Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques

Prasannavenkatesan Theerthagiri | Jyothiprakash Vidya

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

Correspondence
Prasannavenkatesan Theerthagiri, Department of Computer Science and Engineering, GITAM School of Technology, GITAM University, Bengaluru, India.
Email: vprasann@gitam.edu

Abstract
Cardiovascular diseases are one of the most common chronic illnesses that affect people's health. Early detection of cardiovascular diseases' can reduce mortality rates by preventing or reducing the severity of the disease. Machine learning algorithms are a promising method for identifying risk factors. This article proposes a recursive feature elimination-based gradient boosting algorithm in order to obtain accurate heart disease prediction. The patients' health record with important cardiovascular disease features has been analysed for the evaluation of the results. Several other machine learning methods were also used to build the prediction model, and the results were compared with the proposed model. The results of this proposed model infer that the combined recursive feature elimination and gradient boosting algorithm achieves the highest accuracy (89.7%). Further, with an area under the curve of 0.84, the proposed algorithm was found superior and had obtained a substantial gain over other techniques. Thus, the proposed gradient boosting algorithm will serve as a prominent cardiovascular disease estimation and treatment model.

KEYWORDS
cardiovascular disease, classification of cardiovascular diseases, feature ranking, gradient boosting, machine learning, recursive feature elimination

1 | INTRODUCTION

As the effects of societal ageing worsen, health monitoring has become increasingly important. The disease prediction framework can help medicinal experts in anticipating heat alignment because of the clinical information of patients. Subsequently, by implementing a prediction framework utilizing advanced algorithms and investigating different health-related issues, it can have the capacity to predict more probabilistically that the patients will be diagnosed with any health problems (Shi et al., 2019).

One of the most significant parts of healthcare monitoring is heart monitoring. A heart disease prediction system can provide valuable insights to medical professionals in making decisions about the state of heart of patients. Aberrant cardiac rhythms can be caused by an abnormal site of origin or irregular conduction of the electric signal. Arrhythmias are the medical term for these diseases. Some arrhythmias can result in significant consequences and even death (Shi et al., 2019). Medical professionals may neglect to take exact decisions while diagnosing a patient’s heart alignment (Kakulapati et al., 2017a); in this way, a heart alignment prediction method that utilizes machine learning (ML) algorithms aid such cases to get precise outcomes (Prasannavenkatesan, 2021).

Disease prediction, disease categorization, and medical image recognition algorithms are all examples of machine learning techniques that have been widely applied in medicine (Chang et al., 2019). Gradient boosting (GB), a contemporary and efficient method, is presented and enhanced in this article. The gradient boosting decision tree is the source of the gradient boosting learning method (Theerthagiri, 2021). GB performs well as an ensemble classifier in terms of generalization. Furthermore, GB provides a regularization term to regulate the model's complexity,
which prevents overfitting (Usha Ruby et al., 2020). In several machine learning disciplines, GB has outperformed the competition (Wang et al., 2021). As a result, the performance of GB in classifying single cardiac diseases is investigated.

The goal of this research is to create a clinically useful categorization system for cardiac disease. This work presents a hierarchical technique based on the weighted gradient boosting algorithm to achieve this goal. Preprocessing is the most common method for obtaining usable cardiovascular patient datasets (Aggrawal & Pal, 2021). Following that, numerous types of characteristics are extracted. Following that, recursive feature elimination (RFE) is used to choose features. Finally, the feature vectors are fed into a hierarchical classifier, which produces predicted labels. The medical field is an application field of information mining because it has a large number of information assets (Bakhsh, 2021). They realize that it is valuable to include selection and feature reduction. Feature determination is concerned with distinguishing some relevant features sufficient to learn objective thoughts (Prasannavenkatesan et al., 2021).

To choose features for hyperparameter optimization, a stochastic gradient boosting approach is used. To reduce the mean square error (MSE), the features are clustered together. Multiple experimental scenarios are examined, and the findings are compared to several earlier studies and typical ML algorithms to prove the usefulness of the proposed technique (Seth & Pandey, 2009).

The proposed technique is unique because it uses the gradient boosting technique to classify cardiovascular diseases. Although the gradient boosting algorithm is well-known, it is processed with the weights of each feature of the dataset for heart disease prediction and classification. The proposed approach is unique because it uses a hierarchical classifier and RFE to choose the best feature from all other features.

