A novel ensemble approach for predicting heart disease with dimension reduction

SukanyaWattal*, Swati Gupta, Harkiran Kaur

Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology,
Patiala, Punjab 147004, India

⁎ *Corresponding authors* wattalsukanya1512@gmail.com, swattal_be17@thapar.edu

Heart disease (often used interchangeably as cardiovascular disease) is one of the leading causes of mortality globally. More people die annually from cardiovascular diseasesthan from any other cause. Conventional methods used for predicting the existence of heart disease have certain limitations, including low predictive accuracy, and high time and space complexities. Thus, the current global situation creates a demand for an urgent solution to this widely and rapidly growing chronic disease. A novel hybrid approach is proposed and implemented in this paper that combines the gradient boostingensemble approach together with dimension reduction algorithm, principal component analysis, to improve the performance of classification results efficiently. Dimension reduction significantly reduces features to be inputted in the ensemble machine learning model which in turn reduces the time and space complexity of the model.Thus, a reduced number of features are fed to the ensemble machine learning model, which combines outcomes of various machine learning techniques, resulting in the design of the better predictive model.Results indicate that heart disease can be more efficiently predicted by the proposed model with an overall accuracy of 93.44%. Also, the proposed model is more practical and scalable as compared to other models as it minimizes the number of input features from 13 to 4 and thus reducing the space required to store the data and producing an optimal outcome.

Keywords: Heart disease; principal component analysis; dimension reduction; classification; gradient boosting; ensemble ML.

1. Introduction

Heart disease (Cardiovascular Disease (CVD)) is one of the most commonly arising fatal diseases inthe world, these days.According to the World Health Organization(WHO), 23.6 million people are likely to die from heart disease and stroke by 2030 [1].Heart diseases can be categorized asblood vessel diseases (Coronary Artery Disease(CAD)),heart rhythm problems (arrhythmias) and heart defects a person is born with (congenital heart defects) [2].The occurrence of heart disease is mostly found in people of ages 50 years and above, with men and post-menopause women being affected equally likely [1].Taking note of the current worldwide situation, there must be a highly accurate system that detects this widely growing chronic disease at an early stage among patients.

Many risk factors are associated with heart diseases including two classes, viz. preventable and non-preventable risk factors. Non-preventable risk factors include age, gender, family history and heart disease you are born with whereas the preventable risk factors include unhealthy cholesterol level, physical inactivity that may also lead to obesity, heavy smoking [3], poor diets and drinking too much alcohol. Other factors such as diabetes and hypertension have a great impact on the proper functioning of our heart [4].Exercising helps to maintain the blood glucose level in diabetic patients;and adopting a healthy lifestyle, having a balanced diet and quitting smoking can reduce the risk of heart disease [5].

Angiography is one of thefinesttests,which is most widely used for the evaluation of coronary arterydisease (CAD) [6].It uses X-rays to study whether our bloodvessels are narrow, enlarged, blocked or deformed.However, emission of radiations, high cost and time are major constraints for using angiography for the prediction of heart disease [7].Further,this test needs to be performed under the physician’s supervision and hence is prone to human error and istime-consuming. Therefore,the need of the hour is to develop an efficient and fast diagnosis system for predicting the existence of heart disease in humans.

Conventional methods including complex deep learning algorithms used for the prediction of heart disease use large space for storing the healthcare data and involve high computational time. Also, usage of irrelevant features for prediction may produce wrong results leading to low accuracy of the model [8].The proposed techniques address all these limitations by using dimension reduction algorithms (as we reduced the information required for the prediction of the disease) and gradient boosting ensemble algorithm. These proposed techniques henceimprove the accuracy of the model significantly,andprominently reduces the space occupied by the data and the time required by the model to process the data.

Healthcare informatics offer large pools of datasets to extract useful informationfrom them using modern approaches such as data mining and machine learning and make required predictions and decisions [9].This paper presents a novel method, a hybrid of ensemble approach and dimension reduction, which distinguishes the heart disease affected people from healthy ones efficiently.For this proposed model, dimension reduction is performed on the input features of the dataset, before applying machine learning classification models for predicting the heart disease. Thedimension reduction step introduced in the first phase significantly reduces the number of features and thus improves the overall time and space complexity of the classification model;andtheapplication of gradient boosting ensemble model enhances the prediction accuracy, precision, recall and f1-score ofthe heart disease prediction model.

For avoiding the time consuming and costly disease diagnosing processes such as angiography for CAD, researchers are prompted to introduce alternative methods such as the usage of machine learning algorithms for diagnosing diseases. The healthcare industry provides data with a huge number of features. Hence various machine learning techniques have been proposed to predict heart diseases, calculating drug synergy scores and cancer research.

