Mathematical Biosciences and Engineering

MBE, 19(5): 4892–4910.
DOI: 10.3934/mbe.2022229
Received: 14 December 2021
Revised: 07 February 2022
Accepted: 22 February 2022
Published: 14 March 2022

http://www.aimspress.com/journal/MBE

Research article

Prediction of atherosclerosis using machine learning based on operations research

Zihan Chen[1], Minhui Yang[2], Yuhang Wen[3], Songyan Jiang[3], Wenjun Liu[4,*] and Hui Huang[5]

[1] Changwang School of Honors, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] School of Teacher Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
[4] School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5] Department of Ultrasound, Affiliated Hospital of Nanjing University of CM, Nanjing 210029, China

* **Correspondence:** Email: wjliu@nuist.edu.cn; Tel: +862558731160.

Abstract: *Background:* Atherosclerosis is one of the major reasons for cardiovascular disease including coronary heart disease, cerebral infarction and peripheral vascular disease. Atherosclerosis has no obvious symptoms in its early stages, so the key to the treatment of atherosclerosis is early intervention of risk factors. Machine learning methods have been used to predict atherosclerosis, but the presence of strong causal relationships between features can lead to extremely high levels of information redundancy, which can affect the effectiveness of prediction systems. *Objective:* We aim to combine statistical analysis and machine learning methods to reduce information redundancy and further improve the accuracy of disease diagnosis. *Methods:* We cleaned and collated the relevant data obtained from the retrospective study at Affiliated Hospital of Nanjing University of Chinese Medicine through data analysis. First, some features that with too many missing values are filtered out of the 34 features, leaving 25 features. 49% of the samples were categorized as the atherosclerosis risk group while the rest 51% as the control group without atherosclerosis risk under the guidance of relevant experts. We compared the prediction results of a single indicator that had been medically proven to be highly correlated with atherosclerosis with the prediction results of multiple features to fully demonstrate the effect of feature information redundancy on the prediction results. Then the features

that could distinguish whether have atherosclerosis risk or not were retained by statistical tests, leaving 20 features. To reduce the information redundancy between features, after drawing inspiration from graph theory, machine learning combined with optimal correlation distances was then used to screen out 15 significant features, and the prediction models were evaluated under the 15 features. Finally, the information of the 5 screened-out non-significant features was fully utilized by ensemble learning to improve the prediction superiority for atherosclerosis. *Results:* Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), which is used to measure the predictive performance of the model, was 0.84035 and Kolmogorov-Smirnov (KS) value was 0.646. After feature selection model based on optimal correlation distance, the AUC value was 0.88268 and the KS value was 0.688, both of which were improved by about 0.04. Finally, after ensemble learning, the AUC value of the model was further improved by 0.01369 to 0.89637. *Conclusions:* The optimal distance feature screening model proposed in this paper improves the performance of atherosclerosis prediction models in terms of both prediction accuracy and AUC metrics. Code and models are available at https://github.com/Cesartwothousands/Prediction-of-Atherosclerosis.

Keywords: atherosclerosis; machine learning; random forest classifier; ensemble learning, operation research; information redundancy

1. Introduction

According to the World Health Organization, cardiovascular disease has become the most important cause of death worldwide. Atherosclerosis is a kind of slow narrowing of the arteries that influences the blood flow from the heart to the brain, also the cause of most cardiovascular diseases [1–3]. The subclinical latency period for atherosclerosis is 30 to 50 years, with a long asymptomatic period [4]. There are many factors that influence the onset of atherosclerosis, the influence factors are highly relevant and interact with each other. Therefore, doctors often diagnose patients based on a few typical features. And it's difficult to diagnose it accurately at an early stage when the typical features do not change significantly. The traditional statistical model is not effective enough when applied to this situation. In order to identify atherosclerosis earlier, machine learning methods are widely used to build models to predict and prevent the disease. Currently, algorithms like support vector machines, decision trees naive Bayes and several others enhanced the relevant diseases prediction and enabled the continuation of human life worldwide [5–8]. For instance, Couturier et al. decided to use a three-step method based on cluster-supervised classification and frequent itemset search to predict whether the patient was likely to develop atherosclerosis based on the relevance of his lifestyle habits and social environment [9]. Artificial neural networks (ANN) and random forests (RF) are also very widely used in atherosclerosis research. They are applicable to most of the datasets, having superior performance, and are less difficult to use and reduce the computational burden [10].

There are many academics who have contributed to the prediction of cardiovascular disease using different machine learning methods. We review the relevant literature and collect the following research methods and results which are shown in Table 1.

Table 1. Results and methods of previous studies.

Author	Method	Result
V. Sree Hari Rao [11]	In-built imputation algorithm and particle swarm optimization (PSO)	The best accuracy was 99.73%
Jiang Xie [12]	Weight learning approach	The prediction accuracy improved from 11.53 to 16.76% after weighted learning
Wenming He [13]	Kernel extreme learning machine (KELM) optimized by an improved salp swarm algorithm (SSA)	The classification accuracy obtained by STSSA-KELM was 84.40%
Andrew Ward [14]	Trained ML models	Bettr AUC was 0.835; AUC after incorporating additional EHR data was 0.790
Oumaima Terrada [7]	K-medoids and k-means clustering for classification, Artificial Neural Network (ANN) and K-Nearest Neighbor (KNN)	The best accuracy was 96%, the best Matthews's correlation coefficient was 0.92
Soodeh Nikan [15]	Ridge expectation maximization imputation (REMI) technique, conditional likelihood maximization method	The best accuracy was 88.04%
Jiang Xie [16]	Subset Learning (S-learning)	The best AUC was 0.83
Mohan Priya [17]	Fast correlation-based filter	About 99.47%
Antonis I. Sakellarios [18]	Gradient Boosted Trees (GBT) algorithm	The best accuracy was 68%, the best AUC was 0.59
Brajesh Kumar [19]	Support vector machine	The AUC with Hungarian dataset was 79.6%, The AUC with Cleveland dataset was 79.0%, The AUC with Z-Alizadeh Sani dataset was 91.2%, The AUC with Statlog dataset was 79.6%.

