



ScienceDirect®

Informatics in Medicine Unlocked

Volume 24, 2021, 100584

An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach

Mohammed Nasir Uddin, Rajib Kumar Halder  

Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh

Received 1 February 2021, Revised 20 April 2021, Accepted 21 April 2021, Available online 5 May 2021, Version of Record 11 June 2021.

 What do these dates mean?



Show less 

 Outline |  Share  Cite

<https://doi.org/10.1016/j.imu.2021.100584> 

[Get rights and content](#) 

Highlights

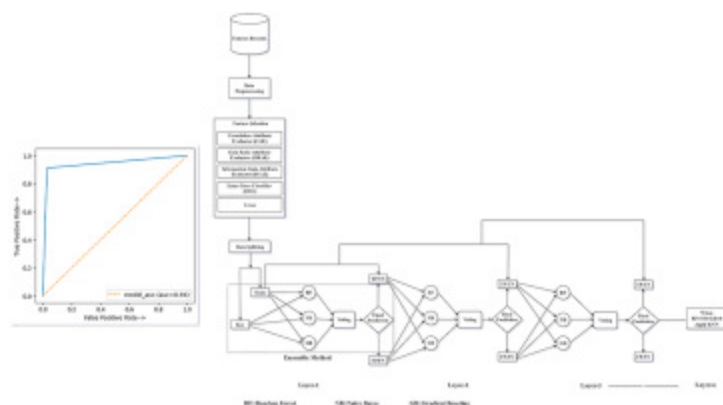
- An intelligent agent is designed to predict cardiovascular disease.
- A layer to layer prediction system is employed that is capable of increasing its current knowledge at each level.
- Correctly classification probability of cardiovascular disease of the proposed model is pointed out by the AUC curve.
- Successful results facilitate physicians in making quicker decisions.

Abstract

Cardiovascular disease is defined as a set of conditions related to the disorder of the heart and blood vessels. Predicting and diagnosing cardiovascular disease is significant to ensure the appropriate treatment of this disease. Machine learning approaches are generally utilized to automatically detect the hidden patterns in vast amounts of data without human intervention. In the early stage of cardiovascular disease, a machine learning model can aid physicians in making the right decision about the medication. This research aims to develop an intelligent agent to predict cardiovascular disease to investigate what steps should be taken before any untoward incident occurs. This paper proposes an ensemble method-based multilayer dynamic system (MLDS) that can improve its current knowledge in every layer. The proposed model applies Correlation Attribute Evaluator (CAE), Gain Ratio Attribute Evaluator (GRAE), Information Gain Attribute Evaluator (IGAE), Lasso, and Extra Trees classifier (ETC) for feature selection. Finally, Random Forest (RF), Naïve Bayes (NB), and Gradient Boosting (GB) classifiers combinedly construct the ensemble method for classification in the model. The K Nearest Neighbor

(KNN) algorithm is applied to find the test data's neighborhood data points while the base classifiers mentioned are failed to classify correctly in any layer. To test the proposed model's efficiency, we have used a realistic dataset (70,000 instances) collected from Kaggle. The proposed model has achieved 88.84%, 89.44%, 91.56%, 92.72%, and 94.16% accuracy based on the train and test data's different splitting ratios (50:50, 60:40, 70:30, 80:20, and 87.5:12.5). Our proposed model has achieved a 0.94 AUC value. AUC=0.94 means it has a 94% probability of correctly classifying positive and negative classes, Whereas the splitting ratio is 87.5:12.5. The Cleveland, Hungarian, and Cleveland-Hungary-Switzerland-Long Beach datasets have also been applied to train the model, and the model achieved 98.88%, 99.53%, 99.98%, 98.36%, 96.66%, 97.77%, 99.56, and 94.37% accuracy depending on the different splitting ratios of these datasets. The proposed model has been compared to five other models, indicating that the proposed model can effectively predict cardiovascular disease.

Graphical abstract



[Download : Download high-res image \(134KB\)](#)

[Download : Download full-size image](#)



Next



Keywords

Machine learning; Cardiovascular disease; Feature selection; Ensemble model; Classification

1. Introduction

World Health Organization (WHO) reported that around 17.9 million people die each year due to cardiovascular disease, where people under the age of 70 account for one-third of all premature deaths [1]. The report also stated that of the 17.3 million deaths caused by heart disease in 2008, approximately 6.2 million were due to stroke, and an estimated 7.3 million were due to coronary heart disease. WHO predicted that around 23.6 million people would die due to heart disease and stroke-related disease by 2030 [2]. Cardiovascular disease shows a range of symptoms, including tightness in the chest, pressure in the chest, discomfort in the chest (angina), pain in the chest, shortness of breath, numbness, weakness, or coldness in the legs or arms if the blood vessels are shrunk in those parts of the body. Some other symptoms of this disease include nausea, fatigue, cold sweats, and discomfort in the neck, elbows, throat, jaw, left shoulder, upper abdomen or back. However, major causes of cardiovascular disease are age, smoking, sugar, obesity, depression, hypertension, high blood pressure, cholesterol, poor diet, and physical inactivity [3]. Cardiovascular disease is also caused by coronary artery damage, damage to the whole or part of the heart, or inadequate supply of nutrients and oxygen to the heart. There are several types of cardiovascular disease such as coronary heart disease, stroke, hypertensive heart disease, inflammatory heart disease, rheumatic heart disease, etc. [4]. Some types of cardiovascular diseases such as hypertrophic cardiomyopathy, dilated cardiomyopathy, right ventricular arrhythmogenic cardiomyopathy are genetically inherited. Monitoring cardiovascular symptoms for a patient is necessary to seek medical advice from healthcare professionals. If the cardiovascular disease is diagnosed at the early stage, the risk of death due to this disease might be reduced. However, manually analyzing symptoms is difficult due to redundancy, multi-attribution, incompleteness, and a close association with time in medical data. Further, medicating a patient appropriately after manually analyzing massive quantities of heart disease-related data is a significant challenge. To address this issue, machine learning (ML) technique aids in creating predictive models that can process and analyze large amounts of complex medical data and predict the absence or presence of cardiovascular disease for a patient with higher accurate results. In ML approach, a computer program is trained to perform a particular task to learn from its previous experience and predict the outcome for testing data based on the training data [5]. Machine learning techniques enable a machine to make proper decisions based on a build-in analytical model when unseen data is provided. Machine

learning techniques consume relatively less time for accurate prediction. Therefore, a ML-based intelligent cardiovascular disease prediction system can assist healthcare practitioners in making faster decisions, enabling them to offer medical treatments to many patients within a short time; hence ML model potentially saves millions of lives. Machine learning techniques have already gained significantly higher precision in classification-based problems [6]. Information abstraction has been achieved using various machine learning techniques, including feature selection, classification, and clustering [7,8]. Latest studies have used machine learning algorithms to predict cardiovascular disease.

Javeed et al. [9] suggested the Random Search Algorithm (RSA) and optimized model for the Random Forest to improve the diagnosis of cardiovascular disease. The proposed model applied the RSA to select features and refine the Random Forest classifier for the accurate prediction/classification of heart disease. The data was collected from the UCI machine learning repository. The total number of instances in the dataset is 303. Two hundred ninety-seven of them have complete information on attributes, while details of six instances are missing. However, the limitation of this research work is that no intelligent technique was used to select the subset of features in this work. The RSA algorithm generates completely random locations. Assuming that there are N features in a dataset, and the RSA-based feature selection technique generates a total of N-1 subsets of features to achieve optimal accuracy. It is time-consuming due to the uncertainty about which subset of features the model can give optimal accuracy.

Karen et al. [10] proposed a classification model for heart disease prediction. The features were selected using the Chi-Squared feature selection technique, and then Principal Component Analysis (PCA) was applied to find principal components. They used six classifiers for the classification task. They performed four types of experiments: i) classified the raw data with all the six classifiers, ii) applied Chi-Square feature selection technique to obtain effective features and validate the features with the classifiers, iii) used the reduced dataset obtained by Chi-Square and then applied PCA before classification, and iv) the final experiment was the direct use of PCA from raw data. They used Cleveland (283 instances), Hungarian (294 instances), and Cleveland-Hungarian (577 instances) datasets for this experiment. The limitation of this research work is that it is difficult to identify how many principal components to keep in practice because original features of a dataset will turn into principal components, which are the linear combination of original features after implementing PCA on the dataset. So, it has a possibility to miss some information compared to the original list of features during the selection of principal components.

Amin Ul et al. [8] proposed a hybrid intelligent framework for predicting heart disease. The researchers used three feature selection algorithms (Relief, mRMR, and LASSO), the K-fold cross-validation method, and seven classifiers (LR, K-NN, ANN, SVM (kernel RBF and kernel linear), NB, DT, and RF). They recorded the accuracy of the different classifiers based on the features extracted from various feature selection algorithms. To carry out this experiment, they used Cleveland (303 instances) dataset collected from the UCI machine learning repository where 297 instances have complete attributes information while six instances have missing details. The disadvantage of the model is that the training algorithm has to rerun from scratch k times. As a result, large datasets are not suitable for this framework because it takes much more time to complete a single computational task.

Domor et al. [11] proposed an improved ensemble learning approach to predict heart disease's risk. The proposed model used the mean-based splitting technique to divide the whole dataset into smaller subsets and applied the Classification and Regression Tree (CART) algorithm to classify each partition. An accuracy-based weighted aging ensemble (WAE) is used to generate a homogeneous ensemble from different CART models. In this research work, the authors used two heart disease datasets, the Cleveland dataset (303 instances) and the Framingham dataset (4238 instances). The limitation of this research work is that no optimization algorithm was deployed to select effective attributes for the model. This system cannot handle noisy data and has a chance to create a noisy decision tree.

Louridi et al. [12] proposed a machine learning model to identify cardiovascular disease using 303 records with 13 attributes. They used Support Vector Machine (SVM), KNN (K Nearest Neighbor), Bayes Naïf (BN) for classification and found the highest accuracy of 86.8% by SVM-linear kernel. The limitation of this research work is that the researchers only handled the missing value in preprocessing unit, but feature selection plays a vital role in getting better accuracy and reducing the execution time. It helps to select effective features that have a significant impact on the target value.

Xiao-Yan et al. [13] proposed an ensemble method-based heart disease prediction model using 1025 instances with 13 independent attributes collected from Kaggle. Two feature selection algorithms (linear discriminant analysis, principal component analysis) are used to select the effective features. In this model, KNN, SVM, DT, RF, NB are used to construct the ensemble method. Both boosting and bagging techniques to classify heart disease. Their proposed model obtained the highest accuracy of 98.6% for the bagging ensemble learning method with a decision tree. This research work's limitations

are i) High time complexity in the training phase, ii) this system has a possibility to miss some information compared to the original list of features during the selection of principal components.

A.Geetha et al. [14] proposed a cardiovascular disease pre-diction model using machine learning approach. They used Cleveland (303 instances) heart disease dataset collected from UCI machine learning repository. KNN algorithm is used to classify heart disease. This model obtained highest 87% accuracy. This research work's limitations are: i) No feature selection algorithm is used in this model. Authors selected the effective features manually. This model is not suitable for the dataset which contains a large number of attributes. ii) In this model, a single classifier is used to make decisions, but it is better to make decisions based on multiple classifiers than a single classifier.

All the above existing models are one-layer filtering systems. These models are unable to expand their current knowledge from their re-sources. We have proposed a multilayer dynamic system that can continue the classification process from one layer to another by enhancing its knowledge to get the optimal result.

Three conventional approaches are used to build MLDS: features selection, ensemble technique, classification via classifiers. Feature se-lection plays a significant role to reduce the data dimensionality. Feature selection is necessary for the classification because irrelevant features often affect the performance of classifiers. Feature selection improves the accuracy of a classifier and reduces the model's execution time [8]. Two methods for feature selection are widely used, where one is the filter method, and another is the wrapper method [15]. In the filter system, features are chosen by various statistical tests based on their scores, which calculate the significance of features by their correlation with the dependent variable or the target variable. The wrapper methods search for a subset of features by evaluating the usefulness of a subset of features with the dependent variable [16]. In this work, we applied five feature selection algorithms: Correlation Attribute Evaluator (CAE), Gain Ratio Attribute Evaluator (GRAE), Information Gain Attribute Evaluator (IGAE), Lasso, Extra Trees classifier (ETC). These feature selection techniques are described in Section 2. In ensemble learning, multiple classifiers are trained simultaneously, and the outcomes from these classifiers are integrated in a different manner. This integration aims to complement the limitation and utilize different mechanisms. The ensemble learning method is used to increase robustness, accuracy, better generalization and decrease error rate. An ensemble method is built in two phases. At the initial phase, all the base learners are trained whereas each of these learners is produced concurrently, and the generation of a learner has an impact on the other learner. In the following phase, decisions from these base learners are aggregated in different ways: bagging, boosting,

stacking, voting etc. There are two voting schemes-hard voting and soft voting [17]. In hard voting, each classifier does the voting individually for a class and the target class with a majority of these votes, which is the mode of distribution, is accepted. In soft voting, each classifier defines the probability value for a particular target class on each data point. The target label with the greatest sum of probabilities is accepted [18]. Classification algorithms are used to predict the target class where predefined labels are assigned to instances by properties. Supervised learning technique is used for classification: Training set→Classification algorithms → Unseen data→Prediction result.