The rest of this article is laid out as follows. The next part provides background information on past efforts as well as an analysis of their shortcomings. With preprocessing, feature selection, and the hierarchical classification approach, Section 3 outlines the proposed approach employed in this study. In Section 4, performance measurements are used to assess the proposed approaches. This section explains the findings and draws some parallels with past research. Section 5 concludes by summarizing all of the works and drawing conclusions.

2 | BACKGROUND

Artificial intelligence and deep learning algorithms are extremely beneficial for using massive data to predict individual outcomes, especially when coupled to Electronic health record (EHR) (Shi et al., 2019). This study used machine learning to increase the prediction accuracy of traditional cardiovascular diseases (CVDs) risk variables in a large United Kingdom (UK) population. The effectiveness of machine learning techniques on longitudinal EHR data for ten-year cardiovascular event prediction was compared to a gold standard reached through pooled cohort risk (Zhao et al., 2019).

A classification approach with three basic steps was developed in this study (Shi et al., 2019). The wavelet approach is used to filter the electrocardiogram (ECG) signal during the preprocessing phase. Then fiducial points are used to find all heartbeats. Feature engineering is a technique for extracting different types of features from time and time-frequency domains. Then, to choose features, this study used RFE. To get the final findings, a hierarchical classifier based on the extreme gradient boosting (XGBoost) classifier and threshold is used in the classification step (Shi et al., 2019).

The authors devised a prediction approach to categorize hypertension patients based on physical examination markers (Chang et al., 2019). The important elements from the patients’ many clinical assessment signs are retrieved in the first stage. The essential features retrieved in the first stage are used in the second stage to forecast the patients’ outcomes. The authors then suggested a model that incorporated recursive feature removal, cross-validation, and a prediction model. XGBoost is believed to successfully forecast patient outcomes by employing their best features subset (Chang et al., 2019).

This work (Chen et al., 2020) proposed a wrapper gene selection strategy with a recursive feature removal approach for efficient classification. The ensemble technique was used for several gene selection strategies, and the top-ranking genes in each methodology were chosen as the final gene set. Multiple gene selection techniques were combined in this study, and the ideal gene subset was obtained by prioritizing and ranking the most essential genes picked by the gene selection approach. Consequently, the authors concluded that selecting a more discriminative and compact gene subset yielded the best results (Chen et al., 2020).

The scientists used machine learning algorithms to forecast a patient's stage of cardiac disease (Kakulapati et al., 2017b). They chose the optimal features using the stochastic gradient boosting technique and RFE. An ensemble of weak prediction models, often using decision trees, was used to create a calculation model. It provides a stage-by-stage approach to boosting, simplifying, and optimizing a subjectively variational failure problem.

The authors of (Padmanabhan, Yuan, Chada, & Van Nguyen, 2019) presented an AutoML approach for automating the process of developing an Artificial Intelligence (AI) model that performs well on any dataset. This study for cardiovascular disease prediction automates data preprocessing, feature extraction, hyper-parameter tweaking, and algorithm selection. The authors claimed that their AutoML model had removed a significant technical hurdle, allowing doctors to employ AI approaches more widely.

For the best feature detection of the Single Proton Emission Computed Tomography (SPECT), Statlog Heart Disease (STATLOG) datasets, recursive feature removal with cross-validation, and stability selection were utilized, and their results were compared (Akyol & Atila, 2019). The
approaches of RFE with cross-validation (RFECV) and stability selection (SS) were used to enhance the productivity of tree-based and probability-based machine learning techniques in this research. The feature with the lowest score is therefore removed. The RFECV adapts to the RFE and adjusts the number of characteristics picked automatically. The SS method returns details about the output variable's properties. This technique, according to the authors, is most useful in determining the treatment strategy for professionals in the area (Akyol & Atila, 2019).

To estimate response variables more correctly, the gradient boosting approach fits new models sequentially during learning. The primary concept behind this technique is to build new base-learners with the highest correlation with the ensemble's negative gradient of the error function (Friedman, 2001). Breiman (Breiman, 2001) invented Random Forest (RF), an ensemble learning system based on random decision trees. The main distinction between RF and decision trees is that when breaking a node, RF looks for the best feature among the random subsets of characteristics. In contrast, decision trees look for the most significant feature. As a result, there is a lot of variety, which leads to a better model. The Bayes' theorem-based NB classifier was utilized, with each pair of classified characteristics being independent of one another. In order to discover the most probable categories, it employs probability theory. When the input has high dimensionality, this approach is appropriate (Theerthagiri et al., n.d.).

