Transactions on Machine Intelligence 6(4) 2023 198-207

ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/

Diagnosis of Atherosclerosis of the Coronary Arteries of the Heart with Data Mining and Machine Learning Techniques

S. Asadzadeh[1], M.A. Shaygan[2,*]

[1] Department of Computer Engineering, Islamic Azad University, Shiraz Branch, Shiraz, Iran.
[2] Assistant Professor, Department of Computer Engineering, Islamic Azad University, Shiraz Branch, Shiraz, Iran

ARTICLE INFO

Article History:
Received 2 September 2023
Received in revised form 8 October 2023
Accepted 5 November 2023
Available online 11 November 2023

Keywords:
Data Mining, Classification, Coronary Artery Disease, Angiography

ABSTRACT

Cardiovascular diseases are the leading cause of death in the world. In this regard, the rapid and timely diagnosis of heart diseases and the prediction of certain risk events associated with the cardiovascular system are among the top priorities of researchers. Due to the risks of invasive diagnostic methods in coronary artery disease, such as angiography, providing a suitable and non-invasive method for timely diagnosis, increasing accuracy, reducing errors in decision-making, reducing treatment costs and improving the quality of services provided by physicians has been the main goal of this research. In the implementation of this practical research, the Cleveland medical data set, consisting of 270 samples with 76 features, and Z-AlizadehSani data set, consisting of 303 samples with 54 features, available in the UCI standard data repository, were used. Initially, preprocessing and feature selection, followed by modeling, data processing and analysis was performed by examining the effect of disease parameters on coronary artery stiffness using a combination of machine learning algorithms. The proposed system, based on accuracy, sensitivity, specificity, and AUC indices, was able to achieve the best performance with the lowest error compared to similar research. Based on the results obtained, the proposed model can prevent potential adverse effects and damages of some invasive procedures such as angiography in patients who do not need it. Moreover, the system can help physicians triage patients who definitely need these diagnostic procedures in order to receive timely treatment with the highest precision.

1. INTRODUCTION

Coronary artery disease, or arterial stiffness in the coronary vessels, is the main cause of heart attacks. Early diagnosis of coronary artery disease is crucial due to the risks it poses to a person's health [1]. Coronary angiography, a type of invasive test as part of a process called cardiac catheterization [2], can diagnose and treat heart and vascular diseases. However, angiography can lead to complications such as bleeding and artery damage, heart attack, stroke,

* Corresponding Author: ma.shayegan@iau.ac.ir
Assistant Professor, Department of Computer Engineering, Islamic Azad University, Shiraz Branch, Shiraz, Iran

http://dx.doi.org/10.47176/TMI.2023.198

kidney damage from contrast dye, tissue damage from x-rays, and even death [3]. Therefore, providing a suitable and non-invasive method for diagnosing coronary artery disease is highly important in the field of health.

Data mining can be effective in preventing, predicting, diagnosing, and treating diseases and providing post-discharge care [4]. Medical data mining has great potential for detecting hidden patterns in data, which can be used for clinical diagnosis [5].

The aim of this study was to use machine learning algorithms to diagnose coronary artery stiffness in a timely manner, with increased accuracy and reduced error in decision-making, as well as to reduce treatment costs and improve the quality of services provided by physicians to patients. One of the achievements of this study is the development of an intelligent system for screening coronary artery stiffness, which is highly important because it prevents potential adverse effects and damages of invasive procedures such as angiography in patients who do not need it and also saves cost in diagnostic tests for these individuals.

The article is organized as follows: In section 2, the theory and background of the research are presented, and in section 3, the details of the proposed method are described in 6 steps. Finally, section 4 presents the evaluation and results of this study.

2. THEMATIC RESEARCH LITERATURE

The prevalence of coronary artery disease plays a crucial role in the mortality of the human population, making timely diagnosis with non-invasive methods extremely important for doctors and researchers. Many researchers have attempted to assist doctors in diagnosing the disease using non-invasive methods such as exercise testing, clinical information, electrocardiograms, and blood tests with different algorithms and approaches.