The objectives of this research work are:

1. To develop an intelligent agent to predict cardiovascular disease in which the learning knowledge flows from layer to layer that im-proves the model's performance.
2. To identify the valuable features from the dataset using multiple feature selection techniques to improve the classification accuracy.
3. To select an optimal ratio between training and testing data for analyzing prediction accuracy.

The significant contributions of this research work to achieve the objectives are summarized as follows:

1. In MLDS, we have introduced the process of passing learning knowledge from one layer to another layer. We have implemented a model that can enhance its learning knowledge from its immediate previous layer (classification performance of MLDS in every layer is given in Section 3).
2. We have applied three classification algorithms to implement the ensemble method in every layer to boost predictive efficiency and find the best result, in addition to move the learning knowledge from one layer to another layer.
3. We have used multiple feature selection techniques to select effective and useful features for the class attribute. We have selected those features which are common within a fixed range in the maximum number of algorithms from the five feature selection algorithms (details are presented in Section 3).
4. In the proposed model, a larger publicly accessible dataset (Kaggle cardiovascular dataset (70,000 instances)) has been used to train the model. The model has shown better performance compared to existing studies as the small number of

records and a single dataset are not always enough to test the efficiency of a model. We have also applied the datasets used in other existing research work on our proposed model and compared our proposed model performance with the existing models. The performance comparison with other systems is given in Section 3.

5. We have divided the whole dataset (70000 instances) into five partitions (50:50, 60:40, 70:30, 80:20, 87.5:12.5) and applied each of these partitions to measure the predictive accuracy (details are presented in Section 3).

The remaining of the paper is organized as follows: In section II, we give details about the methodology. In section III, we describe the evaluation, validation methods and different experiments used in this paper. Section IV contains the discussion about MLDS based on experimental results. Section V and Section VI ends with a conclusion and feature work.

2. Methodology

A real dataset including 70000 records with 11 independent features obtained from Kaggle has been used in this research. These data were collected at the moment of medical examination and information given by the patient. Details of all the features of this dataset with some necessary statistical calculations are shown in Table 1. In the data pre-processing section, the gender column has been transferred from categorical word to numerical value, i.e., gender = 1 for female instead of “f” and gender = 2 for male instead of “m”. The patient's age has been converted from days to a year. Missing value handling is an essential part of data analysis because attributes in a dataset give valuable in-formation. If any value is missing, it impacts decision-making. We have used is null() function to detect the missing values. There are some important functions to handle the missing value: dropna() function is used to drop the missing values, and fillna() function is used to fill NA/NaN values using the specified method. DBSCAN algorithm is used to remove the outliers. Next, feature selection is done by applying the Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator, Lasso, and Extra Tree classifier to select effective features to improve accuracy and decrease the searching time of classification. We have selected those features which are com-mon within a fixed range in the highest number of algorithms among the five feature selection algorithms. To select the good proportion between training and testing data, we have used the “train-test-split” method to divide the whole dataset into multiple proportions and applied proposed MLDS to each of the proportion, and tested which ratio gives us the best results (Details are given in Section 3). We have divided the whole dataset into five partitions: 50:50 (35000 training and 35000 testing out of 70000), 60:40 (42000 training and 28000 testing out of 70000), 70:30 (49000

training and 21000 testing out of 70000), 80:20 (56000 training and 14000 testing out of 70000) and 87.5:12.5 (61250 training and 8750 testing out of 70000). Finally, three classifiers, Random Forest (RF), Naïve Bayes (NB), and Gradient Boosting (GB), have been applied to perform classification. The layer to layer prediction process occurs in this model. In each layer, three classifiers classify the same testing dataset based on the same training dataset. The classifier which classifies more correctly than others is accepted to be reported its accuracy for the related layer. After completing classification in each layer, the correctly classified data (TP+TN) based on comparing with pre-defined target values in the original Kaggle dataset is added to previous training data to enter into the next layer. The incorrectly classified data (FP+FN) will also participate in the next iteration as new testing data. This process will be continued until three classifiers can't improve performance (TP and TN=0). When all three classifiers are being failed to classify correctly, the proposed model tries to find the optimal number of nearest neighborhood data points by utilizing K Nearest Neighbor (KNN) algorithm. This approach is called a multilayer dynamic system (MLDS). These steps are illustrated in [Fig. 1](#). The total accuracy for MLDS can be calculated with the formula presented below:

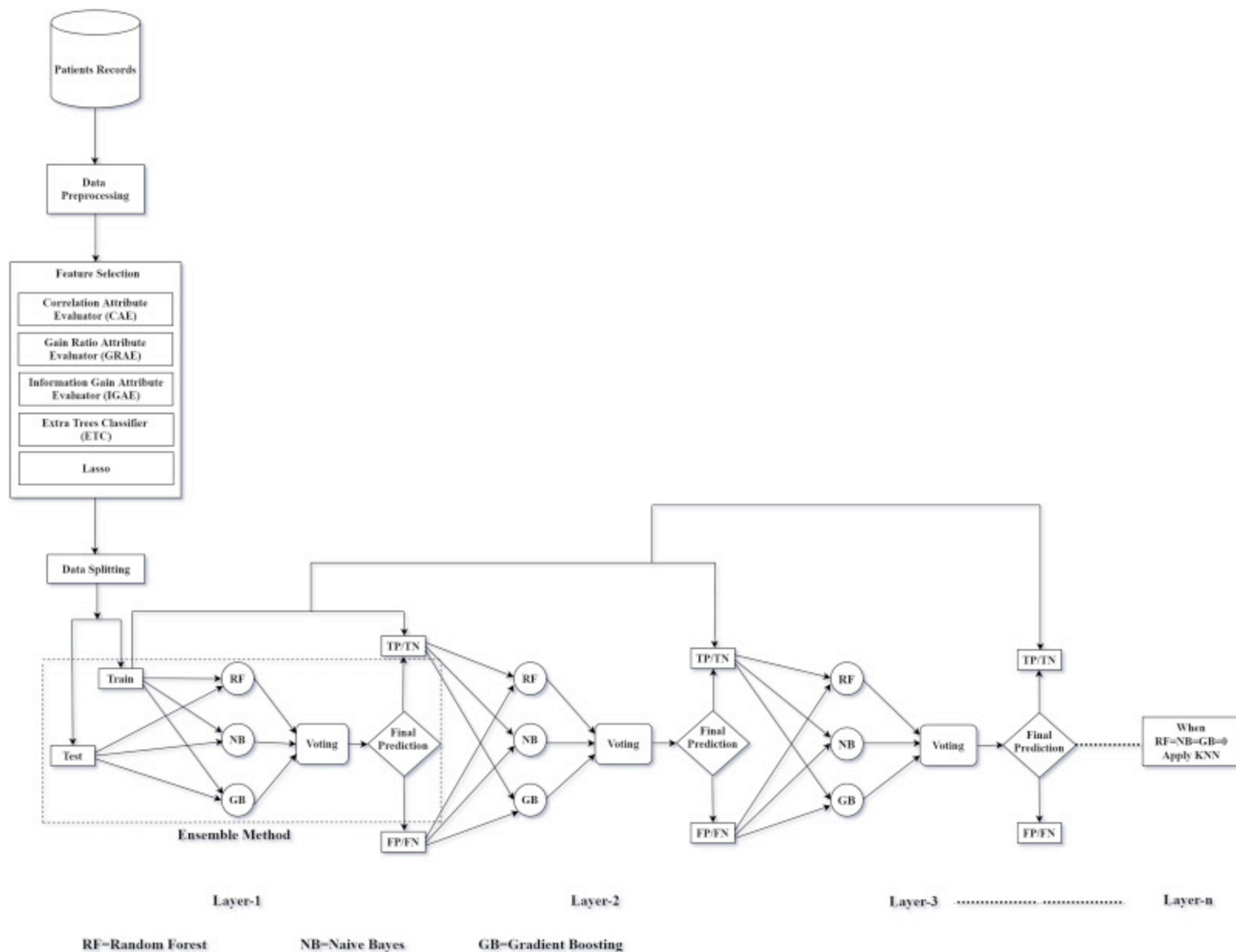
Number of train data in layer i: Number of train data in layer(i-1)+Number of TP, TN in layer (i-1)

Test data in layer i: Number of FP, FN in layer (i-1)

Table 1. Kaggle cardiovascular disease dataset attributes description with some statistical calculation.

Serial Number	Variable Description
1	age-int (days); Min: 10798, Max: 23713, Mean: 19468.866, StdDev: 2467.252
2	Height-int (cm); Min: 55, Max: 250, Mean: 164.359, StdDev: 8.21
3	Weight-float (kg); Min: 10, Max: 200, Mean: 74.206, StdDev: 14.396
4	gender-categorical code; (f = female, m = male)
5	ap_hi-int; Min: -150, Max: 16020, Mean: 128.817, StdDev: 154.011
6	ap_lo-int; Min: -70, Max: 11000, Mean: 96.63, StdDev: 188.473

Serial Number	Variable Description
7	Cholesterol; (1 = normal, 2 = above normal, 3 = well above normal)
8	gluc; (1 = normal, 2 = above normal, 3 = well above normal)
9	Smoke-binary; (1 = smoker, 0 = non-smoker)
10	Alco-binary; (1 = yes, 0 = no)
11	active-binary; (active = 1, inactive = 0)
	Target- binary; (1 = Presence = 1, 0 = absence of cardiovascular disease)



[Download : Download high-res image \(511KB\)](#)

[Download : Download full-size image](#)

Fig. 1. Proposed multilayer dynamic system (MLDS) to predict cardiovascular disease.

Where i = layer number; $i = 1, 2, 3, \dots, n$.

When Random Forest (RF) = 0, Naïve Bayes (NB) = 0, Gradient Boosting (GB) = 0, KNN = 0

$$\text{Total Accuracy} = \frac{\sum_{i=1}^n (TP, TN)}{\text{Total Number of Test Data}} \quad (1)$$

2.1. Random Forest (RF)

Random Forest is an example of an ensemble method involving the collection of decision trees. In the Random Forest algorithm, samples are drawn randomly, and decision trees are built for the random sample, and the process is repeated [19]. It avoids the missing values and outliers by following steps: data analysis and data pre-processing and corrects the overfitting to their training dataset [20]. This ensemble classifier in-corporates several decision trees to get the best result. Decision Trees mainly apply bootstrap aggregating or bagging [21]. For example, a given data, $X = \{x_1, x_2, \dots, x_n\}$ with responses $Y = \{y_1, y_2, \dots, y_n\}$ which repeats the bagging from $b = 1$ to B . The unseen samples x' is made by averaging the predictions $\sum_{b=1}^B f_b(x')$ from every individual trees on x' :

$$j = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (2)$$

The uncertainty of prediction on tree is made through its standard deviation:

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B-1}} \quad (3)$$

2.2. Naïve Bayes (NB)

Naïve Bayes classifier or the Bayesian theorem is another classification technique utilized to predict a target class. It depends on probabilities in its calculations [22,23]. Based on Bayesian theory, each quantity has a statistical distribution by which a test sample can be categorized. It basically follows the bag of words (BOW) feature extraction to eliminate the word location in the document and does not consider the correlation between attributes [24,25]. For example, every instance of data D is

allotted to the class of highest subsequent probability. The model is trained through the Gaussain function with prior probability $P(X_f) = \text{priority} \in (0:1)$

$$P(X_{f1}, X_{f1}, \dots, X_{fn} | c) = \prod_{i=1}^n P(X_{fi} | c)$$

$$P(X_{fi} | c_i) = \frac{P(c_i | X_f) P(X_f)}{P(C_i)} \quad (4)$$

$c \in \{\text{begin, malignant}\}$

At last, the testing data is categorized based on the probability of association:

$$c_{nb} = \arg \max P(c_k) \prod_{i=1}^n P(X_{fi} | c_k), \text{ for } k = 1, 2$$

2.3. Gradient boosting (GB)

For regression and classification issues, the gradient boosting machine learning method is used. In the form of an ensemble of decision trees that are constructed in a stage-wise process, it may generate a prediction model [26]. In gradient boosting, decision trees are generally used. The main advantage of gradient boosting is that the residual last time is reduced in each calculation. The residual gradient direction can be reduced to create a new model to decrease the residual [27]. In boosting, every new tree is a fit on a modified version of the original dataset.

