ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 269-277 (2020)

www.astesj.com

Special Issue on Innovative Research in Applied Science, Engineering and Technology

ASTESJ

ISSN: 2415-6698

Supervised Machine Learning Based Medical Diagnosis Support System for Prediction of Patients with Heart Disease

Oumaima Terrada[1], Soufiane Hamida[1], Bouchaib Cherradi[1, 2], Abdelhadi Raihani[1,*], Omar Bouattane[1]

[1]*Signals, Distributed Systems and Artificial Intelligence laboratory (LSSDIA), ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, 28820, Morocco*

[2]*STIE Team, CRMEF Casablanca-Settat, provincial section of El Jadida, El Jadida, 24000, Morocco*

ARTICLE INFO

Article history:
Received: 20 July, 2020
Accepted: 01 September, 2020
Online: 17 September, 2020

Keywords:
Atherosclerosis
Machine learning techniques
Artificial intelligence

ABSTRACT

Application in the field of medical development has always been one of the most important research areas. One of these medical applications is the early prediction system for heart diseases especially; coronary artery disease (CAD) also called atherosclerosis. The need for a medical diagnosis support system is to detect atherosclerosis at the earlier stages to optimize the diagnosis, avoid the advanced cases, and reduce treatment costs. Earlier, the datasets are collected from specific medical sources and have evaluated against computer applications. In this paper, a supervised machine learning medical diagnosis support system (MDSS) for atherosclerosis prediction is presented that able to obtain and learn automatically knowledge from each patient's clinical data. Therefore, we used three Machine Learning (ML) classifiers for the proposed MDSS for atherosclerosis. Thus, this work is accomplished using databases collected from the UCI repository (Cleveland, Hungarian) and Sani Z-Alizadeh dataset. The performance metrics were computed utilizing Accuracy, Recall and Precision. Furthermore, F1-score and Matthews's correlation coefficient these measures were also calculated to greatly increase the proposed system performance. Additionally, 10-fold cross-validation methods have been used for proposed model performance evaluation that achieved 94% as the best accuracy average. Consequently, the proposed model can be used to support healthcare and facilitate large-scale clinical diagnostic of atherosclerosis diseases.

1. Introduction

As stated by the World Health Organization (WHO), heart disease is one of the leading causes of death when the heart is unable to pump oxygenated blood through the body [1]. There are other forms of Cardiovascular Disease (CVD), including coronary artery disease (CAD), also called atherosclerosis. This disease narrowed arteries and buildup of plaque caused by cholesterol in the blood. This ailment occurs due to narrowed or blocked blood vessels and coronary arteries because of the plaque accumulation. This plaque is made of cholesterol, calcium and other substances. As the buildup increases, the plaque reduces blood flow to the coronary arteries. Therefore, the flow in the myocardium decreases. This can cause symptoms such as angina. The pain can be in the chest, shoulder, abdomen, arms, and neck. During this pain, the oxygenated blood decreases. This situation called myocardial ischemia. When the coronary artery has near completely narrowed, the myocardium tissue dies and leading a heart attack (myocardial infarction) [2,3].

Here, it seems important to establish and develop a medical diagnostic support system (MDSS) to automate the classification and prediction of CVD. However, medical diagnostic research requires greater precision and efficiency to make the best clinical decisions. Although classical MDSS has proven its ability to solve most diagnostic problems, it offers a lower accuracy factor and is unable to make a correct diagnosis [4,5].

Recently, therapy systems and medical diagnostics using Machine Learning (ML) is a wide-ranging section of artificial intelligence (AI). These technologies have influenced scientific fields such as finance, applied sciences, biology and medical

*A. Raihani, University Hassan II of Casablanca, abraihani@yahoo.fr

applications [6–14]. Subsequently, several works have been proposed to develop (MDSS) in order to predict and classify patients with heart diseases to improve health care [15–23].

In this case, we propose a new MDSS using some selected ML algorithms. The main goal is to classify and predict the patient's health issue based on the principal chosen features by analyzing the heart disease databases. Atherosclerosis risk factors have been identified from the knowledge and the expertise of medical experts and doctors. These risk factors are known as uncontrollable risk factors and controllable risk factors. The identification of these factors is based on several features. Uncontrolled Atherosclerosis risk factors contain family history, age and gender [3].

The remainder of this paper is structured as follows: In the second part (Section II), we review some related work in the literature. In the third part (Section III), we have presented and explained our proposed system process. In particular, we present the global flowchart of the proposed MDSS and the selected machine learning algorithms; in addition to used CAD datasets. The fourth part (Section IV) describes the evaluation parameters used to assess and compare our MDSS performance with similar measures. In the fifth part (Section V), we showed the details of implementation and presented the results and discussions. The last part (Section VI) concluded this work and gave certain proposed perspectives.

2. Related work

In this part, we have presented several selected works from literature review on automatic heart disease diagnosis. These works used the same well-known databases and that we will consider later for the performance comparison.

In [15] , The authors applied neural network integration methods to build new models by linking predicted values from previous models. Compared to the ML algorithm, the accuracy rate presented 89.01%. Another work published in [16], the authors suggested a clinical decision support system (CDSS) using Weighted Fuzzy Rules (WFR) for predicting heart disease. They used two scenarios of evaluation; the first scenario automatizes the approach for the WFR generation while the second scenario develops a fuzzy rule-based CDSS. They tested their CDSS using the Cleveland's heart disease database. Compared to the system based on a neural network, the best precision value obtained by this method is 62.35%.

