Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnosis of atherosclerosis based on pulse wave and ensemble learning method: Weighted-Ensemble model

Rongbin Chen
Nanjing University of Information Science and Technology
Wenjun Liu (wjliu@nuist.edu.cn)
Nanjing University of Information Science and Technology
Hui Huang
Affiliated Hospital of Nanjing University of CM
Songtao Bai
Nanjing University of Information Science and Technology

Research Article

Keywords: Atherosclerosis, Classification, Random forests, Weighted-Ensemble

Posted Date: April 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1517698/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Purpose: Atherosclerosis (AS) is closely related to cardiovascular disease (CVD). Nowadays, many scholars have conducted research on CVD, but the diagnosis of AS can only be diagnosed based on traditional medical methods. Pulse wave velocity (PWV) can evaluate potential AS. As a new technology, ultrafast pulse wave velocity (ufPWV) can accurately evaluate PWV. This research aims to screen out relevant features through feature engineering methods and build a Weighted-Ensemble model based on these features to predict AS, which can assist doctors in making more effective diagnosis of atherosclerosis.

Methods: In this paper, the traditional statistical analysis method and Random forests (RF) are used to ensemble selection characteristics. This paper improves the ensemble model and applies it to the prediction of AS. Based on the idea of bagging, the base model of RF is changed, and all decision trees in RF are replaced with prediction models with better generalization ability. In order to make full use of models with high generalization ability, this study will also introduce a boosting strategy to weight each base model to form a Weighted-Ensemble model with high generalization ability. The accuracy (ACC), sensitivity (TPR), specificity (TNR) and AUC are four evaluation criteria to evaluate the model.

Results: There are 37 features in the data set. Based on the statistical analysis of the data set, a total of 16 characteristics affect the diagnosis of AS. According to the importance of the features obtained by RF model, the 10 most important features are used to construct a Weighted-Ensemble model. The results show that the ACC, TPR, TNR and AUC of the Weighted-Ensemble model are 0.91, 0.93, 0.89 and 0.91 respectively. Compared with each model, the ACC, TPR and AUC of the Weighted-Ensemble model are better than other models. The analysis shows that the Weighted-Ensemble model is superior to traditional machine learning methods in distinguishing patients as normal or atherosclerosis.

Conclusions: The accuracy of the algorithm based on the Weighted-Ensemble model in predicting AS has been confirmed in this paper, which implicates that the Weighted-Ensemble model can be successfully used in the atherosclerosis diagnosis and decision-making system.

Introduction

According to reports recently published by Chinese Center for Cardiovascular Diseases, the number of people suffering from cardiovascular disease (CAD) in China is about 330 million, of which 13 million are cerebral stroke, 11 million are coronary heart disease, 8.9 million are heart failure, and 245 million are hypertension [1]. As a heart and vascular disease, CVD is the leading cause of death worldwide [2]. Common diseases such as angina pectoris, myocardial infarction, hypertension, and cardiac insufficiency are multiple heart diseases. In most people's understanding, CVD occurs suddenly. However, CVD has traces to follow before it occurs. The incidence rate of CVD is increasing year by year. The main cause is AS. AS is one of the most serious chronic diseases, which will affect human health. It is the main

pathological basis of ischemic cardiovascular and cerebrovascular diseases such as coronary heart disease, cerebrovascular disease and thromboembolic disease [3].