Over the past two decades, many doctors and researchers have tried to combine machine learning methods and imaging biomarkers to predict atherosclerosis. Lin [20] combined discriminative feature

selection and a semi-supervised graph-based regression to detect changes of plaque. However, atherosclerosis is affected by many factors, the analysis of imaging biomarkers often leads to the ignore in early period. Terrada used different machine learning model like ANN and KNN [7] to achieve a high accuracy performance (The best accuracy was 96%). While, they didn't do additional feature selection for features which include a variety of demographic indexes, physical and chemical indexes, imaging and echo biomarkers. They are medically important, but highly relevant and interact with each other. Rao proposed *N2Genetic optimizer* [11] to improve the performance and their prediction accuracy can achieve 99.73%. Hathaway [21] tried to combined deep learning method and routine atherosclerosis prediction by using simple office-based clinical features. However, they focus on using computing optimization in different features while didn't consider the high relevant of features they use. These features can enhance the information redundancy of the model, and it is necessary to reduce the redundancy through feature selection. According to Jamthikar [22], supervised ML-based algorithms. were made up of five components: (i) data partitioning, (ii) feature engineering, (iii) training model, (iv) prediction model and (v) performance evaluation. Skandha [23] made a major breakthrough on the accuracy of the prediction through applying deep learning to training and prediction model. They have inspired us to focus our efforts on feature engineering and to further integrate it with operations research.

This paper presents an operations research-based machine learning approach for atherosclerosis prediction. The focus is on combining statistical analysis and machine learning methods to reduce information redundancy and further improve the accuracy of disease diagnosis. First, we remove the features with more missing values and fill in the features with fewer missing values among the 34 features, leaving 25 features. Then t-test and chi-square test are used for continuous and discrete features respectively, and the features that fail the statistical test are removed, leaving 20 features Next, machine learning combined with optimal correlation distance is used to filter out 15 significant features, and the prediction model is evaluated under these 15 features. Finally, the information of the screened 5 non-significant features is fully utilized by ensemble learning to improve the predictive advantage of atherosclerosis. The experiments show that the prediction performance is significantly improved by reducing the redundancy of information.

The rest of this paper is presented below. Section 2 briefly describes related work, including data sources and processing, selecting prediction performance metrics and prediction models, and using random forests to filter features. Section 3 focuses on the feature selection inspired by operations research, and proposes an optimal distance model based on Dijkstra. Experimental results are presented in Section 4. The paper concludes with some conclusions in Section 5.

2. Methods

2.1. Data sources and preprocessing

The data in this paper are obtained from a summation study conducted from January 2016 to December 2017 at Affiliated Hospital of Nanjing University of Chinese Medicine, while the study here was approved by the hospital's ethics committee after written informed consent was obtained from all patients. After screening, we select 34 characteristics. In order to better study the pathology of atherosclerosis, we divide total samples under the guidance of relevant experts, with 49% categorized as atherosclerosis risk group and 51% as a control group without atherosclerosis risk at the end.

In consultation with a medical professional and based on relevant tests, we refine the group classification to assess patients at risk for hypertension, cardiovascular, chronic kidney and hyperlipidemia (HL), which constitute the risk group for this study. Patients with hyperlipidemia (HL) are defined as having a low-density lipoprotein (LDL) level (≥ 3.36 mmol/L), and/or a total cholesterol (TC) level (≥ 5.17 mmol/L), and/or a triglyceride (TG) level (≥ 1.69 mmol/L). Healthy controls are mainly selected from the same period of physical examination, but people with abnormal hemoglobin or history of previous cardiovascular events are excluded, and People with cancer, diabetes or autoimmune diseases are the same as the former.

Table 2. Twenty-five features affecting atherosclerosis and statistical values of them.

Features	Average value	Standard deviation	p-value
Sex			0.056
Age	55.7540	14.4806	< 0.001
BMI	23.8016	3.4114	< 0.001
Triglycerides	1.5360	1.4969	< 0.001
Total cholesterol	4.7253	1.9817	0.004
Glucose	5.3104	1.3383	< 0.001
Uric acid	530.9794	195.3352	0.754
Haemoglobin	130.9871	22.9056	< 0.001
white blood cell count	6.2310	3.0117	0.022
Red blood cell count	4.5991	5.2696	< 0.001
Platelet count	195.4817	59.5907	0.477
Glutathione aminotransferase	24.6704	19.1175	0.440
Glutathione transaminase	22.7894	12.2302	0.896
HDL	1.3621	0.3759	< 0.001
LDL	2.5349	0.7251	< 0.001
Systolic blood pressure	134.3746	21.3627	< 0.001
Diastolic blood pressure	78.2170	12.3026	0.003
LCCA-IMT	0.0627	0.0563	< 0.001
RCCA-IMT	0.0584	0.0365	< 0.001
LCCA-RI	0.7167	0.0693	0.023
RCCA-RI	0.7272	0.0893	0.017
LCCA-BS	6.5665	2.7185	< 0.001
LCCA-ES	8.9234	2.4416	< 0.001
RCCA-BS	6.0710	1.4666	< 0.001
RCCA-ES	8.4066	2.2626	< 0.001

Through data analysis, there are 8 discrete features and 26 continuous features in the dataset. Since most of the features have missing values, the missing situation of each feature needs to be analyzed specifically before doing missing value processing. After preliminary statistics, we find that 2 features do not have missing values, and the remaining 30 features have different degrees of missing. In general, the treatment of missing values is to remove the features with more than 30% missing proportion, and filling the missing values of all features will bring some problems such as presence of

biased information on certain extreme pathologies, large filling errors. Therefore, features with more missing values are selectively eliminated and 25 features are retained as shown in Table 2. Considering that the missing proportion is not high in the remaining characteristics and the difficulty of each filling method, this paper uses statistical methods to fill the data with fewer missing values. Specifically, continuous variables are filled with the median if they are skewed, otherwise they are filled with the mean; for discrete variables, the plural is chosen to fill the missing values. At last, we got 622 samples with all the features. Under the guidance of relevant experts, 304 samples are seen as atherosclerosis risk group and 318 samples are seen as a control group without atherosclerosis risk. In addition, the data is randomly divided into a training set of 0.7 and a test set of 0.3 in preparation for the cross-validation.

2.2. Statistical analysis

In this paper, when the data is statistically processed, continuous type characteristics are expressed as 'mean $\pm$ standard deviation', and for the presence of atherosclerosis risk, we take independent sample *t*-test for the difference comparison between these two groups; while discrete type characteristics were expressed as counts and percentages, the difference comparison between groups was done using the chi-square test with $p < 0.05$ used as the inspection standard. In the full sample after analysis, the basic situation is shown in Table 2.