In [17], the authors applied Fast Decision Tree (FDT) and C4.5 tree pruning methods. This approach aims to integrate the machine learning analysis results in different CAD databases. The outcomes showed that the classification accuracy is 78.06% which is higher than the average classification accuracy of separate datasets of 75.48%. Recently in 2017, the authors in[18] proposed a Hybrid Neural Network-Genetic (HNNG) to improve the neural network by strengthening its initial weights based on a genetic algorithm. The highest accuracy rate is 93.85% using Z-Alizadeh Sani data set and the Cleveland's heart disease database.

Other approaches have covered the medical diagnosis issue of heart diseases. In [19], the authors have depicted the CDSS performances for heart failure risk prediction. This system based on two methods, Fuzzy Analytic Hierarchy Process (Fuzzy_AHP) and artificial neural network (ANN). The result shows that compared to the traditional ANN method, the average prediction accuracy of this method reaches 91.10%. More recently, in 2018 the authors of [20] presented the design and implementation of the MDSS for heart diseases. This system is developed using the Fuzzy_AHP method and Fuzzy Inference System (FIS). The results of the developed method indicate the possibility of having a heart disease. From the experimental results it has been proven that the AI and ML methods in the medical field have given good results. In [24], used ML methods, which are Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), ANN, and K-Nearest Neighbours (KNN) algorithms. These ML methods used to improve CAD diagnosis. The reached average accuracy is higher than 80%. As well, specificity and sensitivity results are around 70% to 90%.

Too recently in [22], the authors developed a new method, called Hybrid Feature Selection (2HFS) applying Gaussian Naive Bayes (GNB), Random Forest (RF), Decision tree (DT) and Gradient Boosting (XGBoost) classifiers. In this study, authors have used Nasarian CAD database and they have also tested this approach with Long Beach VA, Hungarian and Z-Alizadeh Sani databases to achieve accuracies of 83.94%, 81.58% and 92.58% respectively.

This work aims to propose a new MDSS for diagnosis of patients with atherosclerosis. The proposed approach is based on five some selected ML algorithms: ANN, RF, Adaptive Boosting (AdaBoost), DT, and XGBoost. The study simulates the execution of the different algorithms configurations in order to evaluate the performance of the resulted models, and then choose which the best was; using performance evaluation methods to improve each one. The actual work is an improvement of our earlier research [11–14].

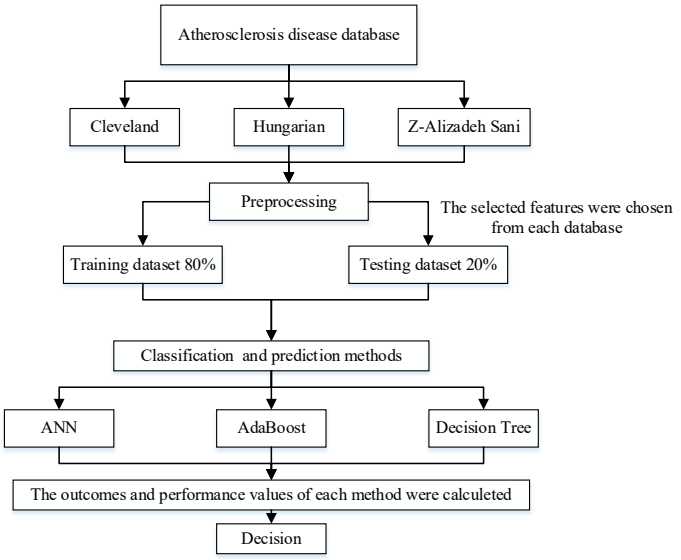

Figure 1: Flowchart of the proposed MDSS using ML algorithm.

3. Materials and methods

3.1. Global overview of the proposed MDSS

In this work, we proposed an MDSS using ML technique. This system based on three supervised ML algorithms. These classifiers have been applied to find the best prediction based on the chosen

important features by analyzing the atherosclerosis databases. Figure 1 shows the flowchart of the proposed work using ML algorithms.

3.1.1. Artificial Neural Network (ANN)

Artificial neural network (ANN) is inspired by the biological neural network to imitate human neurophysiology. At present, researchers have integrated statistical methods and numerical analysis into neural networks to give a mathematic model [25].

Where $\{x_1, x_2, \ldots, x_n\}$ represent the n inputs, ($W_{i,n}$) represents the weights and (y_i) are the outputs of the neural network using sigmoid function as a nonlinear activation function ($f(.)$) for each neuron. The activation function is given by equation (1):

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The ANN algorithm is achieved using the following equations

Network equation:

$$net_i = \sum_{j=1}^{m} w_{i,j} * x_i + b_i \tag{2}$$

Predicted outputs equation:

$$y_i = f_i(net_i) \tag{3}$$

Slope equation:

$$S_i = \frac{\partial f_i}{\partial net_i} \tag{4}$$

Error (e_i) using for the actual output (t_i) and the predicted output (o_i) equation:

$$e_i = t_i - o_i \tag{5}$$

The last step in the ANN algorithm is to check if the standard stop error is reached. This means that the actual error (e_{i+1}) is smaller than the last error and that the approximation of the total error function is valid.