Yi et al. [4] Found that smoking can lead to the prevalence of arteriosclerosis. In 1994, Vanderwal et al. [5] showed that T-cells and mast cells at the site of plaque rupture produce many types of molecules-inflammatory cytokines, proteases, coagulation factors, free radicals, and vasoactive molecules, which can make AS lesions unstable. In 1994, Moreno et al. [6] found that macrophages are markers of unstable atherosclerotic plaque and may play an important role in the pathophysiology of acute coronary syndrome. Studies by Hansson team [7] and Amento team [8] showed that the above reactions may lead to plaque activation and rupture, thrombosis and ischemia. Zhu et al. [9] showed that compared with healthy adults, cIMT in high-risk groups of AS was related to patient age and carotid artery. In addition, cIMT was related to age. In other words, patients with AS are related to age. Ross et al. [10] comprehensively described the development process of AS in literature and proposed that AS is an inflammatory disease. Literature [11, 12, 13] thought that C-reactive protein, the most iconic factor in the process of inflammation, is considered to be a highly sensitive detection index in the occurrence and development of AS. Nofer et al. [14] pointed out that HDL3 can reduce the production of IP3, thereby inhibiting thrombin, and then causing thrombosis. Literature [15, 16] showed that inflammation can oxidatively modify LDL, and the oxidized LDL further promotes the inflammatory process of the arterial intima. In addition to inflammation, hypertension and infection are also important causes of AS. Kranzhofer et al. [17] considered that angiotensin II in hypertensive patients will increase, and angiotensin II will stimulate the growth of vascular smooth muscle, thereby forming AS. Kuvin and Kimmelstiel [18] pointed out that infection can cause AS to form. Geisel et al. [19] found that compared with cIMT and ABI, coronary artery calcification provides the best risk identification, especially in the medium-risk group, IMT thickening can reflect the presence of early atherosclerosis. Ohkuma et al. [20] showed that Brachial-ankle pulse wave velocity was significantly associated with cardiovascular events and was independent of traditional risk factors. Kawada et al. [21] used the plaque score (PS) to describe the severity of plaque formation and found that PS can be used to assess the presence of advanced atherosclerosis. At the same time, Kawada et al. found that metabolic syndrome is significantly associated with carotid AS. Li et al. [22] found that the PWV-BS and PWV-ES values of hypertensive patients were significantly increased in the study of Ultra-Fast imaging technology to determine the pulse wave velocity of hypertensive patients and related influencing factors, which can be used as an index to evaluate the elastic function of carotid artery. Bos et al. [23] found that the traditional cardiovascular risk factors are related to intracranial carotid atherosclerosis by studying the prevalence and risk factors of intracranial carotid atherosclerosis in the general population, but the distribution of risk factors is different between men and women, and the risk of men is higher than that of women. Mirault et al. [24] found that ultrafast imaging can evaluate the carotid pulse wave propagation velocity and its changes in the cardiac cycle. The difference between the early and late pulse wave propagation velocity increases with age. Yang et al. [25] found that age, body mass index and blood pressure were the main factors affecting ufPWV. Carew et al. [26] found that antioxidant LDL antibody has pathogenic effect on aortic lesions when studying the role of LDL in the process of atherosclerotic lesions.

AS has a strong relationship with cholesterol content, so it can be judged by the content of total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C) obtained by blood test [27]. The formed AS damage can be observed by the following features in B-mode ultrasound images: the bulge of intimal medium; Total plaque area; Total plaque volume [28]. The degree of arterial stiffness is the physical property of the arterial vessel itself, so it can be measured by the speed at which the pulse wave travels along the arterial vessel, that is, the pulse wave velocity. At present, a large number of studies have confirmed that PWV can accurately reflect the degree of arterial stiffness. Literature [29] studied more than 3000 people over 60 years old and concluded that there is a statistically significant and strong correlation between arterial stiffness and atherosclerosis. Therefore, clinical measurement of arterial stiffness can be used as a means to detect AS. For patients with known or unknown AS, the conventional methods are biochemical detection, image detection and physical detection, but these methods are not only time-consuming and expensive, but also cause harm to patients, which not only causes economic losses to patients, but also brings hidden dangers to patient's health.

The research team of Vienna Medical University in Austria found that April protein can reduce subendothelial lipid deposition and prevent the formation of AS [30]. Yusuf et al. [31] pointed out that most patients with AS are obese. At the same time, smoking will aggravate AS, and people with too much pressure are also easy to cause AS. Therefore, you can control the total calories of food, eat more foods rich in vitamin C and plant protein, and do not overeat. Carry out sports appropriately and maintain a positive and optimistic attitude towards life and work

Nowadays, for many diseases, many scholars not only use traditional and conventional medical diagnosis methods, but also use machine learning methods for auxiliary diagnosis. Many scholars have applied machine learning algorithm to predict other diseases closely related to atherosclerosis, and achieved good research results.

In 2017, Xu et al. [32] used a logistic regression algorithm to detect 7360 CAD patients and non-CAD patients. They found seven factors that are closely related to CAD: age, gender, Serum creatinine (Scr), smoke, angina, diabetes, Low Density Lipoprotein (ldl). And gave its specific formula

$$\begin{aligned} f(x) = & -5.2782 + 0.0549 * age + 0.6743 * sex + 0.007 * Scr \\ & + 0.4776 * smoke + 0.5516 * angina + 0.8641 * diabetes \\ & + 0.2651 * ldl \end{aligned} \quad (1)$$

According to the sigmoid function, the patient's disease probability can be obtained. Xu et al. set the probability threshold to 0.79. The specificity of the model is 0.709 and the sensitivity is 0.658, but the accuracy of the model is very low. Investigating the reason, the author only considered the linear relationship between the target variable and each factor and did not consider its non-linear relationship. In 2017, Tan et al. [33] used the CNN-LSTM model to detect 6120 CAD and 32000 non-CAD patients. They used two layers of CNN (two layers of pooling layer, two layers of convolution layer) and three layers of LSTM, and the last layer of full connection layer, a total of eight layers of cnn-lstm to extract signals from

