Computers in Biology and Medicine 136 (2021) 104672

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison

Md Mamun Ali [a], Bikash Kumar Paul [a,b,c], Kawsar Ahmed [b,c,*], Francis M. Bui [d], Julian M. W. Quinn [e], Mohammad Ali Moni [e,f,**]

[a] Department of Software Engineering (SWE), Daffodil International University (DIU), Sukrabad, Dhaka, 1207, Bangladesh
[b] Group of Biophotomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh
[c] Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail, 1902, Bangladesh
[d] Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada
[e] Healthy Ageing Theme, Garvan Institute of Medical Research, Darlinghurst, NSW, 2010, Australia
[f] WHO Collaborating Centre on EHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, NSW 2052, Australia

ARTICLE INFO

Keywords:
Cardiovascular disease
Machine learning
Random forest
Decision tree
KNN

ABSTRACT

Machine learning and data mining-based approaches to prediction and detection of heart disease would be of great clinical utility, but are highly challenging to develop. In most countries there is a lack of cardiovascular expertise and a significant rate of incorrectly diagnosed cases which could be addressed by developing accurate and efficient early-stage heart disease prediction by analytical support of clinical decision-making with digital patient records. This study aimed to identify machine learning classifiers with the highest accuracy for such diagnostic purposes. Several supervised machine-learning algorithms were applied and compared for performance and accuracy in heart disease prediction. Feature importance scores for each feature were estimated for all applied algorithms except MLP and KNN. All the features were ranked based on the importance score to find those giving high heart disease predictions. This study found that using a heart disease dataset collected from Kaggle three-classification based on k-nearest neighbor (KNN), decision tree (DT) and random forests (RF) algorithms the RF method achieved 100% accuracy along with 100% sensitivity and specificity. Thus, we found that a relatively simple supervised machine learning algorithm can be used to make heart disease predictions with very high accuracy and excellent potential utility.

1. Introduction

Cardiovascular diseases (CVD) are currently the number one cause of death worldwide and the World Health Organization 2020 estimated this to be around 17.9 million deaths every year [1]. Early-stage detection of CVD is an important way of reducing this toll. Of the many techniques of improving that this disease detection and diagnosis is, data mining. These techniques that relate allow hidden knowledge to be extracted and to identify relationships among attributes within the dataset, and is a promising strategy for CVD classification [2–4].

The delivery of high quality clinical services that are affordable to patients is a critical challenge facing health organizations. The delivery of good service requires both correct diagnosis of patients and identification of effective treatment, while avoiding inaccurate diagnoses [5]. Early-stage detection of CVD also minimizes cost and reduces CVD mortality. Data mining techniques can do the job efficiently at a very low cost using a classification algorithm, which plays a key role in clinical research [6]. Here we investigated how such cheap and simple algorithms might be of enough utility to use clinically, and point the way to improved services.

* Corresponding author. Group of Biophotomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail, 1902, Bangladesh.
** Corresponding author. WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, NSW, 2052, Australia.
E-mail addresses: kawsar.ict@mbstu.ac.bd, k.ahmed.bd@ieee.org, kawsarit08050@gmail.com (K. Ahmed), m.moni@unsw.edu.au (M.A. Moni).

https://doi.org/10.1016/j.compbiomed.2021.104672
Received 1 May 2021; Received in revised form 19 July 2021; Accepted 19 July 2021
Available online 21 July 2021

2. Background study

Researchers have applied different data mining methods such as association rules, classification and clustering to build a model for the prediction of heart disease. Shiva Kazempour Dehkordi1 & Hedieh Sajedi proposed a prediction model based on the prescription using the data mining method [7]. They proposed an algorithm called Skating to enhance the accuracy of the system. Skating is an ensemble method similar to Boosting and Bagging. They compared four classification algorithms such as DT, Naïve Bayes (NB), K-Nearest Neighbours (KNN) and Skating in a different label. They showed that the most accurate given classifier is staking. This classification algorithm gave 73.17% accuracy. However, this is a comparatively low performing method compared to other classification algorithms and methods. For example, Jan et al., in 2018 implemented an ensemble data mining approach using two benchmark datasets collected from a UCI repository (namely Cleveland and Hungarian) where the ensemble of five different classification algorithms such as RF, neural network, NB, classification via regression analysis and support vector machines (SVM) were employed [8]. They observed in that study that the lowest performing algorithm was regression methods, while in contrast, RF provided a very high accuracy of 98.136%.

In 2011, Jyoti Soni et al. applied DT with a genetic algorithm to improve the classification performance, and this was compared with other two algorithms such as NB and classification via cluster methods [9]. They found 99.2% accuracy for the proposed system. Hend Mansoor et al., in 2017 analyzed performances of LR and RF classification algorithms for estimating the risk exposure of CVD patients [10]. They showed that the LR Model showed superior performance than the RF classification algorithm. LR Model provided 89% accuracy, while RF was giving 88% accuracy. Austin et al., in 2013 compared the performance of conventional classification trees with regression trees [6]. Conventional LR showed excellent success in estimating the potential existence of HD.

Le et al., in 2018 employed three classification methodologies for the listed 58 attributes in the dataset collected from UCI Machine Learning Repository [11]. They showed that a support vector machine (SVM) with a linear kernel gave a superior performance, with 89.93 accuracy. Tarawneh and Embarak proposed a hybrid approach using 12 features and compared the performance with KNN, J48, GA, DT, artificial neural network (ANN), SVM and NB [12]. The proposed hybrid method generated 89.2% accuracy, which is the best performance compare to other applied algorithms. Chitra and Seenivasagam in 2013 proposed use of a classifier called cascaded neural network (CNN) that to increase the accuracy to predict heart disease [13]. A CNN has a cascade architecture in which the network is supplemented with cached neurons one at a time, which do not change after adding to the hidden network by neurons. The result of the proposed method was compared with SVM, where SVM giving 82% and CNN 85% accuracy and 0.87 and 0.775 specificity respectively. Having considered these parameters, they suggested CNN since the classifier predicted heart disease with higher accuracy and the model with CNN classifier is more accurate than SVM.