For the Cleveland and Statlog project heart datasets, the authors suggested a model to predict heart disease categorization based on feature selection (Reddy et al., 2019). According to them, the RF algorithm's accuracy is good for feature selection (8 and 6 features) based on classification models. This study also associated sensitivity and specificity with higher scores (Reddy et al., 2019).

The gradient boosting decision tree was used in Reference (Zhang et al., 2019) to estimate blood pressure rates based on human physiologic data obtained by the Eimo device. To pick ideal parameters and avoid overfitting, they employed the cross-validation approach. Also, it has been suggested that when considering the features of age, body fat, ratio, and height, this method displayed a higher accuracy rate with reduced error rates compared to other algorithms (Zhang et al., 2019). This work (Patro et al., 2020) had proposed a framework for the prediction of risk factors of heart disease using several classifier algorithms. They have revealed that the support vector machine performs better prediction accuracy, precision, sensitivity, and F1 score (Patro et al., 2020).

Onan et al. discussed the student evaluations of teaching using opinion mining and recurrent neural network (RNN)-based model (Onan, 2020). Their approach, the deep learning paradigm, a sentiment classification technique, was implemented for teacher assessment evaluations. Further, in Reference (Onan, 2021), they developed a deep learning-based solution to sentiment analysis using Twitter product evaluations. It combines convolutional neural network and long-short term memory (CNN-LSTM) architecture with weighted Glove word embedding (Toçoğlu & Onan, 2020). The advantages of ensemble methods, feature extraction methods on data analysis, and sentiment analysis has been presented in these articles (Padmanabhan, Yuan, Chada, & Nguyen, 2019); it analyses based on different base learning algorithms with the comparison of different feature selections correlation, consistency, information gain, and chi-square-based feature selection and different ensemble learning methods namely boosting, bagging, dagging, and random subspace (Hasan & Bao, 2021; Mohan et al., 2019).

### 3 | PROPOSED FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUE

This section describes the proposed RFE, gradient boosting-based machine learning classification technique, feature ranking, and classification/prediction metrics, which are used to evaluate the performance of the proposed model. The hypothesis considered for this work is as follows.

**Hypothesis 1.** Integrating the RFE algorithm with gradient boosting learning algorithms will improve accuracy and receiver operating characteristic (ROC) values.

#### 3.1 | Dataset of cardiovascular disease

The performance of the RFE-based gradient boosting (RFE-GB) algorithm is analysed using the cardiovascular disease dataset retrieved from the Kaggle repository (Cardiovascular Disease Dataset, 2021). The cardiovascular disease dataset consists of 70,000 patient data records with 11 features and a target classifying CVD or non-CVD patients. The 11 attributes are gender, age, height, weight, blood pressure (BP)-Systolic, BP-Diastolic, glucose, cholesterol, smoking behaviour, physical activities, and patients’ alcohol intake. Table 1 summarizes the sample CVD dataset features and their values.

#### 3.2 | Proposed methods

The RFE algorithm effectively selects the features from the training dataset that are most relevant in target variable prediction. It is an effective method for removing features from a training dataset in preparation for feature selection (Han et al., 2021; Yin & Zhang, 2014). RFE is prominent
because it is good at identifying the features in a training dataset that are more or less important in predicting the target variable. RFE is a wrapper-style feature selection algorithm that internally employs filter-based feature selection. RFE operates by looking for a subset of features in the training dataset, beginning with all features and successfully deleting them until the target number of features (Hasan & Bao, 2021; Park et al., 2018).

This work builds a model with the predictors, and an importance score is being computed for each predictor. The predictors with a minor significance are removed. Then, the model is rebuilt, and the score is computed again. Here, the number of predictor subsets and their size are specified to evaluate a tuning parameter. The optimal subset can be used to train the model. Thus, the RFE algorithm resulted in the group of top-ranked features that can be considered for selecting features (Rani et al., 2021). The dataset has been tested with several subsets of features. It selects the popular features from the cardiovascular disease dataset to classify cardio and non-cardio patients with reduced errors.

Figure 1 illustrates the workflow of the proposed RFE-GB methodology for classifying the CVD and Non-CVD patients. The cardiovascular disease dataset has been pre-processed, where missing values are replaced with the mode values. Then normalization was carried out as the dataset contains various measuring units. After normalizing the data, all the attributes of the original data will be in the same order of magnitude.