Algorithm: Gradient Boosting

```

1.  $F_0(x) = \operatorname{argmin}_{\rho} \sum_{i=1}^N L(y_i, \rho)$ 
2. For  $m=1$  to  $M$  do:
3.  $\hat{y}_i = - \left[ \frac{\partial L(y_i, F(x))}{\partial F(x)} \right]_{F(x)=F_{m-1}(x)}, i = 1, \dots, N$ 
4.  $\alpha_m = \operatorname{argmin}_{\alpha, \beta} \sum_{i=1}^N [\hat{y}_i - \beta h(x_i; \alpha_m)]^2$ 
5.  $\rho_m = \operatorname{argmin}_{\rho} \sum_{i=1}^N L(y_i, F_{m-1}(x_i) + \rho h(x_i; \alpha_m))$ 
6.  $F_m(x) = F_{m-1}(x) + \rho h(x, \alpha_m)$ 
7. end For
8. end

```

[Download : Download high-res image \(146KB\)](#)

[Download : Download full-size image](#)

2.4. K Nearest Neighbor (KNN)

KNN is a supervised machine learning algorithm. It is used for both classification and regression. Features similarity is used to predict the values of new data points. The Euclidean or Manhattan or Hamming methods are used to calculate the distance between test data and each row of training data. Usually, to find the similarity, it works with distance [12].

Algorithm: K Nearest Neighbor

```

Input: x, S, d
Output: class of x
1. For  $(x', l') \in S$  do
2.   Compute the distance  $d(x', x)$ 
3. end For
4. Sort the  $|S|$  distances by increasing order
5. Count the number of occurrences of each class  $I_j$  among the k nearest neighbors
6. Assign to x the most frequent class

```

[Download : Download high-res image \(159KB\)](#)

[Download : Download full-size image](#)

Description of five features selection algorithms (Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator, Lasso):

- i. Information Gain Attribute Evaluator: It is a filter-based feature selection method. For each attribute A_i , the information gain between the attribute and the class Y is determined as given by equation [28]:

$$IG_i = H(Y) - H(Y|A_i) \quad (5)$$

Here $H(Y)$ is the entropy of the class Y. Entropy is a mathematical function, corresponds to the information quantity contained or delivered by a source of information [28]:

$$H(Y) = - \sum_{y \in Y} P(v_i) \log_2 p(v_i) \quad (6)$$

- ii. Gain Ratio Attribute Evaluator: Gain Ratio is used to penalize node proliferation. It is significant when data is uniformly distributed and small when all data belongs to one branch [28]. To full fill this aim, gain ratio evaluates the features by dividing the information gain of the predicted attribute to the entropy of the observed attribute as given by equation:

$$GR = \frac{IG}{H(Y)} \quad (7)$$

- iii. Correlation Attribute Evaluator: Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average [29].

Capabilities of Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator:

Class -- Binary class, Missing class values, Nominal class.

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes.

Minimum number of instances: 1.

- iv. Lasso: The least absolute shrinkage and selection operator eliminates the zero features from the feature's subset. By updating the absolute value of the feature's coefficient, Lasso selects the features. The features having high values of

coefficients will be included in selected feature subsets. LASSO performs excellently with low coefficient feature values [8].

- v. Extra Trees classifier: It produces a large number of unpruned decision trees from the training dataset. In the case of regression, forecasts are made by integrating the prediction of decision trees or majority voting in the case of classification. Each tree is provided with a random sample of K number of features from which each decision tree must select the best feature.

3. Result analysis

We have selected age, cholesterol, weight, gluc, ap_lo, ap_hi as effective features because these six features are common from serial 1 to 6 in four algorithms (Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator, Lasso) out of five algorithms shown in Table 2. Effect of age, cholesterol, weight, gluc, ap_lo, ap_hi on cardiovascular disease.

Table 2. Rank wise features chart using five feature selection algorithms.

Correlation Attribute Evaluator	Gain Ratio Attribute Evaluator	Information Gain Attribute Evaluator	Lasso (using python	Extra Trees classifier (using
(using weka)	(using weka)	(using weka)	libraries)	python libraries)
0.23816 age	0.072691 ap_hi	0.170065 ap_hi	0.014558 age	0.290111 age
0.22115 chol	0.054584 ap_lo	0.106194 ap_lo	0.005504 wt	0.177682 ap_hi
0.18166 wt	0.034366 chol	0.044944 age	0.000142 ap_lo	0.176635 wt
0.08931 gluc	0.01393 age	0.036573 chol	0.000141 ap_hi	0.171074 height
0.06572 ap_lo	0.008123 wt	0.025608 wt	0 chol	0.109448 ap_lo
0.05448 ap_hi	0.008007 gluc	0.006092 gluc	0 gluc	0.045414 chol

Correlation Attribute Evaluator	Gain Ratio Attribute Evaluator	Information Gain Attribute Evaluator	Lasso (using python libraries)	Extra Trees classifier (using python libraries)
(using weka)	(using weka)	(using weka)	libraries)	python libraries)
0.03565 active	0.001284 active	0.000918 active	- 0 smoke	0.011002 gluc
0.01549 smoke	0.000519 height	0.000332 height	- 0 alco	0.005456 gender
0.01082 height	0.000402 smoke	0.000173 smoke	- 0 active	0.004683 active
0.00811 gender	0 alco	0 alco	- 0 gender	0.004346 alco
0.00733 alco	0 gender	0 gender	- 0.00104 height	0.004143 smoke

age: Adults aged 65 and older are more likely to suffer from cardiovascular disease than younger people. Aging can lead to changes in the heart and blood vessels that may increase the risk of cardiovascular disease in a person [30]. Cholesterol: If there is so much cholesterol in our blood, it builds up in the walls of the arteries, activating a mechanism called atherosclerosis, a type of cardiac disease. The arteries are reduced and blood flow to the muscle of the heart is slowed or blocked [31].

gluc: High blood glucose from diabetes can damage blood vessels and the nerves that control our heart and blood vessels. The longer once have diabetes, the greater risk of developing heart disease [32]. ap_hi: By making blood vessels more rigid and damaging the inner lining, high blood pressure damages the blood vessels. The damaged lining increases the risk of deposition of fat, preventing the flow of blood. Because of blood vessel resistance, the heart must work harder to provide the body with oxygen-rich blood adequately [33].

weight: In several respects, obesity leads to heart failure. More body fat contributes to a higher blood flow, which allows it more difficult for the heart to pump all the excess fluid. This causes damaging changes in the structure and function of the heart over the years that can eventually lead to heart failure [34]. Hypertension and enlarged left ventricle (left ventricular hypertrophy) are also associated with excess weight, increasing the risk of heart failure [35]. ap_lo: Low blood pressure, which causes insufficient blood flow to the organs of the body, can lead to strokes, heart attacks and kidney failure [36].

A performance matrix is used to measure the performance of a machine learning model. Matrix module of “scikit-learn” library provides necessary functions to compute performance evaluation metrics. The performance of the model is computed with the help of a confusion matrix. Four outcomes are generated from the confusion matrix, namely TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) for different proportions of the dataset (50:50, 60:40, 70:30, 80:20, 87.5:12.5 respectively).

$$\begin{bmatrix} 16813 & 680 \\ 3224 & 14283 \end{bmatrix} \begin{bmatrix} 13429 & 510 \\ 2446 & 11615 \end{bmatrix} \begin{bmatrix} 10106 & 369 \\ 1402 & 9123 \end{bmatrix} \\ \begin{bmatrix} 6707 & 260 \\ 758 & 6275 \end{bmatrix} \begin{bmatrix} 4262 & 123 \\ 388 & 3977 \end{bmatrix}$$

The following equations are used for the calculation of the accuracy, precision, True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (8)$$

$$\text{Precious} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \quad (10)$$

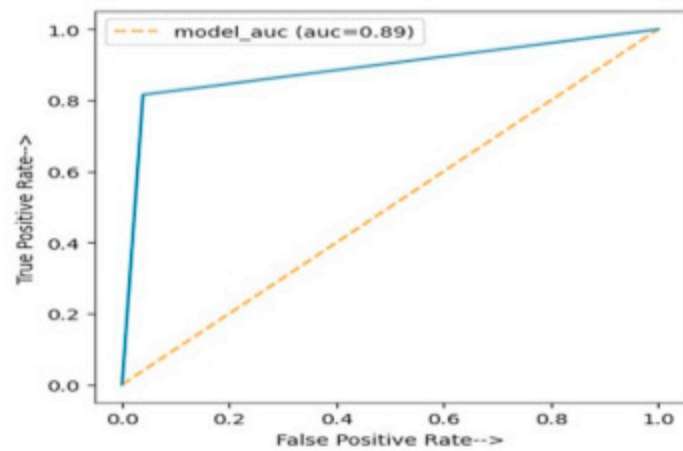
$$\text{False Positive Rate} = \frac{FP}{FP+TN} \quad (11)$$

$$\text{True Negative Rate} = \frac{TN}{TN+FP} \quad (12)$$

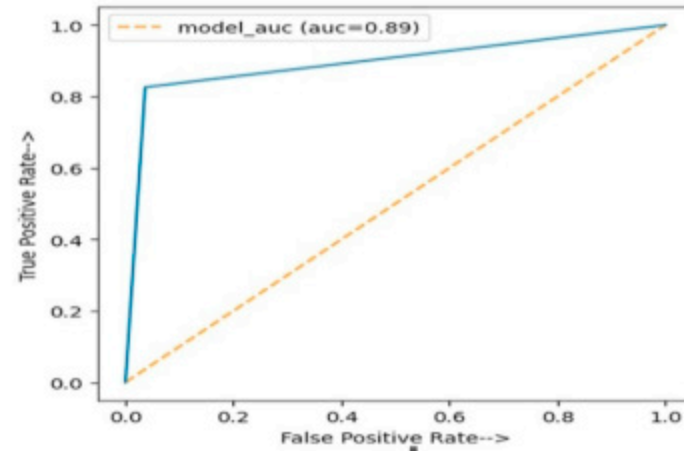
$$\text{False Negative Rate} = 1 - \text{TPR} \quad (13)$$

ROC curve shows the performance of a classification model at all classification thresholds. ROC is a curve of probability. The ROC curve is plotted with TPR against the FPR, where TPR is on the y-axis and FPR is on the x-axis. ROC curves are shown in [Fig. 2](#) for different proportions of the dataset (50:50, 60:40, 70:30, 80:20, 87.5:12.5 respectively). The blue dotted line is the ROC curve which plots the (x, y) = (FPR, TPR) points at all classification thresholds. A ROC curve to the top and left is a better model, which means the proposed model is much better to classify testing datasets in their respective comparison to another

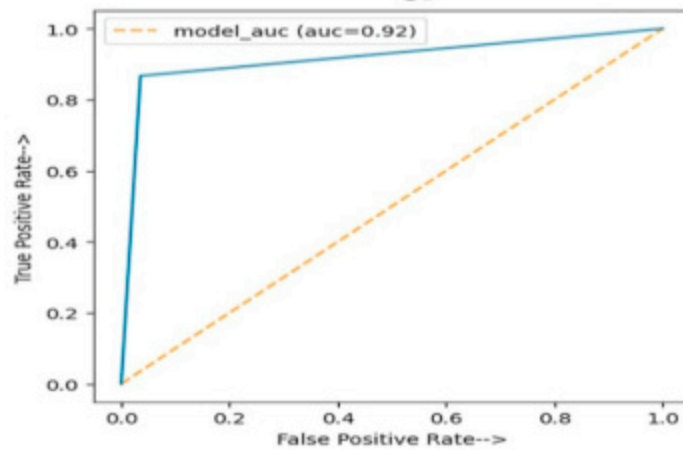
pro-posed model. AUC represents the degree or measures of separability. AUC measures the whole two-dimensional area from (0,0) to (1,1) below the whole ROC curve (just like integral calculus). A model with higher AUC has a better chance to predict 0s as 0s and 1s as 1s. An excellent classifier has AUC near to 1, which means it has a good measure of separability. From Fig. 2 (e), we can see that the AUC value of this proposed MLDS is maximum when the dataset is divided into the ratio 87.5:12.5. The value of AUC is 0.94 means that there is a 94% probability of proposed MLDS to be able to distinguish between absence or presence class for cardiovascular disease prediction correctly. The orange dots line represents the AUC of this proposed model. This pro-posed MLDS is better than another model because it has AUC=0.94 is near to 1. The classification performance of the proposed MLDS with other machine learning algorithms has been compared. Same training datasets have been used to train XGBoost (XGB), Linear Regression (LR), Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), Decision Tree (DT) classification models.