Latha and Jeeva implemented an ensemble classification technique using the Cleveland dataset and ensembled Majority vote with MP, RF, BN and NB using the feature selection method to improve the accuracy of the classifier [14]. The performance was evaluated from six sets of attributes. They built different ensembled models and compared the performance to find the best ensemble model. They found that the Majority vote with MP, RF, BN and NB employing attribute selection method offered the best performance with 85.48% accuracy and they proposed this ensembled method to predict heart disease. However, it is now possible to find methods giving better accuracy than their proposed model. Mohan et al., in 2019 introduced a heart disease prediction model using hybrid machine learning techniques [15]. In this study, they employed Rattle, a Graphical User Interface tool for Data Mining using R, to classify HD based on the dataset collected from the Cleveland UCI repository. This produced an increased performance level, which had 88.7% accuracy through the prediction model for HD with the hybrid RF with a linear model (HRFLM). They compared the proposed model with different classification algorithms and showed that their implemented model gave a better result than other classification algorithms.

Table 1
Details of features.

SN	Attribute name	Description
1	age	Age in years
2	sex	Male = 1; Female = 0
3	cp	chest pain type (4 values)
4	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholesterol in mg/dl
6	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	resting electrocardiographic results
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0–3) colored by fluoroscopy
13	thal	1 = normal; 2 = fixed defect; 3 = reversible defect
14	Target (Class)	0 = no disease and 1 = disease

In this section, some of the research studies are introduced that which were conducted to predict heart disease outcome using machine learning and data mining approaches. It is clear from the above discussion that the accuracy obtained in individual research work is not currently satisfactory. Better performance can be obtained with some of the algorithms compared to others. The research study has been successfully able to identify three algorithms that show 100% accuracy by 10-fold cross-validation. The study thus aims to find those classifiers that are able to predict heart disease efficiently enough to be clinically useful.

3. Experimental setup

3.1. Data collection

In this study, a heart disease dataset was processed to design our expected model. The dataset was gathered from Kaggle [16]. There are 14 attributes in this dataset. Table 1 depicts the details of all features.

The dataset contains 1025 patient records including 713 males and 312 females of different ages where 499 (48.68%) patients are normal and 526 (51.32%) patients have heart disease. Among the patients, who have heart disease, 300 (57.03%) patients are male and 226 (42.97%) patients are female.

3.2. Data preprocessing

We used Weka version 3.8.3 as a data mining tool to conduct the study and Python version 3.8.5 for exploratory data analysis (EDA) and visualization. Data preprocessing is mandatory for any machine learning or data mining approach, since the performance of a machine learning methodology depends on how well the dataset is prepared and structured. A ReplaceMissingValues filter was applied to handle missing values then, another filter was applied, which is known as the Interquartile Range (IQR), to detect outlier and extreme values at the phase of pre-processing. The IQR is a method to measure the variability about the median of a dataset. The outlier is a data point that does outside the expected range of the data and can be assumed for the purposes of the analysis to be due to recording errors or other irrelevant phenomena [17]. For machine learning (ML) or data mining methods [18] it is important to remove such outliers to get a better analytical or statistical result. For outlier detection, data is partitioned into three quartiles, Q_3, Q_2 and Q_1. Here Q_1 and Q_3 are the boundaries of data. We calculated the

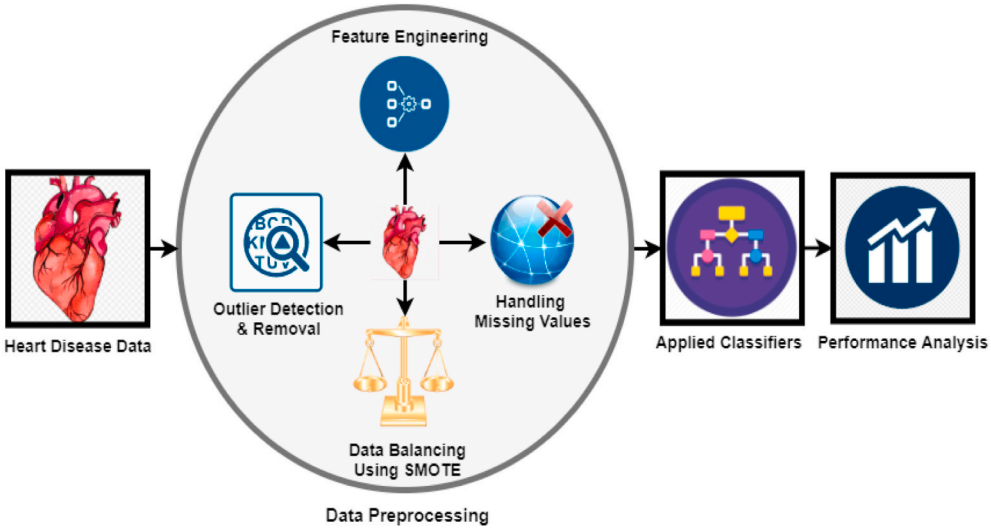

Fig. 1. Experimental methodology.

value of *IQR* by $IQR = Q_3 - Q_1$. Then lower boundary B_l and upper boundary B_u *were* calculated using the following equations [19]:

$$B_l = Q_1 - 1.5*IQR \tag{1}$$

$$B_u = Q_3 + 1.5*IQR \tag{2}$$

Here, a result lower than B_l and greater than B_u is considered as an outlier. Synthetic minority oversampling technique (SMOTE) was also applied to balance the imbalanced dataset. Thus, some exploratory data analyses (EDA) was performed (such as box plot) to confirm that the dataset is free of outliers, and the data was represented as IQR and heatmap to detect correlations among the features, and a KDE plot for both diseased and non-diseased individuals according to age distribution.

3.3. Performance evaluation metrics

Six (06) classification algorithms were applied to the dataset to find the best performer algorithm comparing the accuracy and other statistical variables by 10-fold cross-validation. The algorithms applied were multilayer perceptron (MP), K-nearest neighbours (KNN), random forest (RF), decision tree (DT), logistic regression (LR) and AdaboostM1 (ABM1). These algorithms were compared based on their performance evaluation metrics. A brief overview of these performance evaluations is described in this subsection.

A confusion matrix was obtained to calculate the sensitivity, specificity and accuracy of the result for each algorithm. The below mentioned formulas were used to calculate all the parameters [11,13]:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Specificity} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$\text{TPR} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{7}$$

Here, *TP* and *TN* represent true positive and true negative respectively and *FP* and *FN* demonstrate false positive and false negative; *TPR* represents the true positive rate and *FPR* the false positive rate. Sensitivity relates to the percentage of actual positives that the classifier accurately defines as data and reflects the number of positive predictions that the classifier correctly identifies [20]. Specificity is the ability of the classifier to correctly distinguish negative outcomes [20]. Accuracy is the percentage of correctly classified instances by a classifier [11,13,20].