Several methods have been proposed to predict heart diseases. Das et al. [10] used a Statistical Analysis System (SAS) based software 9.1.3 for diagnosing heart disease and appliedthe multi-layer feedforward neural network type of architecture for prediction where three independentneural network models were used for classification. Further, averaging ensemble technique was applied to the results produced by the three models to generate the outcome leading to an accuracy of 89.01%. KhemphilaandBoonjing [11] introduced an Artificial Neural Network(ANN) algorithm for heart disease classification and calculated information gain for selectedfeatures which contributed significantly to the heart disease dataset. Application of Multi-Layer Perceptron (MLP)with back-propagation learning algorithm along with feature selectionincreased the computational efficiency,decreased the time of computation and also increased the prediction accuracy to 80.99% for data validation and hence, improved the existing methodologies. Paul et al. [12] proposed a heart disease diagnostic system which used Fuzzy Decision Support System (FDSS), built from the generated fuzzy knowledge base, for determining the level of risk of heart disease. Significant attributes were selected from the feature set which formed the basis of generation of weighted fuzzy rules using genetic algorithms. Verma et al. [13] proposed a hybrid classification model for the prediction of heart disease constituting machine learning algorithms such as MLP, Multinomial Logistic Regression (MLR), Fuzzy Unordered Rule Induction Algorithm (FURIA) and C4.5 which produced an accuracy of 88.4%. Dwivedi [14] compared the heart disease prediction accuracy of some well-known classification algorithms such as Naïve Bayes (NB), classification tree, K-Nearest Neighbours (KNN), Artificial Neural Networks (ANN), Logistic Regression (LR) and Support Vector Machine (SVM), among which LR outperformed with an accuracy of 85%. Samuel et al. [15] applieda hybrid of ANN and Fuzzy Analytic Hierarchy (Fuzzy-AHP) modelson an online clinical dataset with 297 samples for the heart failure risk prediction. Upon comparison with the conventional ANN, their proposed model turned out to be better in terms of prediction accuracy, Receiver Operating Characteristic(ROC) plot and performance plot. Shah et al. [16] introduced a model for the heart disease prediction which was based on Probabilistic Principal Component Analysis(P-PCA), used for feature extraction, and Radial Basis Function (RBF) kernel-basedSVM, used for classification with a prediction accuracy of 82.18%. In 2018, Paul et al. [17] introduced an adaptive weighted fuzzy-rule based system based on genetic algorithm and a modified dynamic Multi-Swarm Particle Swarm Optimization (MDMS-PSO) for calculating the risk level of heart disease which produced an accuracy of92.31%.

Amin et al. [18] presenteda heart disease prediction model using significant features and compared the performance measures of seven classification algorithms: KNN, SVM, LR, Neural Network, NB, Decision Tree(DT) and Vote (a hybrid of LR and NB) where highest prediction accuracy of 87.4% was achieved in case of Voting classification technique. Recently, Ali et al. [19] introduced a model which used a hybrid of two SVMs; one for selecting the significant features and the other for the prediction of heart disease, both of which were individually optimized by the proposed Hybrid Grid Search Algorithm (HGSA)and further a prediction accuracy of 92.22% was achieved. In the most recent study, Javeedet al. [20] used Random Forest (RF) model optimized by grid search algorithm for differentiating the people who had heart disease from those who did not with a prediction accuracy of93.33%, 3.3% higher than the accuracy of conventional RF model. Singh et al. [21] proposed an ensemble-based approach for the prediction of drug synergy in cancer which combined the results of techniques such as RF, Fuzzy Rules Using Genetic Cooperative-Competitive Learning method (GFS.GCCL), Adaptive-Network-Based Fuzzy Inference System (ANFIS) and Dynamic Evolving Neural-Fuzzy Inference System (DENFIS) through weighted averaging ensemble technique. On comparison of the performance measures such as accuracy, Root Mean Square Error (RMSE) and coefficient of correlation of the proposed model with the existing models, it was found that the proposed model outperformed the others.Bernardini et al. [22] used Multiple Instance Learning Boosting algorithm (MIL-Boost) which is a supervised machine learning approach for prediction of Type 2 Diabetes and KNN for pre-processing of data. On comparing the results with conventional machine learning methods like DT, RF, KNN, Boosting, SVM Lin, SVM Gauss and SVM Lasso, MIL-Boost turned out to be better (recall from 0.70 up to 0.83).

The major limitation in the above methods is the presence of high dimensional data. This may cause classifiers to have low accuracy, because of redundant data, and high computation time. Hence, to address these limitations, this research study has proposedand developed a method to predict heart disease with high accuracy.

Contributions: The main contributions in this paper are as follows:

(1) Selectingthefeatures influencing the existence of heart diseasefrom the standardised datasets [23, 24].
(2) Design and development of an ensemble model to predict the existence of heart disease using machine learning classification models.
(3) Applying dimension reduction algorithms namely Principal Component Analysis(PCA), Kernel-PCA (K-PCA) and Linear Discriminant Analysis(LDA) on the Cleveland heart disease dataset [23] and attainment of a reduced number of features required to be fed to the ensemble machine learning model;improve its accuracy,and improve its time and space complexity.
(4) Improving the ensemble model designed in (2) by using the reduced set of features achieved during dimension reduction in (1).
(5) Analysing the performance of the ensemble model and the proposed hybrid ensemble model (with dimension reduction) to assist the heart disease prediction.

The paper has been structured as follows. The comparison between the existing methods and the proposed method is described in Sec. 2. The dataset and the evaluation metrics used in this paper are described in Sec. 3. Experimental results and discussions, comparisons between the performance metrics of models before and after dimension reductionand receiver operating characteristic (ROC) curve are described in Sec. 4.

2. Materials and method

Numerous experiments using machine learning techniques have been conducted to predict the existence of heart disease in humans. In this section, we review the existing methods and comparisons are drawn between theexisting predictive methods and the proposed method.

2.1. Existing methods

Machine learning in the medical field is gaining fervent attention. Consequently, many research teams are making their datasets available for public use to facilitate the growing interest in this field. The Cleveland heart disease dataset has been made available at UCI (University of California, Irvine) Machine Learning Repositoryusing which several researchers have made contributions for the prediction of heart disease.Some of the well-known classifiers, such as random forest, KNN [25], SVM [26],DT and LRhave been used for heart disease diagnosis. To reduce the influence of noise on data, ensemble learning methods are often applied. An ensemble learning algorithm is executed through multiple sets of machine learning algorithms and collects the predictive results provided by different classifiers [27].

Researchers showed experimentally that class-imbalanced dataset hampers the fulfilment of traditional classifiers [28]. Class-imbalanced datasets are the ones which have a large count of instances of one class and a lesser count of instances of the other. This imbalance leads to a biased classification towards the class in the majority and therefore the traditionally used classification algorithms like Decision Tree (DT) do not fit on such data [28].Hence, ensemble methods like Gradient Boosting Classifier (GBC), AdaBoostandRandom Forest are significant in such scenarios as in these methods; more than one DT is constructed for prediction of results.