In the study [6] (Kumar et al.) the classification and prediction of heart disease were discussed using the UCI Cleveland database. The algorithms examined included ANN, KNN, and CNN, with CNN performing the best.

In the study [7] (Yong et al.) proposed an optimal classification algorithm, Co-SVM, using clinical datasets of heart patients to classify the disease with decision trees, neural networks, support vector machines, and Co-SVM. The empirical results showed that the proposed Co-SVM algorithm was more accurate than the other three classic algorithms. In the study [8] (sayadi et al.) used machine learning techniques to detect early heart disease on the well-known Z-Alizadeh Sani dataset, using the Pearson correlation feature selection method to identify the most effective features. Then, machine learning techniques such as decision trees, deep learning, logistic regression, random forest, support vector machines, and Xgboost were used based on a semi-random classification framework. The results showed that logistic regression and SVM had the same performance with an accuracy of 95.45%, sensitivity of 95.91%, feature of 99.61%, and F1 score of 96.90%. In the study [9] (maleki et al.) used a new feature selection method called Shahin Harris that combined decision trees and KNN to evaluate their proposed method using two medical data sets of 303 heart patients from the Cleveland and Z-Alizadeh-Sani datasets. The results showed that feature selection using the Shahin Harris algorithm combined with the machine learning method led to an increase in accuracy, with an accuracy rate of 98.0% using decision trees in the Z-Alizadeh-Sani dataset and a rate of 78.0% using KNN. In the study [10] (Dezhallod et al.) used the optimized beetle algorithm and machine learning KNN to diagnose CAD in a dataset of 680 heart patients with an accuracy rate of 98.99%. In addition, In the study[11] (Abdar et al.) used the genetic algorithm, particle swarm optimization algorithm, and support vector machine to diagnose heart disease with an accuracy rate of 83.09%. Finally, In the study[12] (Altashi et al.) used a combination of Shahin Harris algorithm and machine learning algorithms for feature selection and optimization methods such as grey wolf optimizers and SVM to achieve an accuracy rate of 98.93% in diagnosing coronary heart disease.

Based on the studies conducted, it has been determined that selecting features has a significant impact on the performance of machine learning and regardless of the evaluation criteria of machine learning models, determining effective features is critical. Metaheuristic algorithms have not been successful in reducing the number of features and selecting effective ones, although this initiative can increase the accuracy of diagnosis. Although each individual machine learning algorithm has successful performance, none of them alone has the highest accuracy when applied to various problems. Therefore, it seems that combining multiple algorithms with each other will lead to increased accuracy and decision-making with less error to address the problems of previous methods.

3. RESEARCH METHODS

The implementation of this research was conducted in 6 steps:

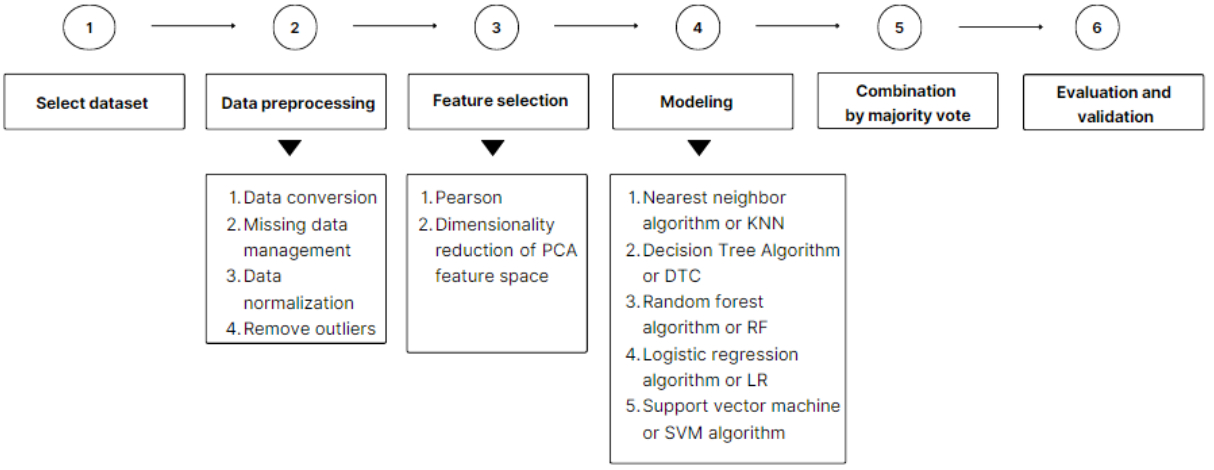


Fig. 1. Proposed Algorithm

3.1. Dataset Selection

In this applied research, the medical data sets Cleveland [13], consisting of 270 samples with 76 features, and Z-AlizadehSani [14], consisting of 303 samples with 54 features in the UCI standard data repository were used. After preprocessing the data and feature selection, only 20 features from the Cleveland dataset and 26 features from the Z-AlizadehSani dataset were selected because other features were practically unusable due to having values that were too unavailable or data imbalance. Some of the data samples from both datasets are shown in Table 1. 70% of the data was used for training, 15% for validation, and 15% for testing. In the first step, the proposed model was trained using the training data. Then, in the next step, the validation data was used to evaluate the system's performance before testing it. If the system performs satisfactorily on the validation data, the implemented system is used to predict the labels of the test data.

Table 1. Part of medical data samples from two datasets (a: Cleveland, b: Z-AlizadehSani

Sample	AGE	Gender	Chest pain	blood pressure
1	63	Male	typical angina	145
2	67	Male	asymptomatic	160
3	67	Male	asymptomatic	120
4	37	Male	non-anginal	130
5	41	Female	pain	130

(a)

Sample	AGE	Gender	Height	Gender
1	53	90	175	Male
2	67	70	157	Female
3	54	54	164	Male
4	66	67	158	Female
5	50	87	153	Female

(b)

3.2. Data Preprocessing

Most machine learning algorithms require complete and clean databases. This is because any impurities (such as missing data, outliers, and qualitative data) can affect the desired results. Therefore, data preprocessing is one of the most critical implementation stages in machine learning systems. The better this stage is done in data mining, the higher the quality of the output of the algorithms and data mining techniques. In the following, the most important activities performed in the data preprocessing of this research are introduced.

3.2.1. Data Transformation

The datasets used in this research consist of quantitative and qualitative data. Since the data mining algorithms used require numerical data and their learning structure is based on learning from numerical matrices, One Hot Encoding technique was used to convert qualitative data to numerical values understandable by the algorithms. Table 2 shows an example of these transformations (such as the gender and chest pain type feature columns).

Table 2. An example of transforming qualitative data to quantitative values using One Hot Encoding technique.

Feature title						
Type of chest pain				**Gender**		**Sample**
No sign	Non-anginal pain	Uncommon angina	Typical angina	Male	Female	
0	0	0	1	0	1	1
0	0	1	0	1	0	2
0	1	0	0	0	1	3
1	0	0	0	1	0	4

Data Transformation

Feature title		
Type of chest pain	**Gender**	**Sample**
Typical angina	Female	1
Uncommon angina	Male	2
Non-anginal pain	Female	3
No sign	Male	4

3.2.2. Missing Data Handling

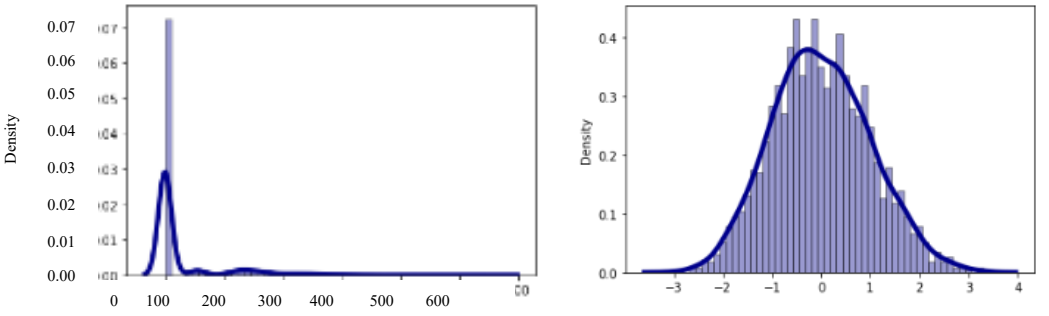
In datasets, some of the features related to one or more samples may lack valid values. This can have various reasons, such as noisy recorded data, lack of recording, or invalid values. In the preprocessing stage, some features were completely lost and had a large number of missing values. Therefore, to maintain the integrity of the dataset, these features were removed. To fill and replace missing values, the k-nearest neighbor algorithm with k=7 Euclidean distance was used to identify the nearest neighbors of the query point, and by calculating the distance between the query point and neighbors, the nearest value was predicted and missing values were imputed.