Different statistical values were used to compare the efficiency of different algorithms such as kappa statistics, precision, recall, f-measure, Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC) and precision-recall (PRC). Kappa statistics estimate inter-rater agreement from identified and expected accuracy for qualitative attributes [21]. Precision is a valid evaluation metric particularly when the proposed ML model is required to validate based on the predicted and actual result [20,21]. It calculates the percentage of expected positives that are actual positives. As a result, it is reliant on TP and FP values. When it is required to determine the number of positives that may fairly be predicted, recall is another useful evaluation metric [20,21], representing the proportion of positives successfully categorized. Recall is measured using TP and FP values. F-Measure maintains a balance between precision and recall for a classifier. The F-Measure score is a number between 0 and 1 that represents the statistically significant measures of precision and recall [20,21]. In machine learning, the MCC is used to assess the validity of binary and multiclass classifications. It accounts for true and false positives and negatives and is often recognized as a balanced metric that may be applied even when the classes are of considerably different sizes. The MCC is essentially a correlation coefficient number ranging from −1 to +1 [22]. These parameters estimate their values using the following equations [22–24]:

$$\text{Kappa, k} = \frac{PrPr\,(a) - Pr(e)}{1 - Pr(e)} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F} - \text{Measure} = \frac{2*precision*recall}{precision + recall} \tag{11}$$

$$\text{MCC} = \frac{TP*TN - FP*FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

ROC is used to determine how much a model is capable of

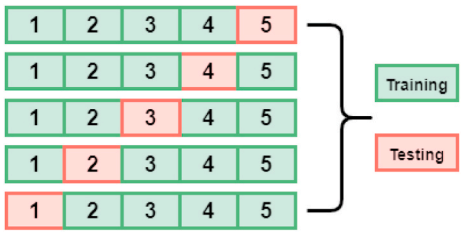

Fig. 2. Graphical representation of K-fold cross-validation.

distinguishing classes. PRC is the ratio of precision and recall.

In this study, K-fold cross-validation was used to train and test the model. In this approach, the data set is divided into a number of groups. K refers to the number of groups, also known as 'fold'. At the same time, cross-validation is an approach to evaluate a machine learning model. K-fold cross-validation is such a technique, where the data set is split into *k* number of groups and the model is trained by *(k-1)* groups and the other group participates to test or evaluate the trained model. In this approach, the model is trained k number of times and each time, different fold participates to evaluate the model. It indicates that each fold participates to train and test a model in K-fold cross-validation.

Fig. 2 represents a 5-fold cross-validation approach. The figure depicts that the dataset is split into 5 folds or groups, where 4 groups participate in model training and another one-fold participates to evaluate the training in each iteration. In our study, 10-fold cross-validation is employed. Cross-validation is performed to protect against overfitting in a predictive model.

3.4. Supervised machine learning algorithms

Different kinds of supervised machine learning algorithms were employed in this study. In supervised machine learning algorithms, the labeled training dataset is employed first of all to practice the fundamental algorithm. This qualified model is then loaded into a non-labeled testing research dataset to categorize it into related categories [25]. A quick overview of these proposed supervised machine learning formulas for disease detection is given in the corresponding subsection.

3.4.1. K-nearest neighbor (KNN)

KNN is one of the oldest and easiest classification algorithms [26,27] or statistical learning techniques [28]. K refers to the number of the nearest neighbours used, which can be defined directly in the object builder or simply calculated using the upper limit provided by the stated value [24]. Similar cases are thus subject to similar classifications [29] and a new instance is categorized by measuring its similarity to each of the current instances [30]. When an unidentified sample is received, the nearest neighbor algorithm will scan the pattern space for the k training samples next to the unfamiliar sample. From the test instance based on their distance, predictions from several neighbours can be calculated and two distinct methodologies are introduced to transform the distance into a weight [28,31]. The algorithm has a number of advantages such as it is analytically tractable and very easy to implement [28]. The classifier is very effective and performs well in disease prediction especially in HD prediction since it works with a single instance. In this study, the value of n_neighbors 2 and leaf_size 40 were the best fit parameter for the dataset.

3.4.2. Random forest

RF is a method for classifying data by ensemble learning based on DT [32]. It creates a large number of trees and also produces a forest of decision trees, while it is under the training stage [33]. Every tree, a member of the forest, forecast class label for every single instance at the testing period. When a class label is predicted by each tree, then majority voting is used to decide the final decision for each test data [34]. The class label that obtains the largest number of votes is considered as the most appropriate label applied to the test data. For every data in the data collection, this cycle is replicated. The best fit random state value for this study was 123, which gave the best performance for the applied dataset.

3.4.3. Decision tree (DT)

DT is one of the oldest and most common machine learning algorithms. A DT designs the logic of the decision in such a way that evaluates and matches results for the classification of data items into a structure as like a tree [25]. Usually a DT has multiple levels of nodes, the topmost level is known as root or parent node and others are child nodes. Evaluation of input variables or features is represented by all internal nodes that contain at least one child node. Depending on the evaluation outcome, the classification techniques branch to the correct child node, where the evaluation and branching process continues before the leaf node is reached [34]. The leaf or terminal nodes refer to the outcomes of the decision. DT is recognized as easy to understand and learn and is a basic component of many protocols for medical diagnosis [35]. The maximum depth for this classification algorithm was defined 7 and the classifier by this maximum depth value produced the best result for the applied dataset in this study.

3.4.4. AdaboostM1 (ABM1)

ABM1 is one kind of ensemble learning based supervised machine learning classifier, which is widely used. It employs an adaptive enhancement approach and produces improved classification results by integrating multiple weak classifiers into a strong classifier [36]. In the initial stage, the same weight is allocated to all the observations. The weights of the observations change with the coefficient of weak classifiers, and the coefficient of the applied classifiers is estimated using the value of the estimation error. So, the value of error generated by a classifier is considered as the coefficient of the classifier. Consequently, the weight of misclassified observations can be raised by the ABM1 algorithm and the weight of correctly identified observations can be reduced. In the subsequent iterations, it will enforce higher weight on the incorrectly classified observations more. Finally, all the weak classifiers developed are combined to form a stronger classifier using a linear combination method [37] to produce accurate classification performance. The value of n_estimators was defined as 200, with this classifier providing the best performance in this study.

