Intelligent Systems with Applications 16 (2022) 200121

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique

Prasannavenkatesan Theerthagiri

Department of Computer Science and Engineering, GITAM School of Technology, GITAM University, Bengaluru, India

ARTICLE INFO

Keywords:
Heart diseases
Feature ranking
Machine learning
Recursive feature elimination
Gradient boosting

ABSTRACT

Background: Heart disease is one of the most frequent chronic ailments people suffer. Early identification can lower death rates by avoiding or lowering cardiovascular disease (CVD) severity. For detecting risk indicators, machine learning algorithms are a potential way.
Methods: To acquire accurate cardiac disease prediction, this work introduces a recursive feature elimination-based gradient boosting (RFE-GB) approach. The outcomes were evaluated using the patients' health records, including crucial CVD characteristics. The prediction model was built using many additional machine learning approaches, and the results were compared to the suggested model.
Results: The combined recursive feature removal and gradient boosting approach deliver the maximum accuracy, according to the findings of this suggested model (88.8 %). Furthermore, the presented RFE-GB method was determined to be superior and had acquired a significant gain over previous strategies, with an area under the curve of 0.84.
Conclusion: The developed RFE-GB method will thus be a useful model for predicting and treating CVD.

1. Introduction

Health monitoring has grown increasingly vital as the impacts of social ageing intensify. The disease prediction system can assist medical professionals in forecasting heat alignment based on patient clinical data. As a result, by developing a prediction framework based on advanced algorithms and analysing many health-related concerns, it will be able to anticipate the patients' likelihood of being diagnosed with any health conditions (Shi et al., 2019).

Heart monitoring is one of the most important aspects of healthcare monitoring. A heart disease prediction system can help medical practitioners make informed judgments regarding the health of their patients' hearts. An improper place of origin or uneven conduction of the electric signal can produce aberrant heart rhythms. These disorders are known medically as arrhythmias. Some arrhythmias can have serious effects, including death. Medical experts may overlook important details while assessing a patient's heart alignment; as a result, a heart alignment prediction approach based on machine learning algorithms might help in such circumstances (Kakulapati et al., 2017; Prasannavenkatesan, 2021).

Machine learning approaches that have been widely used in medicine include illness prediction, disease classification, and medical image recognition algorithms (Chang et al., 2019; Theerthagiri, 2021; Theerthagiri & Ruby, 2022). This study introduces and improves gradient boosting (GB), a modern and efficient approach. The gradient boosting learning approach is based on the gradient boosting decision tree. In terms of generality, GB does well as an ensemble classifier. Additionally, GB includes a regularisation term to control the model's complexity and avoid overfitting. GB has beaten the competition in a number of machine learning areas (Shi et al., 2019; Wang et al., 2021). As a result, the ability of GB to classify single heart illnesses is being studied.

This study aims to develop a therapeutically relevant heart illness classification system. This paper provides a hierarchical approach based on the weighted gradient boosting algorithm to achieve the aim. The most prevalent strategy for getting usable cardiovascular patient datasets is preprocessing. After that, a variety of attributes are extracted. Then, to pick features, recursive feature elimination is utilised. The feature vectors are then input into a hierarchical classifier, which generates predicted labels. Because it has a significant quantity of information assets, the medical area is an application field for information mining. They understand the importance of including selection and feature reduction. Feature determination is the process of identifying certain significant aspects in order to learn objective reasoning (Aggrawal and Pal, 2021; Bakhsh, 2021; Prasannavenkatesan et al.,

E-mail address: vprasann@gitam.edu.

https://doi.org/10.1016/j.iswa.2022.200121
Received 1 March 2022; Received in revised form 17 August 2022; Accepted 1 September 2022
Available online 6 September 2022

2021).

To identify the best feature from all other features, the proposed method employs a hierarchical classifier and recursive feature elimination. The hypothesis investigated in this study is that combining the RFE technique with gradient boosting learning algorithms will increase accuracy and receiver operating characteristic (ROC) values. The following are the work's novel contributions:

- The proposed approach is unique because it classifies cardiovascular illnesses using the gradient boosting technique.
- A stochastic gradient boosting technique is utilised to choose features for hyperparameter optimization.
- The features are grouped together to minimise the mean square error. To demonstrate the utility of the new approach, numerous experimental scenarios are explored, and the results are compared to several previous research and standard ML algorithms.
- The gradient boosting algorithm is being used to predict and classify heart disease using the weights of each feature in the dataset.

The rest of this paper is laid out as follows. The next section contains literature details and a critique of previous initiatives. Section 3 covers the proposed approach used in this work, including pre-processing, feature selection, and the hierarchical classification approach. Performance measures are used to evaluate the proposed procedures in Section 4. This section describes the findings and makes some comparisons to previous studies. Section 5 closes by summarising and drawing conclusions from all of the works.

2. Background

Machine learning and pattern recognition are particularly useful for predicting individual outcomes from large amounts of data, especially when combined with EHRs. In a broad UK population, this study (Shi et al., 2019) employed machine learning to improve the prediction accuracy of conventional CVD risk indicators. Machine learning approaches were compared to a gold standard attained by pooled cohort risk for ten-year Cardiovascular event prediction using longitudinal EHR data (Zhao et al., 2019).

This study (Shi et al., 2019) created a categorization technique with three fundamental phases. During the preprocessing step, the ECG signal is filtered using the wavelet method. Then all heartbeats are located using fiducial points. Feature engineering is a method for extracting various characteristics from time and time-frequency domains. The study then utilised recursive feature elimination to choose features. A hierarchical classifier based on the XGBoost classifier and threshold is utilised in the classification stage to obtain the final results.