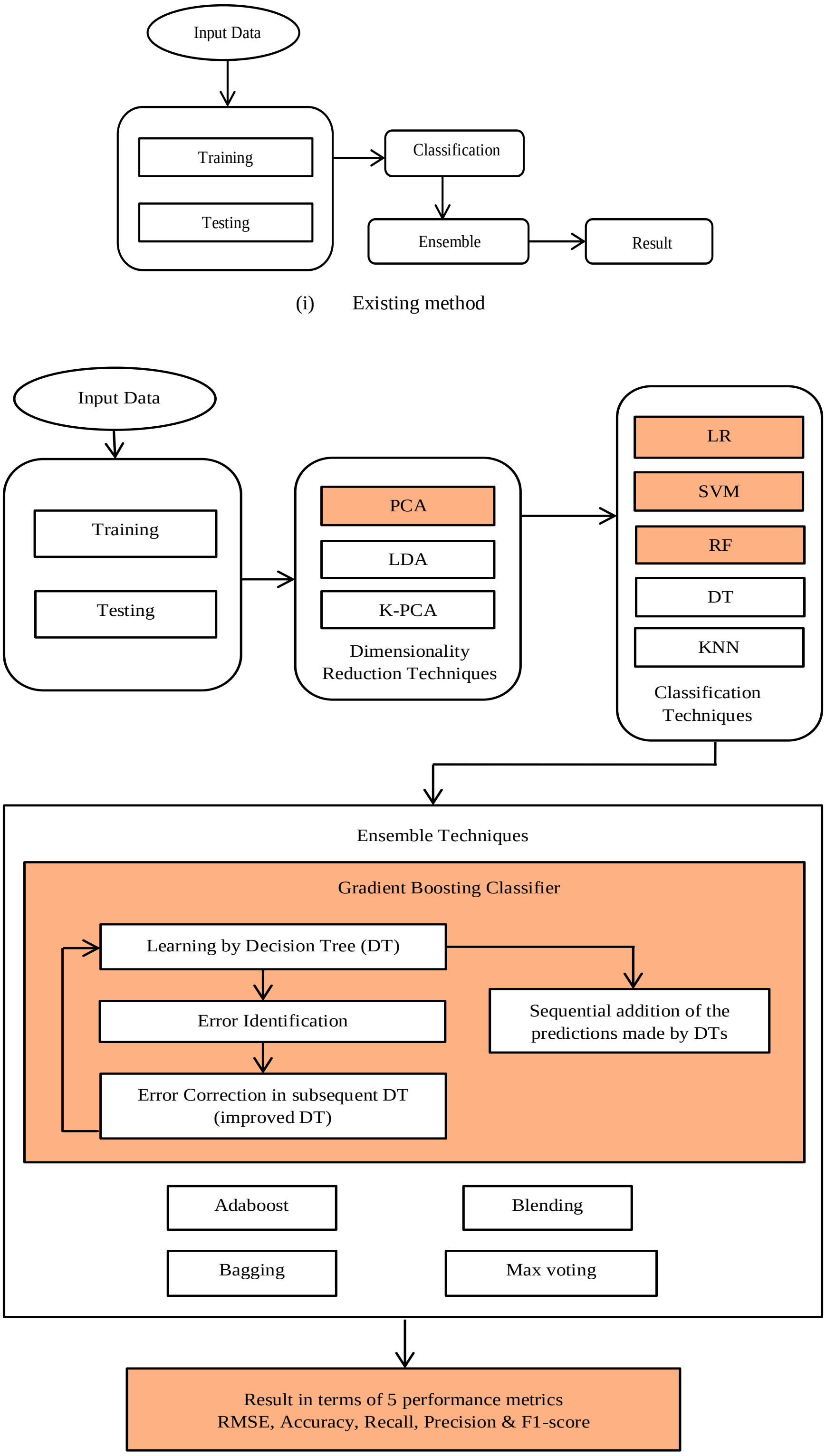
The aforementioned existing methods have been presented in Fig. 1 (i). The dataset is divided into training and testing data, various classification techniques are applied whose results are combined by ensemble algorithms to produce the outcome. However, these studies classify the datasets with no constraint on the time and space complexity. Therefore,we used dimension reduction as the key module to reduce time and space complexity as shownin Fig. 1 (ii). The dimension reduction module maps high dimensional data into low dimensional space by preserving essential features as possible [29] and removing the non-influencing features. Its essence lies in the idea of facilitating data visualization, improving the quality of data features, shortening the training time, reducing overfitting and improving the prediction accuracy of conventional learning methods like classification, clustering etc. Gradient Boosting Classifier is applied for predicting the heart disease and the results are returned in terms of RMSE, accuracy, recall, precision and f1-score. The highlighted parts in Fig. 1 (ii) are the major contributions in the proposed model.

2.2. Proposed methodology

In this research study, the prediction model has been upgraded in terms of ensemble approach, dimension reduction module; improving accuracy, and improving the time and space complexity of ensemble approach on the reduced set of dimensions.Considering the field of Artificial Intelligence (AI), the improvement in performance metrics like prediction accuracy; and time and space complexity increases the overall value of the proposed heart disease prediction system.

2.2.1. Dimension reduction

With an increase in the number of features in a dataset, the information brought to people also increases proportionately but at the same time, a dataset with a large number of features may include a huge amount of noise and redundancy. Due to the high dimensionality of a dataset, certain algorithms struggle to train models effectively [30].


Input Data
Training
Testing
Classification
Ensemble
Result
Input Data
Training
Testing
PCA
LDA
K-PCA
Dimensionality Reduction Techniques
LR
SVM
RF
DT
KNN
Classification Techniques
Ensemble Techniques
Gradient Boosting Classifier
Learning by Decision Tree (DT)
Error Identification
Error Correction in subsequent DT (improved DT)
Sequential addition of the predictions made by DTs
Adaboost
Blending
Bagging
Max voting
Result in terms of 5 performance metrics
RMSE, Accuracy, Recall, Precision & F1-score

(i) Existing method

(ii) Proposed method

Fig. 1. Comparison between existing and proposed methods.

