

Two-level boosting classifiers ensemble based on feature selection for heart disease prediction

Kaushalya Dissanayake¹, Md Gapar Md Johar²

¹School of Graduate Studies, Management and Science University, Shah Alam, Malaysia

²Software Engineering and Digital Innovation Centre, Management and Science University, Shah Alam, Malaysia

Article Info

Article history:

Received Nov 7, 2022

Revised Jun 28, 2023

Accepted Jul 2, 2023

Keywords:

Boosting classifiers

Ensemble model

Feature selection

Heart disease

Machine learning

ABSTRACT

Heart disease is a prevalent global health concern, necessitating early detection to save lives. Machine learning has revolutionized medical research, prompting the investigation of boosting algorithms for heart disease prediction. This study employs three heart disease datasets from the University of California Irvine (UCI) repository: Cleveland, Statlog, and Long Beach, with 14 features each. Recursive feature elimination with a support vector machine (SVM) is utilized to identify significant features. Five boosting algorithms (gradient boosting algorithm (GB), adaptive boosting algorithms (AdaBoost), extreme gradient boosting algorithm (XGBoost), cat boost algorithm (CatBoost) and light gradient boosting algorithms (LightGBM)) are integrated into an ensemble model to achieve the best classification performance. The proposed model demonstrates superior accuracy, precision, recall, f-measure, and area under the curve (AUC) compared to individual boosting models, achieving 93.44%, 83.33%, and 79.75% accuracies for Cleveland, Statlog, and Long Beach datasets. This approach offers an accurate and efficient method for heart disease prediction, which is crucial for clinical decision-making and disease management.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kaushalya Dissanayake

School of Graduate Studies, Management and Science University

Shah Alam, Malaysia

Email: kaushi.dula@gmail.com

1. INTRODUCTION

Heart disease has become one of the most common diseases and leading causes of mortality worldwide. According to the World Health Organization (WHO), it is projected to have caused roughly 17.9 million deaths annually [1], accounting for nearly 15% of all natural deaths. As the American Heart Association (AHA) points out, several symptoms might indicate a heart problem, such as sleep problems or swollen legs, irregular heartbeat, and even unexpected growth of weight daily (approximately 1-2 kg) [2]. Unfortunately, many of these symptoms are related to various other disorders, which occur in the aging population, obstructing the acquisition of a precise diagnosis, which can end in death within a short period.

Some conditions that increase the risk of heart disease are lifestyle-related, such as smoking, obesity, cholesterol, and hypertension. However, other non-lifestyle risk factors, including age, history of the family, and high fibrinogen level, must be considered in addition to lifestyle risk factors. Furthermore, heart disease can be developed in the absence of any of the risk factors or apparent symptoms listed above. As a result, heart disease was among the most prevalent worldwide, adding a high impact to the mortality rate, making it one of the most challenging illnesses to treat.

One of the most extensively used and non-invasive diagnostic methods for cardiovascular disease is the electrocardiogram (ECG), which depicts the heart's electrical activity. Even though it can be conducted quickly

and easily, an ECG has several limitations as a predictive tool for predicting the development of future cardiac disease. Stress tests (nuclear cardiac stress test, exercise stress test), angiography, and cardiac magnetic resonance imaging (MRI) are other types of typically used tests by clinicians to diagnose cardiovascular disease [3]. Based on the different risk factors, the manual prediction to identify the likelihood of getting heart disease is complex.

Many researchers and practitioners are currently looking for ways to improve the accuracy of cardiovascular disease diagnosis by utilizing new technologies such as data mining, machine learning, and artificial intelligence (AI) [3]–[7]. Using these technologies, a few user inputs and attempts have been used to identify and uncover important patterns and information from clinical datasets. Still, novel, more powerful machine-learning techniques will assist us in identifying patterns and extracting usable information from clinical data. Clinical datasets are inherently unpredictable and irregular, making it difficult to use machine learning algorithms without an appropriate pre-processing activity, such as feature selection. Feature selection is the process of removing unnecessary and redundant features from data sets to efficiently reduce feature dimensions and enhance efficiency and classification accuracy [8]. Additionally, it serves as a denoising function, preventing the machine learning model from overfitting.

Feature selection is usually used to find a subset of features highly correlated with pattern recognition problems such as classification learning problems. Filter, wrapper, and embedded are the three types of feature selection algorithms. Filter techniques, which are independent of any classification algorithm, assess the performance of features using training data. Wrapper approaches are frequently used to evaluate the features using a specific learning algorithm. The embedded approaches conduct feature selection by relying on internal factors of the classification model that have been learned [9], [10]. Various learning algorithms have been presented to improve feature selection with the evolution of feature selection approaches. Different algorithms will choose distinct subsets of features throughout the feature selection phase, resulting in distinct outcomes. Based on software metrics, classification algorithms are often used to discover defective software modules. These models are trained to utilize data gathered before doing software testing and operations. A learning algorithm can rarely develop such a comprehensive model in real-world applications. Therefore, practitioners and researchers devote much time and effort to developing the most accurate and feasible model [11].

Researchers have looked at various classifiers for predicting heart disease, both individual and meta. Meta classifiers (e.g., hybrid or ensemble) should be accommodated when an individual classifier is unable to offer satisfactory performance [12]. To forecast the final classification results, a meta-classifier trains multiple distinct classifiers, which makes them more resilient and appropriate for ailment prediction than single classifiers. The combination of multiple classifiers might be heterogeneous (use various classifiers) or homogenous (use the same classifiers). Meta classifiers have been showing remarkable performance at classifying things in many other areas. However, there is still much research about combining different combinations of techniques and base classifiers [13]–[16].