3.1.2. Adaptive Boosting (AdaBoost)

Adaptive Boosting [26,27] as known AdaBoost, is an ML algorithm proved by Yoav Freund and Robert Schapire. This method can be used in combination with many ML algorithms to improve performance for binary classification. AdaBoost structure can be briefly defined as follows.

For each learner (t), AdaBosst calculated the weighted classification error as using the following equation:

$$e_t = \sum_{n=1}^{N} d_n^{(t)} II(y_n \neq h_t(x_n)) \tag{5}$$

With (y_n) is the true class label, (x_n) is predictor vector for observation (n), (h_t) is the hypothesis (learner predictor), ($d_n^{(t)}$) is the observation weight in step (t), (II) is the indicator function and the AdaBoost trains learners sequentially.

AdaBoost can increase weights for each misclassified observation and reduces weights for each observation correctly classified.

After training phase, AdaBoost computes prediction using the following equation:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{6}$$

with:

$$\alpha_t = \frac{1}{2} log \frac{1-\varepsilon_t}{\varepsilon_t} \tag{7}$$

Where (α_t) are the weak hypothesis weights in the ensemble.

AdaBoost training step can be considered as the exponential loss minimization using the following equation.

$$\sum_{n=1}^{N} w_n e^{(-y_n f(x_n))} \tag{8}$$

Where the true class is $y_n \in \{-1, 1\}$, (w_n) is the weight normalized to add up to 1 and $f(x_n) \in (-\infty, +\infty)$ is the predicted classification score.

3.1.3. Decision Tree (DT)

Decision Tree (DT) is a supervised machine learning algorithm. This method is usually used in binary classification problems. The objective is to construct a set of choices in a tree graphic form consisting of nodes and branches based on each collected attribute[28].

The decision tree algorithm is achieved using the following equations:

Probability ($P(T)$) to estimate that an observation (j) is in node (n) is defined with the following expression:

$$P(T) = \sum_{j \in X} w_j \tag{9}$$

Information gain (G(T, X)) for each tree's node to classify all input data is defined with the following expression:

Where: (w_j) is weight of the observation (j).

$$G(T, X) = E(X) - E(T|X) \tag{10}$$

The entropy (E (T)) is defined with the following expression:

$$E(T) \equiv \sum_{i=1}^{c} -p_i log_2 p_i \tag{11}$$

Where: (p_i) is the probability of the class i with i = 1, …, c with (c) is the total number of classes. In the case of binary classification c=2.

3.2. Databases description

3.2.1. Cleveland dataset

Cleveland dataset is collected by David Aha for machine learning repository [29]. It is obtained from the Cleveland Clinic Foundation database of the University of California Irvine. This database consists of 76 attributes of which only 14 attributes are commonly used in most published researches: 13 inputs and one output. In this proposed work, only 270 instances are used from the 303 records patients owing to some missing values. It is noted that this dataset performs with 54% healthy subjects and 46% CAD patients. The healthy subjects are marked 0 while the unhealthy ones are designated by the value 1. Table 1 summarizes all used Cleveland Features.

Table 1: Cleveland dataset features and their descriptions.

No	Features	Description	Scale
1	Age	Age in years	29 - 77
2	GD	Gender	Female (0), Male (1)
3	CP	Chest pain type	Typical angina (1), Atypical angina (2), Non-angina pain (3), Asymptomatic (4)
4	trestbps	Resting blood pressure on admission to the hospital (mm/Hg)	94 - 200
5	chol	Serum cholesterol (mg/dl)	126 - 564
6	Fbs	Fasting blood sugar is greater than 120 mg/dl	No (0), Yes (1)
7	Restecg	Resting electrocardiographic results	Normal (0), Having ST-T wave abnormality (1), Showing probable or definite left ventricular hypertrophy by Estes' criteria (2)
8	Thalach	Maximum heart rate achieved (ppm)	71 - 202
9	Exang	Exercise induced angina	No (0), Yes (1)
10	Oldpeak	ST depression induced by exercise relative to rest	0 - 6,2
11	slope	The slope of the peak exercise ST segment	Up sloping (0), Flat (1), Down sloping (2)
12	ca	Number of major vessels colored by fluoroscopy	0-3
13	Thal	The heart status	Normal (3), Fixed defect (6), Reversible defect (7)
14	num	Diagnosis of heart disease	Healthy (0), Patient has heart disease (1)

3.2.2. Hungarian dataset

The Hungarian dataset is collected by Andras Janosi, at the Hungarian Institute of Cardiology, Budapest [29]. This database contains 10 features. Through the 294 dataset samples, 262 samples were commonly used, 34 simples have been rejected because of missing values. The Hungarian simples are segregated in 62.21% healthy subjects and 37.78% with heart disease.

3.2.3. Z-Alizadeh Sani dataset

The Z-Alizadeh Sani dataset is randomly collected at Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre. This dataset is built for CAD diagnosis, containing 303 samples with 54 features for each patient. The selected features include the main data on the patient's physical examinations, echocardiograms (ECGs), physical examinations, laboratory tests, demographic characteristics, and symptoms [18,23].

Alizadehsani et al [23] have classified patients into two outputs classes: 71% of patients suffered from CAD and 29% healthy. This dataset also contains stenosis prediction outputs of three coronary arteries i.e., LAD, RCA, and LCX. In this study, we have manually selected 17 features as the most important features according to the atherosclerosis risk factor [2,30].