To classify hypertension patients, the authors created a prediction technique based on physical examination indicators (Chen et al., 2020). The first step extracts the relevant elements from the patients' many clinical assessment signals. In the second step, the crucial features gathered in the first stage are used to anticipate patients' outcomes. The authors then proposed a model that included recursive feature elimination, cross-validation, and a prediction model. By utilising their best features subset, extreme gradient boosting (XGBoost) is thought to correctly estimate patient outcomes.

For efficient classification, the work (Chen et al., 2020) provided a wrapper gene selection technique with a recursive feature reduction approach. The ensemble technique was employed for numerous gene selection procedures, and the top-ranking genes in each methodology were picked as the final gene subset. Many gene selection strategies were integrated with this work, and the optimum gene subset was found by prioritising and ranking the most important genes identified by the gene selection strategy. As a result, the researchers concluded that choosing a more discriminative and compact gene subset produced the greatest outcomes.

The researchers employed machine learning algorithms to predict the stage of heart illness in a patient (Kakulapati et al., 2017). They used the stochastic gradient boosting approach and Recursive Feature Elimination to choose the best features (RFE). A calculating model was created utilising an ensemble of weak prediction models, several of which employed decision trees. It offers a step-by-step method for boosting, simplifying, and optimising a subjectively variable failure issue (Padmanabhan et al., 2019, Akyol & Atila, 2019).

The gradient boosting strategy applies new models progressively during learning to better estimate response variables. The main idea behind this method is to create new base learners that have the best correlation with the ensemble's negative gradient error function (Friedman, 2001). Random Forest, an ensemble learning-based random decision tree, was devised by Breiman (2011). The primary difference between RF and decision trees is that RF seeks for the best feature among the random subsets of characteristics when breaking a node, whereas decision trees look for the most important feature. As a result, there is great variability, resulting in a superior model. Each pair of classified attributes was independently categorised using the Bayes' theorem-based NB classifier. It makes use of probability theory to find the most likely groupings. This technique is useful when the input has a high dimensionality (Theerthagiri et al., 2021).

The authors proposed a strategy to predict heart disease classification based on feature selection for the Cleveland and Statlog project heart datasets (Satish Chandra Reddy et al., 2019). Thus, according to the authors, the accuracy of the random forest approach for feature selection (8 and 6 features) based on classification models is good. In this study, sensitivity and specificity were also linked to higher scores.

The work (Zhang et al., 2019) employed a gradient boosting decision tree to predict blood pressure rates using human physiologic data from the EIMO device. They used the cross-validation method to choose optimum parameters and avoid overfitting. Also, when comparing the aspects of age, body fat, ratio, and height, it has been claimed that this approach has a greater accuracy rate and lower error rates than other algorithms. The paper (Patro et al., 2020) provided a system for predicting heart disease risk variables utilising different classifier algorithms. The support vector machine performs better in terms of precision, prediction accuracy, sensitivity, and F1 score.

Elavarasan et al. (2020) introduced a novel hybrid feature extraction method with an aggregation of random forest recursive feature elimination and a correlation-based filter. It identifies an optimal subclass of features from the data collected on soil, climate and ground water characteristics in order to construct a crop-yield forecasting machine learning model that improves performance and accuracy. This technique uses the filter method to reduce the size of the feature set by removing the noisy features. Based on the correlation heuristic evaluation function, the features are arranged and characterized based on their significance. The wrapper method was used to identify the ideal feature subgroup (Elavarasan et al., 2020).

Mahendran and Vincent P M (2022) proposed a deep learning-based classification model to classify Alzheimer's disease patients. A feature selection model was embedded that takes the DNA methylation data set of the patients and pre-processes the data to improve the ability to classify. The p-value, which is greater than 0.01, is eliminated as poor-quality samples. In the downstream analysis using Beta values and M-values, unmethylated and methylated levels of interrogated CpG sites are identified. To eliminate the noise, a penalty will be added to the various parameters of the model using LASSO regression and SVM. This method was evaluated using two cross-validation approaches, 5-fold and leave-one-out cross-validation (Mahendran and Vincent P M, 2022).

Mosavi et al. (2021) have developed a number of machine learning models for mapping groundwater salinity based on dichotomous predictions. The Karaj watershed is the research case, and the Iranian Water Resources Management Company provided the data for 114 groundwater monitoring wells between 2003 and 2017. Using feature selection methods and k-fold cross-validation methodology, the salinity of the groundwater is predicted. The variables that contributed the most to

groundwater salinity mapping were those related to soil type, groundwater extraction, precipitation, land use, and elevation. The areas with the highest groundwater salinity are indicated in the results (Mosavi et al., 2021).

The locations of 339 groundwater resources and the geographical groundwater potential conditioning variables were utilised to construct ensemble models. Iran's Dezekord-Kamfiruz watershed serves as the research area. The Iranian Water Resources Management Company is supported in determining the location of the groundwater resources. The critical features were located using the recursive feature elimination (RFE) approach. The groundwater potential modelling uses 12 variables out of 15 total. The modelling outcomes showed that the Bagging models outperformed the Boosting models in terms of performance. Additionally, the drainage density, elevation, valley depth, and distance from the stream were the modelling method' most significant predictive factors for the topographic position index (Mosavi et al., 2021; Hosseini et al., 2020).

Similar to this, Choubin et al. (2020) used well-known machine learning techniques to study the susceptibility evaluation of snow avalanche dangerous locations. It was conducted at the Taleghan watershed in Iran. An artificial intelligence (AI) model was created to anticipate the risk of mass wasting brought on by snow avalanches. The recursive feature elimination (RFE) approach is used to identify the important characteristics, which are then utilised to calibrate the model. The topographic location index and distance to stream were the most significant variables that contributed the most to the production of the susceptibility maps, according to sensitivity analysis (Theerthagiri and Vidya, 2022).

3. Proposed feature extraction and classification technique

This section explains the proposed recursive feature removal, gradient boosting based machine learning classification approach, feature ranking, and classification/prediction metrics for evaluating the proposed model's performance.