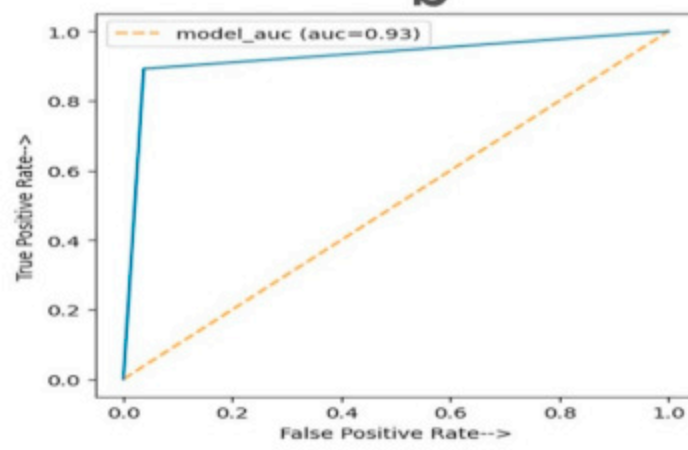
a



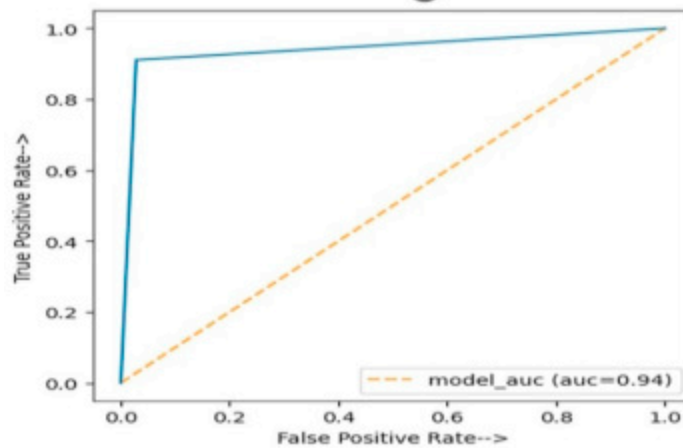
b



c



d



e

Download : [Download high-res image \(674KB\)](#)
Download : [Download full-size image](#)

Fig. 2. ROC curve and AUC of proposed model: (a) 50:50, (b) 60:40, (c) 70:30, (d) 80:20, (e) 87.5:12.5.

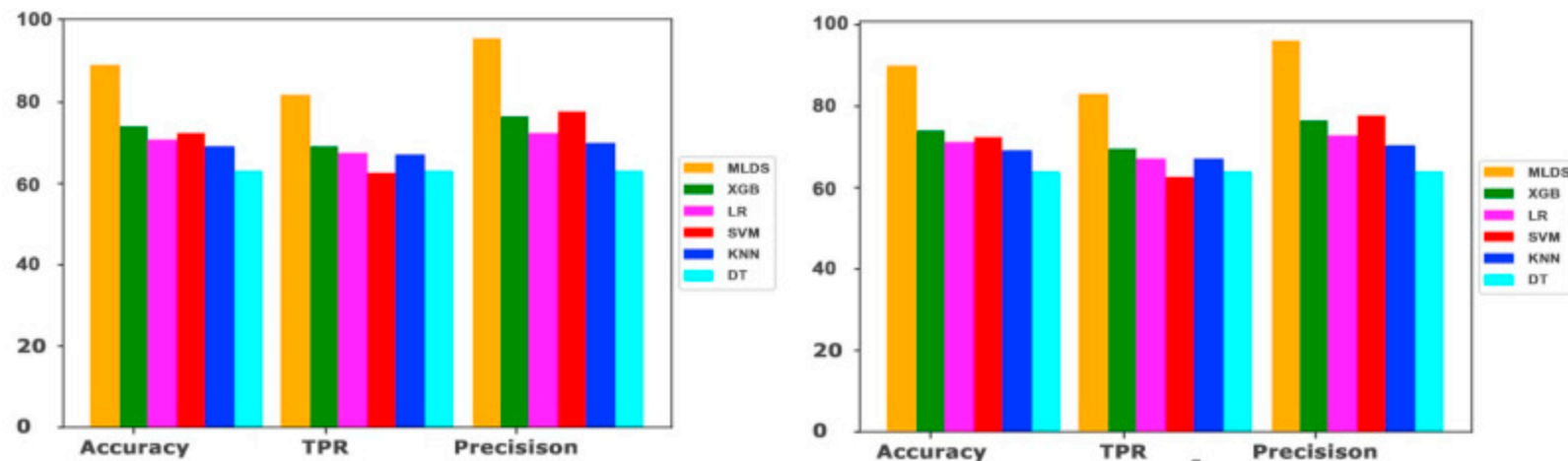
The performance of our proposed MLDS has been compared with other existing classification algorithms based on test partition as shown in [Table 3](#), and its graphical representations are shown in [Fig. 3](#), [Fig. 4](#). The evaluation of our proposed MLDS has been performed with the confusion matrix and compared with the other classification algorithm's confusion matrix based on test partition, as shown in [Table 4](#). Its graphical representations are shown in [Fig. 5](#), [Fig. 6](#).

Table 3. Performance evaluation metrics result of proposed (MLDS), XGB, LR, SVM, KNN, DT based on test partition.

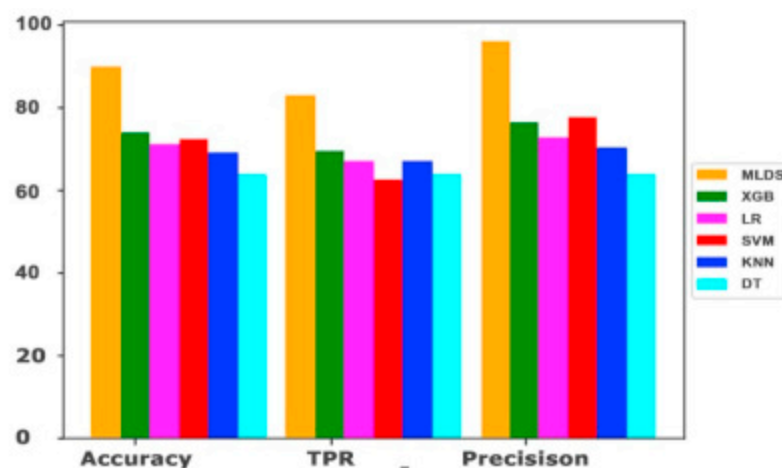
Model Name	Parameter (%)	Train and Test Dataset Ratio				
		50:50	60:40	70:30	80:20	87.5:12.5
Proposed MLDS	Accuracy	88.84	89.44	91.56	92.72	94.16
	Precision	95.45	95.79	96.11	96.02	97
	TPR	81.58	82.60	86.67	89.22	91.11
	FPR	3.88	3.65	3.52	3.73	2.80
	TNR	96.11	96.34	96.47	96.26	97.19
XGB	FNR	18.41	17.39	13.32	10.77	8.88
	Accuracy	73.66	73.62	73.70	73.52	73.87

Model Name	Parameter (%)	Train and Test Dataset Ratio				
		50:50	60:40	70:30	80:20	87.5:12.5
LR	Precision	76.13	76.13	76.18	75.83	75.72
	TPR	68.97	69.16	69.14	69.40	70.10
	FPR	21.63	21.87	21.71	22.31	22.37
	TNR	78.36	78.12	78.28	77.68	77.62
	FNR	31.02	30.83	30.85	30.59	29.89
	Accuracy	70.56	70.70	70.54	70.7	71.10
	Precision	72.16	72.61	72.34	72.32	72.26
SVM	TPR	67.00	66.88	66.74	67.51	68.29
	FPR	25.86	25.43	25.64	26.08	26.08
	TNR	74.13	74.56	74.35	73.91	73.91
	FNR	32.99	33.11	33.25	32.48	31.70
	Accuracy	71.97	72.01	72.09	72.46	72.88
	Precision	77.33	77.53	77.40	77.47	77.55
	TPR	62.20	62.32	62.58	63.71	64.21
KNN	FPR	18.24	18.21	18.35	18.70	18.49
	TNR	81.75	81.78	81.64	81.29	81.50
	FNR	37.79	37.67	37.41	36.28	35.78
	Accuracy	68.92	68.87	69.15	69.22	69.23

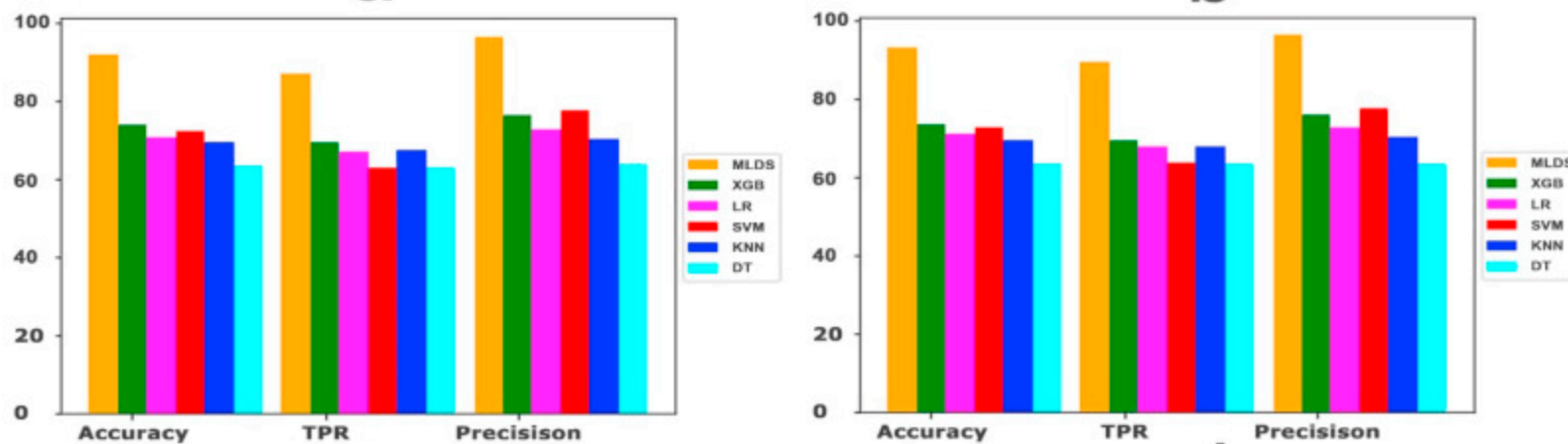
Model Name	Parameter (%)	Train and Test Dataset Ratio				
		50:50	60:40	70:30	80:20	87.5:12.5
DT	Precision	69.72	69.86	70.01	70.12	69.62
	TPR	66.94	66.88	67.25	67.51	67.99
	FPR	29.09	29.10	28.93	29.03	29.53
	TNR	70.90	70.89	71.06	70.96	70.46
	FNR	33.05	33.11	32.74	32.48	32.00
	Accuracy	62.84	63.53	63.14	63.14	63.37
	Precision	62.86	63.69	63.39	63.30	63.10
	TPR	62.83	63.67	62.64	63.34	63.98
	FPR	37.14	36.60	36.34	37.06	37.24
	TNR	62.85	63.39	63.65	62.93	62.75
	FNR	37.16	36.32	37.35	36.65	36.01