As seen in Table 2, using the idea of label encoding, we quantified the gender index as 1 for male and 0 for female. Also, age, BMI (body mass index), triglycerides, total cholesterol, glucose, hemoglobin, white blood cell count, red blood cell count, HDL (High-density lipoprotein), LDL (Low-density lipoprotein), systolic blood pressure, diastolic blood pressure, LCCA-IMT (Left common carotid artery intima-media thickness), RCCA-IMT (Right common carotid artery intima-media thickness), LCCA-RI (Resistance indices of blood flow in the left common carotid artery), RCCA-RI (Resistance indices of blood flow in the right common carotid artery), LCCA-BS (Pulse wave conduction velocity at the beginning of systole in the left common carotid artery), LCCA-ES (Pulse wave conduction velocity at the end of systole in the left common carotid artery), RCCA-BS (Pulse wave conduction velocity at the beginning of systole in the right common carotid artery) and RCCA-ES (Pulse wave conduction velocity at the end of systole in the right common carotid artery) passed the statistical test. We finally selected 20 features while the remaining features were screened down because they did not meet the statistical requirements.

2.3. Predictive performance metrics

We assess our model through a range of relative metrics. The ROC curve is the working characteristic curve of the model being measured, and is a visualization of the relationship between the relevant indicators of the continuous variables of sensitivity and specificity of the model in the form of an image. AUC means the area under the ROC curve. It is a metric often used to evaluate the merits of a dichotomous model, with higher values indicating better prediction.

However, the AUC value only evaluates the overall training effect of the model, and does not reflect how to divide the categories to make the best prediction. In our experiments, we use the KS (Kolmogorov-Smirnov) statistic [24] to evaluate the classification effectiveness of the model. For dichotomous classification problems, the KS value, like the AUC value, uses two metrics, TPR and FPR, to measure the overall predictive effectiveness of the model. And KS uses the maximum of the

difference between TPR and FPR to indicate the optimal classification threshold as shown in Figure 1. Moreover, we use sensitivity-specificity curve and precision-recall curve to Calibrate model.

$$\mathrm{KS}_{\max} = \max\left(TPR - FPR\right) \tag{1}$$

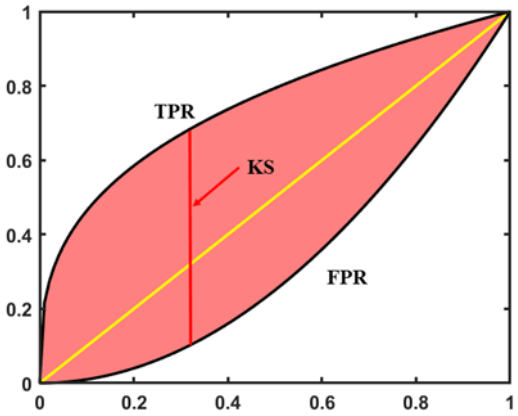


Figure 1. Schematic diagram of the KS curve.

2.4. Predictive model based on RFC

Random Forest (RF) [25] is an optimized version of the bagging algorithm and it was used in many different areas including banking, stock markets, pharmaceuticals and e-commerce. In healthcare, ingredients combined correctly in a drug can be identified by this method, and diseases can be accurately identified based on the patient's medical history. As a result, we build predictive models about the risk of atherosclerosis based on RF for intelligent classification.

2.5. Random forest classifier

We start by splitting 70% of the data into a training set and 30% into a test set, and ensure that all samples are equally likely to be selected for the training set. If the predictions obtained after such multiple equal-likelihood sampling are stable, it means that the current model is feasible. Once the dataset has been partitioned, the random forest is used to train the model.

Single decision trees can be difficult to achieve high accuracy, mainly because solving an optimal (minimum generalization error) decision tree is an NP-hard (unable to exhaust all possible tree structures) problem, and often results in a locally optimal solution. Models constructed from a single tree are often not stable enough, and changes in the sample can easily cause changes in the tree structure.

We use the idea of bagging algorithm to divide the training set into several training subsets randomly, and then build decision trees on each subset separately. In the process of building each decision tree, the idea of feature subspace is introduced into it, that is, for each node of the decision tree in choosing to divide the features to achieve the best, its candidate feature set is not all the features at the corresponding node, but a subset composed of some randomly selected features from it. The final prediction result is derived from the voting result of each decision tree.

We select $\boldsymbol{T}$ sample sets containing $\boldsymbol{m}$ training samples by self-sampling, and then train a base

learner based on each sample set before combining them. When combining and regressing the predicted outputs together, the *Bagging-like* approach usually uses simple voting for the classification task, i.e., a simple average weighted summation method is performed. However, if the number of votes is consistent, it is simplest to select a class at random, or of course to further examine the confidence of the learner votes to determine the final classification. Their exact process is as follows, for a given dataset containing n training samples $D=\{x_1, x_2, \cdots, x_n\}$,

1) A new training set containing n samples is formed using Bootstrap with put-back sampling;
2) Repeat 1) for T times to obtain T training sets;
3) T base classifiers trained independently on each training set using classification algorithm;
4) For each test sample, T predictions are obtained using the above T classifiers;
5) For each test sample, a majority vote is used to obtain the final prediction.

The exact process is shown in Figure 2.

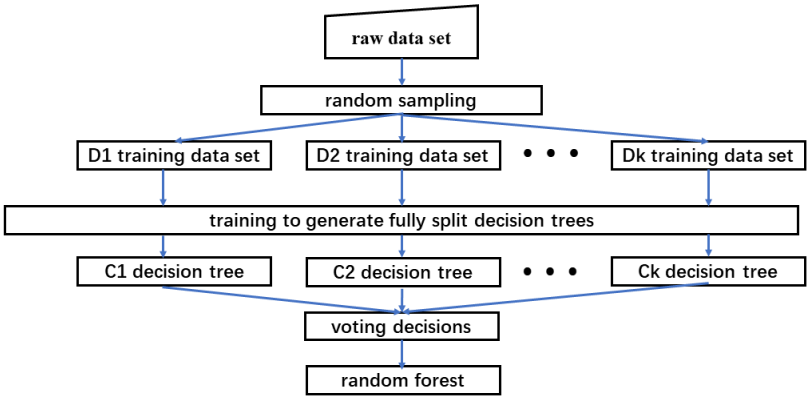


Figure 2. Flow chart of the random forest model.