3.2.3. Data Normalization

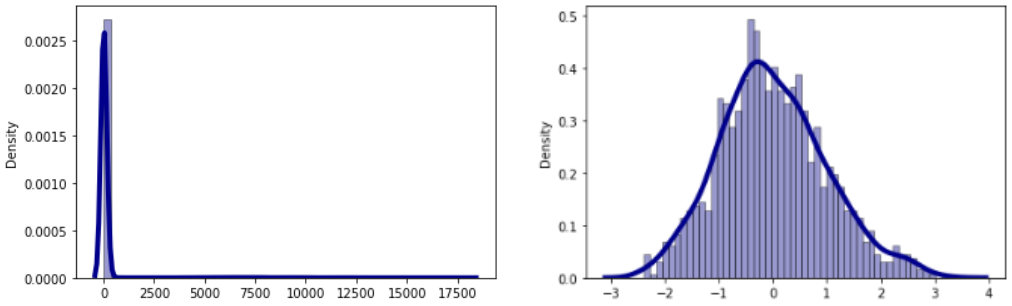
By examining the data distribution, it was determined that the collected data has skewness with a non-normal distribution. Therefore, standard normalization method (Method Z-Score) was used to normalize the data distribution, and the data was unscaled with (μ=0) zero mean and standard deviation equal to (σ =1) one (Figure 1, Formula 1).

$$\mathbf{Z} = \frac{\mathbf{X} - \boldsymbol{\mu}}{\boldsymbol{\delta}} \tag{1}$$

where μ is the mean value and σ is the standard deviation.



(a) Cleveland data set

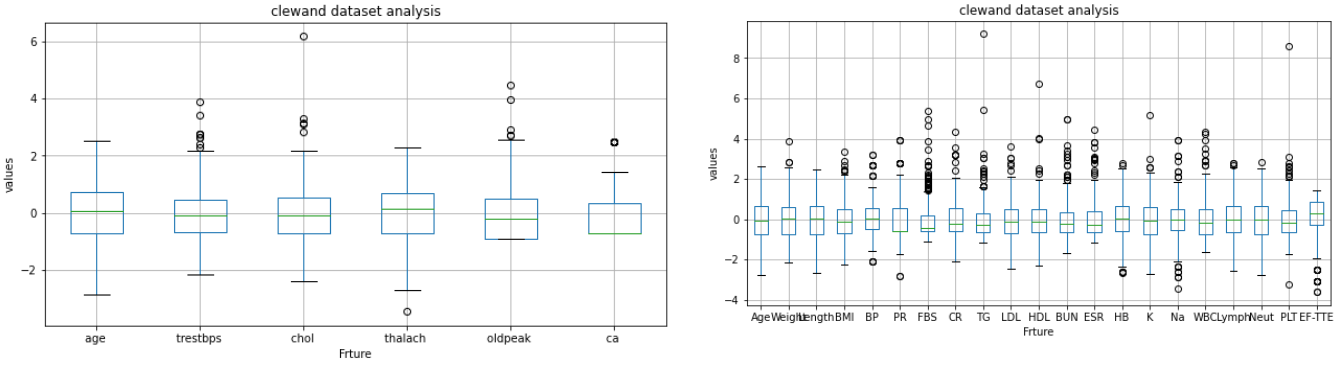


(b) Z-AlizadehSani dataset

Fig. 2. Standard normalization of data distribution

3.2.4. Outlier Data Removal

After normalizing the data, by drawing a box plot (Figure 2), it was observed that some data does not match the behavior or pattern of other data, which is called outlier data. In this study, the Tukey method was used to detect and clean the dataset from these inconsistent data.