Ensemble learning methods like boosting and bagging (or bootstrap aggregation) are commonly utilized in classification tasks that involve manipulating training data [17]. Bagging is a common ensemble classifier technique in which several predictors are made separately and combined using model averaging methods, like the majority vote or the weighted average, to make a single prediction. By contrast, boosting is an approach in which models are constructed sequentially rather than individually, and subsequent predictors are used to compensate for mistakes introduced by earlier predictors [18].

It is challenging to manually predict the possibility of developing heart disease based on the various risk factors. Still, novel, more powerful, accurate machine learning techniques using a limited number of features will need to assist in identifying patterns and extracting usable information from clinical data to fill the gap and drawbacks in existing individual algorithms. In addition, clinical decision support systems have been enhanced with intelligent technologies to aid clinicians in offering a second choice of the decision on heart disease diagnosis to reduce human mistakes.

This study aims to propose a novel ensemble approach combining different boosting algorithms to predict heart disease more accurately, using a limited number of features. The objective is to reduce the misclassification rates produced by individual boosting algorithms and enhance the overall prediction accuracy. Prior research has proposed various algorithms for heart disease prediction; however, these studies did not utilize both feature selection and boosting ensemble algorithms in a single model. Therefore, this study introduces a novel approach that leverages both techniques to enhance the accuracy of heart disease prediction. By incorporating feature selection into the ensemble approach, the model aims to identify the most informative features that contribute to heart disease prediction, while the ensemble algorithm works to combine the outputs of individual boosting algorithms to achieve improved prediction accuracy. The proposed approach is expected to provide a more reliable and efficient method for heart disease prediction, which could have significant implications for clinical decision-making and disease management.

The following sections of this research article outline the methodology, results, discussion, and conclusion of the findings. The method section provides a detailed description of the proposed ensemble approach,

including the selection of boosting algorithms, feature selection techniques, and the evaluation metrics used to measure the performance of the model. The results section presents the performance of the proposed model compared to other state-of-the-art algorithms for heart disease prediction. The evaluation of the model includes a comparison of the accuracy, sensitivity, specificity, and area under the curve (AUC) metrics. The discussion section interprets the results and explains the significance of the findings. This section also highlights the strengths and limitations of the proposed approach, as well as its implications for clinical decision-making and disease management. Finally, the conclusion summarizes the study's main findings and provides a brief overview of the critical contributions of the proposed ensemble approach for heart disease prediction.

2. METHOD

This section contains information on the methods and materials (for example, datasets) that were utilized in the experiment. In addition, it contains information regarding datasets and a conceptual pathway for detecting heart disease, which is represented in Figure 1, including three significant phases. The key phases of the framework are featuring selection, boosting algorithms training phase, including the two-level ensemble, and evaluation phase.