2.6. Measurement of feature importance

The Gini index demonstrates the chance of misclassifying a randomly selected sample in the sample set, and a smaller Gini index indicates a lower probability of being misclassified, i.e., a higher sample purity (when all samples in the set are of one type, the Gini index is 0). The Gini index is calculated as follows.

$$Gini(p)=1-\sum_{k=1}^{K} p_k^2=\sum_{k=1}^{K} p_k(1-p_k) \tag{2}$$

where p_k denotes the chance that the selected sample belongs to the k*th* category. If the certain node has the smallest Gini index, it is also the least likely to make a mistake. Therefore, we use this node as the root node of the decision tree. In the consequent forest, the Gini index represents the fineness of the model, negatively correlated with purity and characteristics.

Let the nodes in the *j*th decision tree in which a feature appears be the set M. Then the importance of the feature under the *j*th decision tree is:

$$A_{ij}^{(Gini)}=\sum\nolimits_{m\in M}(Gini(m)-Gini(i)-Gini(r)) \tag{3}$$

where *Gini*(*i*) and *Gini*(*r*) refer to the Gini index corresponding to the two new nodes of node m after

branching, respectively. Finally, the importance of each feature under the T decision tree is calculated as follows.

$$A_i = \frac{\sum_{i=1}^{T} A_{ij}^{(Gini)}}{\sum_{j=1}^{c}\sum_{i=1}^{T} A_{ij}^{(Gini)}} \tag{4}$$

where A_i is the magnitude of the importance of the object corresponding to the *i*th feature, after normalization. The importance of selected features is shown in Table 3.

Table 3. Fifteen selected features and importance ranking obtained from random forest.

	%IncMSE	IncNodePurity
Age	0.006542	3.688916
BMI	0.00203	3.493642
Triglycerides	0.000509	2.862369
Total cholesterol	0.018734	6.030064
white blood cell count	0.003258	3.490238
HDL	0.045132	15.41302
LDL	0.084491	20.03696
Systolic blood pressure	0.00936	4.092086
Diastolic blood pressure	0.006558	3.329568
LCCA-IMT	0.014924	6.732724
RCCA-IMT	0.041935	13.04469
LCCA-BS	0.010674	4.803951
LCCA-ES	0.00215	3.041142
RCCA-BS	0.017512	7.513578
RCCA-ES	0.007113	5.189474

3. Feature selection (FS) inspired from operations research

3.1. Correlation distances and redundancy between features

To improve the identification accuracy of this atherosclerosis risk predictive model, feature quantities need to be removed that are not relevant to the classification target. We screened out the relevant features by statistical tests followed by feature redundancy needs to be considered. Feature redundancy refers to the correlation between features. Redundant feature quantities that have a high correlation with other feature quantities will also have a significant impact on this model, and if two features are perfectly correlated, they are mutually redundant features. We will model the optimal correlation distance based on the correlation distance between the features by transforming it into an NP problem in operations research.

We can calculate and select features for the model by using a measure of relevance. Similarity measurement can be achieved by calculating the distance between individual features.

Estimating similarity measures between samples is often done in many research questions, also known as correlation coefficients. This is often accomplished by calculating the distance between

samples, and the method used to calculate the distance depends on the correctness of the feature classification and feature selection. We define the correlation coefficient as:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E\left(\left(X - E(X)\right)\left(X - E(Y)\right)\right)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{5}$$

The correlation distances are defined as follows:

$$D_{xy} = 1 - \rho_{XY} \tag{6}$$

The Pearson Correlation Coefficient between features is shown in Figure 3.

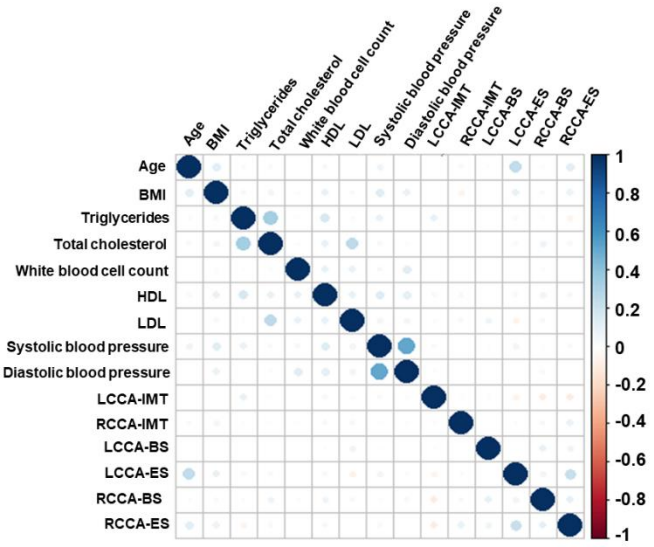


Figure 3. Pearson Correlation Coefficient between 15 features after FS.

3.2. Optimal distance model based on Dijkstra

In the fields of data structures, topological sorting and algorithms, transforming an operations and optimization problem into a graph theoretic problem can help us a lot. Graphs are the basis of graph theory problems, and we study each feature as 20 nodes in the graph set up for this problem, where we can describe the labelled values on each edge in terms of weights, i.e., each edge has a unique corresponding value, and study the graph as a weighted graph. Where the unique weight of each path is the distance associated with Eq (6), we can use a ternary group $G = (V, E, W)$ to represent the entitled undirected graphs to be solved in this paper, where W is the correlation function between the nodes, i.e., the matrix $W_{20\times 20}$ consisting of the correlation distances D_{xy} between the nodes.

We can extract combinations of different nodes through feature filtering to obtain better results for atherosclerosis prediction models. We combine the ideas of computer sampling search and the shortest distance in graph theory problems to build an optimal distance model based on the solution of the Dijkstra algorithm. When we need to select $\boldsymbol{k}$ features, we can assume that the selection of features is done from a certain node. We obtain the optimal set of features from this node by selecting the k-1 features that have less redundancy with each other after composing the feature set. Then by traversing

the computer from each node and comparing them, we can obtain the best set of filtered features when k features need to be selected.