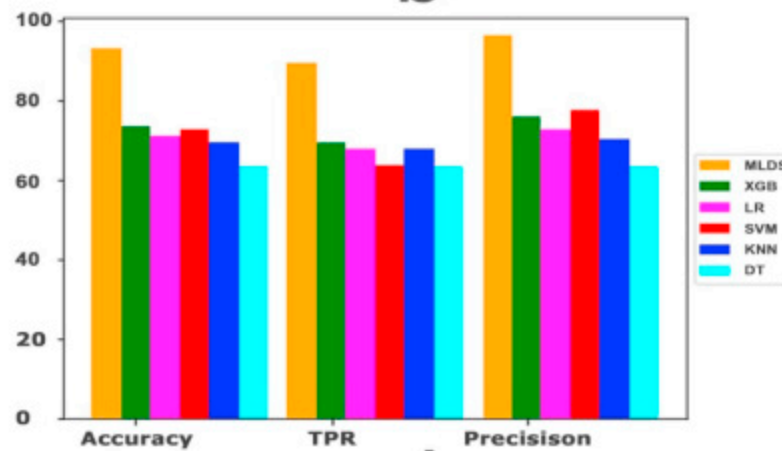
a



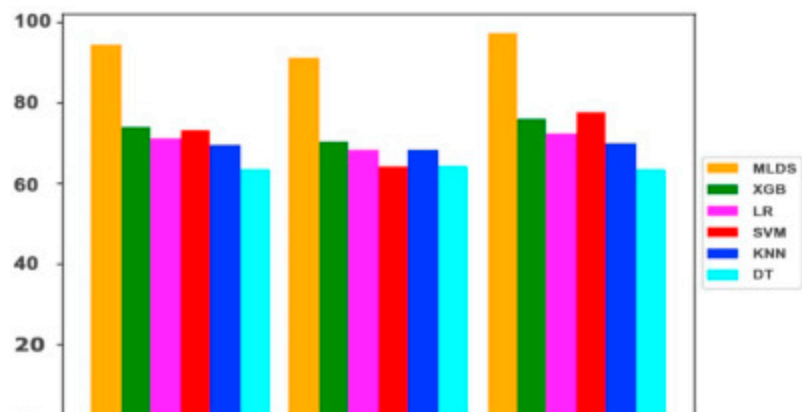
b



c



d

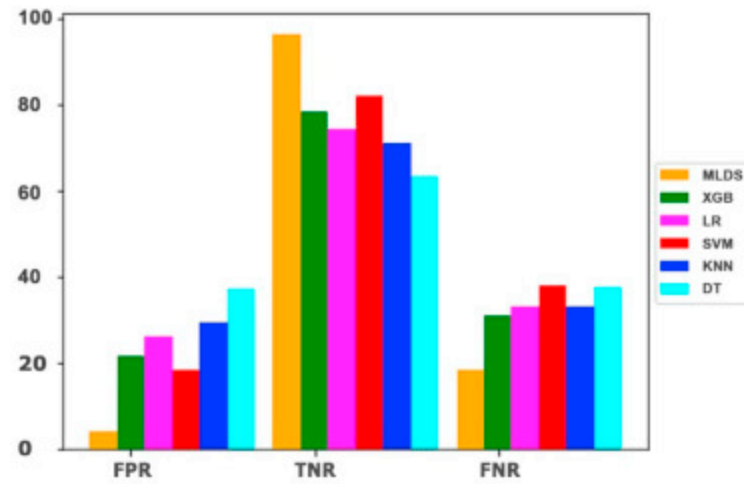




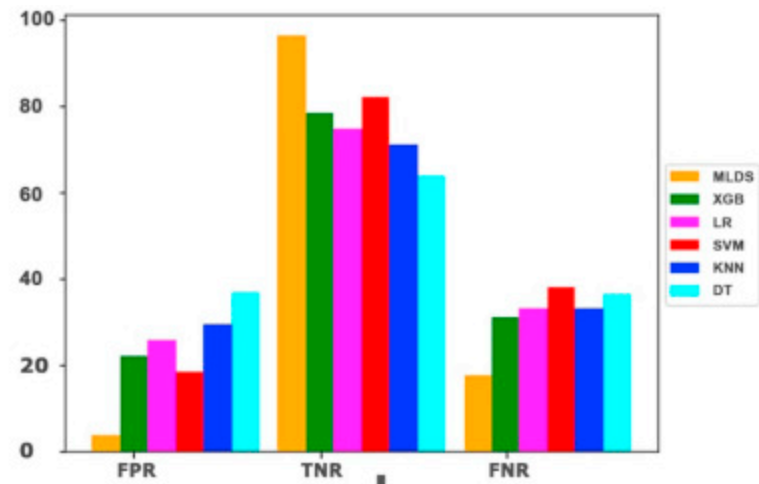
[Download : Download high-res image \(1MB\)](#)

[Download : Download full-size image](#)

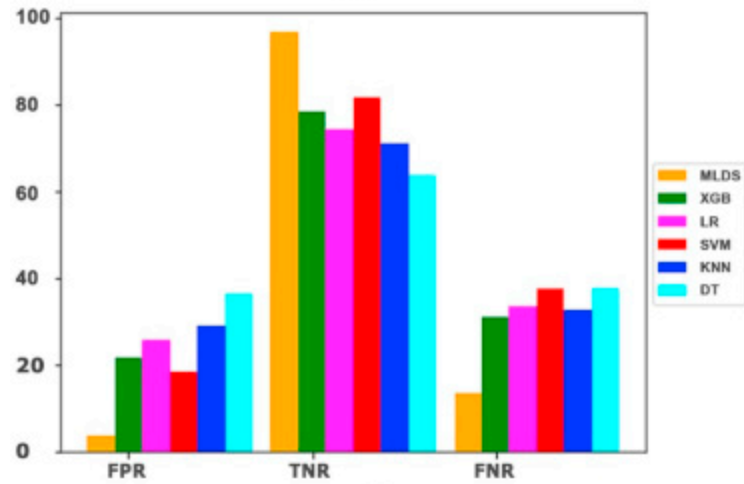
Fig. 3. Results comparison of Accuracy, TPR, Precision: (a) 50:50, (b) 60:40, (c) 70:30, (d) 80:20, (e) 87.5:12.5.



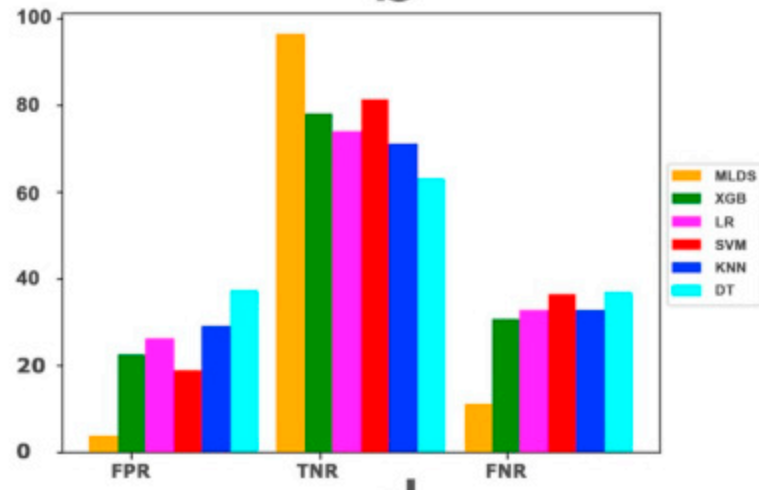
a



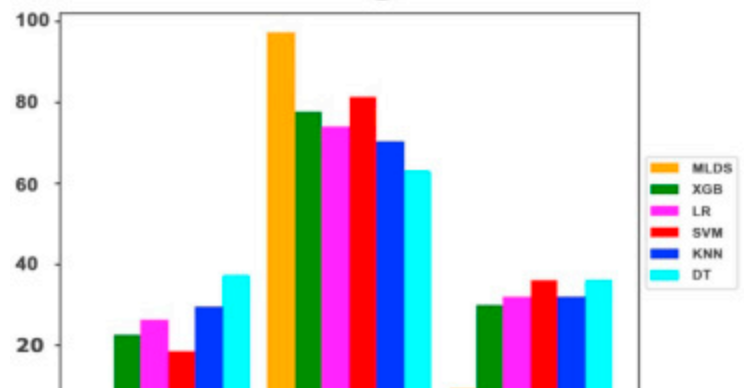
b



c



d





[Download : Download high-res image \(836KB\)](#)
[Download : Download full-size image](#)

Fig. 4. Results comparison of FPR, TNR, FNR: (a) 50:50, (b) 60:40, (c) 70:30, (d) 80:20, (e) 87.5:12.5.

Table 4. Confusion matrix result of proposed (MLDS), XGB, LR, SVM, KNN, DT based on test partition.

Model Name	Split Ratio	Parameter			
		TP	FN	TN	FP
Proposed MLDS	50:50	14283/17507	3224	16813/17493	680
	60:40	11615/14061	2446	13429/13939	510
	70:30	9123/10525	1402	10106/10475	369
	80:20	6275/7033	758	6707/6967	260
XGB	87.5:12.5	3977/4365	388	4262/4385	125
	50:50	12075/17507	5432	13709/17493	3784
	60:40	9725/14061	4336	10890/13939	3049
	70:30	7278/10525	3247	8200/10475	2275

Model Name	Split Ratio	Parameter			
		TP	FN	TN	FP
LR	80:20	4881/7033	2152	5412/6967	1555
	87.5:12.5	3060/4365	1305	3404/4385	981
	50:50	11731/17507	5776	12968/17493	4525
	60:40	9404/14061	4657	10393/13939	3546
	70:30	7025/10525	3500	7789/10475	2686
SVM	80:20	4748/7033	2285	5150/6967	1817
	87.5:12.5	2981/4365	1384	3241/4385	1144
	50:50	10890/17507	6617	14302/17493	3191
	60:40	8763/14061	5298	11400/13939	2539
	70:30	6587/10525	3938	8552/10475	1923
KNN	80:20	4481/7033	2552	5664/6967	1303
	87.5:12.5	2803/4365	1562	3574/4385	811
	50:50	11720/17507	5787	12403/17493	5090
	60:40	9404/14061	4657	9882/13939	4057
	70:30	7079/10525	3446	7444/10475	3031
DT	80:20	4748/7033	2285	4944/6967	2023
	87.5:12.5	2968/4365	1397	3090/4385	1295
	50:50	11000/17507	6507	10996/17493	6497

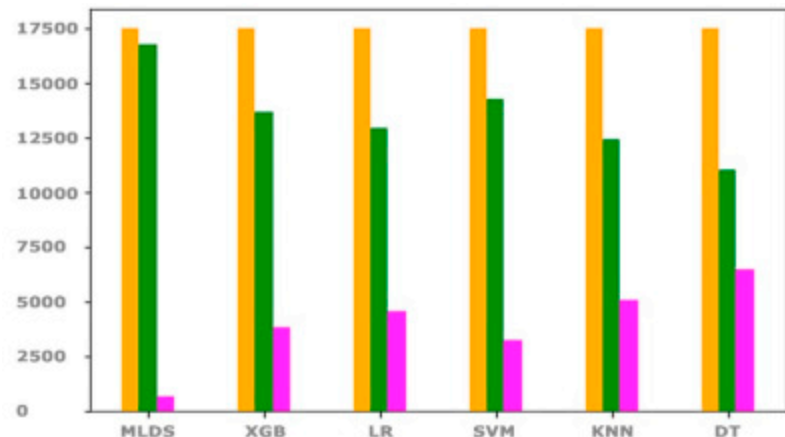
Model Name	Split Ratio	Parameter			
		TP	FN	TN	FP
	60:40	8954/14061	5107	8836/13939	5103
	70:30	6593/10525	3932	6668/10475	3807
	80:20	4455/7033	2578	4385/6967	2582
	87.5:12.5	2793/4365	1572	2752/4385	1633

TP=Total number of accurately identified people who are affected by cardiovascular disease.

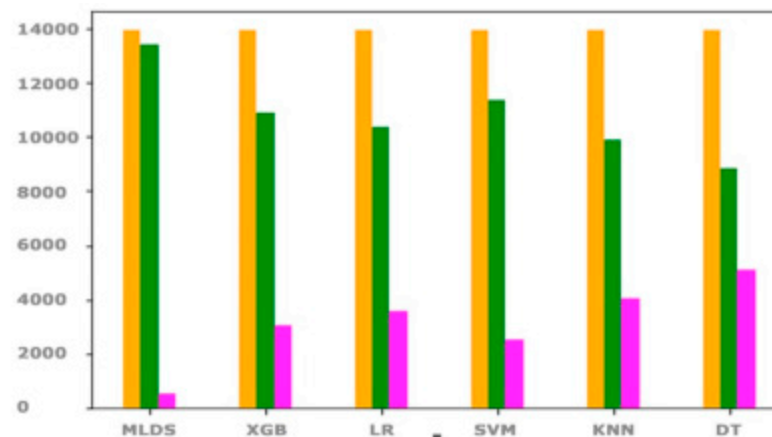
TN=Total number of accurately identified people who are not affected by cardiovascular disease.

FP=Total number of incorrectly identified people who are not affected by cardiovascular disease.

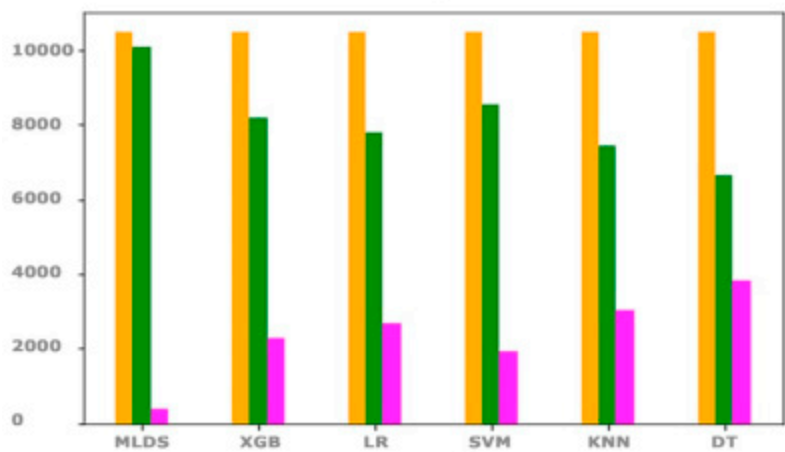
FN=Total number of incorrectly identified people who are affected by cardiovascular disease.