3.4.5. Logistic regression (LR)

LR is a strong classifier among supervised machine learning algorithms [38] and is an extension of the general regression modeling as applied to a dataset, represents the probability of occurrence or nonoccurrence of a particular instance [39]. LR identifies the chances of a new observation belonging to a certain class, the result lying between 0 and 1 since it is a probability. Consequently, a threshold is assigned that defines the separation into two classes to implement the LR as binary classification. For instance, a probability value calculated higher than 0.5 is designated as 'class A' otherwise 'class B'. To design a categorical variable, which contains more than two values, the LR model can be generalized as a multinomial logistic regression [40]. The best fit random state value 1234, and the best fit maximum iteration number 100 were found in this study for the applied dataset.

3.4.6. Multilayer perceptron (MLP)

MLP is a well-established neural network-based classification algorithm, which consists of three or more types of layers: an input layer, output layer and one or more hidden layers between input and output layers [41]. Every layer contains a number of 'neurons' connecting all the layers with each other. MLP is a universal multivariate non-linear mappings calculator that results from the capacity of training data to learn and generalize [2] from training data using backpropagation learning methods [42]. The construction of MLP classifiers consists of adequate input variables and specification of the type of network,

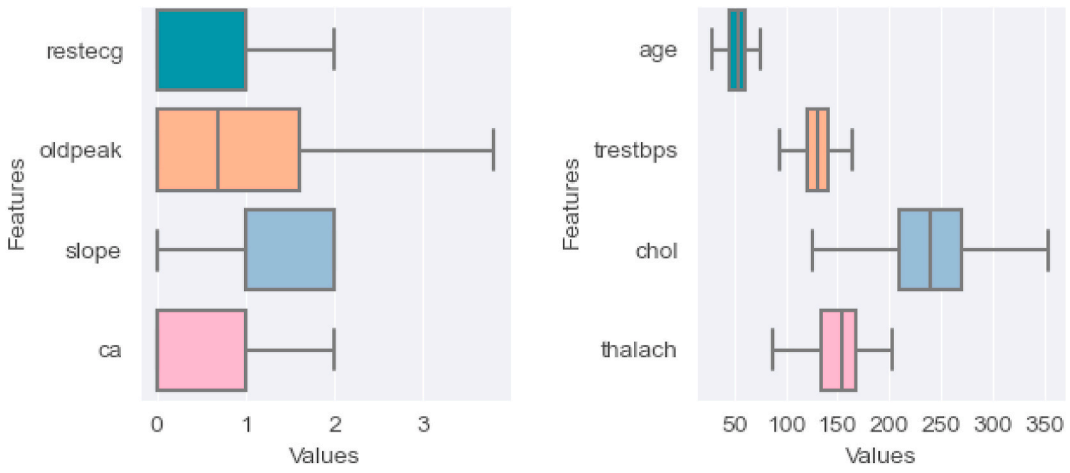

Fig. 3. Box plot for outlier detection and removal.

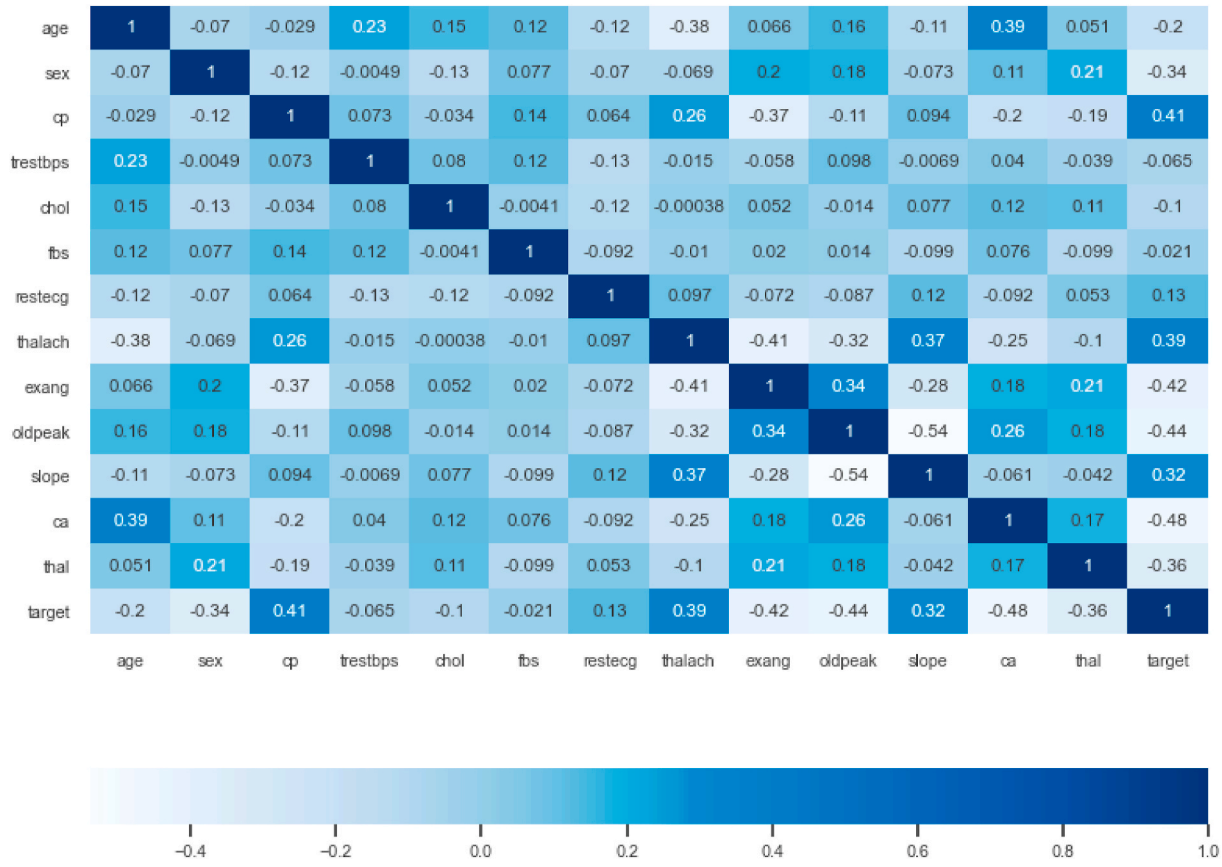

Fig. 4. Heatmap showing correlations among all the features of the dataset.

relevant data pre-processing and partitioning, the configuration of network infrastructure, specification of success parameters, specification of training algorithm (optimization of relation weights), and finally evaluation of model [43]. The default configuration produced the best result by this classifier in this study.