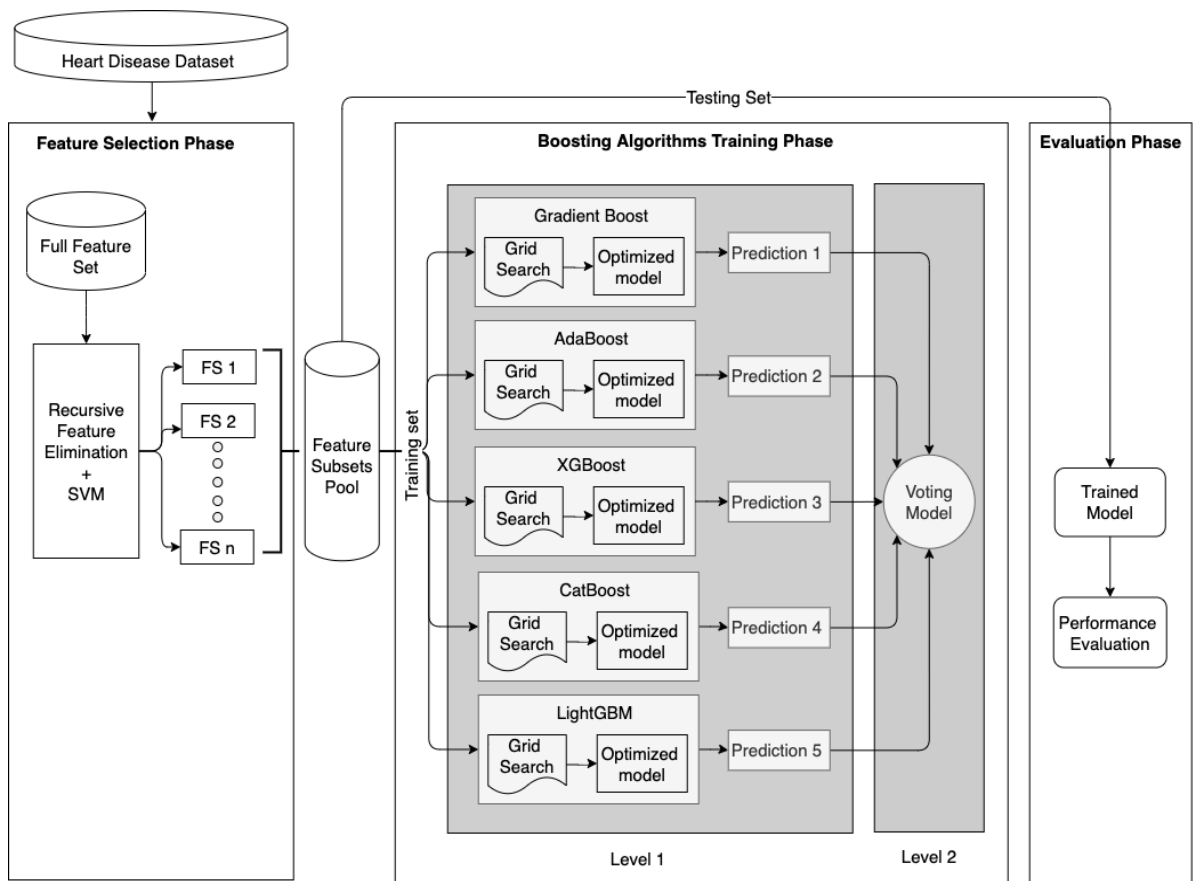


Figure 1. Proposed framework for heart disease prediction

2.1. Heart disease datasets

The three data sets, Cleveland, Statlog, and VA Long Beach, used for this experiment were collected from the University of California Irvine (UCI) machine learning repository, which is generally accessible [1]–[3]. These datasets have been chosen because other researchers regularly use them in this field. The properties and characteristics of each dataset are summarized in Table 1.

Table 1. Summary of the features and properties of each dataset

Dataset	No of features	No of instances	No of heart disease absence instances	No of heart disease presence instances
Cleveland	13	303	169	134
Statlog	13	270	150	120
VA Long Beach	13	200	51	149

2.1.1. Cleveland

The dataset comprised 303 heart disease patients, and a normal was obtained at the Cleveland Clinic Foundation. The 76 features are contained in the original dataset; however, this study only used 14 features, as done in previous studies. In addition, because the initial dataset contains five integer values ranging from 0 (absence of heart disease) to 4 (severe heart disease), the class label feature is normalized into two different classes of heart disease, namely 1 (presence) and 0 (absence).

2.1.2. Statlog

The 270 instances without any missing values related to the 14 features were included in the dataset in 120 patients who suffered from heart disease. The existence or absence of heart disease in the patient is indicated by the "target" feature. The target feature represents two distinct values, where 1 indicates the presence of cardiac disease while 0 indicates the absence of heart disease.

2.1.3. VA long beach

VA Long Beach is a processed dataset created by Robert Detrano that is accessible in the UCI repository. The 200 instances are included with 13 input features. Furthermore, 149 people have been recognized as having heart disease, while the remainder is in normal condition. In addition, 149 people were determined to be suffering from cardiovascular disease, while the remainder were judged to be in good health (cardiovascular disease is not found).

2.2. The heart disease prediction framework

Figure 1 illustrates a conceptual model for the proposed heart disease prediction model. The procedure is divided into three phases: feature selection, boosting algorithm training phase, and evaluation phase. The process for accurately defining a collection of features most helpful in predicting heart disease is covered in the first phase. A detailed discussion of the technique for feature selection may be found in section 2.3.

The second phase involves the development of a two-level ensemble technique. This stage is responsible for creating a two-level ensemble modelling phase using a combination of five homogeneous boosting ensembles, i.e., gradient boosting algorithm (GB), adaptive boosting algorithms (AdaBoost), extreme gradient boosting algorithm (XGBoost), cat boost algorithm (CatBoost) and light gradient boosting algorithms (LightGBM). The purpose is to combine boosting classifiers to reduce each classifier's overfitting and accuracy problem when making the final prediction.

The suggested two-level ensemble approach is finally evaluated in the third phase. K-fold cross-validation is used to construct the evaluation technique; the value of k is set to 10. The experiment commonly employs five performance measures: Accuracy, precision, recall, F-Measure, and receiver operating characteristic-area under the curve (ROC-AUC) score. Section 4 further describes the experimental results of the proposed model.

2.2.1. Feature selection phase

The performance of the classifier may be impaired as a result of the irrelevant and duplicated input features. So, selecting a subset of features from a large range of rigorous and accurate data might be a challenge. This study employed a wrapper-based technique called recursive feature elimination (RFE), where its search method is optimized using a classifier, support vector machine (SVM). In the RFE technique, the model is iteratively trained, and the weights of the algorithm are used as the criteria for each iteration to remove the least significant feature. Smaller and smaller groupings of features are evaluated sequentially using the RFE approach to choose the best features. The feature selection process is completed by using the steps outlined in Algorithm 1.

A lot of experiments were done by changing the number of features in subsets that are used in each one. A support vector machine classifier with the highest performance accuracy is used to find the optimum feature set. The chosen feature subsets feed into the feature subsets pool for the splitting criterion training and testing. Subsampling is used to evaluate the performance of each feature subset, with the portion of data obtained from the original dataset called training set. The remaining data instance is utilized for testing purposes.