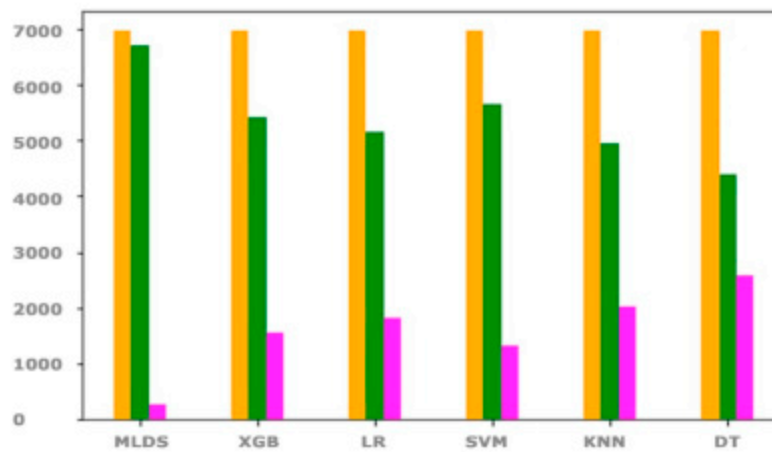
a



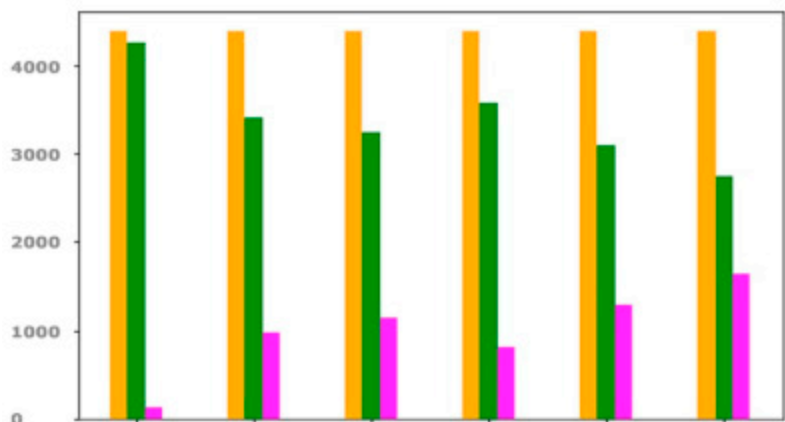
b



c



d



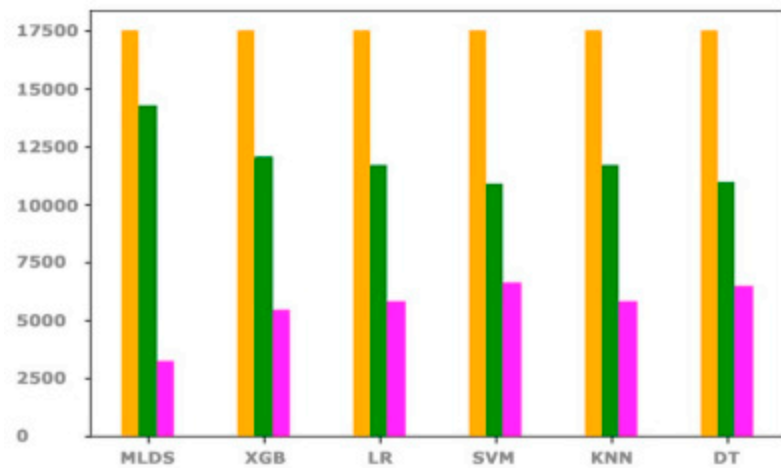
e



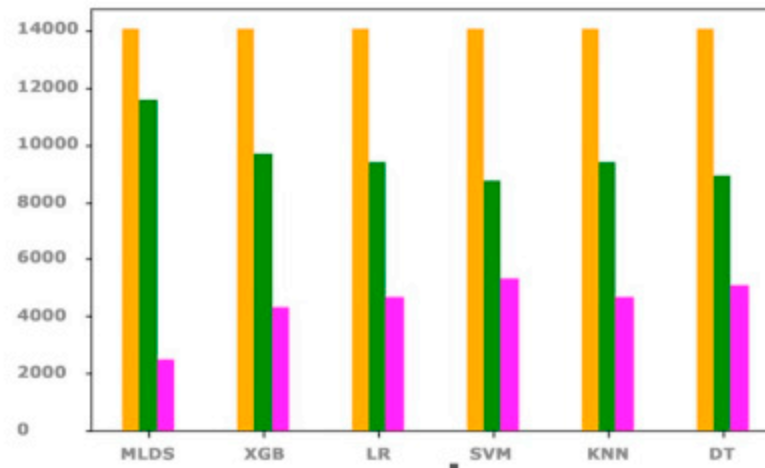
[Download : Download high-res image \(944KB\)](#)

[Download : Download full-size image](#)

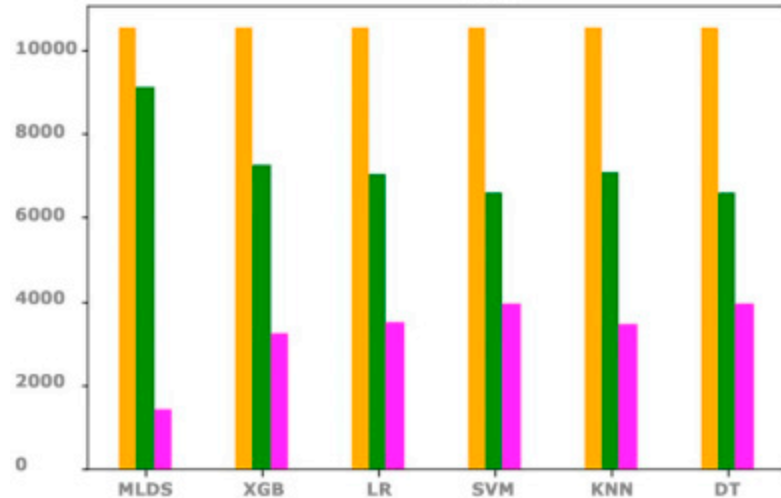
Fig. 5. Comparison of TN, FP: (a) 50:50, (b) 60:40, (c) 70:30, (d) 80:20, (e) 87.5:12.5.



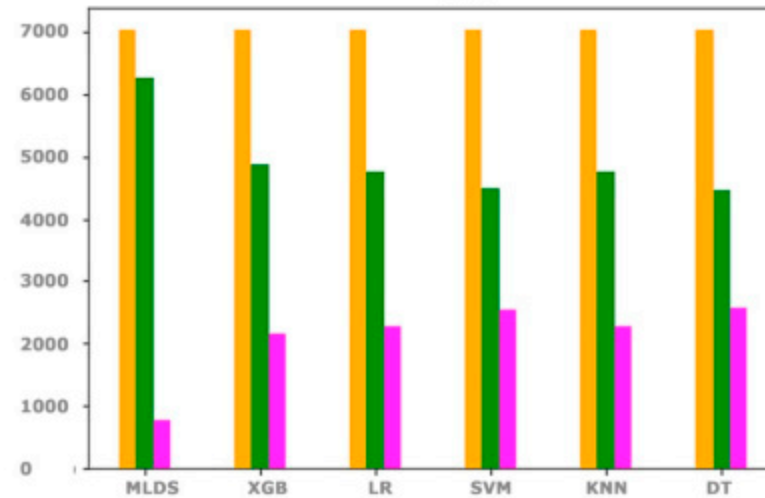
a



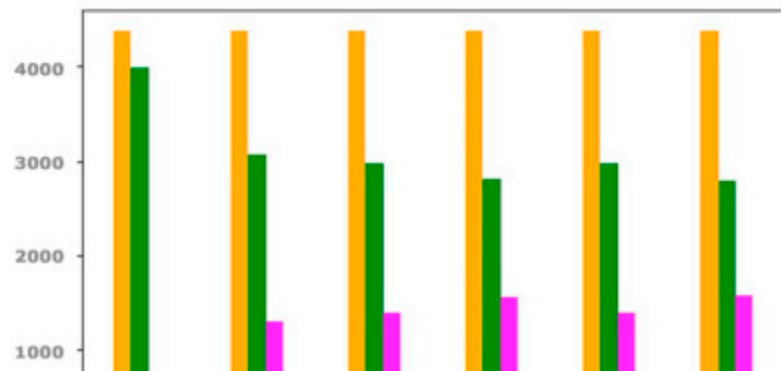
b

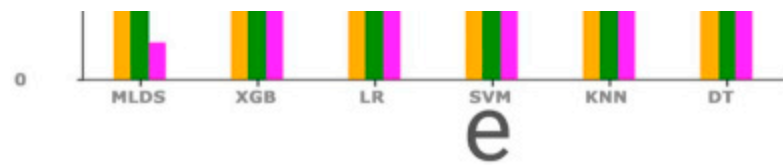


c



d





[Download : Download high-res image \(1013KB\)](#)

[Download : Download full-size image](#)

Fig. 6. Comparison of TP, FN: (a) 50:50, (b) 60:40, (c) 70:30, (d) 80:20, (e) 87.5:12.5.

In Fig. 5, the Yellow bar indicates the total number of actual True Negative in our testing dataset based on a different partition of the testing set, the Green bar indicates that the True Negative (TN) identified by the proposed MLDS, and the Magenta bar indicates the False Positive (FP) identified by proposed MLDS.

In Fig. 6, the Yellow bar indicates the total number of actual True Positive in our testing dataset based on a different partition of the testing set, the Green bar indicates the True Positive (TP) identified by the proposed MLDS, and the Magenta bar indicates the False Negative (FN) identified by proposed MLDS. Classification performance of MLDS in every layer based on the different splitting ratio of Kaggle cardiovascular disease dataset is shown in Table 5, Table 6, Table 7, Table 8, Table 9.

Table 5. Layer to layer classification result of MLDS (Train/Test Ratio: 50:50).

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
1 (initial)	35000	35000	25682	9318	GB	73.37
2	60682	9318	3052	6266	NB	82

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,	
3	63734	6266		792	5474	RF	84.36
4	64526	5474		147	5327	RF	84.78
5	64673	5327		59	5268	NB	84.948
6	64732	5268		63	5205	RF	85.128
7	64795	5205		32	5173	RF	85.22
8	64827	5173		22	5151	RF	85.282
9	64849	5151		18	5133	RF	85.334
10	64867	5133		16	5117	RF	85.38
11	64883	5117		13	5104	RF	85.417
12	64896	5104		10	5094	GB	85.445
13	64906	5094		15	5079	RF	85.488
14	64912	5079		13	5066	RF	85.525
15	64934	5066		8	5058	NB	85.548
16	64942	5058		12	5046	RF	85.582
17	64954	5046		6	5040	GB	85.6
18	64960	5040		6	5034	RF	85.617
19	64966	5034		3	5031	RF	85.625
20	64961	5031		1	5030	RF	85.628

Layer No. i	Train Data, $x_i = x_{i-1} + (TP, TN)_{i-1}$	Test Data, $y_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^N \frac{(TP, TN)_i}{initial, y}$
21	64970	5030	3	5027	GB	85.637
22	64973	5027	4	5023	RF	85.648
23	64977	5023	3	5020	RF	85.657
24	64980	5020	2	5018	RF	85.662
25	64982	5018	0	5018	RF = GB = NB = 0	85.662
26	64982	5018	456	4562	KNN	86.965
27	65438	4562	291	4271	KNN	87.797
28	65729	4271	152	4119	KNN	88.231
29	65881	4119	116	4033	KNN	88.562
30 (final)	65997	4003	99	3904	KNN	88.84

Table 6. Layer to layer classification result of MLDS (Train/Test Ratio: 60:40).

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
1 (initial)	42000	28000	20556	7444	GB	73.41
2	62556	7444	2444	5000	NB	82.14
3	65000	5000	786	4214	RF	84.95

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
4	65786	4214	109	4105	RF	85.339
5	65895	4105	46	4059	RF	85.503
6	65941	4059	35	4024	NB	85.628
7	65976	4024	21	4003	RF	85.703
8	65997	4003	26	3977	RF	85.796
9	66023	3977	16	3961	RF	85.853
10	66039	3961	9	3952	RF	85.885
11	66048	3952	10	3942	GB	85.921
12	66058	3942	14	3928	RF	85.971
13	66072	3928	11	3917	RF	86.01
14	66083	3917	7	3910	GB	86.035
15	66090	3910	5	3905	RF	86.053
16	66095	3905	4	3901	RF	86.067
17	66099	3901	0	3901	RF=GB=NB=0	86.067
18	66099	3901	357	3544	KNN	87.342
19	66456	3544	141	3403	KNN	87.846
20	66597	3403	210	3193	KNN	88.596
21	66807	3193	91	3102	KNN	88.921

Layer No. i	Train Data, $\mathbf{x}_i = \mathbf{x}_{i-1} + (TP, TN)_{i-1}$	Test Data, $\mathbf{y}_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^N \frac{(TP, TN)_i}{initial, y}$
22	66898	3102	81	3021	KNN	89.210
23 (final)	66979	3021	65	2956	KNN	89.442

Table 7. Layer to layer classification result of MLDS (Train/Test Ratio: 70:30).