3.5. Feature importance

Feature importance and its visualization is an important and widely used analysis method in the field of machine learning. It is particularly applied in areas such as biomedicine and social sciences due to their simplicity and interpretability of feature ranking or risk analysis [44]. Feature importance and ranking are found based on the coefficient value of each feature [45,46]. Though most of the supervised algorithms provide feature importance or coefficient score, but MLP and KNN do not generate any feature importance or coefficient score [47,48]. Apart from these two classifiers, feature importance or coefficient scores are identified and represented in corresponding sections.

4. Results & discussion

4.1. Result of exploratory data analysis (EDA)

For this dataset, exploratory data analyses were performed to better

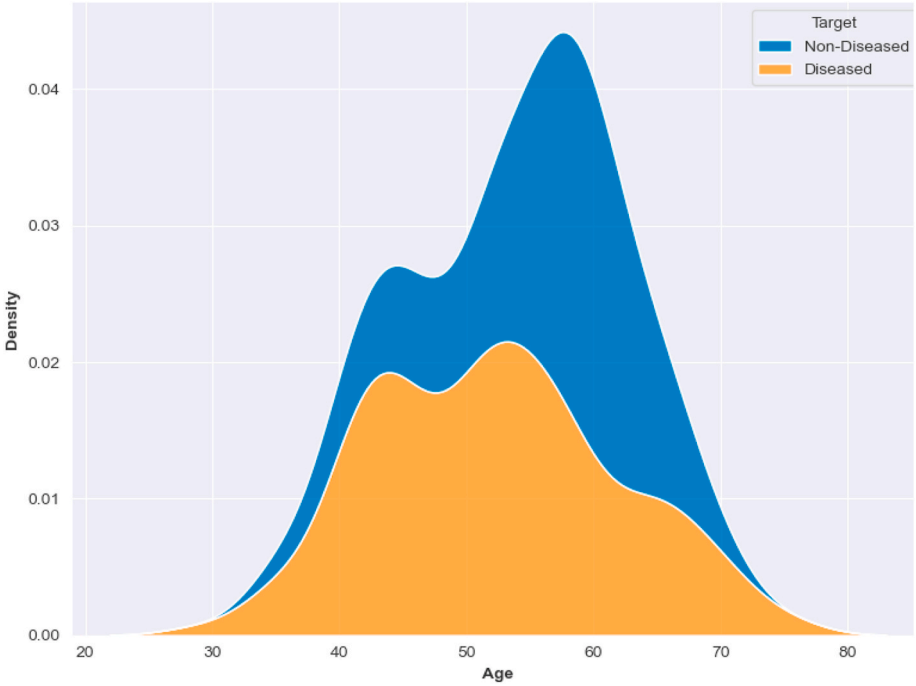

Fig. 5. KDE plot for both heart diseased and non-diseased people according to age distribution.

Table 2
Classification results of Different Classification Algorithm.

Classifier Algorithm	Sensitivity	Specificity	Accuracy
LR	0.820	0.950	89.627
ABM1	0.910	0.979	95.021
MLP	0.984	0.976	97.951
KNN	1.000	1.000	100.000
DT	1.000	1.000	100.000
RF	1.000	1.000	100.000

Table 3
Evaluation by kappa and MCC.

Classifier Algorithm	Kappa	MCC
LR	0.782	0.786
ABM1	0.897	0.898
MLP	0.959	0.959
KNN	1.000	1.000
DT	1.000	1.000
RF	1.000	1.000

understand the features of the dataset. The results of these analyses are described in the following subsection.

Fig. 3 shows the distribution of the quantitative features of the dataset. The values or pints outside of the boxes and whiskers are outliers. At the initial stage, all the detected outliers are visualized in this figure. Outliers were detected using the inter-quartile range and removed from the dataset. The figure depicts no outliers in this dataset after this filtering. After removing all the outliers, rest of the instances of the dataset are used for further analysis.

Fig. 4 is a heatmap representing the correlated values and correlations between features. All the colored cells represent a correlation between two features and their correlated values with the color of the cell indicating the strength of correlation where a correlation value is less than zero indicates negative correlation and zero value indicates no correlation.

Fig. 5 shows the density distribution of a dataset of diseased and non-diseased patients. It can be seen that the patients between the 50–60 years of age are the affected group according to the applied dataset. The plot thus suggests that age is an important factor for heart disease, and the probability of disease increases with age.

4.2. Result of machine learning analysis

For this study, a heart disease dataset has been processed, outliers detected and removed and a number of different classification algorithms applied, including MLP, KNN, DT, RF, LR and ABM1. These all-classification algorithms were employed with 10-fold cross-validation methods on the dataset. The different cross-validation performance parameters were compared to determine the best performing algorithm for predicting heart disease occurrence. Fig. 1 depicts the whole process.

Table 2 shows the performance outcome parameters of the classification algorithms employed, namely sensitivity, specificity and accuracy. These all show good outcomes, with parameters KNN, RF, and DT

Table 4
Precision, recall and F-Measures.

Classifier Algorithm	Precision	Recall	F-Measure
LR	0.896	0.896	0.896
ABM1	0.950	0.950	0.950
MLP	0.980	0.980	0.980
KNN	1.000	1.000	1.000
DT	1.000	1.000	1.000
RF	1.000	1.000	1.000

Table 5
Value of area under ROC and PRC.

Classifier Algorithm	AUROC	AUPRC
LR	0.918	0.921
ABM1	0.998	0.999
MLP	0.927	0.951
KNN	1.000	1.000
DT	1.000	1.000
RF	1.000	1.000

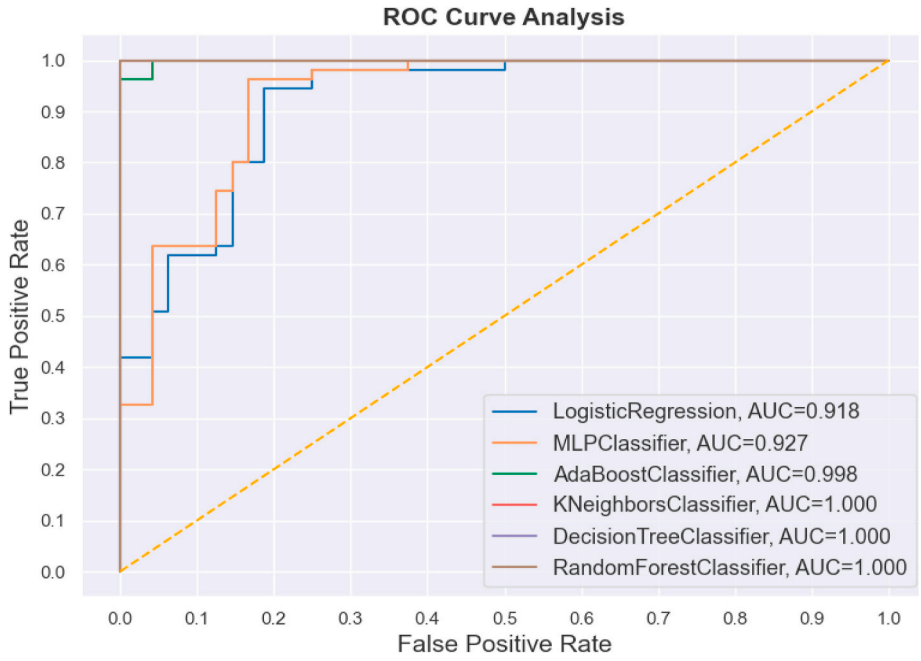

Fig. 6. ROC Curve for different applied algorithms.

