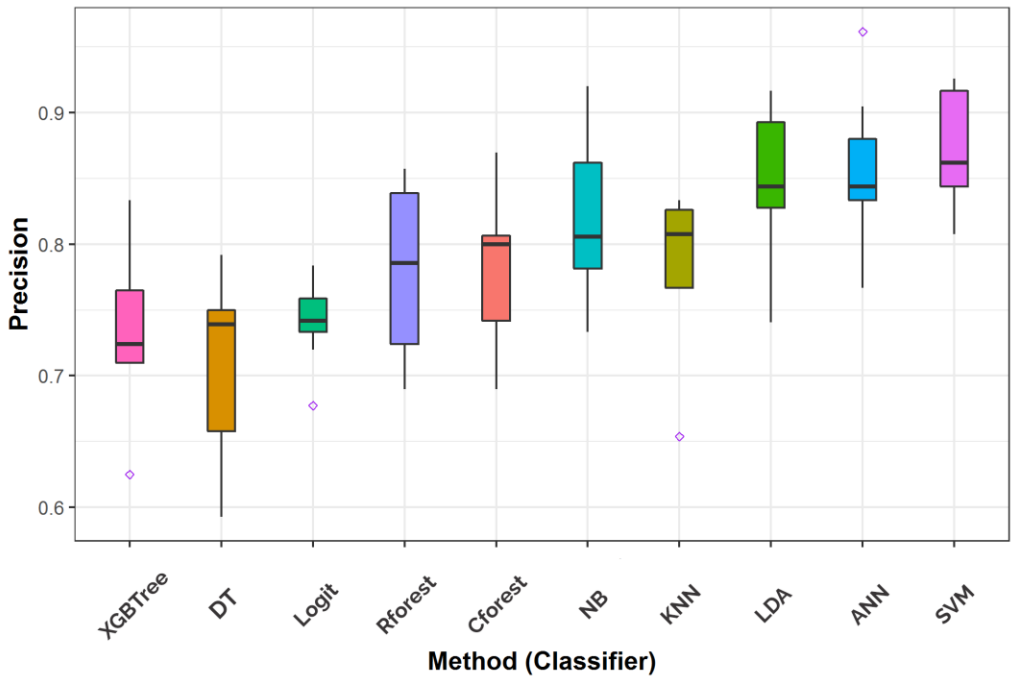

Fig. 7 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on Precision

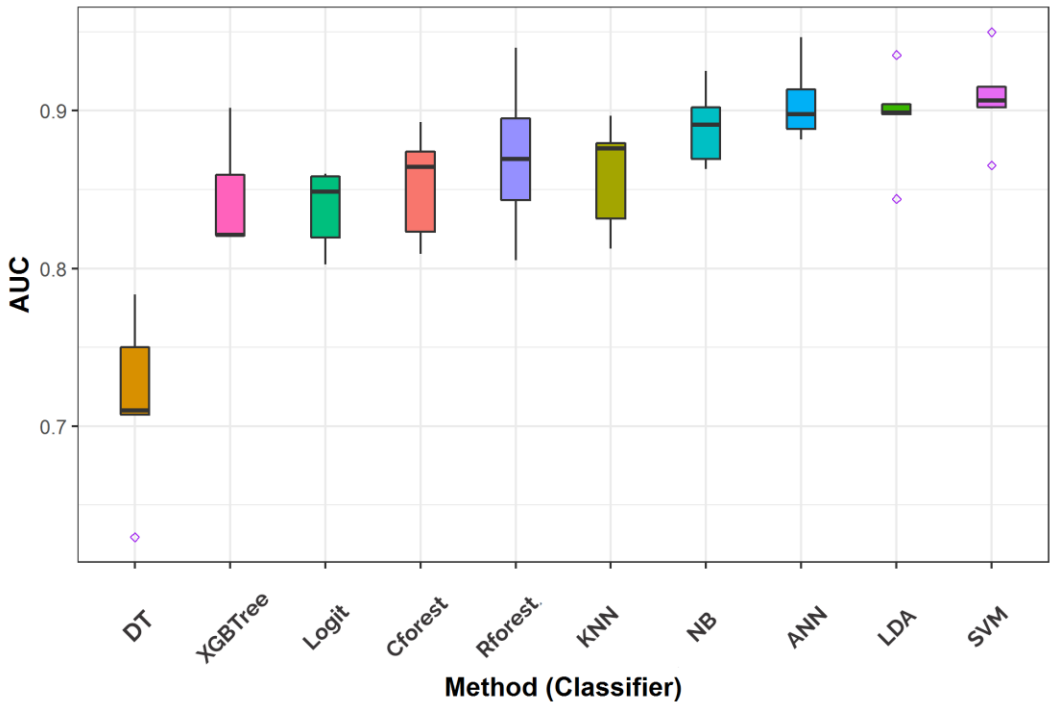

Fig. 8 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on AUC

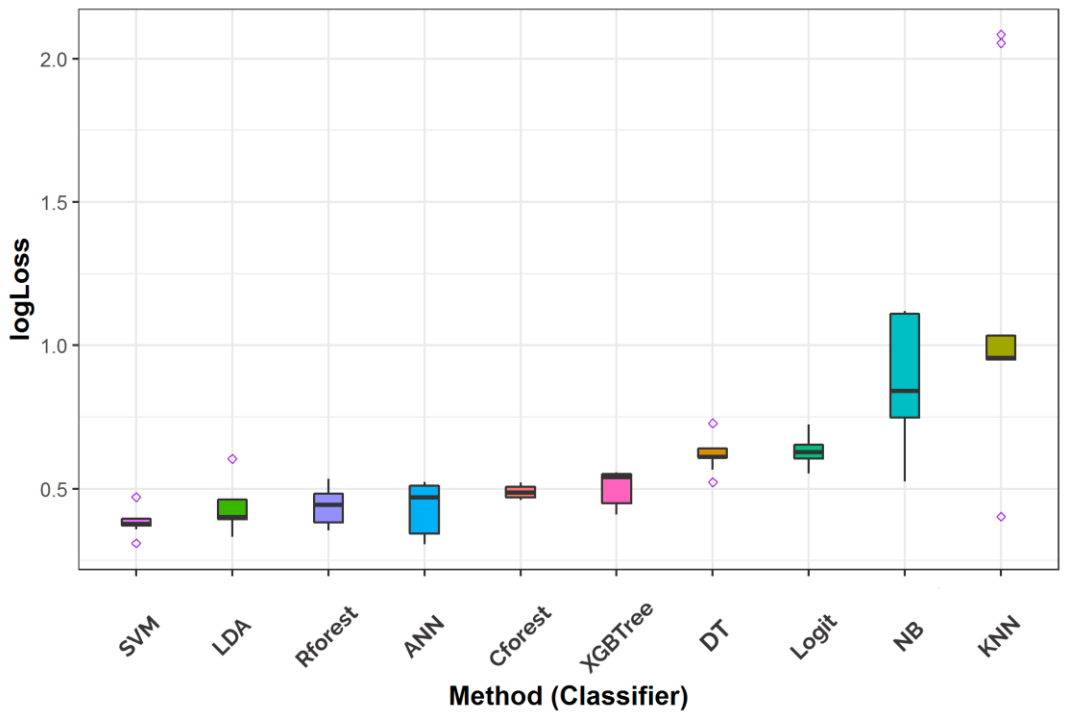

Fig. 9 The box plots of the performances of all the ten classifiers on the test data over 200 replications based on Log-Loss

Table 5: The summary of the prediction performances of all the classifiers on the test data by their median ranks. Models with the best predictive performance in ranks for each model's assessment metric are asterisked.

ML Methods	Median rank					
	Accuracy	Sensitivity	Specificity	Precision	AUC	Log-Loss
Cforest	7	7.5	6	6	7	5
DT	10	10	8	9	10	7
KNN	5	9	4	4	5	10
LDA	3	3	3	2.5	2	2
Logit	7	4.5	9	8	8	8
NB	3	1.5*	7	5	4	9
ANN	3	4.5	1*	2.5	3	4
Rforest	7	7.5	5	7	6	3
SVM	1*	1.5*	2	1*	1*	1*
XGBTree	9	6	10	10	9	6

From the different results obtained in Figures 4 to 9, it is pretty clear that the best model that provided the best prediction performance among all the ten classifiers considered is the SVM. The

SVM outperformed other classifiers, as evident by the assessment metrics results except for the Specificity where SVM came 2nd, as shown in Figure 6.

3.3 Validation Results

As remarked in Section 2, we validated the classification results obtained from the ten models fitted to the Cleveland data using the Statlog data.

The results obtained from the test data showed that the best classifier chosen by all the models' assessment criteria adopted in this study is the SVM. This final chosen model (SVM) prediction performance, as reported in Figures 4 to 10, revealed that the SVM yielded good predictive evaluation results with 85% Accuracy, 87% Precision, 82% Sensitivity, 88% Specificity, 91% AUC, and 38% Log-Loss value.