The feature is selected to focus on the relevant data that helps in analysis and prediction. This feature elimination strategy supports in a cumulative increase of classification accuracy.

In order to rank the features, the ranking criterion as a separating hyperplane has been determined with the largest margin. Then, a set of training samples are considered, the decision function \( f(x) \) is given in Equation (1).

\[
f(x) = w \cdot x + b
\]

where \( w \) is the weight vector which can be obtained by using Equation (2).

\[
w = \sum_{i=1}^{n} a_i y_i x_i
\]

where \( a_i \) are lagrange multipliers, \( x_i \in \mathbb{R}^d \) and \( y_i \in \{-1,1\} \) and \( i = 1, ..., n \).

The square of the element \( k \) of \( w \) is used as the \( k \)'th feature's ranking criterion. Each iteration of RFE is used to train the proposed model, with the lowest-ranking feature being removed because it has the least impact on classification. The remaining features are preserved for the succeeding iteration. This procedure is continued until all of the characteristics have been deleted, at which time they are sorted in the order they were removed. As a result, the attributes that are most important will emerge. Eliminating features one by one will take a long time when the dimension is vast. As a result, more than one feature might be eliminated in each iteration, potentially affecting accuracy and causing the correlation bias problem (Yan & Zhang, 2015). Figure 2 depicts the feature selection and ranking strategy.

Figure 3 shows the cross-validation scores for the various number of CVD patient features. In this graph, the curve line starts with the cross-validation value of 0.697 and growing up for the three and four features. After reaching the cross-validation value of 0.71, its value starts decreasing for the number of features five, six reaches as lower as 0.685. Once, reaching the number of features with six, its cross-validation value again increases. The cross-validation score continues to improve up to eight features. There is a slight deviation, and finally, with 12 features, the curve

| 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 |
Thus, this graph clearly depicts the optimal number of features as four with the highest cross-validation score of 0.71. The four best features based on the proposed RFE-GB algorithm are BP(S), BP(D), cholesterol, and activity.

Gradient Boosting is the boosting technique based on a decision tree that has been used in this proposed work to solve the classification problem of CVD patients. It sums up weak learners using gradient descent to (it finds the local minimum of the differentiable function) minimize the model's loss function. As it is an additive model, it generates the learners during the learning process. The gradient descent optimization procedure determines the impact of a weak learner. Each tree's contribution is calculated by reducing the strong learner's total error (Li, 2016). Consider a gradient boosting technique with X stages and y as the output variable's actual values. Assume an imperfect model $K_x$ at each stage $x (1 \leq x \leq X)$ of gradient boosting. The proposed method incorporates a new estimator, $h_x(j)$, to enhance $K_x$, as shown in Equations (3) and (4).
Thus,

\[ h_x(j) = y - K_m(j) \]  \hspace{2cm} \text{(4)}

Therefore, gradient boosting will fit \( h \) to the residual \( y - K_m(j) \) and give the classification results regarding whether CVD patient or non-CVD patient.

### 3.3 Performance evaluation methods

In this proposed work, 70% of the data are considered as training data, and 30% are taken as testing data from the CVD dataset. To measure the performance of the proposed RFE-GB model, the metrics, namely recall, F1-score, precision, confusion matrix, root mean square error (RMSE), Cohen’s kappa, and MSE, are considered. During error analysis, cohorts of data are identified with higher error rates.

Cohen’s Kappa score is an excellent measure for handling multi-class and imbalanced class problems. Its value ranges from zero to one, and it is derived using Equation (3), where \( p_o \) is the observed class, and \( p_e \) is the expected class of CVD patients. The MSE gives the average error between actual and predicted values. Its value can be determined by taking the average square of the difference between the original and predicted values. Equation (4) gives the way to calculate the MSE, where ‘n’ represents the number of CVD patient records in the dataset. The RMSE gives the square root of the average error between actual and predicted values. Its value can be determined with Equation (7).

\[
k = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \hspace{2cm} \text{(5)}
\]

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\text{actual values} - \text{predicted values})^2 \hspace{2cm} \text{(6)}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{predicted}_i - \text{actual}_i)^2} \hspace{2cm} \text{(7)}
\]

In this work, the Area Under the Curve-Receiver Operator Characteristic (AUC-ROC) measures a classifier’s ability to differentiate between CVD and non-CVD classes. The ROC is a probability curve that plots the true positive rate (TPR) against false positive rate (FPR) at different threshold values. The confusion matrix is an \( n \times n \) matrix that evaluates the performance of a classification model, and it compares the actual...
target values with the values predicted by the machine learning model. Further, the precision, recall, and F1-score are analysed for the proposed RFE-GB model. The precision gives the ratio between the true positive (TP) and all the positives (True Positive and False Positive) as given in Equation (8). The recall is the measure of how the model correctly detects true positives. Its way of calculation is given in Equation (9). The F1-score is the harmonic mean of the recall and precision, and it is presented in Equation (10).