Dimension reduction techniques transform the data from high dimensional feature space to low dimensional feature space; assists in reducing redundancy, noise, the time and space complexity of learning algorithms; and improves the accuracy of classification [8]. Dimension reduction techniques can be categorized as feature selection and feature extraction. In feature selection, features which contribute majorly to the dataset are selected while feature extraction can be considered as applying non-binary weights on the original features to construct new ones [29].These algorithms find a set of features that are most effective for classification to optimize classification performance through mathematical transformation. The most important role played by an appropriate dimension reduction algorithm is to retrieve more effective information from low dimensional data which was earlier hidden in high dimensional spaces [31].In this paper, we have applied feature extraction dimensionality reduction techniques such as PCA, K-PCA and LDA for heart disease prediction.

Principal Component Analysis (PCA) being an ideal algorithm is most commonly used for reducing the number of data attributes.It aids the computational speed of training algorithms and also improves the accuracy manifolds. This algorithm uses mathematical transformations to extract the principal components and prioritizes the features according to their contribution to the characteristics of the data, that is, ranks the features according to their explained variance [32]. The attribute with the highest Eigen-value is returned as the first principal component, the second-highest Eigen-valued feature is returned as the second principal component and so on. These principal components are orthogonal to each other, that is, they are not correlated.

The characteristic property of being a supervised learning algorithm itself differentiates Linear Discriminant Analysis (LDA), an unsupervised learning algorithm, from PCA [32].Considering the working of LDA, maximizing the separability between classes comes under the spotlight, which makes it mandatory to take labelled data as input for the learning process. The number of linear discriminants returned by the algorithm is one less than the number of classes in the labelled dataset. K-PCA is an extension of PCA which maps non-linear data points to a higher dimensional feature space where they can be separated linearly [33].

2.2.2. Ensemble techniques

Ensemble techniques are meta-algorithms that combine the outcomes of various single and unstable machine learning classification models,which otherwise have low classification accuracy rates when applied individually. The process of combining the results of the classifiers is carried out by methods such as averaging, weighted averaging, blending, boosting, max-votingetc. [34].In classification algorithms, features of datasets act as the variables of an equation. The other side of the equation has a target which consists of classes in which the instances will be grouped. Therefore, the dataset is split into training and testing data while training sets will have targets or labels and testing sets will not. Ensemble classifiers are better than single classifiers because they are more stable, decrease variance (bagging), bias (boosting) and predict strongly (stacking) than single classifiers [35].On applying different ensemble techniques on the heart disease dataset, GBCemergedto be the one with the highest prediction accuracy.

Gradient Boosting is a machine learning algorithm which is used for both classification and regression. The idea of gradient boosting originated in the observation byBreiman[36] that boosting can be interpreted as an optimization algorithm on a suitable cost function. Gradient Boosting Classifier typically uses DTs and produces a strong prediction model in the form of the conglomeration of weak learning models [37].It uses a base weak learner and improves the performance by continuously moving hub towards complex observations that were difficult to predict. The constructive strategy of sequential weighted addition of trees while correcting the errors of the previous ones buildsone of the most powerful prediction modelsas compared with other models [38].Gradient boosting ensemble model is a slow learning algorithm which builds a model in a stage-wise fashion with minor changes in every subsequent step leading to improved accuracy. It provides a lot of flexibility, can handle missing data and avoids overfitting [39]using cross-validation. GBCdoes not require cleaning of the dataset and can learn complex non-linear decision boundaries via boosting. Linear classifiers such as LR, linear SVM etc. require careful cleaning and pre-processing of the dataset which makes GBCone of the most convenient techniqueto use among all other models [40]. Dataset cleaning includes missing value imputation, scaling features, outlier detection, selecting features to avoid overfitting due to collinearity and non-linear decision boundaries. A linear classifier’s performance is highly impacted by the imputation method chosen but in case of GBC, the missing value can simply be inputted by a very large or very low value which is not present in the dataset. In the case of DTs, if two features are strongly correlated, one of the features is arbitrarily chosen and the other has no further impact on the model. The limitation of a single DTover GBCis that overfitting can be minimized by limiting the maximum depth of the tree but thereafter it cannot learn a complex decision boundary. GBC learns more complex decision boundaries as compared to RFby successively fitting trees on the error of the previous one. Ensemble techniques such as averaging, weighted averaging, blending, bagging, stacking, max voting and AdaBoost neither sequentially add classifiers nor remove errors in prediction. The above-drawn comparisons show that GBCis better than the other existing classification and ensemble techniques (Table 1).The coloured cells indicate that the data is not available.

Table 1. Comparing proposed ensemble techniquewith other algorithms.

	Ability to fit on the class-imbalanced dataset	**Method of the imputation of missing values in the dataset**	**Ability to learn complex decision boundaries**	**Sequential addition of prediction by classifiers and removal of errors**
GBC	Present	Simple	High	Present
Linear classifiers (LR, Linear SVM, etc.)		Complex		Absent
DT	Absent		Low	Absent
RF	Present		Low	Absent
Max-voting				Absent
Averaging				Absent
Weighted-averaging				Absent
Blending				
Bagging				
AdaBoost	Present			
Stacking				

3. Performance analysis

3.1. Dataset

The Cleveland database [23] contains a total of 76 attributes but all experiments refer using 14 attributes only. The “goal” field is integer-valued from 0 to 4. Experiments conducted so far simply concentrate on distinguishing absence (0) of heart disease from the presence (1, 2, 3 and 4) of heart disease in a person. However, the publicly available datasethas a binary “goal” field (0 = absence and 1 = presence). Security numbers and names of patients were removed from the database and replaced with synthetic values.An additional “ID” field is added in the dataset to uniquely identify the rows of data.