3.3. Features selection

During the preprocessing step that consist essentially of the dataset cleaning (Ignoring inputs with missing values), the prediction inputs are based on the features of each database. Atherosclerosis risk factors have been identified from the expertise of medical experts and doctors. These risk factors are known as uncontrollable risk factors and controllable risk factors. The suitable features are chosen from each dataset as input data based on the related literature [2,30].

The corresponding outputs used for prediction are the binary labels “Diagnosis of heart disease” which reflects the actual condition of the patient considered. These 2 classes are: a patient has atherosclerosis or healthy. Here, a value of 0 means that there is no atherosclerotic disease, this means that the reduction in diameter is less than 50%. A value of 1 indicates the presence of atherosclerotic disease, which means that the diameter is reduced by 50% according to the database collected by UCI data (Cleveland and Hungarian). Regarding the Z-Alizadeh Sani database, the output is divided into two category labels. Therefore, Category 0 specifies that there is no atherosclerotic disease, which means normal. Category 1 indicates the presence of atherosclerotic disease, which indicates CAD.

4. Performance evaluation metrics

In this work, we used many performance methods to improve our proposed MSSD of atherosclerosis disease. These methods represent as following:

- The Recall, the true positive rate (TRR) or the sensitivity calculates the degree of patients having correctly identified the disease.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

- Precision or Positive predictive value, this metric is the positive proportion result in diagnostic tests that is true positive results.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

- Accuracy (ACC) that computes the precision degree.

$$ACC = \frac{TP + FN}{TP + FP + FN + TN} \tag{14}$$

- Just like our case, the Matthews Correlation Coefficient (MCC) is a quality metric used for machine learning binary classification.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TN + FP) * (TN + FN) * (TP + FP) * (TP + FN)}} \tag{15}$$

- F1-score (FS) that shows the precision harmonic means.

$$FS = \frac{2 * TP}{2 * TP + FP + FN} \quad (16)$$

Where FN, TP, FP and TN are respectively false negative, true positive, false positive and true negative. In the ML field confusion matrix is also known as an error matrix. The matrix represents the performance of the algorithm, but it contains two types of information: the predicted value and the actual value. Table 2 explains the confusion matrix for the binary classification [31,32].

Table 2: Confusion matrix.

Predicted diagnostic outcome	**Actual diagnostic outcome**		
	Patient has the disease	**Patient has not the disease**	**Row total**
Negative	FN	TN	TN +FN (Negative test)
Positive	TP	FP	FP +TP (Positive test)
Column total	FN +TP (Patients number have the disease)	TN + FP (Patients number have not the disease)	TP + FP + FN + TN (Total Population)

5. Simulation results and performance comparison

To prove the effectiveness of our proposed classifiers and predictors, many experiments and simulation were performed to empirically identify the best ML models. In this way, three sets of atherosclerosis data are used, and various performance evaluation methods are used to summarize the experimental results in tables to assess the effectiveness of the proposed method. A comparison of the obtained results with previous work was also conducted.

5.1. ML design and implementation

For ANN technique and as any empirical work, many simulations were conducted to select the best hyper parameters. As will be showed later, the best performance is reached for the following architecture configuration presented in table 3.

The learning parameters and neural network architecture used in each dataset in this study relate to the hidden layer, the number of neurons, the value of the learning rate, and the type of activation function in each layer.

Table 3: The proposed ANN architecture specifications and training parameters.

Architecture				
Dataset name		Cleveland	Hungarian	Z-Alizadeh
The layers number		1	1	1
Weights and bias		Randomly initialized		
The neuron number	Input	13	10	17
	Hidden 1	8	8	6
	Output	2		
Activation functions	Input	Tangent-sigmoid (T-S)		
	Hidden 1			
	Output	Linear		
Learning rule		Backpropagation & Levenberg-Marquardt		
Learning rate		0.001		

In DT algorithm, the first step is to calculate the entropy of the output or the target using the equation (11). The next step we obtained the entropy for each branch. The last step, the dataset divided by its branches and repeat the process every branch until all data is classified.

In the second algorithm AdaBoost, we calculate the weighted classification error using equation (5) for each learner. Then we reduce weights for each observation correctly classified by learner t. after finished training, we calculate the prediction for the new obtained data using equation (6). Then we minimize the exponential loss using equation (8).

5.2. Classification and prediction performance evaluation results on testing datasets

The classification techniques described above were implemented to identify subjects with and without heart disease. Those algorithms were compared using standard evaluation metrics: accuracy (ACC), precision, recall, F1-score (FS), Matthews's correlation coefficient (MCC), confusion matrix and Receiver Operating Characteristic curve (ROC).

5.2.1. Confusion matrix results

The MDSS for atherosclerosis is made based on three ML techniques: ANN, AdaBoost and DT algorithms. To validate our model, three databases were used: Cleveland, Hungarian, and Z-Alizadeh Sani database consisting of 270, 262 and 303 patients' records respectively as shown in table 4.

Each database is split on two datasets using interleaved indices: 80% been used for training, and 20% for testing. Then we trained the three classifier algorithms were compared to select the best one. Table 4 shows the results of the confusion matrix obtained after testing 835 patients collected from the Cleveland, Hungary and Z-Alizadeh Sani databases using the ANN, AdaBoost and DT algorithms.

Table 4: Confusion matrix results.