patient's ECG, then the extracted features are fitted. The sensitivity, specificity and accuracy of the model were 0.9985, 0.9984 and 0.9985 respectively. The generalization ability of the model is good, but it does not reveal the relationship between features and target variable. In 2017, Acharya et al. [34] used CNN model to fit 2-seconds segment ECG (model A) and 5-seconds segment ECG (model B) respectively. The model consists of five convolution layers, five pooling layers and one full connection layer. The sensitivity of model A is 0.9372, the specificity is 0.9518, and the accuracy is 0.9495. The sensitivity of model B is 0.9113, the specificity is 0.9588, and the accuracy is 0.9511. The model also fails to reveal the relationship between features and target variable. In 2017, Lih et al. [35] used wavelet packet decomposition to process the ECG of 12308 CAD patients and 3791 normal people and use K-Nearest Neighbor classifier to classify it. The sensitivity of the model is 0.9964, the specificity is 0.9971, and the accuracy is 0.9965. Olaniyi et al. [36] used KNN, DT, naive Bayesian WAC and BPNN to fit Cleveland dataset at the same time, and the accuracy rates were 0.8567, 0.8435, 0.8231, 0.84 and 0.85 respectively. Alizadehsani et al. [37] used SVM to perform CAD prediction on the new data set Z-Alizadeh Sani dataset they extracted in 2016.The researchers used four different kernel functions to fit the model, and the model with the highest accuracy used RBF core, its accuracy rate is 0.8185. In the next few years, scholars successively built different classification models to predict CAD based on the Z-Alizadeh Sani dataset. Arabasadi et al. [38] fused GA and ANN models—GA-ANN. The sensitivity of this model is 0.97, the specificity is 0.92, and the accuracy is 0.9385. The sensitivity of only using the ANN model is 0.86, the specificity is 0.83, and the accuracy is 0.8462.

This paper aims to apply the improved ensemble learning algorithm to the diagnosis of AS. As far as we know, this paper not only applies machine learning method to atherosclerosis detection, but also proposes a new model that ensemble algorithm based on strong classifier. This paper also improved the proposed new model. In addition, we will compare the results of our model with the results of other traditional machine learning methods, such as RF, eXtreme Gradient Boosting (XGboost) and Support Vector machines (SVM). This paper uses AS data to propose a Weighted-Ensemble model based on strong classifiers. Firstly, according to the correlation analysis in statistics, filter out the features in the data set that have no influence on the target variable. Secondly, put the remaining features in the data set into the RF, and screen out the important features according to the Gini index. Finally, we use the selected important features to build our model, and use the fitted model to predict disease. The main innovations and contributions of this paper are as follows. a) The improved machine learning method is applied to the prediction of AS. b) Factors that have important effects on AS were screened from the data set. c) A new model, Weighted-Ensemble model based on strong classifier, is proposed. d) Compared with other machine learning algorithms, our proposed model has higher quasi-prediction accuracy and better generalization ability.

In fact, the results of selecting important features will reveal the relationship between AS and various factors, and doctors and scientists can make more scientific decisions based on these results.

Materials And Methods

Materials and data preprocessing

The data in this paper is a retrospective study conducted by the Affiliated Hospital of Nanjing University of Chinese Medicine from January 2016 to December 2017. 321 valid samples were selected, each with 37 features. After the pathological study of AS and the guidance of experts, 321 samples were divided into 144 AS risk groups and 177 control groups. The classification rules of AS risk group are as follows: a) Experts evaluated the risk of hypertension, chronic kidney disease or cardiovascular and cerebrovascular diseases according to relevant test indicators. b) It is considered that patients with hyperlipidemia (HL) are at risk of AS. HL is defined as the increase of low-density lipoprotein (LDL) level (mmol / L), total cholesterol (TC) level (mmol / L), or triglyceride (TG) level (mmol / L) [39].

The control group was mainly healthy people in the same period. Those who had abnormal hemoglobin, those who had history of cardiovascular and cerebrovascular events, and those who had cancer, Diabetes or autoimmune diseases were excluded.

The 37 features were gender, age, height, weight, body mass index (BMI), low density lipoprotein (LDL), high density lipoprotein (HDL), triglyceride (TG), total cholesterol (TC), glucose (Glu), uric acid (Ua), creatinine (Cre), Urea, systolic blood pressure (SBP), diastolic blood pressure (DBP), LCCA-SDV, LCCA-RI, LICA-SDV, LICA-RI, RCCA-SDV, RCCA-RI, RICA-SDV, RICA-RI, LCCA-IMT, RCCA-IMT, LCCA-BS, LCCA -ES, RCCA-BS, RCCA-ES, white blood cell count (WBC), red blood cell count (RBC), platelet (Pla), alanine aminotransferase (GPT), aspartate aminotransferase (GOT), hemoglobin (Hb), smoking, drinking. Female and male are represented by 0 and 1 respectively, 0 means no smoking or drinking, and 1 means smoking or drinking.