In this paper, all the experiments were conducted in the same experimental environment, in which the CPU was an Intel(R) Core(TM) i5-3230M CPU @ 2.60 GHz, memory was 4 GB.

Table 2 shows the number of samples and the number of attributes in the Cleveland heart disease dataset. The descriptions of the attributes are represented in Table 3.

Table 2. Cleveland Heart Disease Dataset.

Number of samples	303
Number of attributes	76
Number of attributes (public)	14

Table 3. Major attributes used in experiments.

Parameter	**Description**
Age	age of the patient
Gender	1 for male and 0 for female
resting blood pressure	millimetre of mercury (mmHg)
chest pain	4 stands for asymptomatic 3 stands for non-anginal pain 2 stands for atypical angina 1 stand for typical angina
Maxhr	maximum heart rate achieved

Exercise-induced angina	1 = true 0 = false
Thalassemia	thal = 7 means the disorder is reversible; thal = 6 means disorder is fixed; thal = 3 means normal
Cholesterol	Milligrams per decilitre

Fasting blood sugar	if fbs>120 (mg/dl)
	fbs = 1 (true)
	else fbs = 0 (false)

Num Major Vessels	the no. of major blood vessels coloured by fluoroscopy (from 0 to 3)
Slope	the slope of the peak exercise 3 stands for downsloping 2 stands for flat 1 stand for up-sloping
Peak	ST depression induced by exercise
Resting Electrocardiographic Result	2 stands for probable or definite left ventricular hypertrophy 1 stand for having ST-T wave abnormality 0 stands for normal
Goal	heart disease diagnosis 0 stands for less than .5 diameter narrowing 1 stand for greater than .5 diameter narrowing

3.2. Evaluation Metrics

In case of the class-imbalanced dataset, accuracy cannot be used as the sole measure for the efficiency of the applied machine learning model as class imbalance causes biased decision-making towards the majority class which leads to misclassification [41]. Therefore, some additional parameters such as f1-score, precision, recall and RMSEhave been considered for the proposed study to comprehend improvedassessment for the performance of classifiers [28].The tables showing a comparative analysis of the aforementioned parameters for some classification models are shown below.

A confusion matrix is a specific table layout that allows visualization of the performance of a machine learning algorithm. It has two rows and two columns; each row represents the instances in a predicted class and each column represents the instances in an actual class. A confusion matrix reports the count of occurrences of true positive, true negative, false positive and false negative values. It provides a more detailed analysis than mere accuracy. RMSE measures the difference between true values and observed values. RMSD or RMSE is the standard deviation of prediction errors.

	Predicted: No	Predicted: Yes
Actual: No	True Negative	False Positive
Actual: Yes	False Negative	True Positive

Fig. 2. Confusion matrix

$$RMSE = \sqrt{\frac{\sum_{1}^{n}(\text{predicted values} - \text{actual values})^{2}}{\text{n}}}. \tag{1}$$

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}. \tag{2}$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative}. \tag{3}$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}. \tag{4}$$

$$F1\ score = 2 \times \frac{recall \times precision}{recall + precision}. \tag{5}$$

4. Results and discussions

Experiments are conducted by applying five well-known classification techniques on the heart disease dataset, namely LR, SVM, KNN, DT and RF, out of which the best-performing three classifiers, such as LR, SVM and RF, are chosen for further analysis.

Further, ensemble techniques, such as bagging, blending and max voting, are applied on the data classified by the above-mentioned three classifiers. Gradient boosting and AdaBoost ensemble approaches are also applied for classification of the dataset.

To improve the performance of the model, the significant features of the dataset are extracted by applying three dimensionality reduction techniques, such as PCA, LDA and K-PCA. Comparisons are drawn among the applied ensemble techniques based on their performance metrics for achieving the final results.

4.1. Comparison with other models

The Cleveland heart disease dataset is composed of 76 attributes among which only 14 are available for use in research publicly. The details of the dataset and its attributes are shown in Table 2 and 3 respectively.

The dataset is partitioned into training and testing data considering various sizes (40% to 80%) of the training data. According to the Pareto Principle, the ideal size for the division of training and testing data is 80% and 20% respectively [42].The graphical representation in Fig. 3 shows that the best results are achieved with 80% training databased on RMSE (25.61%), accuracy (93.44%), recall (94.12%), precision (94.12%) and f1-score (94.12%). It may seem from the graph that the results are equal in the case of 70% and 80% training data but it is not so. The actual values of metrics for 70% training data are: RMSE equal to 25.68%, accuracy equal to 93.41%, recall equal to 93.62%, precision equal to 93.62% and f1-score equal to 93.62%.

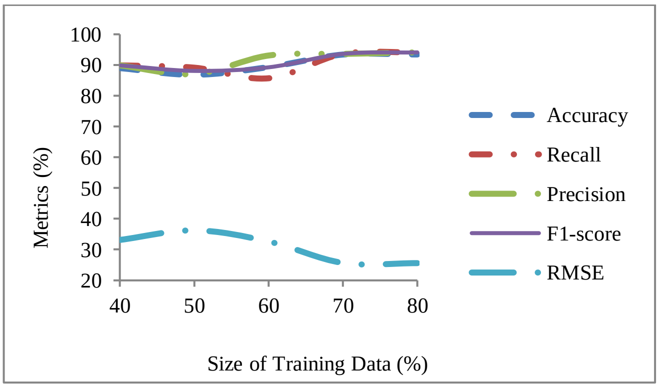


Fig. 3. Impact of varying sizes of training data on the metrics of gradient boosting classifier with data reduced by PCA (PCA-GBC).

Table 4. Evaluation metrics of classification techniques.