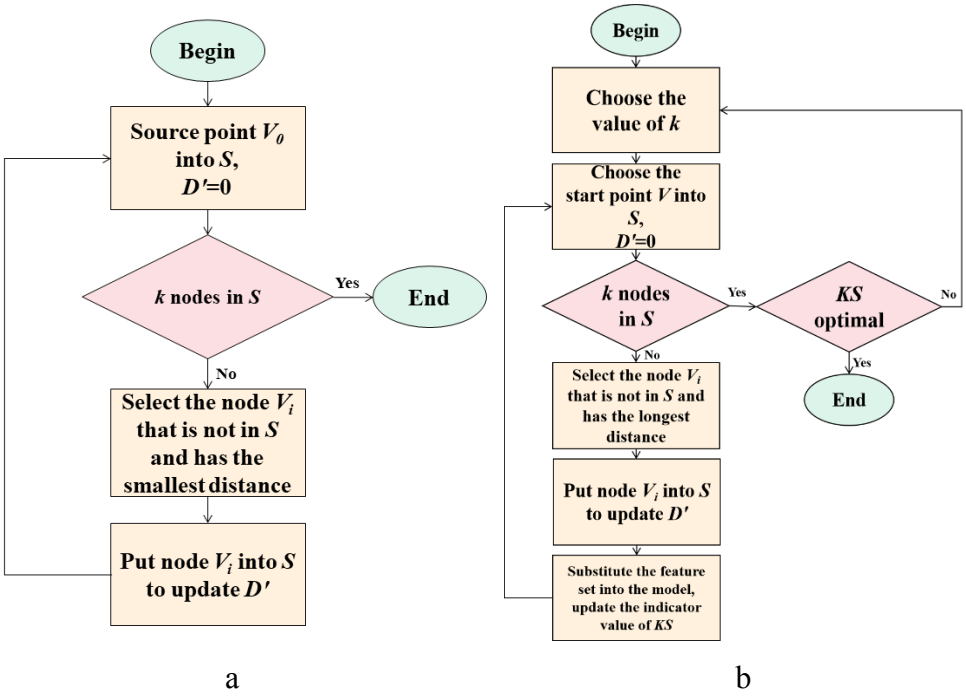


Figure 4. a. Flow chart of the Dijkstra algorithm. b. Flow chart of feature selection based on optimal distance model.

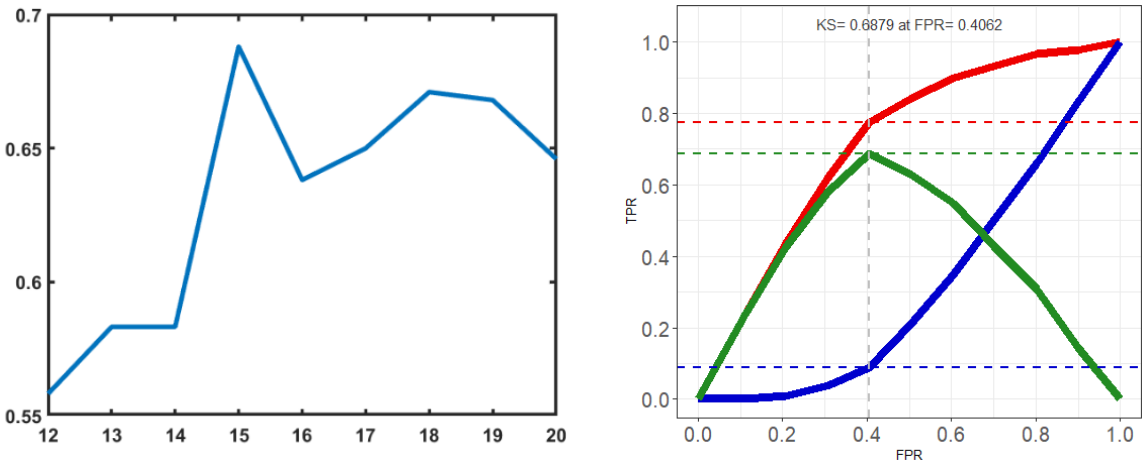
We will choose the Dijkstra algorithm [26] for optimal route solving. We introduce an auxiliary array D while each element $D[i]$ ($1 < i < 20$) is used to represent the currently recorded distance from the starting node to the other nodes. We assume the initial state of D: if we go from the starting ith node to the jth node, then $D = D_{ij}$, i.e., representing the size of the weights on the edges by the path from the ith node to the jth node. The role of Dijkstra algorithm is solving the shortest path in graph theory. In order to pursue the optimal path, we reduce the impact of redundant features on the model by finding the node with the greatest distance from the relevance of this node. Indeed, it is a process of finding the longest path. D [1] is the length of the path from the origin to the node with the greatest distance. We denote all the reached node as the set S, then the shortest path to the next farthest and non-S node t, that is, D [2], and so on, to obtain the objective function max: $\max \sum_{i=1}^{k} D[i]$. The specific process, as shown in flow chart Figure 4a.

Using Dijkstra algorithm, we can obtain the longest path contained the optimal set of features of the ith node when k features are selected. At this point, the computer traversal is then used to obtain the different sets of optimal features when starting from the 1st to the 20th node. Because we want to strengthen the accuracy of the feature screening results, we then perform the selection from 1 to 20 features by traversal and use them in a random forest-based atherosclerosis prediction model, evaluated by KS statistics, and finally obtain the objective function $\max \sum_{k=1}^{20} KS_k$. The final optimal set of features and its KS metric results are obtained. The specific model is as follows.

$$\max \sum_{k=1}^{20} KS_k$$
$$s.t. \begin{cases} G = (V, E, W) \\ D'_k = \max \sum_{i=1}^{k} D[i] \end{cases} \tag{7}$$

At this point, we obtain the set of features that can make the random forest-based atherosclerosis prediction model optimal, and complete the feature screening, the specific steps of which are shown in Figure 4b.

We obtain the optimal feature set by comparing the KS statistics of each feature set based on the optimal distance model solved by Dykstra's algorithm to complete the feature selection. We finally select the optimal special feature set under 15 features, and the variation of KS value for each feature set can be obtained in Figure 5a. Its KS value of 0.688 is better than the rest of the feature classes, KS curve is shown in Figure 5b and the optimal set of features is shown in Table 3.



a b

Figure 5. a. variation in KS of the optimal set for different number of features. b. KS curves under the optimum set.