3.1. Dataset of cardiovascular disease

The RFE-GB algorithm's performance is evaluated using a cardiovascular illness dataset obtained from the Kaggle repository (Cardiovascular disease dataset, 2021). Seventy thousand patient data records with eleven characteristics and a goal classification of CVD or non-CVD patients make up the cardiovascular disease dataset. Gender, age, height, weight, BP-Systolic, BP-Diastolic, glucose, cholesterol, smoking habit, physical activities, and patients' alcohol intake are among the eleven features. The properties and values of the sample CVD dataset are summarised in Table 1.

3.2. Proposed methods

The recursive feature elimination approach extracts the most relevant features from the training dataset for target variable prediction. It is a useful technique for eliminating features from a training dataset before feature selection. RFE is popular because it excels at determining which characteristics in a training dataset are essential in predicting the target variable. RFE is a feature selection wrapper method that uses filter-based feature selection internally. RFE searches the training dataset for a subset of features, starting with all features and successfully removing them until the desired number of features is reached (Han et al., 2021; Yin and Zhang, 2014; Park et al., 2018; Hasan and Bao, 2021).

Table. 1
A sample of cardiovascular disease patient dataset.

Patient/Features	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Age	50	55	52	48	48
Gender	2	1	1	2	1
Height	168	156	165	169	156
Weight	62	85	64	82	56
BP (S)	110	140	130	150	100
BP (D)	80	90	70	100	60
Cholesterol	1	3	3	1	1
Glucose	1	1	1	1	1
Smoke	0	0	0	0	0
Alcohol	0	0	0	0	0
Activity	1	1	0	1	0
Cardio	0	1	1	1	0

This research creates a model with the predictors and calculates an important score for each predictor. Minorly significant predictors are eliminated. The model is then recreated, and the score is calculated once more. To assess a tuning parameter, the number of predictor subgroups and their size are supplied. The model may be trained using the best subset. As a consequence of the RFE method, a collection of top-ranked features can be evaluated for feature selection (Rani et al., 2021). Several subsets of characteristics were tried on the dataset. It uses prominent variables from a cardiovascular disease dataset to categorise cardio and non-cardio patients accurately (Theerthagiri and Vidya, 2022; Choubin et al., 2020).

The suggested RFE-GB approach for categorising CVD and non-CVD patients is depicted in Fig. 1. The missing values in the cardiovascular disease dataset have been replaced with the mode values. Because the dataset comprises several measurement units, normalisation was performed. After normalisation, all of the original data's properties will be in the same order of magnitude. Features are chosen to focus on the important data that aid in analysis and prediction. This feature removal method helps to improve classification accuracy over time.

The biggest margin was used to define the ranking criterion as a separating hyperplane to rank the characteristics. The decision function is provided in Equation after considering a set of training samples (1).

Algorithm:1 RFE-GB Algorithm

(1). Begin
(2). Training the model with CVD dataset using all predictors and features
(3). Calculation of model performance
(4). Calculation of feature importance using bagging, boosting, and voting
(5). For every subset size K_i, i = 1,…S do
  Assume K_i as most essential variables
  Training the RFE-GB model with CVD dataset using K_i predictors
  Calculation of model performance
  The ranking for every predictor is recalculated
  End For
(6). Calculation of performance profile over the K_i
(7). Determination of suitable number of predictors
(8). The RFE-GB model for calculating the optimal K_i
(9). End

$$f(a) = x.a + b \tag{1}$$

where 'x' is the weight vector generated by using Eq. (2)

$$x = \sum_{i=1}^{n} \alpha_i y_i z_i \tag{2}$$

where α_i are lagrange multipliers, $z_i \in R^d$ and $y_i \in \{-1, 1\}$ and i=1,….n.

The kth feature's ranking criterion is the square of the element k of w. The presented model is trained using each iteration of RFE, with the lowest-ranking feature being deleted because it has the least influence on classification. For all iterations, the remaining features are unchanged. This operation is repeated until all of the traits have been eliminated, at which point they are sorted in the reverse order. As a consequence, the most crucial characteristics will emerge. When the dimension is large, eliminating features one by one will take a long time. As a result, more than one feature in each cycle may be removed, decreasing accuracy and producing the correlation bias issue (Yan and Zhang, 2015). The feature selection and ranking techniques are shown in Fig. 2.

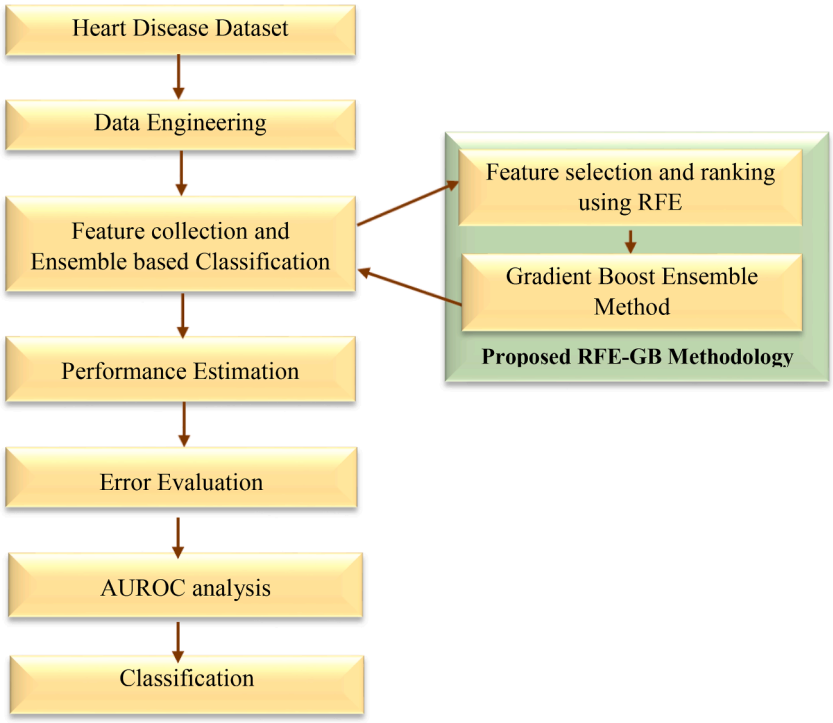

Fig. 1. Proposed RFE-GB algorithm for heart disease prediction.