Datasets	**Methods**	**TP**	**FP**	**FN**	**TN**
Cleveland	ANN	20	4	6	24
	DT	19	5	5	25
	AdaBoost	16	8	23	7
Hungarian	ANN	17	5	3	28
	DT	16	6	8	23
	AdaBoost	15	7	5	25
Z-Alizadeh Sani	ANN	40	1	5	12
	DT	41	9	8	3
	AdaBoost	40	4	5	12

5.2.2. Performance metrics results

During the testing phase, the outcomes are given to the proposed classification system to classify and predict patients with atherosclerosis. The achieved results are calculated using the standards performance metrics: ACC, precision, recall, FS, and MCC. To improve our atherosclerosis prediction system, two further machine learning metrics are used: FS as binary classification accuracy test and MCC as a binary classification quality measure.

The FS and MCC metrics should nearby 1 to assess on the system efficiency. Table 5 shows the obtained evaluation metrics

for ANN, AdaBoost and DT algorithms using Cleveland, Hungarian, and Z-Alizadeh Sani databases.

Table 5: the proposed system performance metrics.

Datasets	Methods	ACC	Precision	Recall	FS	MCC
Cleveland	ANN	**91.41%**	**79.67%**	**70.36%**	**0.75**	**0.60**
	DT	81.48%	79.17%	79.17%	0.80	0.46
	AdaBoost	72.22%	69.57%	66.67%	0.68	0.44
Hungarian	ANN	**90%**	**85%**	**78%**	**0.81**	**0.75**
	DT	73.58%	66.57%	72.73%	0.70	0.46
	AdaBoost	77.36%	75.00%	68.18%	0.71	0.53
Z-Alizadeh Sani	ANN	**94%**	**92.58%**	**97.73%**	**0.94**	**0.75**
	DT	81.97%	83.67%	93.18%	0.88	0.57
	AdaBoost	85.25%	88.89%	90.91%	0.90	0.63

5.3. Cross-validation

In this section, we present the analysis of system performance using the k-factor cross-validation technique. As a result, the databases are divided into k data sets. For each validation, one dataset is used as the test dataset and the rest of the datasets are used as the training dataset.

The principle of cross-validation is that we run a given model several times. In our case, ten times (K = 10), then we average the ten different tests, after that we average the test results of these K experiments. Obviously, this requires more computing time, as we have now conducted K separate learning experiments, but the evaluation of the learning algorithm will be more accurate. In other words, we use all the data for training and all the data for testing. In this case, we use the interleaved analysis method to divide each database into two parts: 80% of the training set and 20% of the test set.

When analyzing the Cleveland training dataset (graph shown in Figure. 2), The Cleveland database average accuracy computation, the ANN algorithm achieved 91.41% compared with the average accuracy of the other algorithms (AdaBoost and DT) are respectively 72.22% and 81.48% as shown in the graph.

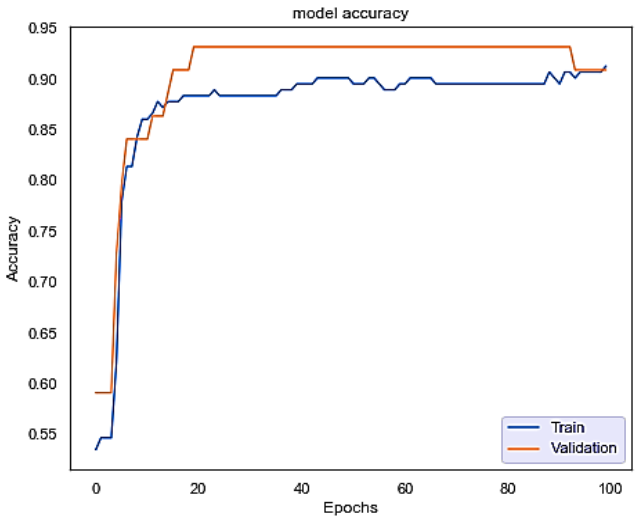

Figure 2: ANN Cross-validation analysis of Cleveland database

In Hungarian training dataset, the average accuracy achieved 90.00% with regard to the graph shown in figure 3. During this database analyze, the ANN algorithm produced a higher accuracy compared with the other algorithms.

Similarly, in Sani Z-Alizadeh training dataset, the ANN algorithm achieved a higher accuracy (94.00%) compared with the other algorithms. in addition, the ANN algorithm average accuracy computation increased nearly by 10% higher than the other algorithms (AdaBoost and DT) are respectively 85.25% and 82.00% (graph shown in Fig. 2).

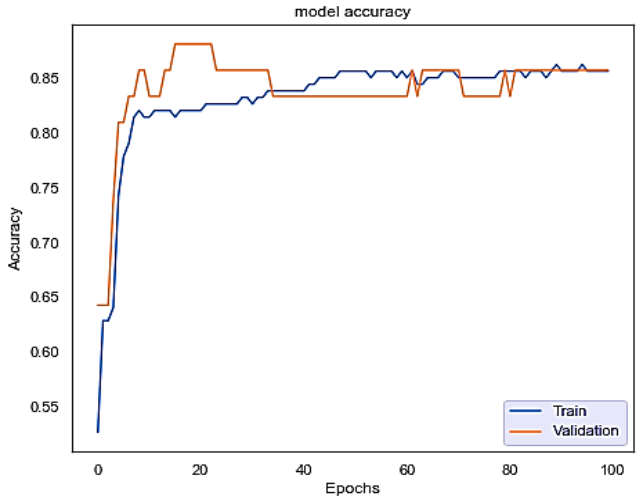

Figure 3: ANN Cross-validation analysis of Hungarian database.