3.3. Ensemble learning (EL)

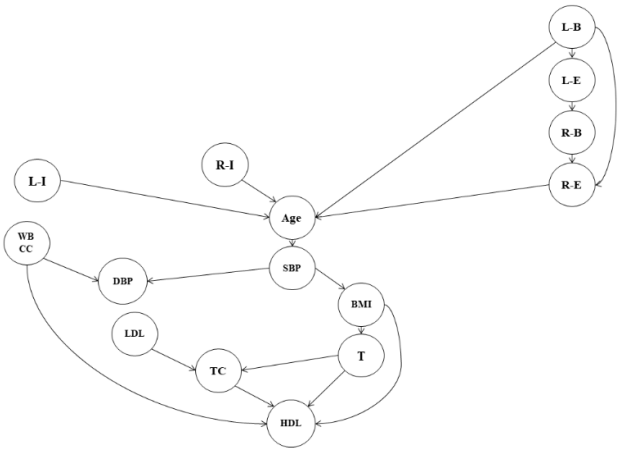
The prediction model above is constructed based on the 15 central feature sets screened by the feature selection model. Although it reduces the complexity of the model, a certain degree of information on variables that are not screened out is lost, which in turn reduces the accuracy of the prediction model.

With the purpose of further enhancing the model accuracy, this paper uses ensemble learning to make the most of the data information embodied in the four non-central variables that were discarded. The basic idea is to model the predictions of the 15 selected central features and the 5 discarded non-central features separately, and then weight the predictions of the two sets of variables to obtain the final prediction results modified by ensemble learning. The result will contain the information of the 5 variables that are initially discarded.

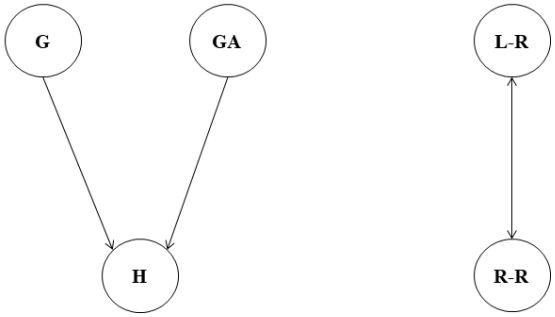
3.4. Bayesian network

Bayesian network [27] is directed acyclic graphs consisting of nodes representing features and directed edges representing the interrelationships between the nodes. The direction of the directed edge points from the parent node to the child node. Dependencies between features are expressed through conditional probabilities, and features without a parent node are expressed informally through their prior probabilities. As Bayesian networks express the causal relationships between characteristic variables in a visual framework structure, they can make the logic of uncertainty between variables clearer and better interpreted.

In our work, Bayesian network is constructed for each of the 15 selected central variable sets and the remaining 5 non-central variable sets. The sparsity of the Bayesian network measures the redundancy of feature information in each of the two feature sets divided.



a



b

Figure 6. a. Bayesian network of 15 selected central features. b. Bayesian network of 5 non-central features.

As can be seen from Figure 6, after the filtering of the feature selection algorithm based on the optimal correlation distance, the Bayesian network structure of the divided two feature subsets is simple and the information redundancy between the features has been effectively reduced.

3.5. Ensemble learning based on the search method

The aim of ensemble learning is to integrate the information described by the set of 15 central features with the set of 5 non-central variables through a certain weighting process, making full use of the information of the 5 non-central features that are screened out, so that the prediction model has a higher accuracy. The prediction model under ensemble learning can be expressed as:

$$r_{learning} = w_1 r_1 + w_2 r_2 \tag{8}$$

where $r_{learning}$ is the prediction result after ensemble learning, r_1 and r_2 are the prediction results for the central and non-central sets of variables, respectively, also w_1 and w_2 are the weights for the central and non-central sets of variables, respectively, which satisfy the following constraint:

$$\begin{cases} w_1 > w_2 \\ w_1 + w_2 = 1 \end{cases} \tag{9}$$

The basic idea of the search method is to search for the set of weights that has the largest value of the model AUC while satisfying the constraint in Eq (9). The optimal combination of weights searched in steps of 0.1 is $w_1 = 0.7$, $w_2 = 0.3$.

4. Results and discussion

The common methods, t-test and chi-square test, used in the paper yielded 20 of the 25 characteristics with statistically significant effects. Some of these 20 features are significantly correlated, so this paper uses the feature selection method to obtain 15 features.

We use AUC values to quantify the optimization of model prediction performance between common methods (abbreviated as CM), FS and EL. After several experiments with randomly assigned test sets and training sets, the model AUC is stable with no overfitting occurred confirming that the random forest-based prediction model works well. Then sensitivity-specificity curve and precision-recall curve are plotted to calibrate model. After FS, AUC is turned out to be 0.8826, resulting in an improvement of 0.052 compared to AUC without FS in the predictive performance of the model. After ensemble learning, the AUC is further increased to 0.896, an improvement of 0.091. The improvement in AUC for both steps is shown in the

Figure **7**a and the sensitivity-specificity curve and precision-recall curve are plotted in

Figure **7**b,c. This shows that the work in this paper is really effective.

In order to show that the method of feature selection in this paper is superior to other dimensionality reduction methods, we select the dimensionality reduction methods of principal component analysis (PCA) and factor analysis (FA), and use the AUC values obtained from the random forest classifier to compare the effects. PCA can regroup the original variables into a new set of uncorrelated composite variables, and a few selected composite variables can better reflect the

information of the original variables. FA is a statistical technique to study the extraction of common factors from a population of variables. Both PCA and FA are more classical methods for dimensionality reduction, but the effect of PCA and FA is much lower than that of feature selection in this paper, as shown in Figure 7d.