Model	**RMSE (%)**	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F1-Score (%)**
Logistic Regression	49.59	75.41	83.87	72.22	77.61
SVM	51.21	73.77	80.64	71.43	75.76
Random Forest	59.59	75.41	80.64	73.53	76.92
Decision Tree	52.79	72.13	74.19	71.87	73.06
KNN	67.75	54.10	74.19	53.49	62.16

Table 5. Evaluation metrics for Ensemble Classifiers before Dimension Reduction.

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Gradient boosting	40.49	83.61	91.18	81.58	86.11
AdaBoost	31.36	90.16	88.23	93.75	90.91
Bagging	38.41	85.24	91.18	83.78	87.32
Blending	52.79	72.13	73.53	75.76	74.63
Max Voting	42.46	81.97	88.23	81.08	84.51

Table 4 shows the evaluation metrics of the classifiers which aidin drawing a performance-comparison of the applied classification algorithms and brings out the best three among them, such as LR, SVM and RF. Ensemble techniques are applied to the results of these classifiers and comparisons aredrawn based on information provided in Table 5. Further, dimensionality reduction techniques are applied on the dataset and the number of features is reduced from 13 to 4 by feature extraction. As inferred from Table 6, the model gives optimal results when the number of features is equal to 4. Though themetrics are same for the number of features equal to 4, 5, 6, 7 and 8, the need to reduce the dimensions of the data makes it important to choose the least possible number of features with a decent performance.

Table 6. Evaluation metrics of PCA-GBC varying with the number of features of the dataset.

No. of Features	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
2	31.36	90.16	91.18	91.18	91.18
3	28.63	91.80	91.18	93.94	92.54
4	25.61	93.44	94.12	94.12	94.12
5	25.61	93.44	94.12	94.12	94.12
6	25.61	93.44	94.12	94.12	94.12
7	25.61	93.44	94.12	94.12	94.12
8	25.61	93.44	94.12	94.12	94.12
9	31.36	90.16	91.18	91.18	91.18
10	31.36	90.16	91.18	91.18	91.18
11	28.63	91.80	91.18	93.94	92.54
12	28.63	91.80	91.18	93.94	92.54
13	28.63	91.80	91.18	93.94	92.54

Table 7. Evaluation metrics of classifiers after applying dimension reduction techniques.

(i) Data reduced by PCA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Logistic Regression	33.87	88.52	91.18	88.57	89.85
SVM	35.61	89.43	90.11	90.13	90.23
Random Forest	31.36	90.16	88.23	93.75	90.91

(ii) Data reduced by LDA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Logistic Regression	40.49	83.61	88.23	83.33	85.71
SVM	42.49	79.61	82.18	81.58	80.11
Random Forest	47.91	77.45	79.41	79.41	79.41

(iii) Data reduced by K-PCA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Logistic Regression	42.46	81.97	85.29	82.86	84.06
SVM	43.41	81.23	82.29	81.88	82.57
Random Forest	44.35	80.33	79.41	84.37	81.82

Table 8. Evaluation metrics after applying ensemble techniques.

(i) Data reduced by PCA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Gradient boosting	25.61	93.44	94.12	94.12	94.12
Adaboost	33.87	88.52	88.23	90.91	89.55
Bagging	33.87	88.52	91.18	88.57	89.85
Blending	33.87	88.52	91.18	88.57	89.85
Max Voting	28.63	91.80	94.12	91.43	92.75

(ii) Data reduced by LDA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Gradient boosting	42.46	81.97	76.47	89.65	82.54
Adaboost	40.49	83.61	79.41	90	84.37
Bagging	40.49	83.61	88.23	83.33	85.71
Blending	40.49	83.61	88.23	83.33	85.71
Max Voting	40.49	83.61	91.18	81.58	86.11

(iii) Data reduced by K-PCA

Model	RMSE (%)	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
Gradient boosting	44.35	80.33	88.23	78.95	83.33
AdaBoost	40.49	83.61	85.29	85.29	85.29
Bagging	46.16	78.69	88.23	76.92	82.19
Blending	44.35	80.33	82.35	82.35	82.35
Max Voting	42.46	81.97	85.29	82.86	84.06

Table 7 and Table 8 summarize the performance measures of the classification techniques and ensemble techniques respectively after applying dimension reduction on the used dataset. On comparing Table 4 with Table 7, it is clear that the applied machine learning methods make better predictions on the dimensionally reduced data. On the contrary, comparing Table 5 and Table 8 show that the performance of some models has decreased on dimensionally reduced data. However, the results produced by Gradient Boosting Classifier (GBC) in Table 8 (i) are unbeatable by any of the metrics in Table 5. Table 8 (i) when compared with the corresponding entries of Table 8 (ii) and (iii), clearly shows that PCA produces results which are better than LDA and K-PCA. This is a similar case when Table 7 (i) is compared with the corresponding entries of Table 7 (ii) and (iii). The proposed ensemble model,i.e.GBC applied on the data reduced by PCA,ranks first in terms of RMSE (25.61%), Accuracy (93.44%), Precision (94.12%) and F1-score (94.12%) as represented in Table 8 (i). Though the recall value of gradient boosting (94.12%) is same as that of max voting, the best overall performer is GBC based on the requirements of the task. The same recall value of GBC can be attributed to class-imbalance in the dataset.

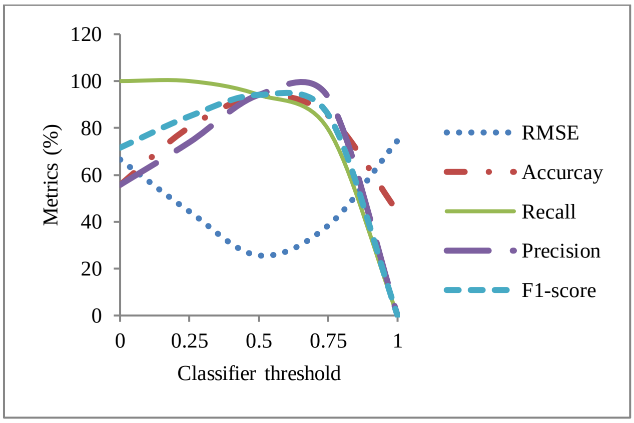


Fig.4. Impact of different thresholds on the metrics of PCA-GBC.