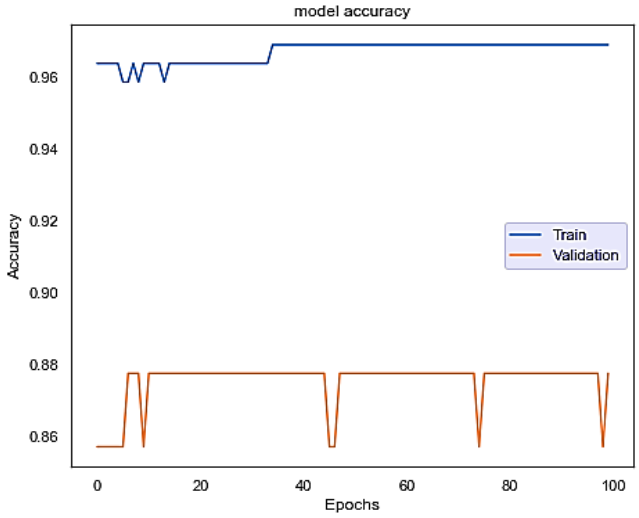

Figure 4: ANN Cross-validation analysis of Z-Alizadeh Sani database.

5.4. Receiver Operating Characteristic Curve (ROC)

In order to increase the prediction of healthy subjects and subjects with CAD, ROC assessment indicators are used to check the performance of our classifier. For each classifier, ROC will apply a threshold in the range [0, 1] to the output field.

In figure 5, the ROC analysis results for Cleveland testing dataset demonstrates that the ANN presents the better classification performance comparing to AdaBoost and DT algorithms. Where has 80% as recall value and the 70.36% as precision value. The ROC analysis results for the Hungarian

database, as shown in figure 6, prove that ANN reached the 85% as best precision with 90% as accuracy value.

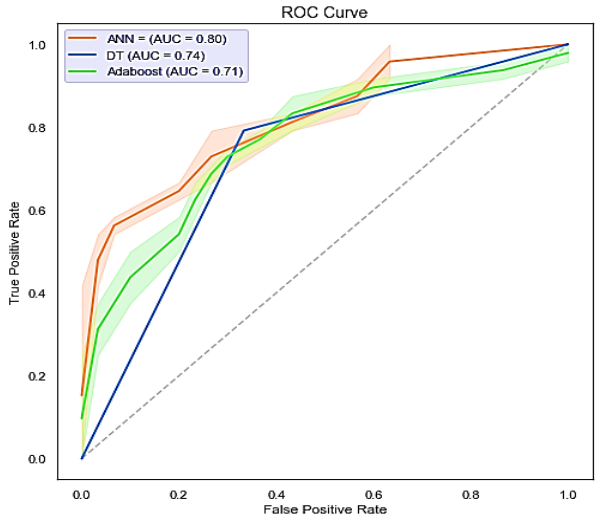

Figure 5: ROC Cleveland Database.

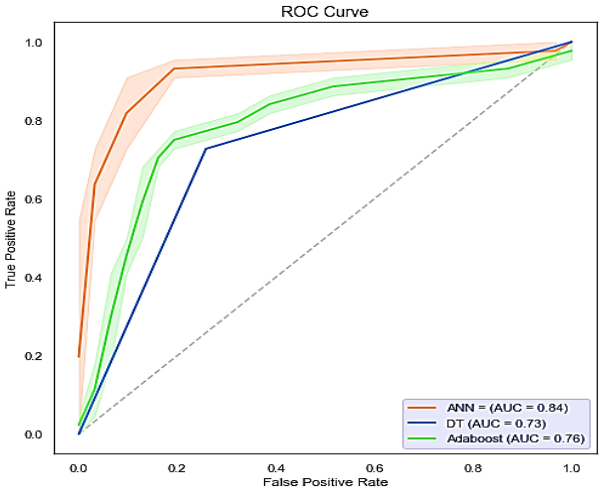

Figure 6: ROC Hungarian Database.

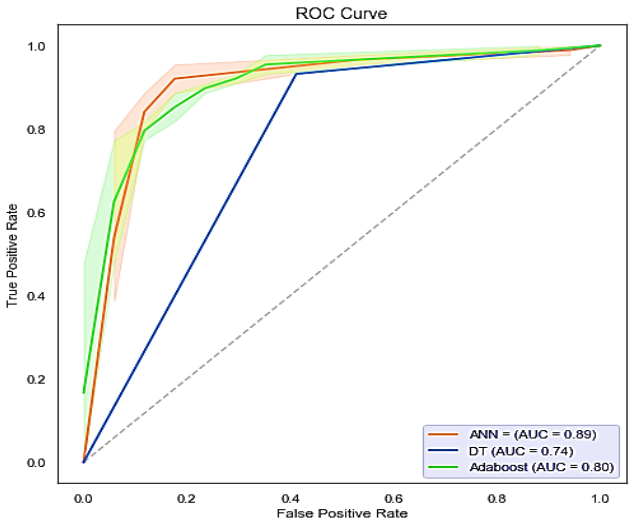

Figure 7: ROC Z-Alizadeh Sani Database.

The Z-Alizadeh Sani database as shown in figure 7, the ROC analysis showed that the ANN method reached 98% as the best recall while the AdaBoost and the DT methods reached respectively the best recall of 90.18% and 93.18%. However, the best results of ROC were obtained when using our proposed ANN method.