Algorithm 1. Recursive feature elimination algorithm

Input: original dataset D with all the features $[f_i, i = 1, 2, 3, \dots, 13]$ is selected.
 each instance $X \in D$ is assigned one of two classes

```
for n (n_features_to_select) = 3 To 12
  1: start the feature selection process with n using a support vector machine classifier.
  2: for each feature, the coef_ is used to determine feature importance, and then the estimator is trained on the feature set.
  3: the less significant features in the original features set are trimmed.
  4: repeat the process until the reduced feature set has the desired number of features.
End for
```

Output: selected feature subsets $S_3, S_4, S_5, \dots, S_{12}$ i.e.; $S_3 = \{f_i, f_j, f_k\}$, $S_4 = \{f_i, f_j, f_k, f_m\} \dots$

2.2.2. Boosting algorithms modelling phase

The suggested two-level ensemble is made up of five different boosting algorithms that are stacked together in parallel: a GB, an AdaBoost, XGBoost, CatBoost, and LightGBM. In contrast to traditional classifier ensembles, which usually use weak individual learners, we explore powerful boosting techniques as the base classifiers in our study. Grid search is used to find the optimum learning hyperparameters for each base classifier by testing all feasible values. The steps used to create the modelling phase are represented in Algorithm 2.

Algorithm 2. Two – tire boosting algorithms ensemble

Input: selected feature subsets $S_3, S_4, S_5, \dots, S_{12}$
 training instances $X_{train} \in D$ is assigned one of two classes
 boosting algorithms B_1, B_2, B_3, B_4, B_5

Function

Boosting algorithm $[B_1, B_2, B_3, B_4, B_5]$
 execute hyperparameter tuning

Return optimized boosting algorithms $[OB_1, OB_2, OB_3, OB_4, OB_5]$

Function

Optimized boosting algorithms $[OB_1, OB_2, OB_3, OB_4, OB_5]$
 C_i [where $i = 1, 2, 3, 4, 5$] = Train algorithms with instances X_{train}

Level 1

testing instances $X_{test} \in D$

P_i [where $i = 1, 2, 3, 4, 5$] = Take predictions on each boosting algorithm

Level 2

Final Prediction = Voting Model (P_1, P_2, P_3, P_4, P_5)

Return Final Prediction

Output: Final Prediction

Output: selected feature subsets $S_3, S_4, S_5, \dots, S_{12}$ i.e.; $S_3 = \{f_i, f_j, f_k\}$, $S_4 = \{f_i, f_j, f_k, f_m\} \dots$

The following section briefly explains the five boosting algorithms used in this study. The principle of correcting prediction errors is an important feature of boosting ensembles. For each model in the ensemble, a correction is made by fitting and adding it sequentially so that the previous model makes an error and will be fixed in the next model sequentially. Decision trees that make one or a few decisions, called "weak learners," are commonly used in this type of analysis [19]. Then, the prediction of the weak learners is aggregated by simple voting or averaging, with their contributions weighted proportionally to their capability or performance. The goal is to make a "strong learner" out of many "weak learners" that were made for a specific purpose.

a. Gradient boosting algorithm

Gradient boosting is an ensemble-based algorithm that may be used to solve problems involving regression or predictive classification modelling. Decision tree models are used sequentially to build ensembles. In this method, trees are added to the ensemble one at a time and fitted to rectify the prediction mistakes caused by prior models. Then, a gradient descent optimization algorithm or arbitrary differentiable loss function is used to fit models. In this way, the "gradient boosting" technique is called that because the loss gradient goes down as the model is built [20]. As a result of the grid search, the learning parameters used as: subsample=0.5, n_estimators=50, max_depth=1, random_state=0, loss='exponential', criterion='friedman_mse' for the algorithm.

b. Adaptive boosting algorithms

In 1996, Robert Schapire and Yoav Freund introduced AdaBoost, an ensemble-boosting classifier. AdaBoost randomly picks a sample of the training data using decision trees with a single level, i.e., trees with only a single split. Additionally, these trees are referred to as decision stumps. Initially, this technique

constructs a model by giving equal weights to all instances in the dataset. The accuracy of the AdaBoost machine learning model is improved by iterative training, which selects the best new set of trained data for each iteration depending on the previous stage's accuracy. The algorithm assigns a higher weight to incorrectly categorized data, raising the possibility that these observations will be classified accurately in the subsequent iteration [21]. Aside from that, it gives different weights to the learned classifier in every iteration based on how well the trained classifier performed in the previous iteration. The highest weight will be given to the classifier with the highest accuracy. This process is continued until all of the data from the training set fits together properly or until the maximum estimator's number has been obtained. For this analysis, set the parameter for the adaptive boost algorithm as follows in response to the grid search: `n_estimators=50`, `learning_rate=0.2`, `random_state=1`, `algorithm='SAMME'`.

c. Extreme gradient boosting machine

Extensive research has gone into the development of XGBoost, a distributed gradient-boosting library that is extremely portable, efficient, and adaptive while retaining maximized accuracies. It is an approach to implementing machine learning algorithms based on the gradient boosting technique. Fast and accurate parallel boosting of trees is provided by XGBoost, also known as glioblastoma (GBM or GDT). The algorithm aims to minimize a cost function by iteratively searching for fine-tuned learning parameters [22]. XGBoost outperforms the gradient boosting algorithm regarding computational efficiency (e.g., processor cache and memory use). Furthermore, it uses a more regularized model, which reduces the model's complexity while enhancing forecast accuracy. As a result of the grid search, we determined the following learning parameters: `learning_rate=0.1`, `max_features=1`, `subsample=0.5`, `max_depth=1` for XGBoost algorithm.