Next, the Receiver Operating Characteristic (ROC) curve is plotted to investigate the impact of different thresholds on the evaluation metrics of the ensemble algorithm as shown in Fig. 4. When threshold = 0, all the samples of patients were classified as 1 i.e. presence of heart disease. Therefore, the recall value reaches 100%. At the same time, accuracy = 55.74% indicates that the number of patients diagnosed with the presence of heart disease accounts for 55.74% of the total testing samples. As the threshold increases, recall decreases smoothly to 0%. Accuracy and RMSE achieve their maximum (93.44%) and minimum value (25.61%)respectively at threshold = 0.5 after which they decrease and increase respectively. Precision and F1-score first increase with an increase in threshold until somewhere between 0.5 to 0.75 threshold value after which they drop down to 0% at threshold = 1.

The confusion matrix of PCA-gradient boosting model, [[25 2],[2 32]], shows an increase in the true negative and true positive values while a decrease in the false negative and false positive values which indicate an accuracy of 93.44%. The results confirm that the proposed method can achieve a remarkable classification performance.

Conclusion

In this study, we examined multiple approaches for prediction of heart disease and proposed a novel and efficient method for the same. The proposed model is a dimensionally reduced ensemble-based heart disease prediction technique, a hybrid PCA-GBC model that outperforms all the other models in terms of accuracy, precision, RMSE, recall and F1-score.In addition to this comparison, the proposed method is more practical and scalable. The results convey that machine learning techniques can replace the manual diagnosis of diseases and hence have a lot of potential in the medical field. Presently, the model predicts the target as 0 or 1, i.e., whether the person is having a risk of heart disease or not, but in the future study, this investigation would extend to focus on the percentage of narrowing blood vessels and the risks associated with it.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research study did not receive any specific grant from funding agenciesin the public, commercial, or non-profit sectors.

References

[1] WHO | Cardiovascular Diseases (CVDs). https://www.who.int/cardiovascular_diseases/about_cvd/en/.

[2] S. Mendis, P. Puska, B. Norrving, Global Atlas on Cardiovascular Disease Prevention and Control (PDF), World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, 2011, 3–18.

[3] M. E. Piché, P. Poirier, I. Lemieux, J .P. Després, Overview of Epidemiology and Contribution of Obesity and Body Fat Distribution to Cardiovascular Disease: An Update, Progress in Cardiovascular Diseases 61 (2) (2018) 103-113, https://doi.org/10.1016/j.pcad.2018.06.004.

[4] David Jiménez-Pavón, C. J. Lavie, S. N. Blair, The role of cardiorespiratory fitness on the risk of sudden cardiac death at the population level: A systematic review and meta-analysis of the available evidence, Progress in Cardiovascular Diseases 62 (3) (2019) 279-287, https://doi.org/10.1016/j.pcad.2019.05.003.

[5] A.Younus, E. C. Aneni, E. S. Spatz, C. U. Osondu, L. Roberson, O. Ogunmoroti, R. Malik, S. S. Ali, M. Aziz, T. Feldman, S. S. Virani, W. Maziak, A. S. Agatston, E. Veledar, K. Nasir, A Systematic Review of the Prevalence and Outcomes of Ideal Cardiovascular Health in US and Non-US Populations, Mayo Clinic Proceedings 91 (5) (2016) 649-670, https://doi.org/10.1016/j.mayocp.2016.01.019.

[6] R. J. Applegate, P. M. Belford, S. K. Gandhi, M. A. Kutcher, R. M. Santos, D.X. Zhao, Chapter 38 - Cardiovascular Disease and Renal Transplantation, Kidney Transplantation, Bioengineering and Regeneration, Academic Press, Massachusetts, 2017, 543-554.

[7] R. D. Anderson, C. J. Pepine, Coronary Angiography, Circulation 127 (17) (2013) 1760-1762, https://doi.org/10.1161/CIRCULATIONAHA.113.002566.

[8] N. Sharma, K. Saroha, Study of dimension reduction methodologies in data mining, International Conference on Computing, Communication & Automation (2015) 133-137, https://doi.org/10.1109/CCAA.2015.7148359.

[9] M. S. Islam, M. M. Hasan, X. Wang, H. D. Germack, M. Noor-E-Alam, A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining, Healthcare (Basel) 6 (2) (2018), https://doi.org/10.3390/healthcare6020054.

[10] R. Das, I. Turkoglu, A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications 36 (2009) 7675-7680, https://doi.org/10.1016/j.eswa.2008.09.013.

[11] A. Khemphila, V. Boonjing, Heart Disease Classification Using Neural Network and Feature Selection, 2011 21st International Conference on Systems Engineering (2011) 406-409, https://doi.org/10.1109/ICSEng.2011.80.

[12] A. K. Paul, P. C. Shill, M. R. I. Rabin and M. A. H. Akhand, Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease, 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (2016) 145-150, https://doi.org/10.1109/ICIEV.2016.7759984.

[13] L. Verma, S. Srivastava, P.C. Negi, A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data, Journal of Medical Systems 40 (2016) 178, https://doi.org/10.1007/s10916-016-0536-z.

[14] A.K. Dwivedi, Performance evaluation of different machine learning techniques for prediction of heart disease, Neural Computing and Applications 29 (2018) 685–693, https://doi.org/10.1007/s00521-016-2604-1.

[15] O. W. Samuel, G. M. Asogbon, K. Sangaiah, P. Fang, G. Li, An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction, Expert System with Applications 68 (2017) 163-172, https://doi.org/10.1016/j.eswa.2016.10.020.