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
1 (initial)	49000	21000	15428	5572	GB	73.46
2	64428	5572	1762	3810	NB	81.857
3	66190	3810	664	3146	RF	85.019
4	66854	3146	78	3068	RF	85.390
5	66932	3068	29	3039	RF	85.528
6	66961	3039	20	3019	RF	85.623
7	66981	3019	19	3000	RF	85.714
8	67000	3000	16	2984	RF	85.790
9	67016	2984	14	2970	NB	85.857
10	67030	2970	12	2958	RF	85.914
11	67042	2958	10	2948	GB	85.961

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
12	67052	2948	8	2940	RF	86
13	67060	2940	8	2932	RF	86.038
14	67068	2932	9	2923	RF	86.08
15	67077	2923	9	2914	GB	86.123
16	67086	2914	8	2906	RF	86.161
17	67094	2906	6	2900	RF	86.190
18	67100	2900	3	2897	RF	86.204
19	67103	2897	4	2893	RF	86.223
20	67107	2893	6	2887	RF	86.252
21	67113	2887	2	2885	RF	86.261
22	67115	2885	2	2883	GB	86.271
23	67117	2883	4	2879	RF	86.290
24	67121	2879	3	2876	RF	86.304
25	67124	2876	4	2872	GB	86.323
26	67128	2872	3	2869	RF	86.338
27	67131	2869	1	2868	RF	86.342
28	67132	2868	3	2865	RF	86.357
29	67135	2865	4	2861	RF	86.376

Layer No. i	Train Data, $x_i = x_{i-1} + (TP, TN)_{i-1}$	Test Data, $y_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^N \frac{(TP, TN)_i}{initial, y}$
30	67139	2861	2	2859	RF	86.385
31	67141	2859	1	2858	GB	86.39
32	67142	2858	1	2857	RF	86.395
33	67143	2857	1	2856	RF	86.4
34	67144	2856	3	2853	RF	86.414
35	67147	2853	3	2850	GB	86.428
36	67150	2850	3	2847	RF	86.442
37	67153	2847	0	2847	RF=GB=NB=0	86.442
38	67153	2847	265	2582	KNN	87.704
39	67418	2582	156	2426	KNN	88.447
40	67574	2426	105	2321	KNN	88.947
41	67679	2321	70	2251	KNN	89.28
42	67749	2251	64	2187	KNN	89.585
43	67813	2187	27	2160	KNN	89.714
44 (final)	67840	2160	389	1771	KNN	91.566

Table 8. Layer to layer classification result of MLDS (Train/Test Ratio: 80:20).

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
2	66290	3710	1230	2480	NB	82.285
3	67520	2480	494	1986	RF	85.814
4	68014	1986	60	1926	RF	86.242
5	68074	1926	20	1906	RF	86.385
6	68094	1906	16	1890	RF	86.5
7	68110	1890	14	1876	NB	86.6
8	68124	1876	10	1866	RF	86.671
9	68134	1866	10	1856	RF	86.742
10	68144	1856	11	1845	RF	86.821
11	68155	1845	10	1835	RF	86.892
12	68165	1835	3	1832	GB	86.914
13	68168	1832	4	1828	RF	86.942
14	68172	1828	4	1824	RF	86.971
15	68176	1824	3	1821	GB	86.992
16	68179	1821	4	1817	RF	87.021
17	68183	1817	2	1815	GB	87.035
18	68185	1815	1	1814	GB	87.042
19	68186	1814	4	1810	RF	87.071

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
20	68190	1810	2	1808	RF	87.085
21	68192	1808	1	1807	GB	87.092
22	68193	1807	1	1806	NB	87.1
23	68194	1806	1	1805	GB	87.107
24	68195	1805	1	1804	RF	87.114
25	68196	1804	1	1803	RF	87.121
26	18197	1803	2	1801	RF	87.135
27	68199	1801	1	1800	RF	87.142
28	68200	1800	1	1799	GB	87.15
29	68201	1799	4	1795	RF	87.178
30	68205	1795	1	1794	GB	87.185
31	68206	1794	2	1792	RF	87.2
32	68208	1792	1	1791	NB	87.207
33	68209	1791	1	1790	RF	87.214
34	68210	1790	0	1790	RF=GB=NB=0	87.214
35	68210	1790	161	1629	KNN	88.364
36	68371	1629	108	1521	KNN	89.135
37	68479	1521	69	1452	KNN	89.628

Layer No. i	Train Data, $x_i = x_{i-1} + (TP, TN)_{i-1}$	Test Data, $y_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^N \frac{(TP, TN)_i}{initial, y}$
38	68548	1452	45	1407	KNN	89.95
39	68593	1407	38	1369	KNN	90.221
40	68631	1369	225	1144	KNN	91.828
41 (final)	68856	1144	126	1018	KNN	92.728

Table 9. Layer to layer classification result of MLDS (Train/Test Ratio: 87.5:12.5).

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
1 (initial)	61250	8750	6465	2285	GB	73.885
2	67715	2285	776	1509	NB	82.754
3	68491	1509	308	1201	RF	86.274
4	68799	1201	36	1165	RF	86.685
5	68835	1165	20	1145	RF	86.914
6	68855	1145	8	1137	GB	87
7	68863	1137	11	1126	RF	87.131
8	68874	1126	4	1122	GB	87.177
9	68878	1122	5	1117	RF	87.234

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
10	68883	1117	4	1113	NB	87.28
11	68887	1113	2	1111	RF	87.302
12	68889	1111	4	1107	RF	87.348
13	68893	1107	6	1101	RF	87.417
14	68899	1101	5	1096	RF	87.474
15	68904	1096	1	1095	RF	87.485
16	68905	1095	2	1093	RF	87.508
17	68907	1093	5	1088	RF	87.565
18	68912	1088	2	1086	RF	87.588
19	68914	1086	2	1084	RF	87.611
20	68916	1084	2	1082	GB	87.634
21	68918	1082	1	1081	GB	87.645
22	68919	1081	1	1080	RF	87.657
23	68920	1080	2	1078	RF	87.68
24	68922	1078	0	1078	RF=GB=NB=0	87.68
25	68922	1078	94	984	KNN	88.754
26	69016	984	64	920	KNN	89.485
27	69080	920	44	876	KNN	89.988

Layer No. i	Train Data, =	Test Data, =			Classifier	Total Accuracy,
28	69124	876	20	856	KNN	90.217
29	69144	856	26	830	KNN	90.514
30	69170	830	21	809	KNN	90.754
31	69191	809	9	800	KNN	90.857
32	69200	800	7	793	KNN	90.937
33	69207	793	12	781	KNN	91.074
34	69219	781	4	777	KNN	91.12
35	69223	777	5	772	KNN	91.177
36	69228	772	3	769	KNN	91.211
37	69231	769	3	766	KNN	91.245
38	69234	766	3	763	KNN	91.28
39	69237	763	3	760	KNN	91.314
40	69240	760	1	759	KNN	91.325
41	69241	759	1	758	KNN	91.337
42	69242	758	1	757	KNN	91.348
43	69243 (preprocessed)	757 (preprocessed)	116	641	KNN	92.674
44	69359	641	56	585	KNN	93.314
45	69415	585	31	554	KNN	93.668

Layer No. i	Train Data, $x_i = x_{i-1} + (TP, TN)_{i-1}$	Test Data, $y_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^N \frac{(TP, TN)_i}{initial, y}$
46	69446	554	29	525	KNN	94
47 final	69475	525	14	511	KNN	94.16

To check the efficiency of our proposed MLDS, we have also implemented other author's models and applied our dataset to their model, and compared the result with our proposed model, as shown in [Table 10](#).

Table 10. Comparison of MLDS with other author's model using same dataset.

Authors	Method	Accuracy (%)
Our study	MLDS	94.16
Javeed et al. [9]	RSA-RF	72.53
Karen et al. [10]	PCA	61.36
Amin Ul et al. [8]	CH-PCA	66.98
	FS: Relief, CL: LR, C = 100	71.95
	FS:mRMR, CL: DT, C = 100	71.73
Domor et al. [11]	FS: LASSO, CL: RF, C = 100	68.36
	CART	74.0
Louridi et al. [12]	(SVM Kernel = Linear)	73.2
Xiao-Yan et al. [13]	Bagging+PCA	73.16
	Bagging+LDA	77

Authors	Method	Accuracy (%)
A.Geetha et al. [14]	Boosting+PCA	75.30
	Boosting+LDA	71.2
	KNN	67.53

Three different datasets Cleveland (303 instances), Hungarian (294 instances), and CHSL (1025 instances), have also been used in this research work. The First two datasets have been collected from the UCI machine learning repository, and the third has been collected from Kaggle. All attributes description of these two datasets is shown in [Table 11](#). Other Researchers commonly used these two datasets. We have applied these datasets to our proposed MLDS, tested the accuracy, and compared it with the other author's system, as shown in [Table 12](#).

Table 11. Cleveland, Hungarian, and CHSL heart disease dataset attributes description.

Attributes	Description	Type
Age	Age	Integer
Sex	Sex	Integer; (1 = male, 0 = female)
Cp	Chest pain type	Integer; (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
Trestbps	Resting blood pressure	Integer
Chol	Serum cholesterol in mg/dl	Integer
Fbs	Fasting blood sugar	Integer; (1 = true, 0 = false)
Restecg	Resting electrocardiographic results	Integer; [0,2]
Talach	Maximum heart rate achieved	Integer

Attributes	Description	Type
Exang	Exercise induced angina	Integer; (1 = yes, 0 = no)
Oldpeak	ST depression induced by exercise relative to rest	Real
Slope	The slope of the peak exercise ST segment	Integer
Number of major vessels	Number of major vessels (0–3) colored by flourosopy	Integer
Thal	Thal	Integer
Num	The predicted attribute	Integer; (0 = no, 1 = yes)

Table 12. Comparison of MLDS with another authors model's using their datasets.

Authors	Dataset Name/number of instance/Splitting ratio	Train and Test Dataset Ratio	Authors Model Accuracy	MLDS Accuracy
Javeed et al. [9]	Cleveland (297 instances)	70:30	93.33	98.88
Karen et al. [10]	Cleveland (283 instances)	70:30	98.7	98.88
Karen et al. [10]	Hungarian (294 instances)	70:30	99	99.53
Karen et al. [10]	CH (577 instances)	70:30	99.4	99.98
Amin Ul et al. [8]	Cleveland (297 instances)	K Fold = 10	89	96.66
Domor et al. [11]	Cleveland (303 instances)	70:30	93	97.77
Louridi et al. [12]	Cleveland (294 instances)	80:20	86.8	98.36

Authors	Dataset Name/number of instance/Splitting ratio	Train and Test Dataset Ratio	Authors Model Accuracy	MLDS Accuracy
Xiao-Yan et al. [13]	Cleveland, Hungary, Switzerland-Long Beach-CHSL (1025 instances)	75:25	98.6	99.56
A.Geetha et al. [14]	Cleveland (294 instances)	67:33	87	94.37

4. Discussion

The use of a multilayer dynamic prediction system increases the accuracy of predictions. Our research work found that a layer's correctly classified data can be used as a new training set for the next layer. In that case, this training set can extract new hidden patterns from incorrectly classified data. That is, accurately classified data of a layer helps to provide new knowledge to the model. The proposed model's average accuracy is significantly better than others when a large dataset is applied (Table 10). In comparison, a small dataset's calculated accuracy is very similar and close to each other as the variance of a small dataset is limited.

5. Conclusion

Cardiovascular disease is considered one of the leading causes of death around the world. Early diagnosis can help to prevent the progression of the disease. Our proposed model has achieved 88.84%, 89.44%, 91.56%, 92.72%, and 94.16% accuracy based on the different proportions (50:50, 60:40, 70:30, 80:20, and 87.5:12.5 respectively) between training and testing set of Kaggle heart disease dataset. Correctly classification probability of cardiovascular disease of the proposed system has been pointed out by the AUC curve. The proposed model has also gained a better accuracy, i.e., 98.88%, 99.53%, 99.98%, 96.66%, 97.77%, 98.36%, 99.56, and 94.37% based on different partitions of Cleveland, Hungarian, Cleveland-Hungarian (CH), and Cleveland-Hungary-Switzerland-Long Beach (CHSL) dataset. The suggested methodology has shown improved performance compared to other machine learning models. Also, the proposed model can be used to predict the risk of cardiovascular disease and help effectively provide clinical advice.

6. Future work

In the future, we will enhance this proposed method by implementing a deep learning model to scan the data to search for features that correlate and combine them to enable faster learning without being explicitly told to do so. We will try to test more features to extract the hidden pattern and improve the detection accuracy. If anyone has cardiovascular disease, we will add a module to our future research to determine the types of heart disease they have.

Data availability

We used datasets from the following: Kaggle cardiovascular dataset available at <https://www.kaggle.com/sulianova/cardiovascular-diseases> e-dataset; UCI machine learning Cleveland and Hungarian heart dis-ease dataset available at <http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>; Cleveland-Hungary-Switzerland-Long Beach (CHSL) dataset at <https://www.kaggle.com/johnsmith88/heart-disease-dataset> and available from the corresponding author upon request.