d. Cat boost algorithm

Gradient-boosted decision trees are the foundation of the CatBoost algorithm. During training, a series of decision trees are constructed sequentially. Each new tree is constructed with reduced loss compared to the previous trees. The initial parameter settings determine the number of trees. The overfitting detector can help you avoid overfitting. When it is activated, the construction of trees is paused [23]. Working with non-numeric elements is supported by CatBoost, saving time while improving training outcomes. As a result of the grid search, we determined the following learning parameters: `verbose=0`, `n_estimators=100` for CatBoost algorithm.

e. Light gradient boosted machine algorithm

It is known as LightGBM or light gradient boosted machine, for short. It is an open-source package that implements the gradient boosting technique quickly and efficiently. LightGBM divides the tree leaf-by-leaf, in contrast to other boosting algorithms, which develop the tree level-by-level. It selects to grow on the leaf that has the greatest delta loss. The leaf-wise approach has a smaller loss than the level-wise technique since the leaf is fixed. However, leaf-wise tree development may increase the model's complexity in the limited number of instance datasets and lead to overfitting [24]. LightGBM enhances the gradient boosting technique by incorporating an auto feature selection mechanism and concentrating on cases with bigger gradients. This can speed up training and improve the accuracy of predictions. We configured the learning settings based on the grid search for LightGBM algorithm: `learning_rate=0.1`, `max_depth=2`, `n_estimators=50`.

This study focuses on the two-level ensemble of homogeneous boosting algorithms, which uses a stacking approach to combine them. There are a variety of ways to combine the base classifiers in practice. However, we are attempting to demonstrate the efficiency of such an architecture for heart disease prediction; we will refer to those five boosting methods as the base classifiers in our analysis. To forecast the result, an initially existing training set is used to train the base classifiers. The results of the base classifiers are then used to train a meta-classifier using the generalized voting model. Finally, the pseudocode is used to implement the two-level ensemble approach depicted in Algorithm 2.

f. Evaluation phase

Some of the assessment measures employed by the researcher to evaluate the efficiency of the suggested boosting ensemble classification model are accuracy, precision, F1-score, AUC, and recall [25]–[28]. All the performance matrices are calculated using the confusion matrix components. Researchers may use the confusion matrix to determine the rate of a classification result's performance based on four primary factors: true negative (TN), false negative (FN), true positive (TP), and false positive (FP). The accuracy of the model demonstrated that it can correctly identify those at high risk of developing heart disease.

Accuracy may be measured using (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Precision in (2), defined as the proportion of individuals diagnosed with heart disease who were actually at risk, is computed as the total of accurately recognized occurrences. The precision is provided by:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

recall, as in (3), is a metric that reflects whether the algorithm correctly identified the proportion of individuals with heart disease who actually had it.

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

The F-score, as in (4), is used to determine the accuracy of the test, and the formula for calculating it is.

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The degree or amount of separability is represented by the AUC. It indicates how well the model can discriminate among classes. The greater the AUC, the more accurate the model is in discriminating between patients with and without ailment.

3. RESULTS AND DISCUSSION

This comparative analysis approach utilized Python as the programming language to generate the analytical model using PyCharm (version 2021.3.3), which was the integrated development environment. This facilitates dataset exploration and enables accurate pattern recognition. The findings of all experiments are discussed in this section. In this paper, we first report the findings of feature selection, followed by a classification result for the identification of heart disease.

3.1. Results of feature selection

Recursive feature elimination with a support vector machine algorithm is used to select optimal feature subsets by running a different number of iterations. According to the predictive accuracy of the proposed model, the best feature subset was selected. Table 2 provides an input feature set derived using the recursive feature reduction method for each dataset. Seven features were selected from Cleveland and Statlog datasets, respectively, and nine features were selected from the VA Long Beach data set using the feature selection algorithm. The sex, chest pain, resting electrocardiographic results, and thallium stress test results were selected from all three datasets as essential features for heart disease prediction.

Table 2. The selected features were obtained using the recursive feature elimination feature selection technique for each dataset

Dataset	No of features	Feature name
Cleveland	7	Chest pain type, Sex, Exercise-induced angina, resting electrocardiographic results, Number of major vessels (0-3) coloured by fluoroscopy, The slope of the peak exercise ST segment, Thallium stress test result
Statlog	7	Resting electrocardiographic results, Chest pain type, Sex, Fasting blood sugar level, Number of major vessels (0-3) colored by fluoroscopy, ST depression induced by exercise relative to rest, Thallium stress test result
VA Long Beach	9	Age, Chest pain type, Sex, resting electrocardiographic results, Fasting blood sugar, ST depression induced by exercise relative to rest, The slope of the peak exercise st segment, Thallium stress test result

The results in Figures 2-4 highlighted the prediction accuracy of each algorithm was improved after applying the feature selection to the process. The best accuracies obtained by the seven features selected from the Cleveland dataset: Chest pain type, Sex, Number of major vessels (0-3) colored by fluoroscopy, Exercise-induced angina, Resting electrocardiographic results, The slope of the peak exercise ST segment, Thallium stress test result on each algorithm. A set of 7 features was obtained from the Cleveland dataset generating maximum accuracy of 93.44% with the proposed approach.