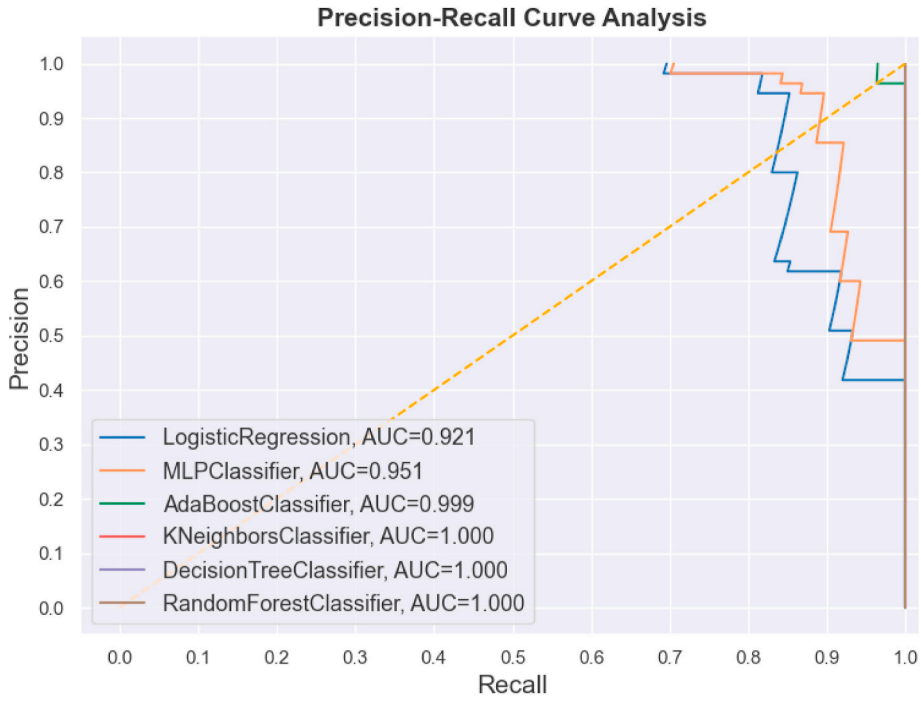

Fig. 7. Precision-Recall Curve for different applied algorithms.

provide maximal accuracy, sensitivity and specificity, followed by MLP which shows better performance than LR and ABM1.

Table 3 represents kappa statistics and MCC values of different classification algorithms. According to the results presented in the table, MLP is a far better performer than ABM1 and LR. KNN, RF and DT are the best performers and show the highest value.

According to Table 4, LR and ABM1 give a poorer performance than MLP, while precision, recall and f-measures are considered. At the same time KNN, RF and DT show very high performance, 100%.

Table 5 shows the value of area under ROC and PRC for all the applied classification algorithms. The area under ROC represents a common area of true positive rate and false positive rate, while the area under PRC represents a common area of precision and recall. Though LR and MLP show the closer result, ABM1 shows higher performance than them. On the other hand, KNN, RF and DT provided the best results.

Fig. 6 represents the ROC curve, which is built by the value of the true positive rate and false positive rate. It is a graphical representation of the area under ROC (AUROC).

Fig. 7 shows the area under the precision-recall curve (AUPRC) of different classification algorithms. The area is found by the AUPRC values. It is a graphical representation of the AUPRC value of Table 5.