Prediction results from the validation data showed that the best-chosen model (SVM) yielded an accuracy of 87.04%, Precision of 85.1%, Sensitivity of 85.8%, Specificity of 88%, AUC of 93%, Logloss of 36.1%, and F1-score of 85.5% as shown in Figure 10. These prediction results are similar to the SVM model results on the test data, as reported earlier.

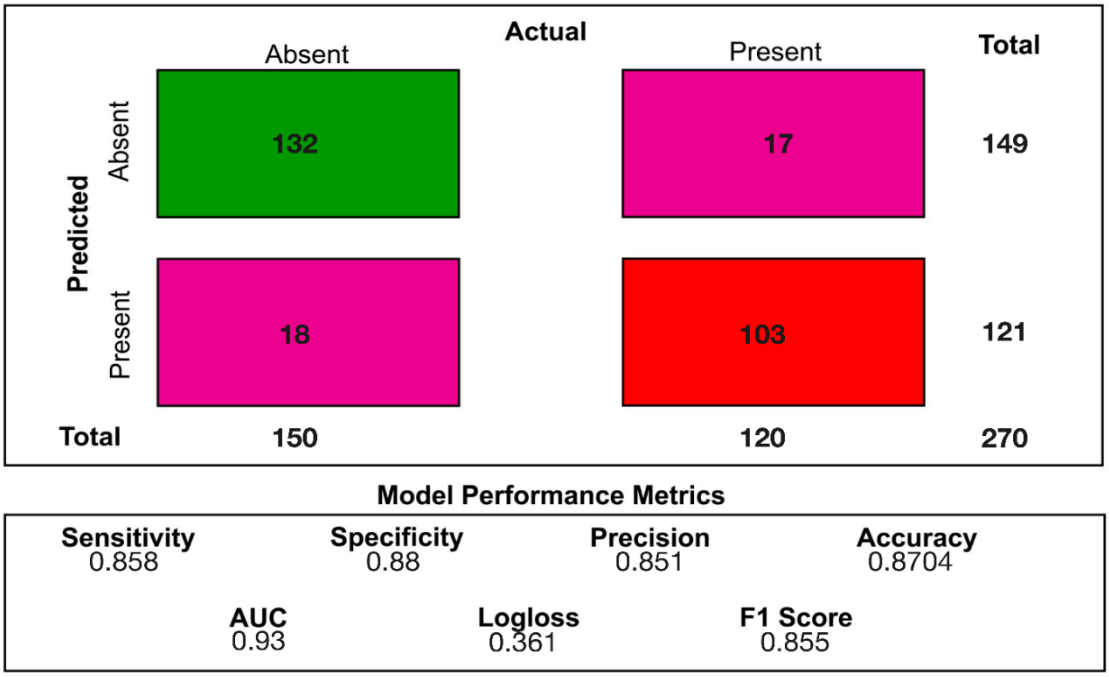

Fig. 10 Prediction performance of the best model (SVM) on validation data (Statlog heart disease data)

4.0 Discussion of results

This study examines the prediction performances of ten selected state-of-the-art machine learning methods for predicting the heart disease status (present or absent) of groups of patients from two real-life publicly available data sets. Results obtained in Figures 4 to 9 and Table 5 showed that choices made by one metric in evaluating the performance of a model were quite different from the choices made by another. Still, overall, the Support Vector Machine (SVM) always performs better than all other classifiers under consideration in the test and validation datasets. Thereby confirmed the study by Pouriyeh et al. (2017) [8] for predicting heart disease with SVM.

Ogundepo and Fokoué (2019) [10] suggested that there is no free lunch of statistical machine learning models. Therefore, we compared ten models based on each model's assumptions and evaluated them using six (6) different classification metrics. Detailed exploratory analysis results from the Chi-square test of association showed that the following bio-clinical categorical variables: *Chest pain type, Exercise Induced Angina, Slope of the peak exercise ST segment, Number of major vessels colored by fluoroscopy, and Thallium stress test* are all strongly associated with the heart disease conditions of the patients in the two data sets ($p < 0.001$).