(a) Cleveland data set

(b) Z-AlizadehSani dataset

Fig. 3. Standard normalization of data distribution

3.3. Feature Selection

Feature selection (dimensionality reduction) process aims to develop a prediction model. Selecting a suitable set of features during the design of machine learning models improves the performance, accuracy, and efficiency of learning methods. It also reduces the risk of overfitting of the model. Here, dimensionality reduction was performed by method Feature mapping with algorithm PCA based on sparseness.

3.3.1. Pearson Correlation Coefficient

One of the most important processes in feature selection is identifying the correlation between features. One of the most famous methods for measuring the dependency between two quantitative variables is calculating the Pearson correlation coefficient. This coefficient measures the distance or relative correlation between two variables, and its value ranges from +1 to -1. If the obtained value is positive, it means that the changes in the two variables occur in the same direction, i.e., with an increase in one variable, the other variable also increases. If the value is negative, it means that the two variables act in the opposite direction. The Pearson correlation coefficient between two random variables is defined as their covariance divided by the product of their standard deviations (Equation 2).

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma X \sigma Y} = \frac{E(X-\mu x)(Y-\mu y)}{\sigma X \sigma Y} \quad (2)$$

In Equation 2, COV represents covariance, σX represents the standard deviation of variable X ،μx, and represents the mean of variable X ،E.

3.3.2. *Feature Space Dimension Reduction (PCA)*

Method PCA focuses on the principal components of the feature matrix that preserve the maximum variance. One of the important parameters of Method PCA ، n_components is assigned in two ways depending on the problem. If its value is greater than 1, most features are returned by Method PCA , and in this case, Method PCA cannot generate the optimal number of features. If the parameter value n_components is between 0 and 1, Method PCA returns the least number of features that preserve this difference. In this study, n_components=99% was considered for Method PCA, meaning that 99% of the variance of the principal features is preserved. The next parameter was whiten = True, which changes the values of each principal component so that they have a unit variance and a zero mean. Finally, effective features were obtained from the entire dataset. Cleveland dataset resulted in 14 features, and Z-AlizadehSani dataset resulted in 22 features.

3.4. Modeling

This research was implemented in Python using supervised learning. In this study, five classification algorithms were used: K- Nearest Neighbor (KNN), Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM). Each algorithm estimated its prediction, and then a voting rule was applied among all predictions to select the class with the highest vote as the final class.

In the KNN algorithm, the number of neighbors k = range(1 , 30) was adjusted to find the optimal parameter A. Then, the Minkowski similarity measure with a value of P = 2 was used, and the distance between the data was calculated using the Euclidean distance to determine similarity.

In the Decision Tree algorithm, the Gini index (Equation 3) was used because we were looking for the shortest decision tree. The Gini index is a statistical measure of distribution. In machine learning, the Gini index is used to measure how often a randomly chosen element would be incorrectly identified if it were randomly labeled according to the distribution of labels in the dataset.

$$Gini = 1 - \sum_{j} p_j^2 \quad (3)$$

In Equation 3, $\sum_j p_j^2$ the sum of the square of the probabilities of all classes is obtained. In decision tree algorithms, the first goal is to select the best feature from among all features for the root node and then select the best features for the inner layers of the tree in a nested manner.

The Random Forest algorithm is a combination algorithm that uses several decision trees. In fact, a collection of decision trees generates a forest, and this forest can make better decisions (compared to a single tree). In the Random Forest algorithm, a forest of 1000 trees was used to generate the forest and determine the Gini index. Then, a subset of data was applied to each of the trees. Finally, the Random Forest algorithm chooses the class with the highest vote by using voting.