Feature importance and the coefficient score is estimated for all the

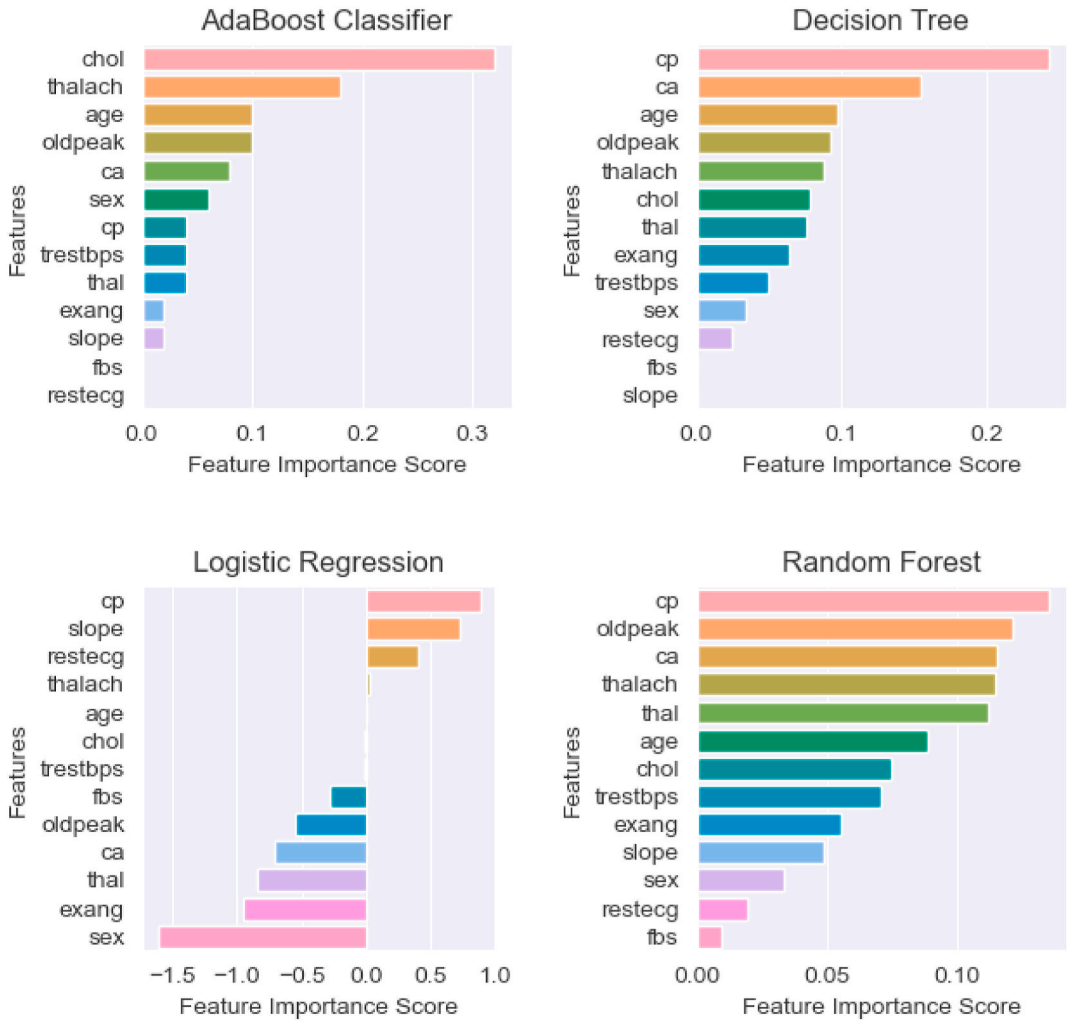

Fig. 8. Graphical representation of feature importance.

Table 6
Feature importance and coefficient scores of different applied algorithms.

Features Name	LR	ABM1	DT	RF
Cp	0.883683	0.04	0.242482	0.135172
Oldpeak	−0.542273	0.10	0.092035	0.121586
Ca	−0.714725	0.08	0.153895	0.115208
Thalach	0.029882	0.18	0.088194	0.114974
Thal	−0.849471	0.04	0.075335	0.112088
Age	0.006131	0.10	0.097665	0.088674
Chol	−0.004922	0.32	0.078611	0.074510
Trestbps	−0.010131	0.04	0.049644	0.070781
Exang	−0.944543	0.02	0.064114	0.055598
Slope	0.733114	0.02	0.000000	0.049065
Sex	−1.610145	0.06	0.034146	0.033472
restecg	0.399593	0.00	0.023878	0.019447
Fbs	−0.285682	0.00	0.000000	0.009424

applied classification algorithms except MLP and KNN since these two algorithms do not produce such any feature importance or coefficient values. These feature importance and coefficient scores are represented in the table according to corresponding features. The table is visualized in Fig. 8 for a better understanding of feature ranking and importance according to classification algorithms.

Fig. 8 is a graphical representation of Table 6. The figure shows the feature ranking based on feature importance and coefficient scores for all the applied classification algorithms except MLP and KNN. The figure also tends to represent the highly responsible attributes for heart disease.

Table 7
Top five features for heart disease according to applied algorithms.

Feature Ranking	LR	ABM1	DT	RF
1st	cp	chol	cp	cp
2nd	slope	thalach	ca	oldpeak
3rd	restecg	age	age	ca
4th	thalach	oldpeak	oldpeak	thalach
5th	age	ca	thalach	thal

Table 7 shows the five most significant features according to feature importance and correlation value. According to the table, it is found that chest pain (cp) is the significant feature or factor for heart disease identification and prediction. Besides age, the number of maximum heart rates achieved (thalach), ST depression induced by exercise relative to rest (oldpeak), and the number of major vessels (0–3) colored by fluoroscopy (ca) are also significant factors rpedicting heart disease.

In summary, we collected a heart disease dataset, preprocessed as necessary, then was performed to better understand the dataset. Then we applied six machine learning algorithms, ABM1, LR, MLP, KNN, DT, and RF, and evaluated their predictions based on accuracy, sensitivity, specificity, kappa statistics, precision, recall, F-Measure, and MCC. ROC curve and Precision-Recall curve. We found the good performance of all the applied algorithms, where KNN, DT, and RF showed the best performance with 100% accuracy indicating that these are the most

efficient at predicting heart disease. We also estimated the feature importance and coefficient values of all the applied algorithms except MLP and KNN since these two algorithms did not generate any feature importance score or coefficient values. The results of feature importance score are indicated in Table 6, where these features are ranked and represented graphically in Fig. 8 according to the feature importance score. This analysis identified highly predictive features for the detection of heart disease that show potential utility to clinicians seeking to predict heart disease occurrence in their patients. It should be noted, however, the quantity of data on heart disease provided by this dataset was not large enough to adequately address all issues and that further data and analysis is needed to show produce a robust prediction method. Nevertheless, in future, we hope to better understand the limits of this methodology and that analysis of additional data will enable highly accurate predictions of heart disease and related conditions using machine learning approaches.

5. Conclusion

Heart disease is life-threatening, which leads to potentially fatal complications such as heart attacks. Due to its potential for accurate disease prediction rate, the importance of data mining and machine learning techniques could be used to predict its occurrence. Here, we used a heart disease dataset to test the utility of ML approaches to heart disease prediction, finding that three classification algorithms KNN, RF and DT performed extremely well with 100% accuracy. In addition, feature importance scores for each feature was estimated for all the applied algorithms except MLP and KNN. These features were ranked based on feature importance score. The study aimed to find the best ML techniques, among a number of algorithms that are well accepted and easy to implement, finding that, at least for this dataset, they performed well. This is an early stage of using ML approaches but suggests that it could prove to be an excellent adjunct to patient care.

Ethics approval and consent to participate

The authors provide the Ethical Approval for this research and the participants have no competing interests.

Consent for publications

There is no competing interest for publication among all the authors who have participated in this research.

Funding

This work was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Declaration of competing interest

We declare that we have no conflict of interest.

Acknowledgement

The Authors are grateful to those who have contributed in this research and also to those who give their valuable data in this research.