This paper specifically analyzes the missing values of all features. There are 12 features with a missing rate exceeding 40%, namely LCCA-RI, LICA-RI, RCCA-RI, RICA-RI, RICA-SDV, smoking history, drinking history, RCCA-SDV, LICA-SDV, LCCA-SDV, SBP, DBP. For features with a missing rate of more than 40%, these features are removed, and the missing rate of the remaining features is less than 20%. The common processing methods of missing value filling include statistical filling (such as mean, median, mode), multiple imputation, K nearest neighbor filling, regression filling and so on. Taking into account the difficulty of operation, we use the mean-filling method for continuous variables, and the mode-filling method for discrete variables.

Feature selection

In this paper, statistical analysis, RF and ensemble feature selection algorithms are used for feature selection. this research starts with a correlation analysis. For continuous features, we express them as meanstandard deviation, and independent sample t-test was used to compare the differences between groups. Discrete features are expressed in counts and percentages, and the difference between groups is compared by chi-square test. Take P = 0.005 as the inspection standard. There are 16 features with

significant differences. Then, the features with significant differences are scored by RF model, and the most important part is selected as the model training features.

Random forests model and Weighted-Ensemble model

The ensemble methods mainly include bagging and boosting [40]. RF is a typical representative of bagging, and its base model is decision tree [41]. The characteristics of decision tree determine that RF can not only achieve the purpose of regression, but also be used for classification. RF will first generate multiple sub training sets according to bootstrap sampling, and then each sub training set will train a decision tree respectively. For the classification problem, the RF uses the voting method, and the final result is determined by the majority principle [42]. Figure 1 shows the RF model.

Each tree in the forest has relevant criteria when generating child nodes. When our goal is classification, the main classification criteria are impulse based criteria, information gain and Gini index. When our goal is regression, the main criterion is mean square error. This paper uses Gini index.

There are two main factors that affect the ensemble learning algorithm. On the one hand, determine whether the base model of ensemble learning is independent. If the base models are highly correlated, then the ensemble model and each base model will not have obvious predictive effects, even worse. This result does not improve the effect of the model, and increases the complexity of the model. On the other hand, the accuracy of each base model will also affect the accuracy of the ensemble learning algorithm. If the accuracy of each base model is low, there is no doubt that the accuracy of the ensemble learning model will be low. Because each base model of the traditional ensemble model is consistent (the same model), based on the features of ensemble learning, this study improves the RF and proposes an ensemble learning model based on strong classifier. The base classifier is no longer a single decision tree, but some classifiers with stronger generalization ability, such as RF, XGboost, SVM, KNN and ANN. Since the base models are different, each base model is naturally independent of each other, which solves the first factor that affects the effect of ensemble learning. Practice has shown that the above base model has always had an ideal effect on the classification effect, and the second factor that affects the effect of ensemble learning is also solved.

Based on the strong classifiers model, according to the idea of boosting, this paper give weight to the base model, so that all the base models can be fully utilized, to improve the generalization ability of the model, and finally get a weighted ensemble model with better recognition ability for atherosclerotic diseases. Figure 2 is the improved ensemble model. Figure 3 is the Weighted-Ensemble model. As shown in Fig. 2, the decision tree in the RF is replaced with a model with better generalization ability. As shown in Fig. 3, the original base model is multiplied by a weight,

$a_1, a_2 \ldots . a_n .\quad a_k = m_k \Big/ \sum_{i=1}^{n} m_i \;\; m_k = \frac{1}{2} * \log\left(\frac{(1-e_k)}{e_k} \right)$, e_k is the error rate of the base model.

Statistical analysis

Firstly, this study divide the AS data set into two parts according to the ratio of 8:2. The former is used as the training set of the model, and the latter is used as the test set of the model. Then, in the training stage, the data of the training set will be used for the training of each base model. The base model adjusts the model structure according to the goal of reducing the loss function. The data used for testing does not participate in the adjustment of the model structure but is used to measure the generalization ability of the model. Thus, the entire model building process can be summarized as: dividing the training set and the test set, using the training set for training the model, and using the test set for the ability of the created model to predict invisible data. It needs to be emphasized that we use the 10-fold cross validation procedure to train the model. Although the cross-validation method will increase the training time of the model, it can provide us with more data to use, so that the model can achieve better results.

To construct a Weighted-Ensemble model for predicting whether the samples in the atherosclerosis data set are atherosclerotic patients, this research used PyCharm Community Edition 3.8. The computer version information used for model training and prediction is intel@ Core i5 @ 1.6 GHz CPU and 4 GB 1600MHz DDR3. To make the model have better generalization ability so that it can well predict the samples in the test, we use the grid search method to adjust the parameters of each base model. For example, the parameters of the RF in the base model are criterion = 'Gini', max_ depth = 3, min_ samples_ split = 3, n_ estimators = 100.The specific parameters of all base models are shown in Table 1. Figure 4 is the base model XGboost in the trained Weighted-Ensemble model.