Logistic Regression is the same as regression that is obtained by applying a logistic function (sigmoid function) to regression. This classifier was created by determining the cost function (Equation 4) for optimizing the algorithm parameters.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & y = 1 \\ -\log(1 - h_\theta(x)) & y = 0 \end{cases} \quad (4)$$

In Equation 4, h_θ the sigmoid function, x prediction of the algorithm, and y the actual label are used. The Support Vector Machine (SVM) is a classification algorithm that in the simplest cases (binary classification), a linear structure model (Equation 5) is used, which is very similar to what is used in multi-layer perceptron neural networks.

$$y_j = f\left(\sum_{i=0}^{m} x_i w_{i,j}\right) \tag{5}$$

where $w_{i,j}$ indicates the weight related to the connection of input i to neuron j, i is the input number and j is the neuron number, x_i is the input number i and y_j is the output of neuron number j.

3.5. Ensemble of Majority Voting Classification Algorithms

Ensemble is a statistical and computational learning method based on creating a set of hypotheses and combining them to better evaluate training samples and achieve the highest accuracy and lowest error [15]. When combining independent and different decision makers, since each of these decision makers will perform better than a random guess at worst, the probability of making the correct decision will be reinforced. Finally, for detecting the class or position of the test sample, the output of all models is aggregated. The combination model makes a decision about the class of the training sample based on the majority vote among the outputs of classifiers. In other words, individual classifiers are combined, considering the majority vote in their decision-making, and the class that was selected by the majority of classifiers is introduced as the final class.

3.6. Evaluation and Statistical Analysis

In this study, 4 common evaluation and validation criteria for prediction models, including accuracy, error, precision, and sensitivity (according to Equations 6, 7, 8, and 9, respectively) were used. In these equations, TN is the number of records whose actual class is negative, and the classification algorithm has correctly classified them as negative. TP is the number of records whose actual class is positive, and the classification algorithm has correctly classified them as positive. FP is the number of records whose actual class is negative, and the classification algorithm has mistakenly classified them as positive. Finally, FN is the number of records whose actual class is positive, and the classification algorithm has mistakenly classified them as negative.

$$\text{Accuracy} = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{6}$$

$$\text{Eror Rate} = \frac{FN + FP}{TN + TP + FN + FP} \tag{7}$$

$$\text{Precision} = \frac{Tp}{Tp + Fp} \tag{8}$$

$$\text{Recall} = \frac{Tp}{Tp + Fn} \tag{9}$$

Equation 10 is an evaluation criterion AUC with curve ROC that is used to evaluate the quality of detection and classification of classifiers. The larger the value of this criterion for a classifier, the better the final performance of the classifier is assessed.

$$\text{AUC} = \frac{\frac{Tp}{Tp + Fn}}{\frac{Fp}{Tn + Fp}} \tag{10}$$

In Tables 3 and 4 as well as Figures 4 and 5, the evaluation criteria and the results of the proposed method have been shown for both datasets used in this study.

Table 3. Classification evaluation criteria in the proposed algorithm (Cleveland dataset)

Model	Accuracy	Eroor Rate	Precision	Recall	AUC
DTC	88%	12%	96%	95%	90%
LR	75%	25%	77%	75%	90%
RF	98%	2%	98%	97%	98%
KNN	85%	15%	98%	75%	89%
SVM	88%	12%	98%	75%	90%
ENSEMBLE	99%	1%	99%	99%	99%