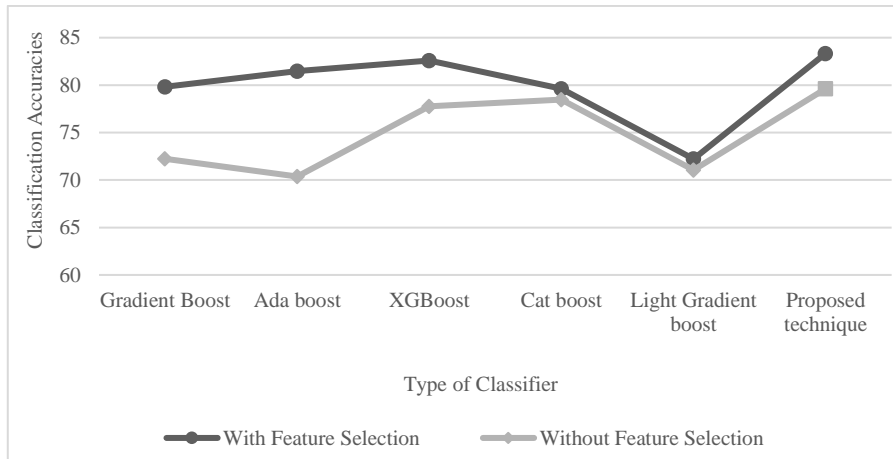


Figure 2. Classification accuracies with feature selection and without feature selection on the Cleveland dataset

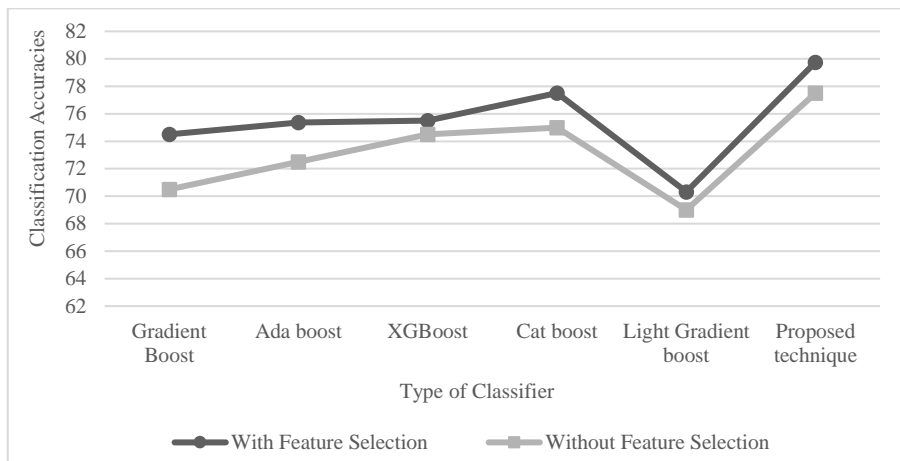


Figure 3. Classification accuracies with feature selection and without feature selection on the Statlog dataset

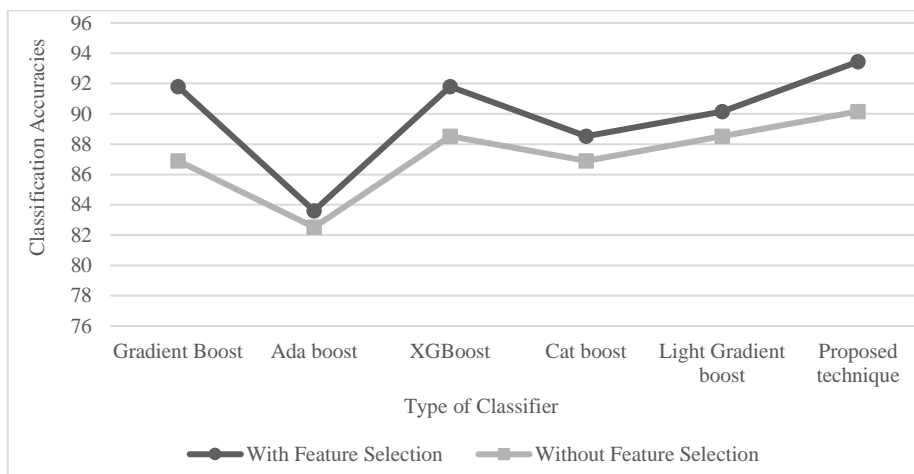


Figure 4. Classification accuracies with feature selection and without feature selection on the VA Long beach dataset

3.2. Result of heart disease classification

The performance of a two-level boosting ensemble is compared to that of other classifiers i.e., GB, AdaBoost, XGBoost, CatBoost, and LightGBM in this section of the paper. Tables 3-5 illustrate the comparison of the experiment results of the proposed model with various other boosting algorithms using the three datasets. The results presented in the table are averaged over ten-fold cross-validation with ten runs. The best predictive accuracy obtained for each data set is highlighted with a boldfaced print.

Table 3. Result of boosting techniques and proposed techniques for selected features by RFE on the Cleveland dataset

Techniques	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
Gradient boost	91.80	95.83	85.19	90.20	91.12
Ada boost	83.61	81.48	81.48	81.48	83.39
XGBoost	91.80	95.83	85.19	90.20	91.12
Cat boost	88.52	91.67	81.48	86.27	87.80
Light gradient boost	90.16	88.89	88.89	88.89	90.03
Proposed technique	93.44	96.00	88.89	92.31	92.97

The proposed technique achieved an accuracy of 93.44%, which was higher than all other algorithms, including GB and XGBoost, which achieved accuracies of 91.80%. Regarding precision and recall, the proposed approach demonstrated superior performance compared to all other algorithms. The precision and recall values for the proposed technique were 96.00% and 88.89%, respectively, which were higher than all other algorithms evaluated in this study. Additionally, the proposed technique achieved a higher F-measure and AUC than all other algorithms on the Cleveland dataset, which suggests that it provides a better balance between precision and recall and can effectively discriminate between positive and negative cases.