5.0 Conclusion

We have investigated the possibility of using machine learning to predict an instance of heart disease in this paper. Out of the ten classification models evaluated on the heart disease datasets, we found the SVM method most suitable for predicting heart disease patients' health condition (present or absent) given a set of bio-clinical variables. We validated the performance level of the SVM on the Statlog data set and obtained noticeable similar predictive results. Therefore, the SVM model could be adopted in the future to analyze data sets with a similar structure for better efficiency. Applying artificial intelligence in predictive medicine will help us flag risk factors so that physicians can work together with patients to reduce the chances of future problems. For instance, patients with a greater risk of heart attacks and irregularities could receive more regular EKGs and cardiologist appointments to ensure the best possible quality of life.

Supplementary Materials

This study used the RStudio IDE with R version 4.0.2 to run statistical analyses. The packages used included the rmarkdown for running the codes chunk by chunk; tidyverse suite of packages

for data analysis and visualization; and caret for classification training. The raw data and scripts employed in this study are all available on GitHub via https://github.com/gbganalyst/Heart-disease-paper.

Statements and Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Hannah R, Max R (2018) Causes of Death. Our World in Data

2. Kochanek KD, Murphy SL, Xu J, Arias E (2019) Deaths: final data for 2017

3. Fida B, Nazir M, Naveed N, Akram S (2011) Heart disease classification ensemble optimization using genetic algorithm. In: 2011 IEEE 14th International Multitopic Conference. IEEE, pp 19–24

4. Anderson RN, Smith BL (2003) Deaths: leading causes for 2001

5. Nahar J, Imam T, Tickle KS, Chen Y-PP (2013) Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications 40:96–104. https://doi.org/10.1016/j.eswa.2012.07.032

6. Patel SB, Yadav PK, Shukla DP (2013) Predict the diagnosis of heart disease patients using classification mining techniques. IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) 4:61–64

7. Yahya WB, Rosenberg R, Ulm K (2014) Microarray-based Classification of Histopathologic Responses of Locally Advanced Rectal Carcinomas to Neoadjuvant Radiochemotherapy Treatment. Turkiye Klinikleri Journal of Biostatistics 6:

8. Pouriyeh S, Vahid S, Sannino G, et al (2017) A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE, pp 204–207

9. Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 16:100203. https://doi.org/10.1016/j.imu.2019.100203

10. Ogundepo EA, Fokoué E (2019) An Empirical Demonstration Of The No Free Lunch Theorem. Math Appl 8:173–188. https://doi.org/10.13164/ma.2019.11

11. Janosi A, Steinbrunn W, Pfisterer M, Detrano R (1988) Heart disease data set. The UCI KDD Archive

12. Dua D, Graff C (2017) UCI machine learning repository

13. Song Y-Y, Ying LU (2015) Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry 27:130. https://doi.org/10.11919/j.issn.1002-0829.215044

14. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics 1189–1232

15. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp 785–794

16. Strobl C, Zeileis A (2009) Party on!–A new, conditional variable importance measure for random forests available in party

17. Hapfelmeier A, Babatunde W, Yahya RR, Ulm K (2012) 26 Predictive Modeling of Gene Expression Data. Handbook of Statistics in Clinical Oncology 471. https://doi.org/10.1201/b11800-31

18. Breiman L (2001) Random forests. Machine learning 45:5–32

19. Zou J, Han Y, So S-S (2008) Overview of artificial neural networks. Artificial Neural Networks 14–22

20. Yahya WB, Oladiipo MO, Jolayemi ET (2012) A fast algorithm to construct neural networks classification models with high-dimensional genomic data. Annals Computer Science Series 10:39–58

21. Yahya WB, Ulm K, Ludwig F, Hapflemeir A (2011) K-SS: A sequential feature selection and prediction method in microarray study. International Journal of artificial intelligence 6:19–47

22. Kouiroukidis N, Evangelidis G (2011) The effects of dimensionality curse in high dimensional knn search. In: 2011 15th Panhellenic Conference on Informatics. IEEE, pp 41–45

23. McLachlan GJ (2004) Discriminant analysis and statistical pattern recognition. John Wiley & Sons

24. Brownlee J (2016) Master Machine Learning Algorithms: discover how they work and implement them from scratch. Machine Learning Mastery

25. Buja A, Stuetzle W, Shen Y (2005) Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November 3:

26. Fawcett T (2006) An introduction to ROC analysis. Pattern recognition letters 27:861–874

27. Prati RC, Batista GE, Monard MC (2004) Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence. Springer, pp 312–321

28. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30:1145–1159

29. Drummond C, Holte RC (2004) What ROC Curves Can't Do (and Cost Curves Can). In: ROCAI. Citeseer, pp 19–26

30. Tharwat A (2020) Classification assessment methods. Applied Computing and Informatics