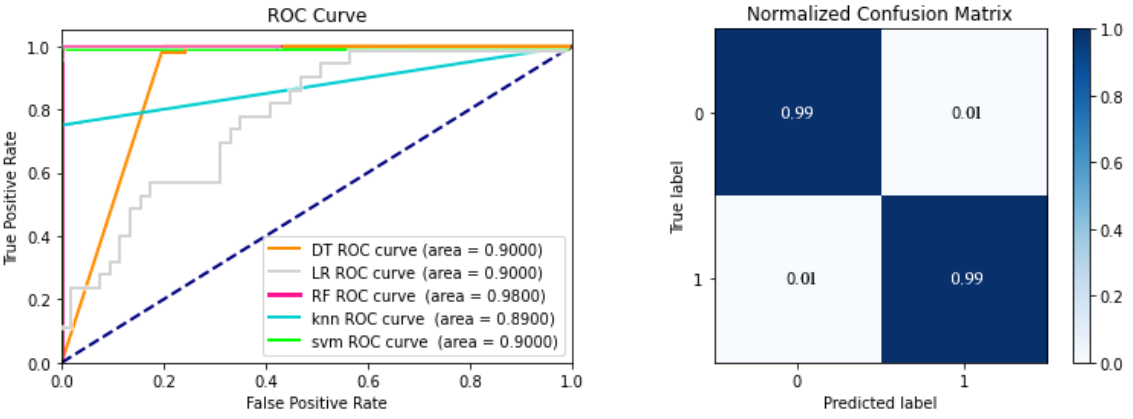


Fig. 4. Confusion matrix and ROC curve and checking the efficiency of the algorithms in the Cleveland dataset Cleveland

Table 4. Classification evaluation criteria in the proposed algorithm (Z-AlizadehSani dataset)

Model	Accuracy	Eroor Rate	Precision	Recall	AUC
DTC	88%	12%	80%	100%	90%
LR	77%	23%	75%	75%	95%
RF	100%	0%	100%	100%	100%
KNN	88%	12%	100%	75%	87%
SVM	88%	12%	100%	75%	100%
ENSEMBLE	100%	0%	100%	100%	100%

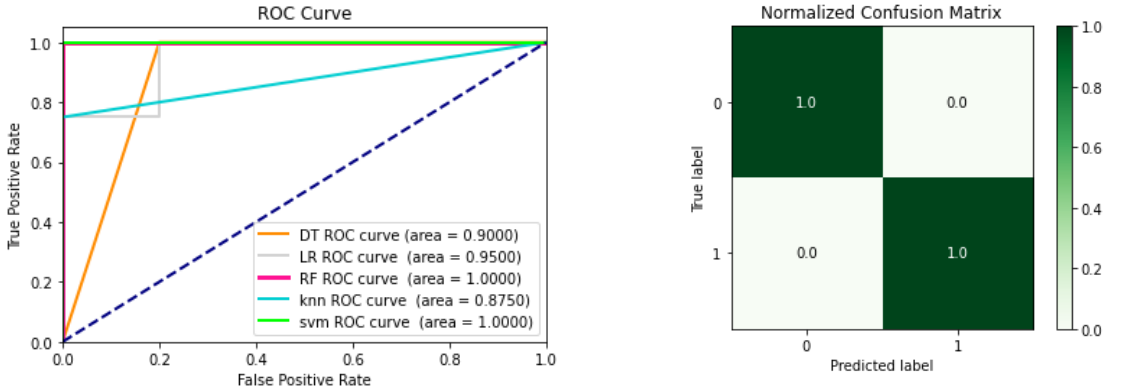


Fig. 5. Confusion matrix and ROC curve and checking the efficiency of the algorithms in the Z-AlizadehSani dataset

The results show that the values of the evaluation criteria for the proposed method were better compared to the studied methods. For example, the error rate for the proposed method in dataset Cleveland and Z-AlizadehSani was 1% and 0% respectively, while it was 23%, 12%, and 5% for the best studied method. The performance of the proposed method in other criteria also demonstrated its efficiency compared to the studied methods, which indicates the effectiveness of using a combined approach of algorithms.

4. RESULTS AND DISCUSSION

Various algorithms have been introduced for classification, but none of them on their own can achieve the desired accuracy. Combining several algorithms together leads to higher accuracy in classification. Each learning algorithm is a specific model with its own assumptions, and using specific assumptions for one method for other problems can lead to errors. Furthermore, learning is an ill-conditioned problem that, given limited data, can lead to convergence to different answers. Although precise settings in the learning process will lead to the highest accuracy on validation data, achieving such precise settings is a very complex task. Additionally, even a model with the highest accuracy

may not have enough accuracy on some data, but combining several learning models properly increases the final accuracy.

This study can assist physicians in triaging high-risk patients to secondary care more effectively, prevent unnecessary treatments, and minimize invasive and expensive procedures such as angiography, as well as help in related healthcare, prediction, and treatment, and save costs for patients and healthcare facilities.