Table 4. Result of boosting techniques and proposed techniques for selected features by RFE on the Statlog dataset

Techniques	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
Gradient boost	79.82	80.65	83.33	81.97	79.17
Ada boost	81.48	83.33	83.33	83.33	81.25
XGBoost	82.58	83.33	83.33	83.33	81.25
Cat boost	79.63	78.79	86.67	82.54	78.75
Light gradient boost	72.22	72.73	80.00	76.19	71.25
Proposed technique	83.33	81.82	90.00	85.71	82.50

The accuracy on the Statlog dataset of the suggested method was 83.33%, outperforming all other algorithms, including AdaBoost and XGBoost, which had accuracy results of 81.48% and 82.58%, respectively. In addition, the proposed technique has precision and recall values of 81.82% and 90.00%, respectively, higher than all other algorithms assessed in this study. The results also showed that AdaBoost, XGBoost, and CatBoost achieved approximately similar accuracy, precision, recall, F-measure, and AUC performance. At the same time, light gradient boost had the lowest performance among all evaluated algorithms.

Table 5. Result of boosting techniques and proposed techniques for selected features by RFE on the VA Long Beach dataset

Techniques	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	AUC (%)
Gradient boost	74.50	75.00	90.00	81.82	50.00
Ada boost	75.36	77.78	93.33	84.85	56.67
XGBoost	75.50	79.41	90.00	84.38	60.00
Cat boost	77.50	78.38	96.67	86.87	58.33
Light gradient boost	70.32	78.12	83.33	80.65	56.67
Proposed technique	79.75	80.56	96.67	87.88	63.33

Based on the results obtained for the VA Long Beach dataset, the proposed technique outperformed all the other boosting techniques. The proposed technique achieved an accuracy of 79.75%, precision of 80.56%, recall of 96.67%, F-measure of 87.88%, and AUC of 63.33%. Among the other boosting techniques, the Cat boost had the highest recall of 96.67% but the lowest AUC of 58.33%. Ada Boost and XGBoost had relatively high precision and recall scores, but their AUC scores were only 56.67% and 60.00%, respectively.

The Cleveland dataset, on the other hand, produces the highest classification performance, with the proposed model achieving 93.44% accuracy, 96.00 % precision, 88.89 % recall, and 92.31 % f-measure. And both Statlog and VA Long Beach data sets performed well, with the proposed algorithm achieving classification accuracy of 83.33% and 79.75%, respectively. As a result, it was evident that the suggested model significantly outperforms previous competition-boosting algorithms. The accuracy of comparing the Cleveland and Statlog datasets against the VA Long Beach data set is relatively high. This results from the nature of the class distribution of three datasets; the VA Long Beach data set is an unbalanced dataset, while the Cleveland dataset and Statlog dataset are fair. In contrast, the Ada boost algorithm takes 83.61 % lowest accuracy in the Cleveland dataset, light gradient boost algorithm takes 72.22 % and 70.32 % accuracies on Statlog and VA Long Beach data sets, respectively.

This study aimed to investigate the effect of feature selection on the classification performance of boosting algorithms in predicting heart disease. The classification accuracies of five boosting algorithms, namely Gradient Boost, Ada Boost, XGBoost, Cat Boost, light gradient boost, and the proposed algorithm, were evaluated on three different datasets: Cleveland, Statlog, and VA Long Beach, before and after applying recursive feature elimination (RFE) technique for feature selection depicted on Figures 2-4. The results showed that after reducing the number of features in the datasets using RFE, the classification accuracies of some algorithms were significantly improved. In contrast, some algorithms remained the same or slightly improved. The proposed boosting ensemble approach improved performance by combining the feature subset selected using RFE on all three datasets. These findings highlight the importance of feature selection in enhancing the classification accuracy of boosting algorithms, especially when dealing with high-dimensional datasets. The suggested algorithm emerged as the best solution for the three datasets, according to the findings noted in the analysis.

4. CONCLUSION

In this study, a two-level boosting classifiers ensemble approach was proposed to efficiently predict heart disease using a limited number of features selected using a feature selection algorithm. RFE classified under the wrapper-based feature selection algorithm, was used to determine the most important features for each dataset (Cleveland, Statlog, and VA Long Beach). The selected features were then combined using the proposed boosting ensemble approach. The results showed that the proposed technique achieved the highest accuracy, precision, recall, F-measure, and AUC for all three datasets, outperforming the individual boosting algorithms and other ensemble techniques. Furthermore, the performance of some algorithms was significantly improved after reducing the number of features in the dataset, indicating the importance of feature selection in machine learning. Overall, the proposed boosting ensemble approach with RFE feature selection can be a promising method for accurately predicting heart disease using limited features.