Table 1
Base models and its parameters

Base Models	parameters
XGboost	base_score = 0.5, booster='gbtree', importance_type='gain', learning_rate = 0.01, max_depth = 6, min_child_weight = 2, n_estimators = 300, subsample = 0.8
RF	criterion='Gini', max_depth = 3, min_samples_split = 3, n_estimators = 100
ANN	activation='logistic', alpha = 1e-05, hidden_layer_sizes= (50, 50, 50)
KNN	n_neighbors = 7
SVM	kernel='rbf', C = 1, gamma = 0.001

Results

Features selection using statistical analysis

A total of 321 patients participated in the trial. Through statistical analysis, there are 16 characteristics that are significantly different. it can be seen from Table 2 that age, BMI, UA, CRE, urea, LCCA-ES and RCCA-ES in AS risk group are significantly higher than those in normal group, and HDL, RBC and Hb are

significantly lower than those in normal group. As can be warned by several indicators, For example, when HDL and Ua are significantly decreased, we have reason to suspect that the patient may have AS, so we can conduct a more comprehensive examination.

Table 2
Normal and AS t-test and chi-square test

Features	Normal	AS	p-value
Gender			
male	71	75	0.032
female	106	69	
Age, years	50.0114.32	59.8413.30	< 0.001
Height, cm	165.297.44	166.067.47	0.366
Weight, kg	62.079.66	68.2611.80	< 0.001
BMI, kg/m2	22.662.79	24.683.57	< 0.001
LDL, mmol/l	2.550.56	2.460.81	0.261
HDL, mmol/l	1.480.38	1.190.34	< 0.001
TG, mmol/l	1.140.72	1.802.30	< 0.001
TC, mmol/l	4.720.87	4.662.08	0.695
Glu, mmol/l	5.060.87	5.811.88	< 0.001
Ua, umol/l	281.0887.45	361.97122.13	< 0.001
Cre, umol/l	67.3815.53	78.318.92	< 0.001
Urea, mmol/l	5.091.49	7.885.80	< 0.001
LCCA-BS, m/s	6.381.44	6.925.11	0.227
LCCA-ES, m/s	8.082.35	9.702.23	< 0.001
RCCA-BS, m/s	5.971.31	5.851.64	0.492
RCCA-ES, m/s	7.662.36	9.202.06	< 0.001
WBC, 10^9/l	5.951.58	6.692.24	0.002
RBC, 10^12/l	4.610.48	4.020.74	< 0.001
Pla, 10^9/l	202.2952.31	197.0569.06	0.461
GPT, u/l	23.9315.31	25.7519.41	0.370
GOT, u/l	23.2911.88	23.2715.95	0.985
Hb, g/l	138.3015.13	124.9220.96	< 0.001
LCCA-IMT, mm	0.050.01	0.070.04	< 0.001
RCCA-IMT, mm	0.050.01	0.060.05	0.003

Features selection using RF

Using the RF model to select the features of the data set, we selected 10 features for model training, which are RBC, BMI, CRE, Ua, HDL, RCCA-ES, Hb, Urea, Age and LCCA-ES. Figure 5 shows the importance of each feature in the atherosclerosis model based on random forest diagnosis.

Features selection using ensemble feature selection algorithms

According to the results in Table 2, the features are reprocessed, and the significantly different features are screened by using the random forest model. Figure 5 shows the results of the ensemble feature selection algorithm.

Results and analysis of classification algorithm

Table 3 shows the classification results of different feature selection methods and models. It can be seen from the table that the performance of the weighted-ensemble model is the best no matter which feature is selected to train the model. The ACC, TPR, TNR and AUC of the weighted-ensemble model based on statistical analysis feature selection on the test set are 86%, 86%, 86% and 0.86 respectively. The ACC, TPR, TNR and AUC of the weighted-ensemble model based on RF feature selection on the test set are 88%, 83%, 92% and 0.87 respectively. The ACC, TPR, TNR and AUC of the weighted integration model based on the Ensemble feature selection algorithms are 91%, 93%, 89% and 0.91 respectively.

The weighted-ensemble model based on Ensemble feature selection algorithms has the best comprehensive performance in AS data set Compared with other models and other feature selection methods, its ACC, TPR and AUC value are the highest, which shows that the weighted integrated model based on integrated feature selection algorithm can effectively predict AS.