6. Discussion and performance comparison

To assess the effectiveness of our proposed method, we conducted experiments on the aforementioned Cleveland, Hungary and Z-Alizadeh Sani databases. We compare our results with some previous work as shown in Tables 6, 7 and 8. We can see from these tables that our proposed system has better prediction performance compared to other classifiers.

In Table 6, we present the results by comparing the accuracy of the proposed system with previous work using the Cleveland database.

Table 6: Classification Cleveland database accuracies.

Author and year	Method	Accuracy (%)
Anooj et al. (2012) [16]	Weighted fuzzy rules	62.35
El-Bialy et al. (2015) [17]	C4.5	78.54
	FDT	77.55
Arabasadi (2017) [18]	Neural Network	84.80
	HNNG	89.40
Das et al. (2009) [15]	Neural Networks Ensemble	89.01
Proposed method	**ANN**	**91.41**

In addition, we have used comparative study to show the performance of the proposed system. In the Cleveland analysis dataset, the proposed system showed that the ANN method can achieve a higher accuracy of 91.41% as shown in Table 6, while the accuracy of previous systems such as Weighted Fuzzy, C4.5, FDT, Neural Network, Set neural networks, HNNG, is 62.35%, 78.54%, 77.55%, 84.80%, 89.01% and 89.40%, respectively.

In the Hungarian database, the proposed system achieves a better precision of 90.00%, as shown in Table 7. Here, the previous systems (like the weighted fuzzy rules method) only obtained 46.93%, HNNG gained 87.10%. Table 7 lists the correctness of the classification of the Hungarian database.

Table 7: Classification Hungarian database accuracies.

Author and year	Method	Accuracy (%)
Anooj P.K.(2012)[16]	Weighted fuzzy rules	46.93
El-Bialy et al. (2015) [17]	C4.5	78.57
	FDT	78.23
Arabasadi (2017) [9]	Neural Network	82.90
	HNNG	87.10
Alizadeh (2018) [23]	SVM, Naïve Bayes, and C4.5	88.77
Proposed method	ANN	90.00

Similarly, when analyzing the Z-Alizadeh Sani dataset, our system obtained the best accuracy, as presented in Table 8. For the accuracy of HNNG achieved 93.85% compared to our system's accuracy of 94.00%. As shown in Table 8, compared to previous research, our proposed system works best for performing efficient classification and prediction.

Table 8: Classification Z-Alizadeh Sani database accuracies.

Author and year	Method	Accuracy (%)
Arabasadi et al. (2017) [18]	HNNG	93.85
	Neural Network	84.62
Abdar et al. (2019) [33]	SVC	92.45
	nuSVM	93.08
	LinSVM	92.09
Nasarian (2020)[22]	2HFS	92.58
Proposed method	**ANN**	**94.00**

7. Conclusion

In this work, we proposed an MDSS for the early prediction of atherosclerosis. Applied to datasets, the proposed system is based on three ML algorithms (ANN, AdaBoost and DT algorithms) to generate functionalities suitable for predicting patients with / without atherosclerotic disease. Using clinical data sets, a total of 835 samples were obtained from the databases in Cleveland, Hungarian and Z-Alizadeh Sani. The experimental results show that compared to other ML techniques, the ANN algorithm has better accuracy. In addition, the Accuracy, Precision, Recall and F1_Score indicators and the ROC graph are used to assess the performance of the proposed algorithm. Finally, a comparative predictive analysis is carried out between the experimental results and the different methods available in the literature (such as weighted fuzzy rules, HHNG and 2HFS). Based on common performance indicators, this comparison shows that our proposed system has the highest accuracy of 94% in predicting and classifying atherosclerosis. Like future research guidelines, the proposed system will include different methods and functions for other heart diseases to improve the accuracy of predictions.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Cardiovascular diseases (CVDs).

[2] R. Ross, "Atherosclerosis — An Inflammatory Disease," New England Journal of Medicine, **340**(2), 115–126, 1999, doi:10.1056/NEJM199901143400207.

[3] P. Libby, P.M. Ridker, G.K. Hansson, "Progress and challenges in translating the biology of atherosclerosis," Nature, **473**(7347), 317–325, 2011, doi:10.1038/nature10146.

[4] R.A. Miller, "Medical Diagnostic Decision Support Systems--Past, Present, And Future: A Threaded Bibliography and Brief Commentary," Journal of the American Medical Informatics Association, **1**(1), 8–27, 1994, doi:10.1136/jamia.1994.95236141.

[5] E. Shortliffe, Computer-Based Medical Consultations: MYCIN, Elsevier, 2012.

[6] O. Daanouni, B. Cherradi, A. Tmiri, Type 2 Diabetes Mellitus Prediction Model Based on Machine Learning Approach, Springer International Publishing, Cham: 454–469, 2020.

[7] S. Laghmati, A. Tmiri, B. Cherradi, "Machine Learning based System for Prediction of Breast Cancer Severity," in 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM), IEEE, Fez, Morocco: 1–5, 2019, doi:10.1109/WINCOM47513.2019.8942575.

[8] S. Hamida, B. Cherradi, A. Raihani, H. Ouajji, "Performance Evaluation of Machine Learning Algorithms in Handwritten Digits Recognition," in 2019 1st International Conference on Smart Systems and Data Science (ICSSD), IEEE, Rabat, Morocco: 1–6, 2019, doi:10.1109/ICSSD47982.2019.9003052.