References

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 02 June 2021].
[2] R.D. Canlas, Data Mining in Healthcare: Current Applications and Issues, School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.
[3] Christoph Helma, Eva Gottmann, Stefan Kramer, Knowledge discovery and data mining in toxicology, Stat. Methods Med. Res. 9 (4) (2000) 329–358.
[4] I.-N. Lee, S.-C. Liao, M. Embrechts, Data mining techniques applied to medical information, Med. Inf. Internet Med. 25 (2) (2000) 81–102.
[5] L. Parthiban, R. Subramanian, Intelligent heart disease prediction system using CANFIS and genetic algorithm, Int. J. Biol., Biomed. Med. Sci. 3 (3) (2008).
[6] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (4) (2013) 398–407.
[7] S.K. Dehkordi, H. Sajedi, Prediction of disease based on prescription using data mining methods, Health Technol. 9 (1) (2018) 37–44.
[8] M. Jan, A.A. Awan, M.S. Khalid, S. Nisar, Ensemble approach for developing a smart heart disease prediction system using classification algorithms, Res. Rep. Clin. Cardiol. 9 (2018) 33–45.
[9] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comput. Appl. 17 (8) (2011) 43–48.
[10] H.M. Islam, Y. Elgendy, R. Segal, A.A. Bavry, J. Bian, Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach, J. Heart & Lung (2017) 1–7.
[11] H.M. Le, T.D. Tran, L.A.N.G. Van Tran, Automatic heart disease prediction using feature selection and data mining technique, J. Comput. Sci. Cybern. 34 (1) (2018) 33–48.
[12] M. Tarawneh, O. Embarak, February. "Hybrid approach for heart disease prediction using data mining techniques, Acta Sci. Nutr. Health 3 (7) (2019) 147–151, 2019.
[13] R. Chitra, V. Seenivasagam, Heart disease prediction system using supervised learning classifier, Bonfring Int. J. Softw. Eng. Soft Comput. 3 (1) (2013), 01-07.
[14] C.B.C. Latha, S.C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Info. Med. Unlocked 16 (2019) 100203.
[15] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554.
[16] https://www.kaggle.com/johnsmith88/heart-disease-dataset [Accessed 02 June 2021].
[17] M.R. Rahman, T. Islam, T. Zaman, M. Shahjaman, M.R. Karim, F. Huq, J.M. Quinn, R.D. Holsinger, E. Gov, M.A. Moni, Identification of molecular signatures and pathways to identify novel therapeutic targets in alzheimer's disease: insights from a systems biomedicine perspective, Genomics 112 (2) (2019) 1290–1299.
[18] Four Techniques for Outlier Detection, https://www.kdnuggets.com/2018/12/four-techniques-outlier-detection.html.
[19] Md Satu, Syeda Atik, Mohammad Moni, A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus, 2019.
[20] S. Asaduzzaman, M.R. Ahmed, H. Rehana, S. Chakraborty, M.S. Islam, T. Bhuiyan, Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective, BMC Bioinf. 22 (1) (2021) 1–17.
[21] T. Akter, M.S. Satu, M.I. Khan, M.H. Ali, S. Uddin, P. Lio, J.M. Quinn, M.A. Moni, Machine learning-based models for early stage detection of autism spectrum disorders, IEEE Access 7 (2019) 166509–166527.
[22] S.M. Vieira, U. Kaymak, J.M.C. Sousa, Cohen's kappa coefficient as a performance measure for feature selection, in: International Conference on Fuzzy Systems, 2010 [online] Available at, http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000005584447. (Accessed 21 August 2019).
[23] Z. Lei, Y. Sun, Y.A. Nanehkaran, S. Yang, M.S. Islam, H. Lei, D. Zhang, A novel data-driven robust framework based on machine learning and knowledge graph for disease classification, Future Generat. Comput. Syst. 102 (2020) 534–548.
[24] X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu, Z. Huo, M. Yu, J. Peng, Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features, Sci. Rep. 9 (1) (2019) 1–13.
[25] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (1) (2019) 1–16.
[26] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (1967) 21–27.
[27] B.V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques, IEEE Comput. Soc. Tutorial (1991), 10012834200.
[28] K.H. Raviya, B. Gajjar, Performance Evaluation of different data mining classification algorithm using WEKA, Indian J. Research 2 (1) (2013) 19–21.
[29] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerg. Artif. Intel. Appl. Comput. Eng. 160 (2007) 3–24.
[30] R.L. De Mantaras, E. Armengol, Machine learning from examples: inductive and Lazy methods, Data Knowl. Eng. 25 (1–2) (1998) 99–123.
[31] S. Vijayarani, S. Sudha, Comparative analysis of classification function techniques for heart disease prediction, Int. J. Innov. Resear. Compute. Commun. Eng. 1 (3) (2013) 735–741.
[32] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[33] S.M.M. Hasan, M.A. Mamun, M.P. Uddin, M.A. Hossain, February. Comparative analysis of classification approaches for heart disease prediction, in: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, 2018, pp. 1–4.
[34] J.R. Quinlan, Induction of decision trees, Mach. Learn. (1986) 81–106.
[35] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Canc. Inf. 2 (2006), 117693510600200030.
[36] K. Li, G. Zhou, J. Zhai, F. Li, M. Shao, Improved PSO_AdaBoost ensemble algorithm for imbalanced data, Sensors 19 (6) (2019) 1476.
[37] C. Zhang, Y. Chen, Improved piecewise nonlinear combinatorial adaboost algorithm based on noise self-detection, Comput. Eng. 43 (2017) 163–168.

[38] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, vol. 398, John Wiley & Sons, 2013.
[39] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (1) (2019) 1–16.
[40] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inf. 35 (5–6) (2002) 352–359.
[41] K. Kwon, D. Kim, H. Park, A parallel MR imaging method using multilayer perceptron, Med. Phys. 44 (12) (2017) 6209–6224.
[42] S. Tajmiri, E. Azimi, M.R. Hosseini, Y. Azimi, Evolving multilayer perceptron, and factorial design for modelling and optimization of dye decomposition by bio-synthetized nano CdS-diatomite composite, Environ. Res. 182 (2020) 108997.
[43] Y. Azimi, Prediction of seismic wave intensity generated by bench blasting using intelligence committee machines, Int. J. Eng. 32 (4) (2019) 617–627.
[44] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Cham, 2018, pp. 655–670.
[45] V.A. Huynh-Thu, Y. Saeys, L. Wehenkel, P. Geurts, Statistical interpretation of machine learning-based feature importance scores for biomarker discovery, Bioinformatics 28 (13) (2012) 1766–1774.
[46] M.M. Ahamad, S. Aktar, M. Rashed-Al-Mahfuz, S. Uddin, P. Liò, H. Xu, M. A. Summers, J.M. Quinn, M.A. Moni, A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients, Expert Syst. Appl. 160 (2020) 113661.
[47] https://datascience.stackexchange.com/questions/44700/how-do-i-get-the-feature-importace-for-a-mlpclassifier [Accessed on 01 June 2021].
[48] https://stats.stackexchange.com/questions/363662/can-you-derive-variable-importance-from-a-nearest-neighbor-algorithm [Accessed on 01 June 2021].