Table 3
Error analysis

evaluation criteria Features selection	ACC	TPR	TNR	AUC
Statistical analysis	Weighted-Ensemble	0.86	0.86	0.86	0.86
	XGboost	0.86	0.90	0.84	0.87
	RF	0.85	0.83	0.86	0.85
	ANN	0.82	0.83	0.81	0.82
	KNN	0.79	0.72	0.84	0.78
	SVM	0.76	0.76	0.76	0.76
RF	Weighted-Ensemble	0.88	0.83	0.92	0.87
	XGboost	0.86	0.86	0.86	0.86
	RF	0.85	0.79	0.92	0.86
	ANN	0.82	0.76	0.86	0.81
	KNN	0.79	0.66	0.95	0.80
	SVM	0.76	0.66	0.84	0.75
Ensemble feature selection	Weighted-Ensemble	0.91	0.93	0.89	0.91
	XGboost	0.85	0.79	0.89	0.84
	RF	0.86	0.79	0.92	0.85
	ANN	0.82	0.86	0.78	0.82
	KNN	0.85	0.86	0.84	0.85

Discussion and future work

Pulse wave velocity can effectively evaluate potential atherosclerosis and is independent of traditional risk factors. This study found that LCCA-ES, and RCCA-EShad significant differences between high-risk group and non high-risk group, while LCCA-BSand RCCA-BS had no significant differences. In addition, from the comparative observation of traditional atherosclerosis risk factors in this study, it can be seen that there are significant differences between CRE, Glu, Hb, HDL, RBC, TG, UA, urea and WBC, indicating that these factors play an important role in the formation and development of atherosclerosis, but these

factors have different characteristics in the mechanism of atherosclerosis. At the same time, age and weight are also an important cause of AS.

It is generally believed that hypertension, diabetes, hyperlipidemia, and smoking are independent risk factors for atherosclerosis. However, this study was conducted to analyze and study the data, because the data collected during the collection process failed to collect these data, leading to the high rate of data deletion and the inability to analyze all the potential pathogenic factors. In the future work, we will focus on the collection of data in this regard.

In conclusion, the occurrence and development of atherosclerosis is not only related to traditional cardiovascular risk factors, but also related to the conduction velocity of pulse wave. Clinically, people with one or more cardiovascular risk factors can be ufPWV detected in order to find the lesions of early as and intervene to reduce the occurrence of cardiovascular events.

This paper improved the RF, changed the decision tree in the RF to a classification model with better generalization ability. Besides, this study added a weight to each base model, which is determined by the classification error of the base model. The improved model is tested on the atherosclerosis dataset to check its performance improvement. Four indicators are used to evaluate the performance of the model in this research, namely accuracy, sensitivity, specificity, and AUC value. The results show that proposed method has better detection performance than other traditional classification models in predicting atherosclerosis. Specifically, using model proposed method to predict atherosclerosis can well predict whether the patient has atherosclerosis only by providing some test data, which helps to eliminate the high cost (there is no need for too many medical tests) and main side effects, To provide doctors and experts with a reliable method to diagnose atherosclerosis.

In addition to traditional machine learning algorithms, deep learning, reinforcement learning, transfer learning and other methods have also developed rapidly in recent years. In the future work, these methods can be making some small changes, so that these models can be used to predict atherosclerosis, so as to improve our prediction performance of atherosclerosis. Meanwhile, future research will continue to tune the parameters in our model until we find better parameters. Finally, continue to conduct more in-depth research on feature selection methods to find more representative features as our training features to improve the performance of the model.

Declarations

Declaration of Competing Interest

The authors declare no conflict of interest.

Author contributions

RC performed the data analyses and wrote the manuscript. WL contributed significantly to analysis and manuscript preparation. HH provided guidance on medical background and samples' attributes. SB

helped perform the analysis with constructive discussions. The writing process of the study is done by all of us.

Acknowledgements

We greatly appreciate the funding from the National Natural Science Foundation of China (11771216) and the Key Research and Development Program of Jiangsu Province (Social Development) (BE2019725).