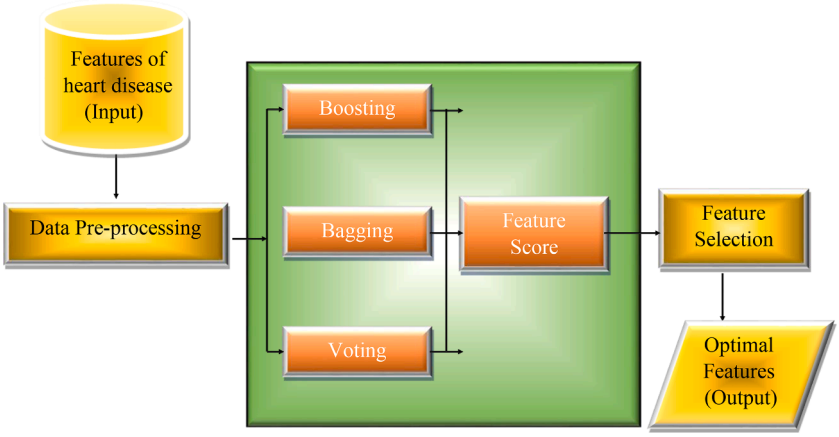

Fig. 2. Feature selection and ranking.

The cross-validation ratings for various CVD patient features are shown in Fig. 3. The curve line in this graph begins with a cross-validation value of 0.697 and grows upward for the three and four features. After obtaining a cross-validation value of 0.71, the value starts to decline for the number of features five and six, eventually reaching a value of 0685. When the number of features reaches six, the cross-validation value rises again. Up to eight features are included in the cross-validation score. There is a tiny variation, and the curve line ultimately reaches the cross-validation value near 0.71 with twelve features. BP(S), BP(D), Cholesterol, and Activity are the four top features according to the proposed RFE-GB algorithm.

Gradient Boosting is a decision tree-based boosting approach that has been employed to tackle the classification problem of CVD patients in this presented study. It sums together weak learners and uses gradient descent to minimise the model's loss function (finding the local minimum of the differentiable function). It creates learners during the learning process since it is an additive model. The gradient descent optimization process determines the impact of a poor learner. The contribution of each tree is calculated by lowering the overall error of the strong learner Li, 2016). Consider a gradient boosting method with X stages and y as the actual values of the output variable. Assume an incomplete model K_x at each gradient boosting stage x $(1 \leq x \leq X)$. As illustrated in Eqs. (3) and ((4), our strategy introduces a new estimator, $h_x(j)$, to improve K_x (4).

$$K_{m+1}(j) = K_m(j) + h_x(j) = y \tag{3}$$

Thus,

$$h_x(j) = y - K_m(j) \tag{4}$$

As a consequence, gradient boosting will fit h to the residual $y - K_m(j)$ and determine if the patient is a CVD patient or not.

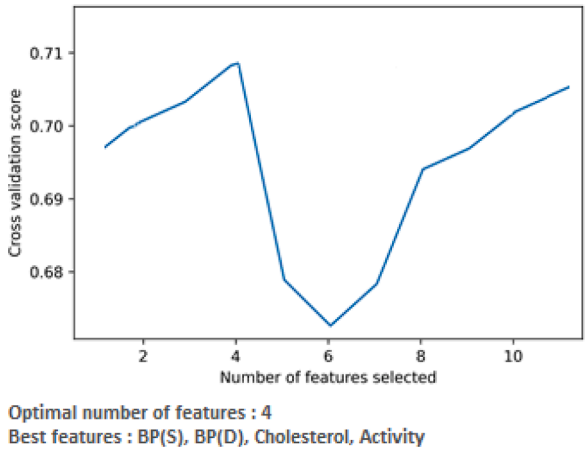

Fig. 3. Feature selection using RFE-GB algorithm.

3.3. Performance evaluation methods

From the CVD dataset, 70% of the data is considered training data and 30% is considered testing data in this suggested study. The metrics recall, F1-score, precision, confusion matrix, RMSE, AUC-ROC, Cohen's kappa, and MSE are used to assess the performance of the proposed RFE-GB model. During error analysis, data cohorts with greater error rates are found.

Cohen's Kappa is an effective metric for dealing with multi-class and unbalanced class issues. It is calculated using Eq. (3), where p_o is observed class and p_e is the predicted class of CVD patients. Its value ranges from zero to one. The average error between actual and anticipated values is calculated using the Mean Squared Error (MSE). The average square of the difference between the original and anticipated values may be used to calculate its value. The MSE is calculated using Eq. (4), where 'n' is the number of CVD patient records in the dataset. The square root of the average error between actual and anticipated values is given by the Root Mean Squared Error (RMSE). The equation can be used to calculate its value (5).

$$k = \frac{p_o - p_e}{1 - P_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left((actual\ values - predicted\ values)^2\right) \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}} \tag{7}$$