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{8}
\]

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{9}
\]

\[
F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}
\]

The chi-square test has been conducted to determine the proposed method's statistical performance. The chi-square test \(X^2\) evaluates the degree of correlation between the features as given Equation (11), where 'O' is observed values and 'P' is the predicted values. The high value of the chi-square test (0.891) for the proposed method indicates that the alternative hypothesis and the chosen features are dependent on others.

\[
X^2 = \sum \frac{(O - P)^2}{p} \tag{11}
\]

In this research, the statistical validity of the chi-square test has been estimated using the confidence interval. The confidence interval quantifies the uncertainty of the prediction and the margin of error in the prediction. It has been observed that all subset of the data set rejects the null hypothesis (\(P < 0.001\)) with a 5% level of significance, implying that these subsets are appropriate for further analysis. The 95% confidence interval has been chosen in this work. The chi-square test of 0.891 indicates that the results are symmetric around the median at the 2.5th percentile and the 97.5th percentiles. Such that the confidence interval of 95% likelihood produces with the range 0.815 to 0.878.

4 | FINDINGS AND DISCUSSIONS

This section presents the cardiovascular disease prediction results of the proposed RFE-based Gradient Boosting (RFE-GB) classification algorithm with the traditional linear discriminant analysis (LDA), K-Nearest Neighbours (KNN), Decision Tree (DT), Naive Bayes (NB), and Multilayer perceptron (MLP) algorithms. The resampling method used in this study to verify the machine learning algorithms’ research results is k-fold cross-validation. The ‘k’ value in this work is set to 10. As a result, it’s often referred to as a 10-fold cross-validation resampling process. The 10-fold cross-validation approach is designed to reduce the prediction model's bias.

| S. No. | Algorithm/metrics | Accuracy | Precision | Recall | F1-score | ROC_AUC |
|-------|-------------------|----------|-----------|--------|----------|---------|
| 1     | Linear discriminant analysis (LDA) | 64.87    | 0.67   | 0.61   | 0.64    | 0.71    |
| 2     | K-Neighbours Classifier (KNN)       | 68.74    | 0.7     | 0.66    | 0.68    | 0.74    |
| 3     | Decision Tree (DT)                  | 63.58    | 0.65   | 0.62   | 0.63    | 0.64    |
| 4     | Naive Bayes (NB)                    | 58.3     | 0.74   | 0.27   | 0.4     | 0.69    |
| 5     | Multi-Layer Perceptron (MLP)        | 76.42    | 0.72   | 0.72   | 0.72    | 0.79    |
| 6     | Gaussian Naive Bayes algorithm [41] | 83       | 0.84   | 0.83   | 0.83    | 0.83    |
| 7     | Backpropagation algorithm [41]      | 74       | --     | --     | --      | --      |
| 8     | Neural Network (Reddy et al., 2019) | 85.29    | 0.773  | --     | --      | --      |
| 9     | Random Forest (Hasan & Bao, 2021)  | 0.712    | 0.68   | 0.67   | 0.68    | --      |
| 10    | AdaBoost (Rani et al., 2021)       | 82.12    | 81.02  | 79.85  | 80.43   | --      |
| 11    | AutoML (Padmanabhan, Yuan, Chada, & Nguyen, 2019) | 0.74 | --     | --     | --      | 0.8     |
| 12    | Proposed RFE-GB                     | 89.78    | 0.86   | 0.84   | 0.83    | 0.84    |
4.1 Performance evaluation

The effectiveness of machine learning prediction algorithms is often measured using a set of classification algorithm-based metrics. The prediction error rates are quantified using this study’s mean square error, RMSE, and Kappa score (Friedman, 2001). The confusion matrix and receiver operating characteristic area under the curve (ROC AUC) are used to analyse the predictions’ true/false positive/negative rate (Theerthagiri et al., n.d.). The machine learning algorithms’ prediction performance is measured by prediction accuracy, precision, recall, and f1 score (Mohanty et al., 2007; Patro et al., 2020).