References

1. National Center for Cardiovascular Diseases (2020) China. Annual report on cardiovascular health and diseases in China 2019 (in Chinese). Journal of Cardiovascular and Pulmonary Diseases,
2. WHO. Global Health Estimates (2016) : Deaths by Cause, Age, Sex, By Country and By Region. World Health Organization, 2018
3. Hansson GK (2005) Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med 352(16):1685–1695
4. Yi M, Chun EJ, Lee MS et al (2015) Coronary CT angiography findings based on smoking status: Do ex-smokers and never-smokers share a low probability of developing coronary atherosclerosis. Int J Cardiovasc Imaging 31(2):169–176
5. Vanderwal AC, Becker AE, Das PK et al (1994) Site of intimal rupture or erosion of thrombosed coronary atherosclerotic plaques is characterized by an inflammatory process irrespective of the dominant plaque morphology. Circulation 89(1):36–44
6. Moreno PR, Falk E, Palacios IF et al (1994) Macrophage infiltration in acute coronary syndromes: Implications for plaque rupture. Circulation 90(2):775–778
7. Hansson GK, Hellstrand M, Rymo L et al (1989) Interferon gamma inhibits both proliferation and expression of differentiation-specific alpha-smooth muscle actin in arterial smooth muscle cells. J Exp Med 170(5):1595–1608
8. Amento EP, Ehsani N, Palmer H et al (1991) Cytokines and growth factors positively and negatively regulate interstitial collagen gene expression in human vascular smooth muscle cells. Arterioscler Thromb 11(5):1223–1230
9. Zhu ZQ, Chen LS, Wang H et al (2019) Carotid stiffness and atherosclerotic risk: non-invasive quantification with ultrafast ultrasound pulse wave velocity. Eur Radiol 29(3):1507–1517
10. Ross R (1999) Atherosclerosis—an inflammatory disease. N Engl J Med 340(2):115–126
11. Koenig W (2001) Inflammation and coronary heart disease: an overview. Cardiol Rev 9(1):31–35
12. Yu H, Rifai N (2000) High-sensitivity C-reactive protein and atherosclerosis: from theory to therapy. Clin Biochem 33(8):601–610
13. Albert MA, Ridker PM (1999) The role of C-reactive protein in cardiovascular disease risk. Curr Cardiol Rep 1(2):99–104

14. Nofer JR, Walter M, Kehrel B et al (1998) HDL3-mediated inhibition of thrombin-induced platelet aggregation and fibrinogen binding occurs via decreased production of phosphoinositide-derived second messengers 1,2-diacylglycerol and inositol 1,4,5-tris-phosphate. Atherosclerosis Thromb Vascular Biology 18(6):861–869
15. Memon RA, Staprans I, Noor M, Arteriosclerosis et al (2000) Thromb Vascular Biology 20(6):1536–1542
16. Pentikäinen MO, Öörni K, Ala-Korpela M et al (2000) Modified LDL–trigger of atherosclerosis and inflammation in the arterial intima. J Intern Med 247(3):359–370
17. Kranzhofer R, Schmidt J, Hagl S et al (1999) Angiotensin induces inflammatory activation of human vascular smooth muscle cells. Arterioscler Thromb Vasc Biol 19(7):1623–1629
18. Kuvin JT, Kimmelstiel CD (1999) Infectious causes of atherosclerosis. Am Heart J 137(2):216–226
19. Geisel MH, Bauer M, Hennig F et al (2017) Comparison of coronary artery calcification, carotid intima-media thickness and ankle-brachial index for predicting 10-year incident cardiovascular events in the general population. Eur Heart J 38(23):1815–1822
20. Ohkuma T, Ninomiya T, Tomiyama H et al (2017) Brachial-ankle pulse wave velocity and the risk prediction of cardiovascular disease: an individual participant data meta-analysis. Hypertension 69(6):1045–1052
21. Kawada T, Andou T, Fukumitsu M (2016) Metabolic syndrome showed significant relationship with carotid atherosclerosis. Heart Vessels 31(5):664–670
22. Li HB, Wang H, YIN L P et al (2017) Determination of pulse wave velocity and related influencing factors in patients with hypertension by Ultra-Fast imaging technology (In Chinese). Chin J Hypertens 25(5):477–481
23. Bos D, Rijk VD, Hofman A et al (2012) Intracranial carotid artery atherosclerosis: prevalence and risk factors in the general population. Stroke 43(7):1878–1884
24. Mirault T, Pernot M, Frank M et al (2015) Carotid stiffness change over the cardiac cycle by ultrafast ultrasound imaging in healthy volunteers and vascular Ehlers–Danlos syndrome. J Hypertens 33(9):1890–1896
25. Yang W, Wang Y, Yu Y et al (2020) Establishing normal reference value of carotid ultrafast pulse wave velocity and evaluating changes on coronary slow flow[J]. Int J Cardiovasc Imaging 36(10):1931–1939
26. Carew TE (1989) Role of biologically modified low-density lipoprotein in atherosclerosis. Am J Cardiol 64(13):G18–G22
27. Brunner D, Weisbort J, Meshulam N et al (1987) Relation of serum total cholesterol and high-density lipoprotein cholesterol percentage to the incidence of definite coronary events: twenty-year follow-up of the Donolo-Tel Aviv Prospective Coronary Artery Disease Study. Am J Cardiol 59(15):1271–1276
28. Steinl DC, Kaufmann BA (2015) Ultrasound imaging for risk assessment in atherosclerosis. Int J Mol Sci 16(5):9749–9769