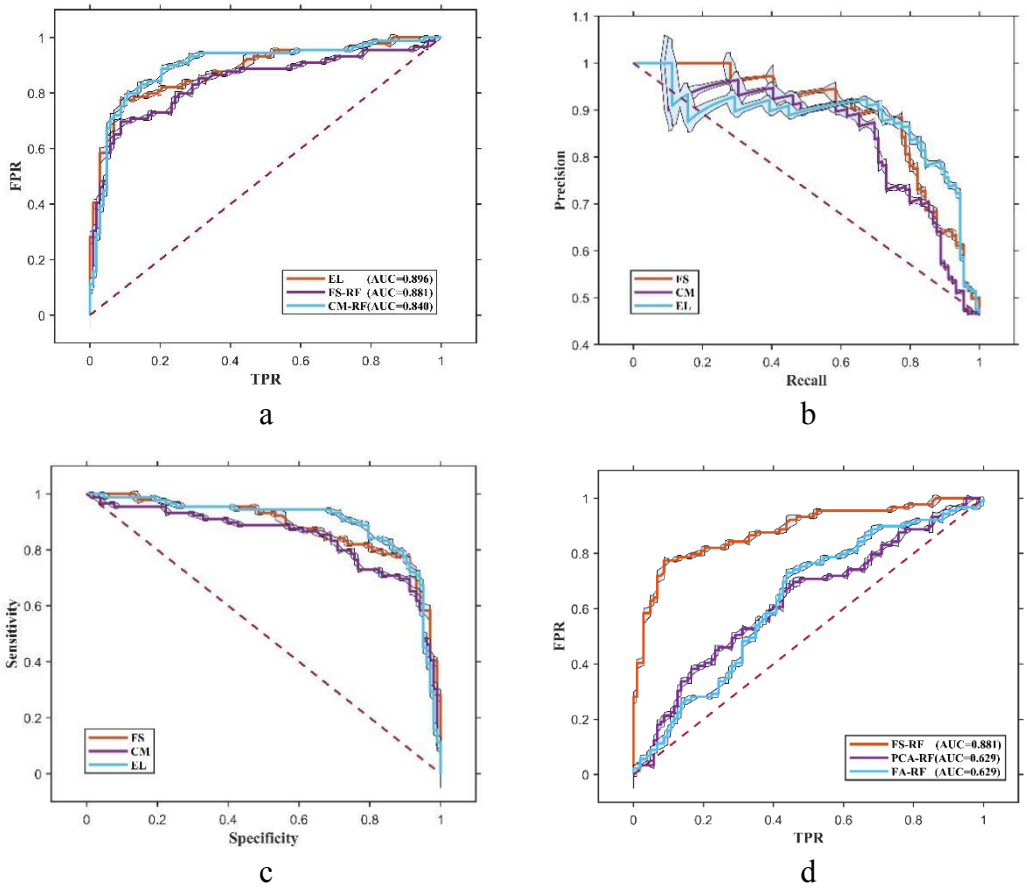


Figure 7. a. AUC of CM-RF, FS-RF and EL. b. Precision-recall curve of CM-RF, FS-RF and EL. c. Sensitivity-specificity curve of CM-RF, FS-RF and EL. d. AUC of our FS-Rf compared to PCA and FA.

Moreover, we review the medical literature related to atherosclerosis. HDL levels correlate strongly with atherosclerosis [28]. While increased IMT is an early clinical manifestation of atherosclerosis [29]. To further demonstrate the advantages of the prediction methods used in this paper, it is necessary to evaluate these single indicators in predicting atherosclerotic outcomes that have been shown to have a direct association with atherosclerosis. The ROC curves for HDL and IMT used to predict atherosclerosis respectively and the 25 feature sets prior to feature selection are shown in Figure 8.

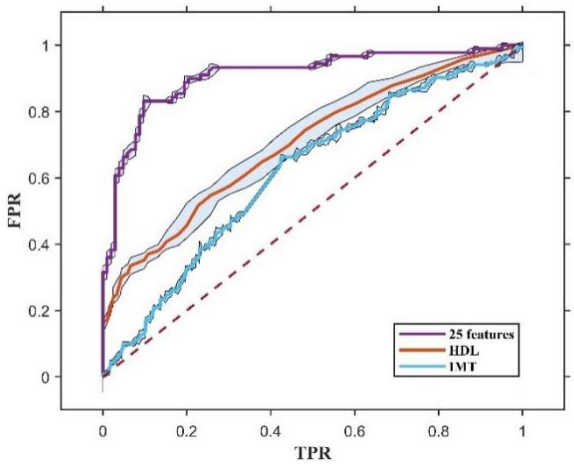


Figure 8. The ROC curves for HDL and IMT used to predict atherosclerosis respectively and the 25 feature sets prior to feature selection.

5. Conclusions

The work in this paper is based on research work on atherosclerosis and the guidance of professionals. The original data set is preprocessed and statistically analyzed through relevant statistical analyses and tests. We build an atherosclerosis prediction model based on random forest classifier with good results. As the result of the comparison between the ROC curves for HDL/IMT and the 25 feature sets prior to feature selection, Even the use of a single medically proven indicator with a strong correlation with atherosclerosis for prediction is far from the predictive effect of our model. Further evidence of the great role that feature redundancy plays in model prediction is provided. Then we transform the data screening problem into an optimization problem based on optimal paths. We obtain an optimized feature set containing 15 features by building an optimal distance model solved based on Dijkstra algorithm, whose model effects are optimized. Finally, it is then boxed and the feature set is discretized. The final model yielded an AUC metric of 0.9170, an improvement of 0.0472 from the initial one. This illustrates that the optimal distance feature screening model proposed in this paper improves the performance of the atherosclerosis prediction model in terms of both prediction accuracy and AUC metrics.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (11771216), the Key Research and Development Program of Jiangsu Province (Social Development) (BE2019725), NUIST Students' Platform for Innovation and Entrepreneurship Training Program (XJDC202110300552), and the Undergraduate Innovation & Entrepreneurship Training Program of Jiangsu Province (202110300098Y).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. C. Sinning, A. Kieback, P. S. Wild, R. B. Schnabel, F. Ojeda, S. Appelbaum, et al., Association of multiple biomarkers and classical risk factors with early carotid atherosclerosis: results from the Gutenberg Health Study, *Clin. Res. Cardiol.*, **103** (2014), 477–485. https://doi.org/10.1007/s00392-014-0674-6
2. J. F. Polak, M. J. Pencina, D. H. O'Leary, R. B. D'Agostino, Common carotid artery intima-media thickness progression as a predictor of stroke in multi-ethnic study of atherosclerosis, *Stroke*, **42** (2011), 3017–3021. https://doi.org/10.1161/STROKEAHA.111.625186
3. M. W. Lorenz, C. Schaefer, H. Steinmetz, M. Sitzer, Is carotid intima media thickness useful for individual prediction of cardiovascular risk? Ten-year results from the Carotid Atherosclerosis Progression Study (CAPS), *Eur. Heart J.*, **31** (2010), 2041–2048. https://doi.org/10.1093/eurheartj/ehq189
4. M. Soni, M. Ambrosino, D. S. Jacoby, The use of subclinical atherosclerosis imaging to guide preventive cardiology management, *Curr. Cardiol. Rep.*, **23** (2021), 61. https://doi.org/10.1007/s11886-021-01490-7
5. A. Hazra, S. K. Mandal, A. Gupta, A. Mukherjee, A. Mukherjee, Heart disease diagnosis and prediction using machine learning and data mining techniques: a review, *Adv. Comput. Sci. Technol.*, **10** (2017), 2137–2159.
6. M. Shouman, T. Turner, R. Stocker, Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients, *Comput. Sci. Conf. Proc.*, **5** (2012), 125–137. https://doi.org/10.5121/csit.2012.2511
7. O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, Classification and prediction of atherosclerosis diseases using machine learning algorithms, in *International Conference on Optimization and Applications (ICOA)*, **5** (2019), 1–5. https://doi.org/10.1109/ICOA.2019.8727688
8. D. Han, K. K. Kolli, S. J. Al'Aref, L. Baskaran, A. R. van Rosendael, H. Gransar, et al., Machine learning framework to identify individuals at risk of rapid progression of coronary atherosclerosis: from the PARADIGM registry, *J. Am. Heart Assoc.*, **9** (2020), e013958. https://doi.org/10.1161/JAHA.119.013958
9. O. Couturier, H. Delalin, H. Fu, G. Edouard, A three-step approach for stulong database analysis: characterization of patients groups, in *Proceeding of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery*, 2004.
10. M. Abdar, W. Ksiazek, U. R. Acharya, R. S. Tan, V. Makarenkov, P. Plawiak, A new machine learning technique for an accurate diagnosis of coronary artery disease, *Comput. Methods Programs Biomed.*, **179** (2019), 104992. https://doi.org/10.1016/j.cmpb.2019.104992
11. V. S. H. Rao, M. N. Kumar, Novel approaches for predicting risk factors of atherosclerosis, *IEEE J. Biomed. Health*, **17** (2012), 183–189. https://doi.org/10.1109/TITB.2012.2227271
12. J. Xie, R. Wu, H. Wang, Y. Kong, H. Li, W. Zhang, A novel weight learning approach based on density for accurate prediction of atherosclerosis, in *Intelligent Computing Theories and Application* (eds. D. S. Huang, K. H. Jo., Z. K. Huang), Springer, (2019), 190–200. https://doi.org/10.1007/978-3-030-26969-2_18
13. W. He, Y. Xie, H. Lu, M. Wang, H. Chen, Predicting coronary atherosclerotic heart disease: an extreme learning machine with improved salp swarm algorithm, *Symmetry*, **12** (2020), 1651. https://doi.org/10.3390/sym12101651