The Area Under the Curve- Receiver Operator Characteristic (AUC-ROC) is used to evaluate a classifier's ability to distinguish between CVD and non-CVD classes in this study. The Receiver Operator Characteristic is a probability curve that shows the True Positive Rate (TPR) versus the False Positive Rate (FPR) at various threshold settings. The confusion matrix is a 'n x n' matrix that compares the actual target values with the values predicted by the machine learning model to evaluate the performance of a classification algorithm. The accuracy, recall, and F1-score of the suggested RFE-GB model are also examined. As shown in Eq. (6), the precision is the ratio between the TP (True Positive) and all the positives (True Positive and False Positive). The recall is a measurement of how well the model detects True Positives. The equation describes how it is calculated (7). Eq. (1) shows the F1-score, which is the harmonic mean of recall and accuracy (8).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{8}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{9}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

4. Discussions and findings

The developed Recursive Feature Elimination based Gradient Boosting (RFE-GB) classification technique is compared to the classic Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Naive Bayes (NB) algorithms in this part. The k-fold cross-validation resampling approach was utilised in this study to check the machine learning algorithms' research outcomes. In this work, the 'k' value is set to 10. As a result, a 10-fold cross-validation resampling procedure is commonly used. The 10-fold cross-validation method is intended to decrease the bias in the prediction model.

The proposed RFE-GB model has been implemented using google colaboratory with the support of python libraries along with an intel i3 processor, 8 GB RAM and 108 GB hard disk.

4.1. Performance evaluation

A set of classification algorithm-based metrics is frequently used to assess the efficacy of machine learning prediction algorithms. This study uses the mean square error (MSE), root mean square error (RMSE), and Kappa score to calculate the prediction error rates. The true/false positive/negative rate of the predictions is examined using the confusion matrix and receiver operating characteristic area under the curve (ROC AUC). Prediction accuracy, precision, recall, and f1 score are used to assess the performance of machine learning algorithms (Friedman, 2001; Theerthagiri et al., 2021; Patro et al., 2020).

This study also determines the accuracy of the aforementioned machine algorithms (whether the patient has cardiovascular disease or not). Based on its hyperparameters, each classification model has a specific illness prediction accuracy and efficiency above other prediction models. In this research, 70% of the dataset is used for training, while 30% of the data samples are used to evaluate classification algorithms. In Table 2, the suggested illness categorization reports of the RFE-GB algorithm are compared to established machine learning methods.

The performance results, including prediction accuracy, precision, recall, and F1-score, are summarised in Table 2. The accuracy, precision, recall, and F1-score of the proposed recursive feature elimination-based gradient boosting technique are 88.84, 0.88, 0.85, and 0.83, respectively. Other machine learning algorithms, on the other hand, provide lesser outcomes than the suggested RFE-GB Algorithm. As a result, the proposed recursive feature elimination-based gradient boosting algorithm achieves an 88.84 percent prediction accuracy, while the other machine learning classification algorithms logistic regression, random forest, multilayer perceptron, support vector machines, and Naive Bayes only achieve 71.36, 72.69, 76.42, 72.6, and 58.3 percent prediction accuracy, respectively.

Table. 2
Performance metrics.

S. No.	Algorithm/metrics	Accuracy	Precision	Recall	F1-score
1	Logistic Regression (LR)	71.36	0.693	0.672	0.702
2	Random Forest (RF)	72.69	0.708	0.702	0.716
3	Multi-Layer Perceptron (MLP)	76.42	0.72	0.72	0.72
4	Support Vector Machines (SVM)	72.6	0.681	0.637	0.695
5	Naive Bayes (NB)	58.3	0.74	0.27	0.43
6	**Proposed RFE-GB**	**88.84**	**0.88**	**0.85**	**0.83**

The Naive Bayes algorithm performs the poorest, with 58.3 percent accuracy, precision, recall, and F1-scores of 0.74, 0.27, and 0.43, respectively. The accuracy of the proposed RFE-GB and other current algorithms is shown in Fig. 4. The proposed recursive feature elimination-based gradient boosting algorithm predicts cardiovascular disease patients about 90% more accurately than the other algorithms (based on patient's age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol intake, and physical activity).

Fig. 4 further shows that the developed RFE-GB algorithm has the best accuracy of 88.84 percent. The presented RFE-GB method searches the training dataset for a subset of features, starting with all of them and successfully removing them until the goal number of features is reached. Several subsets of characteristics from the cardiovascular disease dataset have been evaluated. It picks the most popular attributes from the dataset to categorise cardio and non-cardio patients accurately. Consequently, the suggested recursive feature elimination-based gradient boosting strategy surpasses the competition. As a consequence, the proposed RFE-GB algorithm beats the LR, RF, MLP, SVM, and NB algorithms by a factor of 20.06 to 23.82. Fig. 5 depicts accuracy and recall performance. It can be noted that both measurements have higher scores than others, with 0.88 and 0.85, respectively. Furthermore, the proposed RFE-GB algorithm's F1-score has enhanced outcomes by 2 to 15 percent.

Fig. 6 shows Cohen's kappa scores for proposed and current machine learning algorithms; as the graph shows, the proposed RFE-GB approach has a greater kappa score than conventional methods. Cohen's kappa score ensures the consistency of the classification algorithm's predictions (Theerthagiri et al., 2021). Furthermore, it shows that among the tested methods, the developed RFE-GB approach has the greatest consistency (Kappa score) of 0.56, whereas the LR, RF, MLP, SVM, and NB have values of 0.42, 0.41, 0.44, 0.43, and 0.47, respectively.

The mean square error (MSE) and root mean square error (RMSE) values for several machine learning approaches are shown in Table 3. The LR, RF, MLP, SVM, NB, and proposed RFE-GB algorithms have MSE error rates of 0.28, 0.29, 0.27, 0.28, 0.51, and 0.20, respectively. Compared to the other algorithms, it has the lowest error rate of 0.20 for correctly identifying heart disease incidences (Fig. 7). The proposed RFE-GB method enhances the weak features from the higher-ranked and chosen features in the training and testing datasets. As a consequence, the number of errors is minimised. As shown in Table 3, the presented RFE-GB method has a very low RMSE error rate (0.44), whereas the error rates for other algorithms include LR (0.53), RF (0.54), MLP (0.52), SVM (0.55), and NB (0.26).