REFERENCES

- [1] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.
- [2] M. Kamboj, "Heart disease prediction with machine learning approaches," *International Journal of Science and Research (IJSR)*, vol. 9, no. 7, pp. 1454–1458, 2020, doi: 10.21275/SR20724113128.
- [3] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.
- [4] K. Dissanayake, M. G. M. Johar, and N. H. Ubeysekera, "Data mining techniques in disease classification: descriptive bibliometric analysis and visualization of global publications," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 289–301, Jan. 2023, doi: 10.12785/ijcds/130124.
- [5] N. K. Sharma, B. R. Reddy, M. Monika Chowdary, Y. R. D. P. Swetha, B. R. Varma, and C. Bharat, "A novel heart disease prediction approach using the hybrid combination of GA, PSO, and CNN," in *Lecture Notes in Networks and Systems*, vol. 425, 2022, pp. 403–414, doi: 10.1007/978-981-19-0707-4_37.
- [6] N. Absar *et al.*, "The efficacy of machine-learning-supported smart system for heart disease prediction," *Healthcare (Switzerland)*, vol. 10, no. 6, p. 1137, Jun. 2022, doi: 10.3390/healthcare10061137.
- [7] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023, doi: 10.1016/j.health.2022.100130.
- [8] H. Takci, "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 1, pp. 1–10, 2018, doi: 10.3906/elk-1611-235.
- [9] X. Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, pp. 1–10, Feb. 2021, doi: 10.1155/2021/6663455.
- [10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.
- [11] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.




- [12] A. Abdul, R. M. Isiaka, R. S. Babatunde, and J. F. Ajao, "An improved coronary heart disease predictive system using random forest," *Asian Journal of Research in Computer Science*, pp. 17–27, Aug. 2021, doi: 10.9734/ajrcos/2021/v11i130253.
- [13] H. Kathleen, H. Julia, and J. George, "Diagnosing coronary heart disease using ensemble machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10, 2016, doi: 10.14569/ijacsa.2016.071004.
- [14] J. Emakhu, S. Shrestha, and S. Arslanturk, "Prediction system for heart disease based on ensemble classifiers," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2020.
- [15] R. Gujare, D. Viji, and S. Bhatt, "Enhanced heart disease prediction using ensemble learning methods," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp. 76–85, 2020.
- [16] R. Li *et al.*, "Multilevel risk prediction of cardiovascular disease based on Adaboost+RF ensemble learning," *IOP Conference Series: Materials Science and Engineering*, vol. 533, no. 1, 2019, doi: 10.1088/1757-899X/533/1/012050.
- [17] B. U. Sri, "Effective heart disease prediction model through voting technique," *International Journal of Engineering Technology and Management Sciences*, vol. 4, no. 5, pp. 10–13, 2020, doi: 10.46647/ijetms.2020.v04i05.003.
- [18] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.
- [19] N. S. Rajliwall, R. Davey, and G. Chetty, "Cardiovascular risk prediction based on XGBoost," in *Proceedings - 2018 5th Asia-Pacific World Congress on Computer Science and Engineering, APWC on CSE 2018*, Dec. 2018, pp. 246–252, doi: 10.1109/APWCConCSE.2018.00047.
- [20] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
- [21] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, p. 184, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp184-190.
- [22] M. Karthikeyan, C. R. Myakala, and S. C. Chappidi, "Heart attack prediction using XGBoost," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp. 2392–2399, 2020.
- [23] X. Zhang and G. X. Wu, "Text classification method of Dongba classics based on catboost algorithm," in *The 8th International Symposium on Test Automation & Instrumentation (ISTA 2020)*, 2021, pp. 133–139, doi: 10.1049/icp.2021.1336.
- [24] D. D. Rufo, T. G. Debelee, A. Ibenhal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.
- [25] I. I. M. Manhrawy, M. Qaraad, and P. El-Kafrawy, "Hybrid feature selection model based on relief-based algorithms and regularizer algorithms for cancer classification," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 17, Sep. 2021, doi: 10.1002/cpe.6200.
- [26] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "An efficient prediction system for Coronary heart disease risk using selected principal components and hyperparameter optimization," *Applied Sciences (Switzerland)*, vol. 13, no. 1, p. 118, Dec. 2023, doi: 10.3390/app13010118.
- [27] S. Goyal, "Predicting the heart disease using machine learning techniques," in *Lecture Notes in Networks and Systems*, vol. 517, 2023, pp. 191–199, doi: 10.1007/978-981-19-5224-1_21.
- [28] S. K. Shaw and S. Patidar, "Heart disease diagnosis using machine learning classification techniques," in *Lecture Notes in Networks and Systems*, vol. 383, 2023, pp. 445–460, doi: 10.1007/978-981-19-4960-9_35.

BIOGRAPHIES OF AUTHORS



Kaushalya Dissanayake    is currently Ph.D. research scholar in Computer Science, School of Graduate Studies, Management and Science University, Malaysia. She received M.Sc. in Information Technology (Specialized Cyber Security) from Sri Lanka Institute of Information Technology, Sri Lanka. She completed her B.Sc. from the University of Ruhuna, Sri Lanka. Her research interests are in the field of machine learning, data mining, artificial intelligence, and deep learning. She can be contacted at email: kaushi.dula@gmail.com.



Md Gapar Md Johar    is Senior Vice President System, Technology and Innovation of Management and Science University, Malaysia. He is a professor in Software Engineering. He holds Ph.D. in Computer Science, M.Sc. in Data Engineering and B.Sc. (Hons) in Computer Science and Certified E-Commerce Consultant. He has more than 40 years of working and teaching experience in various organizations include Ministry of Finance, Ministry of Public Enterprise, Public Service Department, Glaxo Malaysia Sdn Bhd and Cosmopoint Institute of Technology. His research interests include learning content management system, knowledge management system, blended assessment system, data mining, RFID, e-commerce, image processing, character recognition, data analytics, artificial intelligent, and healthcare management system. He can be contacted at email: mdgapar@msu.edu.my.