[9] S. Hamida, B. Cherradi, H. Ouajji, "Handwritten Arabic Words Recognition System Based on HOG and Gabor Filter Descriptors," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, Meknes, Morocco: 1–4, 2020, doi:10.1109/IRASET48871.2020.9092067.

[10] S. Hamida, B. Cherradi, H. Ouajji, A. Raihani, Convolutional Neural Network Architecture for Offline Handwritten Characters Recognition, Springer International Publishing, Cham: 368–377, 2020.

[11] O. Terrada, A. Raihani, O. Bouattane, B. Cherradi, "Fuzzy cardiovascular diagnosis system using clinical data," in 2018 4th International Conference on Optimization and Applications (ICOA), IEEE: 1–4, 2018.

[12] O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, "A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors," in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), IEEE, Kenitra: 1–6, 2018, doi:10.1109/ICECOCS.2018.8610649.

[13] O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, "Classification and Prediction of atherosclerosis diseases using machine learning algorithms," in 2019 5th International Conference on Optimization and Applications (ICOA), IEEE, Kenitra, Morocco: 1–5, 2019, doi:10.1109/ICOA.2019.8727688.

[14] O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, "Atherosclerosis disease prediction using Supervised Machine Learning Techniques," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, Meknes, Morocco: 1–5, 2020, doi:10.1109/IRASET48871.2020.9092082.

[15] R. Das, I. Turkoglu, A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, **36**(4), 7675–7680, 2009, doi:10.1016/j.eswa.2008.09.013.

[16] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Journal of King Saud University - Computer and Information Sciences, **24**(1), 27–40, 2012, doi:10.1016/j.jksuci.2011.09.002.

[17] R. El-Bialy, M.A. Salamay, O.H. Karam, M.E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," Procedia Computer Science, **65**, 459–468, 2015, doi:10.1016/j.procs.2015.09.132.

[18] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A.A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, **141**, 19–26, 2017, doi:10.1016/j.cmpb.2017.01.004.

[19] O.W. Samuel, G.M. Asogbon, A.K. Sangaiah, P. Fang, G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Systems with Applications, **68**, 163–172, 2017, doi:10.1016/j.eswa.2016.10.020.

[20] S. Nazari, M. Fallah, H. Kazemipoor, A. Salehipour, "A fuzzy inference-fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases," Expert Systems with Applications, **95**, 261–271, 2018, doi:10.1016/j.eswa.2017.11.001.

[21] M. Abdar, U.R. Acharya, N. Sarrafzadegan, V. Makarenkov, "NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease," IEEE Access, **7**, 167605–167620, 2019, doi:10.1109/ACCESS.2019.2953920.

[22] E. Nasarian, M. Abdar, M.A. Fahami, R. Alizadehsani, S. Hussain, M.E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U.R. Acharya, R.-S. Tan, N. Sarrafzadegan, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach," Pattern Recognition Letters, **133**, 33–40, 2020, doi:10.1016/j.patrec.2020.02.010.

[23] R. Alizadehsani, M.J. Hosseini, A. Khosravi, F. Khozeimeh, M. Roshanzamir, N. Sarrafzadegan, S. Nahavandi, "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries," Computer Methods and Programs in Biomedicine, **162**, 119–127, 2018, doi:10.1016/j.cmpb.2018.05.009.

[24] A. Cuvitoglu, Z. Isik, "Classification of CAD dataset by using principal component analysis and machine learning approaches," in 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE), IEEE, Istanbul: 340–343, 2018, doi:10.1109/ICEEE2.2018.8391358.

[25] D. Hanbay, I. Turkoglu, Y. Demir, "An expert system based on wavelet decomposition and neural network for modeling Chua's circuit," Expert Systems with Applications, **34**(4), 2278–2283, 2008, doi:10.1016/j.eswa.2007.03.002.

[26] Y. Freund, M. Kearns, D. Ron, R. Rubinfeld, R.E. Schapire, L. Sellie, "Efficient Learning of Typical Finite Automata from Random Walks," Information and Computation, **138**(1), 23–48, 1997, doi:10.1006/inco.1997.2648.

[27] Y. Freund, R.E. Schapire, "Adaptive Game Playing Using Multiplicative Weights," Games and Economic Behavior, **29**(1–2), 79–103, 1999,

doi:10.1006/game.1999.0738.

[28] S.R. Safavian, D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Transactions on Systems, Man, and Cybernetics, **21**(3), 660–674, 1991, doi:10.1109/21.97458.

[29] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K.H. Guppy, S. Lee, V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, **64**(5), 304–310, 1989, doi:10.1016/0002-9149(89)90524-9.

[30] M. Rafieian-Kopaei, M. Setorki, M. Doudi, A. Baradaran, H. Nasri, "Atherosclerosis: process, indicators, risk factors and new hopes," International Journal of Preventive Medicine, **5**(8), 927–946, 2014.

[31] A.S. Glas, J.G. Lijmer, M.H. Prins, G.J. Bonsel, P.M.M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," Journal of Clinical Epidemiology, **56**(11), 1129–1135, 2003, doi:10.1016/S0895-4356(03)00177-X.

[32] A.G. Lalkhen, A. McCluskey, "Clinical tests: sensitivity and specificity," Continuing Education in Anaesthesia Critical Care & Pain, **8**(6), 221–223, 2008, doi:10.1093/bjaceaccp/mkn041.

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