Importantly, the accuracy of the above-mentioned machine algorithms is determined in this study (whether the patient has cardiovascular disease or not). Each classification model has a distinctive disease prediction accuracy and efficiency over other prediction models based on its hyperparameters. Seventy percent of the dataset is utilized for training, whereas 30% of the data samples are utilized to test classification methods in this study. The proposed disease classification reports of the RFE-GB algorithm are compared with traditional machine learning models, and it is presented in Table 2.

Table 2 summarizes the performance results, namely prediction accuracy, precision, recall, and F1-score. The proposed RFE-based gradient boosting algorithm has higher accuracy, precision, recall, and F1-score as 89.78, 0.86, 0.84, and 0.83, respectively. In contrast, other machine learning algorithms produce lower results as compared with the proposed RFE-GB Algorithm. Such that the proposed RFE-based gradient boosting algorithm produces the prediction accuracy of 89.78%, whereas the other machine learning classification algorithms LDA, k-nearest neighbours, decision tree, Naive Bayes, multilayer perceptron, Gaussian Naive Bayes algorithm [41], Backpropagation algorithm [41], Neural Network (Reddy et al., 2019), RF (Hasan & Bao, 2021), Adaboost (Rani et al., 2021), and AutoML (Padmanabhan, Yuan, Chada, & Nguyen, 2019) resulted with 64.87%, 68.74%, 63.58%, 58.3%, 76.42%, 83%, 74%, 85.29%, 0.712%, 82.12%, and 0.74% of respective prediction accuracy only. Thus, hypothesis-1 has been proved with the improved overall accuracy for the RFE-GB algorithm as compared with other ML algorithms.
Among that, the Naive Bayes algorithm performs worst with the accuracy of 58.3%, precision, recall, F1-score of 0.74, 0.27, 0.4, respectively. Figure 4 gives the accuracy of the proposed RFE-GB and other existing algorithms. It can be seen that the proposed RFE-based gradient boosting algorithm predicts the cardiovascular disease patients about 90% (based on patient’s age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol intake, and physical activity) more accurately than the other algorithms.

Figure 4 clearly states that the proposed RFE-GB algorithm has the highest accuracy of 89.78. The proposed RFE-GB algorithm operates by looking for a subset of features in the training dataset, beginning with all the features and successfully deleting them until the target number of features. The cardiovascular disease dataset has been tested with several subsets of features. It selects the popular features from the dataset to

**TABLE 3** Error performance metrics

| S. No. | Algorithm/metrics                                      | MSE     | RMSE     |
|--------|--------------------------------------------------------|---------|----------|
| 1      | Linear discriminant analysis (LDA)                     | 0.35129 | 0.59269  |
| 2      | K-Neighbours classifier (KNN)                         | 0.31257 | 0.55908  |
| 3      | Decision tree (DT)                                    | 0.36414 | 0.60344  |
| 4      | Naive Bayes (NB)                                      | 0.41693 | 0.6457   |
| 5      | Multi-layer perceptron (MLP)                          | 0.27579 | 0.52515  |
| 6      | Gaussian Naive Bayes algorithm (Patro et al., 2020)   | –       | –        |
| 7      | Backpropagation algorithm (Patro et al., 2020)        | –       | –        |
| 8      | Proposed RFE-GB                                       | 0.19243 | 0.43866  |
classify cardio and non-cardio patients with reduced errors. As a result, the proposed RFE-based gradient boosting technique outperforms the other classification rate methods. As a result, the proposed RFE-GB algorithm outperforms the LDA, KNN, DT, NB, and MLP algorithms by 13.36% to 31.48%. Figure 5 gives the performance of precision and recall. It can be seen that both metrics have the highest scores as 0.84 and 0.83, than others. Further, the F1-score of the proposed RFE-GB algorithm has 11% to 43% of improved results.

**FIGURE 8** Normalized confusion matrix of ML algorithms.
Cohen's kappa score for the proposed and existing machine learning algorithms is depicted in Figure 6; the proposed RFE-GB method has a higher kappa score than traditional methods, as seen by the graph. Cohen's kappa score assures that the classification algorithm's predictions are consistent (Theerthagiri et al., 2021). Furthermore, it indicates that among the experimented algorithms, the proposed RFE-GB algorithm generates the highest consistency (Kappa score) of 0.57; whereas 0.3973, 0.3125, 0.5112, 0.3770, and 0.3213 are the values for the KNN, Naive Bayes, extra trees, decision trees, and radial base function, respectively.