[16] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, S. A. Hussain, Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis, Physica A: Statistical Mechanics and its Applications 482 (2017) 796-807, https://doi.org/10.1016/j.physa.2017.04.113.

[17] A. K. Paul, P. C. Shill, M. R. I. Rabin, K. Murase, Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease, Applied Intelligence 48 (7) (2018) 1739-1756, https://doi.org/10.1007/s10489-017-1037-6.

[18] M. S. Amin, Y. K. Chiam, K. D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics 36 (2019) 82-93, https://doi.org/10.1016/j.tele.2018.11.007.

[19] L. Ali, A. Niamat, J.A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour and S.A. C. Bukhari, An optimized stacked support vector machines based expert system for the effective prediction of heart failure, IEEE Access 7 (2019) 54007-54014, https://doi.org/10.1109/ACCESS.2019.2909969.

[20] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, R. Nour, An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection, in IEEE Access 7 (2019) 180235-180243, https://doi.org/10.1109/ACCESS.2019.2952107.

[21]H. Singh, P. S. Rana, U. Singh, Prediction of drug synergy in cancer using ensemble-based machine learning techniques, Modern Physics Letters B 32 (11) (2018) 1850132, https://doi.org/10.1142/S0217984918501324.

[22] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, L. Burattini, Early temporal prediction of Type 2 Diabetes Risk Condition from a General Practitioner Electronic Health Record: A Multiple Instance Boosting Approach, Artificial Intelligence in Medicine 105 (2020) 101847, https://doi.org/10.1016/j.artmed.2020.101847.

[dataset] [23] Ronit, Cleveland Heart Disease data set, UC Irvine Machine Learning Repository, v1, 2018, https://www.kaggle.com/ronitf/heart-disease-uci.

[dataset] [24] Sonum, Heart Disease data set, UC Irvine Machine Learning Repository, v1, 2018, https://www.kaggle.com/sonumj/heart-disease-dataset-from-uci.

[25] M. Alex P., S. P. Shaji, Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique, 2019 International Conference on Communication and Signal Processing (ICCSP) (2019) 0848-0852, https://doi.org/10.1109/ICCSP.2019.8697977.

[26] C. Yang, B. An and S. Yin, Heart-Disease Diagnosis via Support Vector Machine-Based Approaches, IEEE International Conference on Systems, Man, and Cybernetics (SMC)(2018) 3153-3158, https://doi.org/10.1109/SMC.2018.00534.

[27] Ying-Tsang Lo , Hamido Fujita, Tun-Wen Pai, Prediction of coronary artery disease based on ensemble learning approaches and co-expressed observations, Journal of Mechanics in Medicine and Biology 16 (1) (2016) 1640010, https://doi.org/10.1142/S0219519416400108.

[28] V. Arora, R. Leekha, R. Singh and I. Chana, Heart sound classification using machine learning and phonocardiogram, Modern Physics Letters B 33 (26) (2019) 1950321, https://doi.org/10.1142/S0217984919503214.

[29] Xuan Huang, Lei Wu, Yinsong Ye, A Review on Dimensionality Reduction Techniques, International Journal of Pattern Recognition and Artificial Intelligence 33 (10) (2019) 1950016, https://doi.org/10.1142/S0218001419500162.

[30] María Alonso, José Malpica, Alex Martinez-Agirre, Consequences of the Hughes phenomenon on some classification Techniques, American Society for Photogrammetry and Remote Sensing Annual Conference (2011).

[31] B. Feng, J. Wang, K. Zhang, Patch-Based and Tensor-Patch-Based Dimension Reduction Methods for Hyperspectral Images, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium (2019) 3364-3367, https://doi.org/10.1109/IGARSS.2019.8898036.

[32] M. A. Akbar, A. Ait Si Ali, A. Amira, F. Bensaali, M. Benammar, An Empirical Study for PCA- and LDA-Based Feature Reduction for Gas Identification, IEEE Sensors Journal 16 (14) (2016) 5734-5746, https://doi.org/10.1109/JSEN.2016.2565721.

[33] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee and Q. M. Gu, A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, Neurocomputing 55 (2003) 321-336, https://doi.org/10.1016/S0925-2312(03)00433-8.

[34] P. Yildirim, K. U. Birant, V. Radevski, A. Kut and D. Birant, Comparative analysis of ensemble learning methods for signal classification, 2018 26th Signal Processing and Communications Applications Conference (SIU) (2018) 1-4, https://doi.org/10.1109/SIU.2018.8404601.

[35] Vadim Smolyakov, Ensemble Learning to Improve Machine Learning Results.https://blog.statsbot.co/ensemble-learning-d1dcd548e936.

[36] L. Breiman, Arcing the Edge, Technical Report 486, Statistics Department University of California, Berkeley CA. 94720.

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

[38] G. I. Webb, Z. Zheng, Multistrategy ensemble learning: reducing error by combining ensemble learning techniques, IEEE Transactions on Knowledge and Data Engineering 16 (8) (2004) 980-991, https://doi.org/10.1109/TKDE.2004.29.

[39] J. H. Friedman, Stochastic gradient boosting, Computational Statistics & Data Analysis 38 (4) (2002) 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.

[40] Prasanth Omanakuttan, Why does Gradient boosting work so well ?. https://blog.goodaudience.com/why-does-gradient-boosting-work-so-well-bf5a1d65a5c4.

[41] AlaaTharwat, Classification assessment methods, Applied Computing and Informatics (2018), https://doi.org/10.1016/j.aci.2018.08.003.

[42] Y.-S. Chen, P.P. Chong, M.Y. Tong, Mathematical and computer modelling of the Pareto principle, Mathematical and Computer Modelling 19 (9) (1994) 61-80, https://doi.org/10.1016/0895-7177(94)90041-8.