Fig. 8 depicts the confusion matrix for several machine learning techniques. This graph shows the percentages of predicted and real values as true positives/negatives and false positives/negatives, respectively. As shown in Fig. 8, the proposed RFE-GB algorithm correctly estimates 88 percent (true positive) of cardio cases with only 12 percent (false positive) misclassification; for non-cardio cases, the RFE-GB algorithm gives 16 percent (false negative) misclassification and

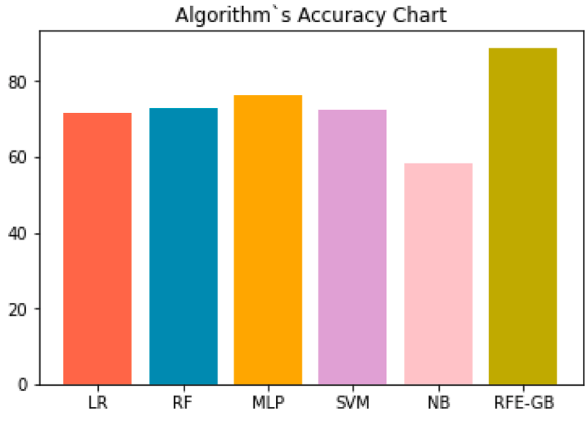

Fig. 4. Prediction accuracy of ML algorithms.

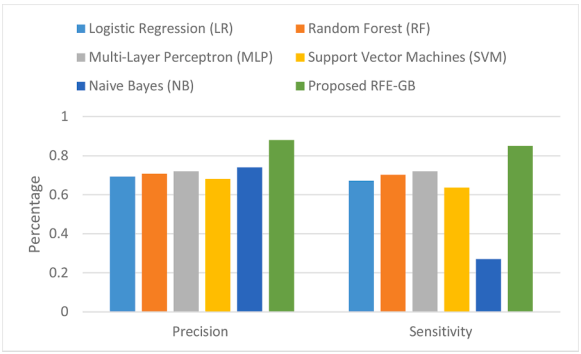

Fig. 5. Classification report.

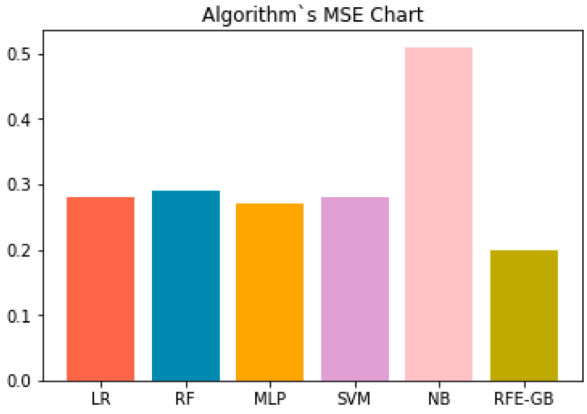

Fig. 6. Cohen's kappa scores.

Table. 3
Error performance metrics.

S.No.	Algorithm/metrics	MSE	RMSE
1	Logistic Regression (LR)	0.28	0.53
2	Random Forest (RF)	0.29	0.54
3	Multi-Layer Perceptron (MLP)	0.27	0.52
4	Support Vector Machines (SVM)	0.28	0.55
5	Naive Bayes (NB)	0.51	0.26
6	**Proposed RFE-GB**	**0.20**	**0.44**

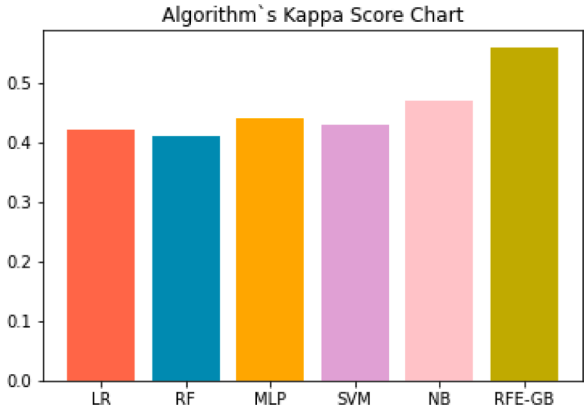

Fig. 7. MSE rates.