Table 3 shows the MSE, and RMSE values for the various machine learning techniques. The MSE error rates of the LDA, KNN, DT, Naive Bayes, MLP, and proposed RFE-GB algorithms are, respectively, 0.35129, 0.31257, 0.36414, 0.41693, 0.27579, and 0.19243. It has the lowest error rate of 0.1924 for predicting correct cardiac disease instances when compared to the other algorithms (Figure 7). The proposed RFE-GB algorithm boosts the weak features from the higher-ranked and selected features in the training and testing dataset. As a result, error rates are reduced. Likewise, the RMSE error rate of the proposed RFE-GB algorithm is also very low (0.43), as illustrated in Table 3; whereas the error rates for other algorithms are LDA (0.59), KNN (0.55), DT (0.60), NB (0.64), and MLP (0.52).

The confusion matrix for various machine learning algorithms is shown in Figure 8. The predicted values and the real values are represented in terms of true positives/negatives and false positives/negatives in this graph. The proposed RFE-GB algorithm accurately estimates 88% (true positive) of cardio cases, with just 12% (false positive) misclassification; for the non-cardio cases, the RFE-GB algorithm gives 16% (false negative) of misclassification and 84% (true negative) of precise classification as illustrated in Figure 8f. Similarly, Figure 8a-e depicts the confusion matrix of LDA, KNN, NB, DT, and MLP algorithms, respectively, with lower true positives/negatives and false positives/negatives rates than the RFE-GB algorithm.

In the form of a ROC area under curve, Figure 9 depicts the relationship between the false positive rate and the true positive rate. As opposed to LDA (0.71), KNN (0.74), NB (0.69), DT (0.64), and MLP (0.79) algorithms, the RFE-GB algorithm generates the highest value of 0.84. It proves hypothesis-1 with the improved overall ROC than other ML algorithms. These results prove that the proposed RFE-GB algorithm accurately classifies cardiovascular disease patients based on their health records.

### 4.2 Result analysis and its implications

The major goal of this research is to provide a machine learning framework for dynamically developing clinical prediction models from the patient survival data to support clinicians in their decision-making. The performance results of the proposed RFE-GB algorithm have been compared and depicted in Figures 4, 5, 6, 7, 8, and 9, and Tables 2 and 3. The proposed RFE-GB algorithm has achieved the highest accuracy of 13.4%–31.5% as compared to the above-mentioned algorithms. Since the proposed RFE-GB algorithm operates by looking for a subset of features in the training dataset, it begins with all the features and successfully deleting them until the target number of features. The cardiovascular disease dataset has been tested with several subsets of features. It selects the popular features from the dataset to classify cardio and non-cardio patients with reduced errors. As a result, the proposed RFE-based gradient boosting technique outperforms the other classification rate methods. The precision and recall of the proposed RFE-GB algorithm have been improved by 12%–21% and 12%–57% than other compared algorithms.

Further, the AUROC is improved by 20%, and MSE reduced by 22% for the proposed RFE-GB algorithm. Thus, during the training and testing phase, the proposed RFE-GB method identifies weaker features with higher ranking and then chooses features for the proposed model. The performance of the RFE-GB approach illustrates its value in clinical prediction forecasting and prognostic health research.

The development of RFE-GB technology is expected to make physicians more accessible and accelerate the clinical diagnosis research discovery process. The research findings clearly imply that the proposed RFE-GB method is a potential technique for physicians to efficiently develop...
competitive models using machine learning models for the diagnosis. Physicians may desire to enhance classification accuracy while keeping the classifier’s sensitivity above a specified level. Similarly, the proposed RFE-GB method improves accuracy while preserving sensitivity.

This discovery is likely to influence how biomedical researchers and clinicians view towards machine learning. Although the focus of this work is on a cardiovascular illness dataset, the primary findings of RFE-GB’s efficiency and efficacy may apply to other biological datasets. Furthermore, as a clinical implication of the proposed technique, the clinicians might be greatly supported in diagnosing CVD patients by data analysis.