5. SUGGESTIONS

The real value of this study is to serve as a criterion for designing future studies and evaluating techniques for diagnosing coronary artery disease in patients who are usually examined by specialists. With increasing data and the use of more powerful processing systems, as well as further research in artificial intelligence, machine learning, and data mining, higher accuracy and minimum error can be achieved on standard data while considering the level of clinical expertise of specialists.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCE

[1] Fadhil, M., Satria, I., & Miftah, M. (2022). Study of feature extraction algorithms on photoplethysmography (PPG) signals to detect coronary heart disease. In *2022 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE. https://doi.org/10.1109/icodsa55874.2022.9862855

[2] Maleki, M., Alizadehasl, A., & Haghjoo, M. (2021). *Practical cardiology: Principles and approaches* (2nd ed.). Elsevier Health Sciences.

[3] Yaqoob, A., Ali, T. S., Barolia, R., & Hasnani, F. B. (2022). Risk factors associated with complications of coronary angiography at a tertiary care hospital in Karachi, Pakistan. *Asian Journal of Allied Health Sciences (AJAHS), 7*(1). https://doi.org/10.52229/ajahs.v7i1.1605

[4] Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques. *ICT Express, 8*(1), 109–116. https://doi.org/10.1016/j.icte.2021.08.021

[5] Shaikh, A. A., Doss, A. N., Subramanian, M., Jain, V., Naved, M., & Mohiddin, M. K. (2022). Major applications of data mining in medical. *Materials Today: Proceedings, 56*, 2300–2304.

[6] Pallathadka, H., Naved, M., Phasinam, K., & Arcinas, M. M. (2022). A machine learning based framework for heart disease detection. *ECS Transactions, 107*(1), 8667–8673. https://doi.org/10.1149/10701.8667ecst

[7] Peng, J., Zhang, X., Wang, L., Zhu, F., Zhou, N., Zuo, Y., … Gao, Y. (2022). Research on application of data mining algorithm in cardiac medical diagnosis system. *BioMed Research International, 2022*, Article 7262010. https://doi.org/10.1155/2022/7262010

[8] Sayadi, M., Varadarajan, V., Sadoughi, F., Chopannejad, S., & Langarizadeh, M. (2022). A machine learning model for detection of coronary artery disease using noninvasive clinical parameters. *Life, 12*(11), 1933. https://doi.org/10.3390/life12111933

[9] Maleki, S., & Zare Mehrjerdi, Y. (2022). Diagnosis of coronary artery disease by bat and Harris hawk meta-heuristic optimization algorithms and machine learning methods. *Journal of Health Administration, 25*(1), 57–68. https://doi.org/10.52547/jha.25.1.57

[10] DezhAloud, N., & Soleimanian Gharehchopogh, F. (2020). Diagnosis of heart disease using binary

grasshopper optimization algorithm and K-nearest neighbors. *Journal of Health Administration, 23*(3), 42–54. https://doi.org/10.29252/jha.23.3.42

[11] Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine, 179*, Article 104992. https://doi.org/10.1016/j.cmpb.2019.104992

[12] Al-Tashi, Q., Rais, H., & Jadid, S. (2019). Feature selection method based on grey wolf optimization for coronary artery disease classification. In F. Saeed, N. Gazem, F. Mohammed, & A. Busalim (Eds.), *Recent trends in data science and soft computing: IRICT 2018. Advances in intelligent systems and computing* (pp. 257–266). Springer.

[13] Alizadehsani, R., Roshanzamir, M., & Sani, Z. (2013). Z-Alizadeh Sani [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5Q31T

[14] Center for Machine Learning and Intelligent Systems. (2010). Cleveland heart disease data details. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heartdisease

[15] Asadzadeh, S., & Ravaei, B. (2023). Diagnosis of breast cancer at the molecular-cellular level with an artificial intelligence approach. *Journal of Modeling in Engineering, 21*(72), 19–30. https://doi.org/10.22075/jme.2022.26164.2261