14. A. Ward, A. Sarraju, S. Chung, J. Li, R. Harrington, P. Heidenreich, Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population, *NPJ Digit. Med.*, **125** (2020), 1–7. https://doi.org/10.1038/s41746-020-00331-1
15. S. Nikan, F. Gwadry-Sridhar, M. Bauer, Machine learning application to predict the risk of coronary artery atherosclerosis, in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, (2016), 34–39. https://doi.org/10.1109/CSCI.2016.0014
16. J. Xie, H. Wang, J. Zhang, C. Meng, Y Kong, S. Mao, et al., A novel hybrid subset-learning method for predicting risk factors of atherosclerosis, in 2*017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2017), 2124–2131. https://doi.org/10.1109/BIBM.2017.8217987
17. M. Priya, P. Ranjith Kumar, A novel intelligent approach for predicting atherosclerotic individuals from big data for healthcare, *Int. J. Prod. Res*., **53** (2015), 7517–7532. https://doi.org/10.1080/00207543.2015.1087655
18. A. I. Sakellarios, V. C. Pezoulas, C. Bourantas, K. K. Naka, L. K. Michalis, P. W. Serruys, et al., Prediction of atherosclerotic disease progression combining computational modelling with machine learning, in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, (2020), 2760–2763. https://doi.org/10.1109/EMBC44109.2020.9176435
19. B. Kumar, H. Mathur, Comprehensive analysis of atherosclerosis disease prediction using machine learning, *Ann. Rom. Soc. Cell Biol.*, **4** (2021), 17962–17975.
20. M. Lin, H. Cui, W. Chen, A. van Engelen, M. de Bruijne, M. R. Azarpazhooh, et al., Longitudinal assessment of carotid plaque texture in three-dimensional ultrasound images based on semi-supervised graph-based dimensionality reduction and feature selection, *Comput. Biol. Med*., **116** (2020), 103586. https://doi.org/10.1016/j.compbiomed.2019.103586
21. Q. A. Hathaway, N. Yanamala, M. J. Budoff, P. P. Sengupta, I. Zeb, Deep neural survival networks for cardiovascular risk prediction: The Multi-Ethnic Study of Atherosclerosis (MESA), *Comput. Biol. Med*., **139** (2021), 104983. https://doi.org/10.1016/j.compbiomed.2021.104983
22. A. D. Jamthikar, D. Gupta, L. Saba, N. N. Khanna, K. Viskovic, S. Mavrogeni, et al., Artificial intelligence framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound, *Comput. Biol. Med*., **126** (2020), 104043. https://doi.org/10.1016/j.compbiomed.2020.104043
23. S. S. Skandha, S. K. Gupta, L. Saba, V. K. Koppula, A. M. Johri, N. N. Khanna, et al., 3-D optimized classification and characterization artificial intelligence paradigm for cardiovascular/stroke risk stratification using carotid ultrasound-based delineated plaque: Atheromatic™ 2.0, *Comput. Biol. Med*., **125** (2020), 103958. https://doi.org/10.1016/j.compbiomed.2020.103958
24. R. H. Lopes, I. D. Reid, P. R. Hobson, The two-dimensional Kolmogorov-Smirnov test, *Prod. Sci.*, (2007), 1–12.
25. G. Biau, E.Scornet, A random forest guided tour, *Test*, **25** (2016), 197–227. https://doi.org/10.1007/s11749-016-0481-7
26. M. Noto, H. Sato, A method for the shortest path search by extended Dijkstra algorithm, in *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions'*, **3** (2000), 2316–2320. https://doi.org/10.1109/ICSMC.2000.886462

27. N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, *Mach. Learn*., **29** (1997), 131–163. https://doi.org/10.1023/A:1007465528199
28. F. Xu, J. Zhang, X. Zhou, H. Hao, Lipoxin A4 and its analog attenuate high fat diet-induced atherosclerosis via Keap1/Nrf2 pathway, *Exp. Cell Res.*, **412** (2022), 113025. https://doi.org/10.1016/j.yexcr.2022.113025
29. F. Polak, J. Y. C. Backlund, M. Budoff, P. Raskin, I. Bebu, J. M. Lachin, et al., Coronary artery disease events and carotid intima-media thickness in Type 1 diabetes in the DCCT/EDIC cohort, *J. Am. Heart Assoc*., **24** (2021), e022922. https://doi.org/10.1161/JAHA.121.022922

AIMS Press