Code availability

The source code files of this research work are available from the corresponding author upon request.

Declaration of competing interest

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

Acknowledgment

This work was financially supported by Jagannath University Dept. of Computer Science and Engineering (Grant number [2020-M180305745](#)).



[Recommended articles](#)

References

- [1] Whoint
Cardiovascular diseases
[online] Available at
https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 ↗, Accessed 26th Jan 2021
[Accessed]
[Google Scholar](#) ↗
- [2] int Who
WHO | about cardiovascular diseases
[online] Available at
https://www.who.int/cardiovascular_diseases/about_cvd/en ↗, Accessed 26th Jan 2021
[Accessed]
[Google Scholar](#) ↗
- [3] C. Sowmiya, P. Sumitra
Analytical study of heart disease diagnosis using classification techniques
IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS), March 2017 (2017),
pp. 23-25, [10.1109/ITCOSP.2017.8303115](https://doi.org/10.1109/ITCOSP.2017.8303115) ↗
[Srivilliputhur, India]
[View at publisher](#) ↗ [Google Scholar](#) ↗
- [4] Whoint
[online] Available at
https://www.who.int/cardiovascular_diseases/en/cvd_atlas_01_types.pdf?ua=1/ ↗, Accessed 26th Jan 2021

[Google Scholar](#) ↗

- [5] Susmita Ray
A quick review of machine learning algorithms. Faridabad, India
International conference on machine learning, big data, cloud and parallel computing (COMITCon), vols. 14–16 (2019), pp. 35-39,
[10.1109/COMITCon.2019.8862451](#) ↗
February 2019
[View at publisher](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [6] Md Razu Ahmed, S.M. Hasan Mahmud¹, Hossin Md Altab, Jahan Hosney, Haider Noori Sheak Rashed
A cloud based four-tier architecture for early detection of heart disease with machine learning algorithms. Chengdu, China
IEEE 4th international conference on computer and communications, vols. 7–10 (2018), [10.1109/CompComm.2018.8781022](#) ↗
1951–5, December 2018
[View at publisher](#) ↗ [Google Scholar](#) ↗
- [7] J. Dhanalakshmi, N. Ayyanathan
An implementation of energy demand forecast using J48 and simple K means. Chennai, India
Fifth international conference on science technology engineering and mathematics, vols. 14–15 (2019),
[10.1109/ICONSTEM.2019.8918883](#) ↗
174–8, March 2019
[View at publisher](#) ↗ [Google Scholar](#) ↗
- [8] Amin Ul Haq, Jian Ping Li, Hammad Memon Muhammad, Nazir Shah, Ruinan Sun
A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms
Mobile Inf Syst, 2018 (2018), pp. 1-21, [10.1155/2018/3860146](#) ↗
Article ID 3860146, 2 (December 2018)
[View at publisher](#) ↗ [Google Scholar](#) ↗

- [9] Javeed Ashir, Shijie Zhou, Yongjian Liao, Qasim Iqbal, Noor Adeeb, Nour Redhwan
An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection
IEEE Access, 7 (7) (2019), pp. 180235-180243, [10.1109/ACCESS.2019.2952107](https://doi.org/10.1109/ACCESS.2019.2952107). November 2019 [↗](#)
[View at publisher ↗](#) [Google Scholar ↗](#)
- [10] A. Gárate-Escamila, A. Hajjam El Hassani, E. Andrès
Classification models for heart disease prediction using feature selection and PCA
Inform. Med. Unlocked, 19 (2020), p. 100330, [10.1016/j.imu.2020.100330](https://doi.org/10.1016/j.imu.2020.100330) [↗](#)
 [View PDF](#) [View article](#) [View in Scopus ↗](#) [Google Scholar ↗](#)
- [11] I. Mienye, Y. Sun, Z. Wang
An improved ensemble learning approach for the prediction of heart disease risk
Inform. Med. Unlocked, 20 (2020), p. 100402, [10.1016/j.imu.2020.100402](https://doi.org/10.1016/j.imu.2020.100402) [↗](#)
 [View PDF](#) [View article](#) [View in Scopus ↗](#) [Google Scholar ↗](#)
- [12] Louridi Nabaouia, Amar Meryem, El Ouahidi Bouabid
Identification OF cardiovascular diseases using machine learning
7th mediterranean congress of telecommunications (CMT) (2019), pp. 24-25, [10.1109/CMT.2019.8931411](https://doi.org/10.1109/CMT.2019.8931411) [↗](#)
fes,` Morocco, Morocco, October 2019
[Google Scholar ↗](#)
- [13] X. Gao, A. Amin Ali, H. Shaban Hassan, E. Anwar
Improving the accuracy for analyzing heart diseases prediction based on the ensemble method
Complexity, 2021 (2021), pp. 1-10, [10.1155/2021/6663455](https://doi.org/10.1155/2021/6663455) [↗](#)
[View in Scopus ↗](#) [Google Scholar ↗](#)
- [14] A. Geetha Devi, Rao Borra Surya Prasada, K. Vidya Sagar
A method of cardiovascular disease prediction using machine learning

Int J Eng Res Technol, 9 (5) (2021), pp. 243-246

IJERTCONV9IS05050 March 2021

[Google Scholar](#) ↗

[15]

Ashrafi Esfahani Hamidreza, Ghazanfari Morteza

Cardiovascular disease detection using a new ensemble classifier. Tehran, Iran

IEEE 4th international conference on knowledge-based engineering and innovation, vol. 22, KBEI (2017),

[10.1109/KBEI.2017.8324946](#). December 2017 ↗

1011–4

[Google Scholar](#) ↗

[16]

Abu Taher Kazi, Yasin Jisan Md Billal Mohammed, Rahman Mahbubur

Network intrusion detection using supervised machine learning technique with feature selection

International conference on robotics, electrical and signal processing techniques, vols. 10–12, ICREST, Dhaka, Bangladesh (2019),

pp. 643-646, [10.1109/ICREST.2019.8644161](#) ↗

January 2019

[Google Scholar](#) ↗

[17]

Shikha Mehta, Priyanka Rana, Shivam Singh, Ankita Sharma, Parul Agarwal

Ensemble learning approach for enhanced stock prediction. Noida, India

Twelfth international conference on contemporary computing, IC3 (2019), pp. 8-10, [10.1109/IC3.2019.8844891](#) ↗

Aguest 2019

[Google Scholar](#) ↗

[18]

Mung Pau Suan, Phyu Sabai

Effective analytics on healthcare big data using ensemble learning

IEEE conference on computer applications, ICCA, Yangon, Myanmar (2020), pp. 27-28, [10.1109/ICCA49400.2020.9022853](#) ↗

February 2020

[Google Scholar](#) ↗

- [19] R. Gayathri Devi, P. Sumanjani
Improved classification techniques by combining KNN and random forest with naive bayesian classifier.
Coimbatore, India
IEEE international conference on engineering and technology, ICETECH (2015), p. 20, [10.1109/ICETECH.2015.7274997](https://doi.org/10.1109/ICETECH.2015.7274997) [↗](#)
March 2015
[Google Scholar](#) [↗](#)
- [20] G. Dinesh Kumar, K. Arumugaraj, D. Santhosh Kumar, V. Mareeswari
Prediction of cardiovascular disease using machine learning algorithms. Coimbatore, India
IEEE international conference on current trends toward converging technologies, vols. 1–3 (2018), pp. 1-7,
[10.1109/ICCTCT.2018.8550857](https://doi.org/10.1109/ICCTCT.2018.8550857). March 2018 [↗](#)
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [21] Senthilkmar Mohan, Thirumalal Chandrasegar, Gautam Srivastava
Effective heart disease prediction using hybrid machine learning techniques
IEEE Access, 7 (19) (2019), pp. 81542-81554, [10.1109/ACCESS.2019.2923707](https://doi.org/10.1109/ACCESS.2019.2923707). June 2019 [↗](#)
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [22] Nasr Mona, Shaaban Essam, Samir Ahmed
A proposed model for predicting employees performance using data mining techniques: Egyptian case study. 1947–5500
Int J Comput Sci Inf Secur, 17 (1) (2019), pp. 31-40
J31121817 anuary 2019
[Google Scholar](#) [↗](#)
- [23] C. Latha, S. Jeeva
Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques
Inform. Med. Unlocked, 16 (2019), p. 100203, [10.1016/j.imu.2019.100203](https://doi.org/10.1016/j.imu.2019.100203) [↗](#)

[View PDF](#)[View article](#)[View in Scopus](#)[Google Scholar](#)

[24] R. Saravanan, Sujatha Pothula

A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. Madurai, India

Second international conference on intelligent computing and control systems, vols. 14–15 (2018), pp. 945-949,

[10.1109/ICCONS.2018.8663155](#)

June 2018

[View in Scopus](#) [Google Scholar](#)

[25] Islam Saiful, Jahan Mst Nusrat, Khatun Eshita

Cardiovascular disease forecast using machine learning paradigms. Erode, India

Fourth international conference on computing methodologies and communication, vols. 11–13 (2020), pp. 487-490,

[10.1109/ICCMC48092.2020.ICCMC-00091](#)

March 2020. M.N. Uddin and R.K. Halder

[Google Scholar](#)

[26] Minh Pham, Jing Lin, Yanjia Zhang

Diagnosing voice disorder with machine learning

IEEE international conference on big data (big data), vols. 10–13, Seattle, WA, USA (2018), pp. 5263-5266,

[10.1109/BigData.2018.8622250](#)

December 2018

[View in Scopus](#) [Google Scholar](#)

[27] Shengyu Lu, Beizhan Wang, Hongji Wang, Qingqi Hong

A hybrid collaborative filtering algorithm based on KNN and gradient boosting

13th international conference on computer science & education, vols. 8–11, ICCSE, Colombo, Sri Lanka (2018), pp. 432-436,

[10.1109/ICCSE.2018.8468751](#)

A guest 2018

[View in Scopus ↗](#) [Google Scholar ↗](#)

[28] M. Trabelsi, N. Meddouri, M. Maddouri

A new feature selection method for nominal classifier based on formal concept analysis

Procedia Comput. Sci., 112 (2017), pp. 186-194, [10.1016/j.procs.2017.08.227](https://doi.org/10.1016/j.procs.2017.08.227) ↗

 [View PDF](#) [View article](#) [View in Scopus ↗](#) [Google Scholar ↗](#)

[29] Correlationattributeeval

Infochim.u-strasbg.fr

[Accessed]

<http://infochim.u-strasbg.fr/cgi-bin/weka-3-9-1/doc/weka/attributeSelection/CorrelationAttributeEval.html> ↗ (2020), Accessed 29th Dec 2020

[Google Scholar ↗](#)

[30] National Institute on Aging

Heart Health and aging

(2021)

<https://www.nia.nih.gov/health/heart-health-and-aging#:~:text=Adults%20age%2065%20and%20older,risk%20of%20developing%20cardiovascular%20disease/> ↗, Accessed 26th Jan 2021

Accessed

[Google Scholar ↗](#)

[31] Ba, Social Distancing Q

Cholesterol and heart disease

[online] WebMD. Available at

<https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk#1> ↗, Accessed 26th Jan 2021

Accessed

[Google Scholar ↗](#)

- [32] Information H, overview D, problems P, diabetes &, diabetes a, center T, Health N. Diabetes, heart disease, and stroke | NIDDK
[online] national institute of diabetes and digestive and kidney diseases
<https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke#:~:text=Over%20time%2C%20high%20blood%20glucose,you%20will%20develop%20heart%20disease.&text=People%20with%20diabetes%20tend%20to,age%20than%20people%20without%20diabetes> ↗, Accessed 26th Jan 2021
Accessed
[Google Scholar](#) ↗
- [33] Cardiosecurcom
High blood pressure – causes and connection to heart attacks | cardiosecur
[online] Available at
<https://www.cardiosecur.com/magazine/special-articles-on-the-heart/high-blood-pressure-causes-and-connection-to-heart-attacks> ↗
[Accessed 26 January 2021]. Accessed
[Google Scholar](#) ↗
- [34] [online] Available at
<https://health.usnews.com/health-care/for-better/articles/2016-10-19/how-obesity-can-affect-your-heart#:~:text=Obesity%20leads%20to%20heart%20failure,eventually%20lead%20to%20heart%20failure> ↗
[Accessed 26 January 2021]. Accessed
[Google Scholar](#) ↗
- [35] Clinic Cleveland. Obesity & heart disease [online] Available at: . [Accessed 26 January 2021]. [Accessed].
[Google Scholar](#) ↗
- [36] MedicineNet
Low blood pressure symptoms, chart, causes, and treatments
[online] Available at

https://www.medicinenet.com/low_blood_pressure/article.htm ↗, Accessed 26th Jan 2021

[Accessed]

[Google Scholar](#) ↗

Cited by (0)

© 2021 The Author(s). Published by Elsevier Ltd.



All content on this site: Copyright © 2024 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the Creative Commons licensing terms apply.