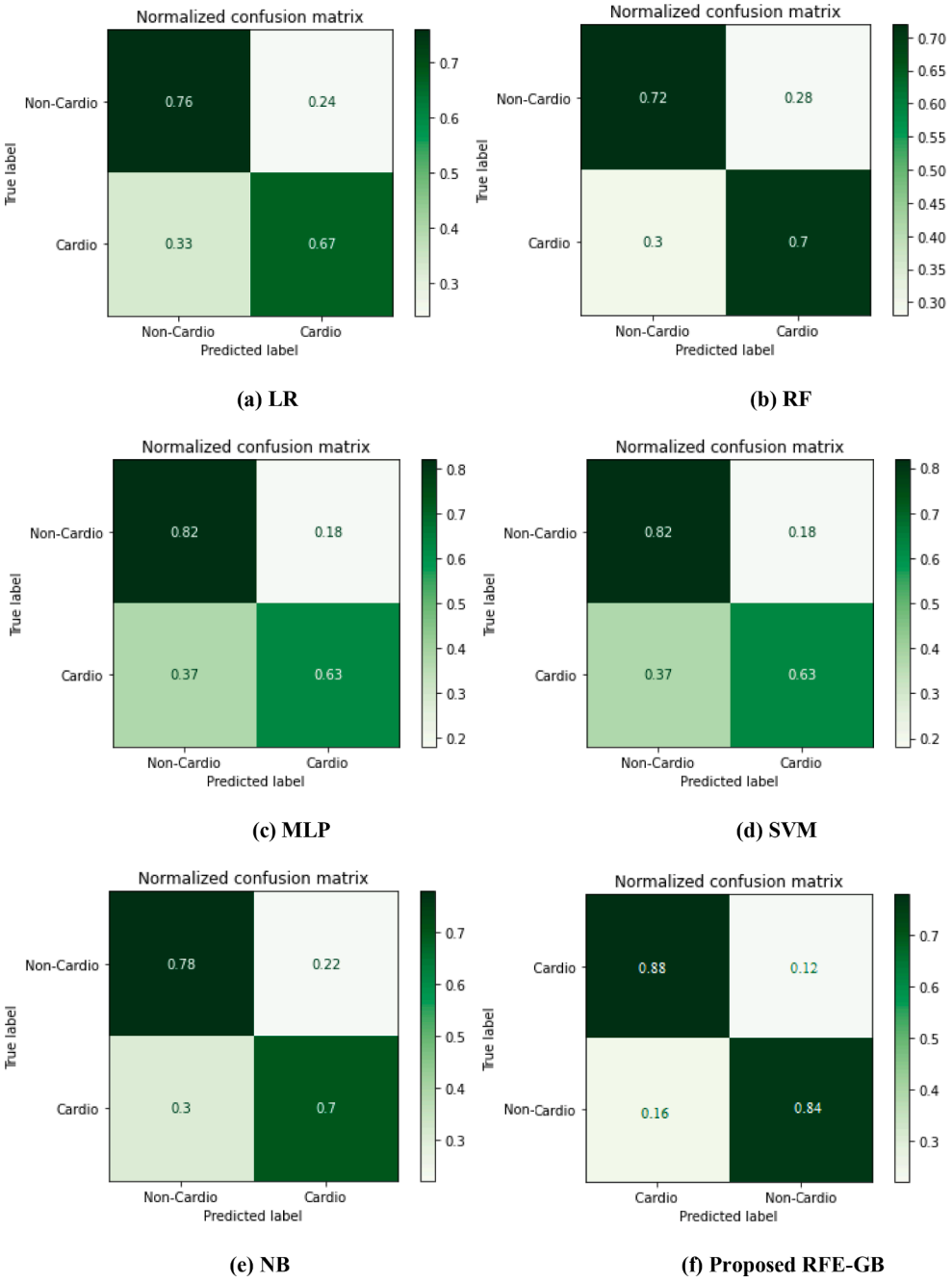

Fig. 8. Normalized confusion matrix of ML algorithms.

84 percent (true negative) precise classification (f). Fig. 8(a)-(e) show the confusion matrix of the LR, RF, MLP, SVM, and NB algorithms, with lower true positives/negatives and false positives/negatives rates than the RFE-GB method, respectively.

Fig. 9 displays the link between the false positive rate and the true positive rate as a ROC area under the curve. The RFE-GB method yields the highest value of 0.85, compared to the LR (0.77), RF (0.768), MLP (0.79), SVM (0.785), and NB (0.69) algorithms. These findings show that the proposed RFE-GB algorithm correctly diagnoses individuals with cardiovascular disease based on their medical information.

5. Conclusion

Integrating machine learning algorithms for medical disease prediction and diagnosis is much needed for promising growth in the healthcare field. This work chooses the most significant features from the cardiovascular disease dataset using a recursive feature elimination-based gradient boosting approach. From the 12 characteristics, the RFE-GB algorithm picks three ideal numbers of features, such as blood pressure, cholesterol, and physical activity. A gradient boosting ensemble technique has been developed to forecast cardiovascular disease cases using these three characteristics. The proposed RFE-GB method has been tested using a variety of measures, and the results have been compared to those of other machine learning algorithms. Compared to LR, RF, MLP, SVM, and NB algorithms, the presented RFE-GB method improves accuracy by 14.12 percent to 30.12 percent.

Furthermore, it has a greater consistency (Kappa score) of 0.56 and a lower error rate MSE of 0.1924 to accurately predict cardiac diseases. With an AUROC score of 85 percent, the developed RFE-GB algorithm successfully estimates 88 percent true positives and 85 percent true

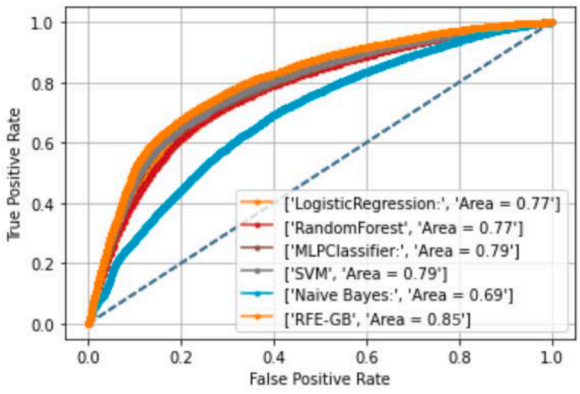

Fig. 9. ROC_AUC Curve.

negatives from 70,000 patient records. The suggested RFE-GB algorithm appears to be capable of identifying and categorising diabetic patients as a result of the findings. Future research will be enhanced with other combinations of machine learning methods for facilitating better support for physicians.

CRediT authorship contribution statement

Prasannavenkatesan Theerthagiri: Conceptualization, Methodology, Data curation, Writing – original draft, Visualization, Investigation, Validation, Writing – review & editing.

Declaration of Competing Interest

None.

References

Aggrawal, R., & Pal, S. (2021). Elimination and backward selection of features (*p*-value technique) in prediction of heart disease by using machine learning algorithms. *Turkish Journal of Computer and Mathematics Education, 12*(6), 2650–2665.

Akyol, K., & Atila, Ü. (2019). A study on performance improvement of heart disease prediction by attribute selection methods. *Academic Platform Journal of Engineering and Science, 7-2*, 174–179.