5 | CONCLUSIONS

It is worth researching much of what is required to effectively forecast and diagnose any disease using machine learning. This work RFE-based gradient boosting algorithm has been proposed to select the most important features from the cardiovascular disease dataset. The RFE-GB algorithm selects three optimal features: blood pressure, cholesterol, and physical activity, from the 12 features. Adopting these three features, a gradient boosting ensemble approach has been developed to predict cardiovascular disease cases. The proposed RFE-GB algorithm has been evaluated with various metrics, and its performance results are compared with exploring different machine learning algorithms. Among that, the proposed RFE-GB algorithm has 13.36%–31.48% of improved accuracy as compared to LDA, KNN, DT, NB, and MLP algorithms. This proposed technique using machine learning approaches in clinical practice is easier to understand and validate for physicians, increasing their adoption.

Further, it produces higher consistency (Kappa score) of 0.57 with a reduced error rate MSE of 0.1924 on the prediction of accurate cardio disease cases. The proposed RFE-GB algorithm accurately estimates 88% true positives and 84% of true negatives from 70,000 patient records with the AUROC score of 84%. As a consequence of the findings, the proposed RFE-GB algorithm appears to diagnose and classify diabetes patients. The proposed algorithm can be analysed and evaluated with the optimization algorithms. Considering it as a limitation, future work will concentrate on improving the results of this work. Further, various combinations of machine learning approaches can be used for the future plans of this research.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

ORCID

Prasannavenkatesan Theerthagiri https://orcid.org/0000-0003-3420-598X

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

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

Cardiovascular Disease Dataset. (2021). Retrieved from Kaggle Repository https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

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, 178.

Chen, Q., Meng, Z., & Ran, S. (2020). WERFE: A gene selection algorithm based on recursive feature elimination and ensemble strategy. *Frontiers in Bioengineering and Biotechnology*, 28, 496.

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, 2183–2189.

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

Kakulapati, V., Kirti, A., Kulkarni, V., & Raj, C. P. (2017a). Predictive analysis of heart disease using stochastic gradient boosting along with recursive feature elimination. *International Journal of Science and Research*, 6(5), 909–912.

Kakulapati, V., Kirti, A., Kulkarni, V., & Raj, C. P. (2017b). Predictive analysis of heart disease using stochastic gradient boosting along with recursive feature elimination. *International Journal of Science and Research*, 6, 2319–7064.

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

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access.*, 19(7), 81542–81554.

Mohanty, R. P., Seth, D., & Mukadam, S. (2007). Quality dimensions of e-commerce and their implications. *Total Quality Management & Business Excellence.*, 18(3), 219–247.

Onan, A. (2020). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, 28(1), 117–138.
Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience, 33*(23), e5909.

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, S. E., 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.

Prasannavenkatesan, T., Jeena Jacob, I., Usha Ruby, A., & Vamsidhar, Y. (2021). Vamshidhar “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.

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. In *Concurrency and computation: Practice and experience* (p. e6248).

Theerthagiri, P., Gopala Krishnan, 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.

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.

Zhang, P., Padhy, N., & Chiranjevi, D. (2020). Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. *Evo-Intelligence, 14*, 941–969.

AUTHOR BIOGRAPHIES

Prasannavenkatesan Theerthagiri is working as the Assistant Professor in the Department of Computer Science and Engineering, GITAM Deemed to be University, Bengaluru, India. He was awarded PhD (Full-Time) Research degree in the year 2021 on the work of wireless communication with machine learning from Anna University, Chennai, India. He was awarded the Mobility grant award by the Republic of Slovenia in the year 2017-2018. He has published his research works in 11 SCI-indexed journals, 15 SCOPUS indexed journals, 10 national/international conferences, 7 book chapters in Springer, Wiley, Emerald, Taylor & Francis etc. publisher. He has 4 years of research experience and 2 years of teaching experience. Also, he has submitted 14 research project proposals on various themes to several external funding agencies like ICMR, DST, MoES, SERB, UN Economic council, etc. He is a Member of Institutions of Engineers India.

Vidyajothiprakash is an Assistant Professor in the Department of Computer Science and Engineering at GITAM University, Bengaluru, Karnataka, India. She is currently pursuing her PhD in healthcare with machine learning algorithms. She has 6 years of teaching experience. Her area of interest includes Internet of Things and Machine Learning.

How to cite this article: Theerthagiri, P., & Vidyajothiprakash, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert Systems, 39*(9), e13064. [https://doi.org/10.1111/exsy.13064](https://doi.org/10.1111/exsy.13064)