Bakhsh, AA. (2021). High-performance in classification of heart disease using advanced supercomputing technique with cluster-based enhanced deep genetic algorithm. *The Journal of Supercomputing, 77*(9), 1–22.

Breiman, L. (2011). Random forests. *Machine Learning, 45*(1), 5–32.

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics, 9*(4), 178.

Chen, Qi, Meng, Z., & Su, R. (2020). WERFE: A gene selection algorithm based on recursive feature elimination and ensemble strategy. *Frontiers in Bioengineering and Biotechnology, 8*, 496.

Choubin, B., Borji, M., Hosseini, F. S., Mosavi, A., & Dineva, A. A. (2020). Mass wasting susceptibility assessment of snow avalanches using machine learning models. *Scientific Reports, 10*(1), 1–13.

Elavarasan, D., Vincent, P. M. D. R., Srinivasan, K., & Chang, C. Y (2020). A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. *Agriculture, 10*(9), 400.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

Han, Y., Huang, L., & Zhou, F. (2021). A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. *Bioinformatics, 37*(15), 2183–2189.

Hasan, N., & Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. *Health and Technology, 11*(1), 49–62.

Hosseini, F. S., Choubin, B., Mosavi, A., Nabipour, N., & Haghighi, A. T. (2020). Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: application of the simulated annealing feature selection method. *Science of the Total Environment, 711*, Article 135161.

Kakulapati, V., Kirti, A., Kulkarni, V., & Raj, C. P. (2017). Predictive analysis of heart disease using stochas- tic gradient boosting along with recursive feature elimination. ACADEMIA, Accelerating the world's research *International Journal of Science and Research (IJSR) ISSN, 6*(5), 909–912 (Online).

Kakulapati, V., Ankith, K., Vaibhav, K., & Charan, P. R. (2017). Predictive analysis of heart disease using stochas-tic gradient boosting along with recursive feature elimination. *International Journal of Science and Research, 6*(5), 909–912.

Cardiovascular disease dataset, retrieved from Kaggle repository, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, 2021.

C. Li "A gentle introduction to gradient boosting." URL: http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf, 2016.

Mahendran, N., & Vincent P M, D. R. (2022). A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Computers in Biology and Medicine, 141*, Article 105056.

Mosavi, A., Sajedi, H. F., Choubin, B., Goodarzi, M., & Rafiei, S. E. (2021). Ensemble boosting and bagging based machine learning models for groundwater potential prediction. *Water Resources Management, 35*(1), 23–37.

Mosavi, A., Sajedi Hosseini, F., Choubin, B., Taromideh, F., Ghodsi, M., Nazari, B., & Dineva, A. A. (2021). Susceptibility mapping of groundwater salinity using machine learning models. *Environmental Science and Pollution Research, 28*(9), 10804–10817.

Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. *Journal of Clinical Medicine, 8*(7), 1050.

Park, D., Lee, M., Park, SE., Seong, J. K., & Youn, I. (2018). Determination of optimal heart rate variability features based on SVM-recursive feature elimination for cumulative stress monitoring using ECG sensor. *Sensors, 18*(7), 2387.

Patro, S. P., Padhy, N., & Chiranjevi, D. (2020). Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. *Evolutionary Intelligence, 14*(2), 941–969. Special Issue.

Prasannavenkatesan, I., Jeena, J., Usha, A., & Vamshidar, Y. (2021). Prediction of COVID-19 possibilities using KNN classification algorithm. *International Journal of Current Research and Review, 13*(06), 156.

Prasannavenkatesan, T. (2021). Probable forecasting of epidemic COVID-19 in using COCUDE model. *EAI Endorsed Transactions on Pervasive Health and Technology, 7*(26), e3.

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments, 7*(3), 263–275.

Satish Chandra Reddy, N., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: heart disease prediction. *International Journal of Innovative Computing, 9*(1), 39–46.

Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., & Liu, C. (2019). A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine, 171*, 1–10.

Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., & Liu, C. (2019). A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine, 171*, 1–10.

Theerthagiri, P. (2021). Forecasting hyponatremia in hospitalized patients using multilayer perceptron and multivariate linear regression techniques. *Concurrency and Computation: Practice and Experience, 33*(16), e6248.

Theerthagiri, P., Gopala, K. C., & Nishan, A. H. (2021). Prognostic analysis of hyponatremia for diseased patients using multilayer perceptron classification technique. *EAI Endorsed Transactions on Pervasive Health and Technology, 7*(26), e5.

Theerthagiri, P., & Ruby, A. U. (2022). RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction. *Expert Systems, early view*, e13048.

Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert Systems, early view*, e13064.

Theerthagiri, P., Ruby A, U., Vidya, J., et al. (2021). *Diagnosis and classification of the diabetes using machine learning algorithms*. Research Square. https://doi.org/10.21203/rs.3.rs-514771/v3. 17 MayPREPRINT available at.

Wang, Z., Yanrui, J., Liqun, Z., & Chengliang, L. (2021). A heart sound classification method based on joint decision of extreme gradient boosting and deep neural network. *Journal of Biomedical Engineering, 38*(1), 10–20.

Yan, Ke, & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical, 212*, 353–363.

Yin, Z., & Zhang, J. (2014). "Operator functional state classification using least-square support vector machine based recursive feature elimination technique.". *Computer methods and programs in biomedicine, 113*(1), 101–115.

Zhang, B., Ren, J., Cheng, Y., Wang, B., & Wei, Z. (2019). Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm. *Special Section On Data-Enabled Intelligence For Digital Health, 7*, 32423–32433.

Zhao, J., Feng, Q. P., Wu, P., Lupu, RA., Wilke, R. A., Wells, Q. S., … Wei, W. Q. (2019). Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event Prediction. *Scientific Reports, 9*(1), 1–10.