29. Popele NM, Grobbee DE, Bots ML et al (2001) Association between arterial stiffness and atherosclerosis: the Rotterdam Study. Stroke 32(2):454–460
30. Tsiantoulas D, Eslami M, Obermayer G et al (2021) APRIL limits atherosclerosis by binding to heparan sulfate proteoglycans. Nature 597(7874):92–96
31. Yusuf S, Hawken S, Ounpuu S et al (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364(9438):37–52
32. Xu H, Duan Z, Miao C et al (2017) Development of a diagnosis model for coronary artery disease. Indian Heart J 69(5):634–639
33. Tan JH, Hagiwara Y, Pang W et al (2018) Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Comput Biol Med 94:19–26
34. Acharya UR, Fujita H, Lih OS et al (2017) Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. Knowl Based Syst 132:62–71
35. Lih OS, Adam M, Tan JH et al (2017) Automated identification of coronary artery disease from short-term 12 lead electrocardiogram signals by using wavelet packet decomposition and common spatial pattern techniques. J Mech Med Biology 17(07):1740007
36. Olaniyi EO, Oyedotun OK, Helwan A et al (2015) Neural network diagnosis of heart disease. International Conference on Advances in Biomedical Engineering (ICABME). IEEE, : 21–24
37. Alizadehsani R, Zangooei MH, Hosseini MJ et al (2016) Coronary artery disease detection using computational intelligence methods. Knowl Based Syst 109(1):187–197
38. Arabasadi Z, Alizadehsani R, Roshanzamir M et al (2017) Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Methods Programs Biomed 141:19–26
39. Arnett DK, Blumenthal RS, Albert MA et al (2019) 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 74(10):e177–e232
40. Dietterich TG (2000) Ensemble methods in machine learning. 10 (2):1–15
41. Breiman (2001) Random Forests Machine Learning 45(1):5–32
42. Bai X, Liu W, Huang H et al (2022) Ultrafast pulse wave velocity and ensemble learning to predict atherosclerosis risk. Int J Cardiovasc Imaging 1–9. DOI: 10.1007/s10554-022-02574-3

Figures

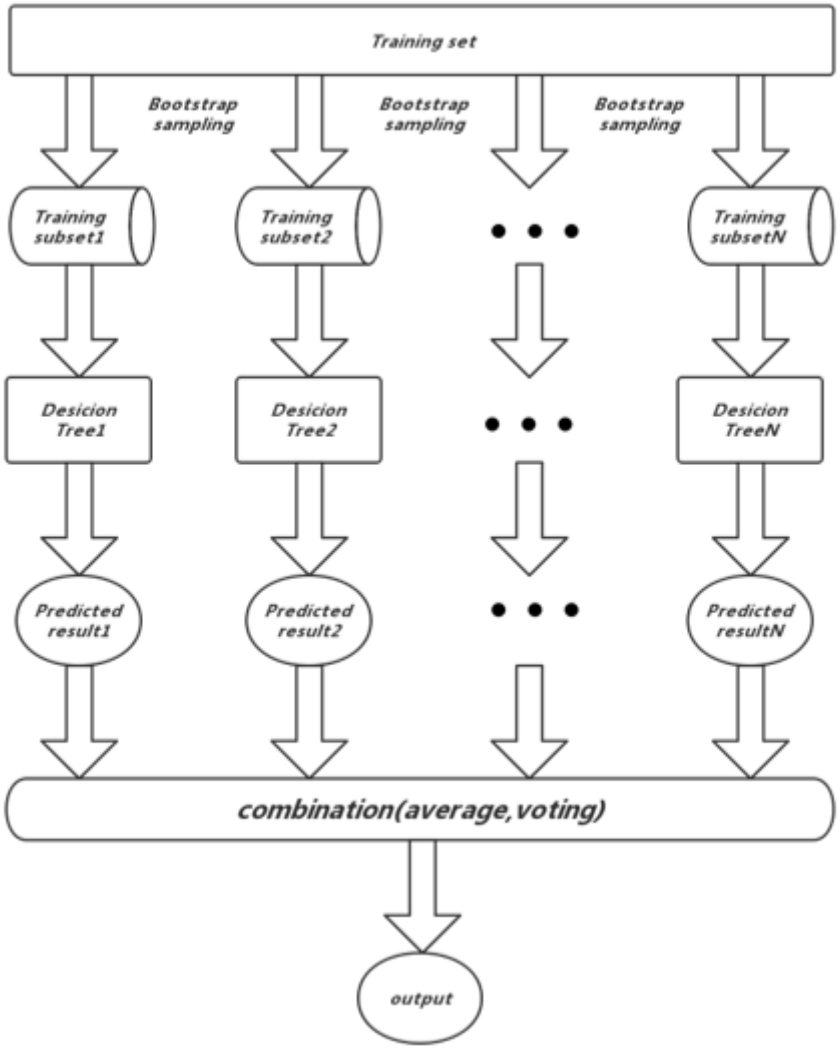

Figure 1

RF model

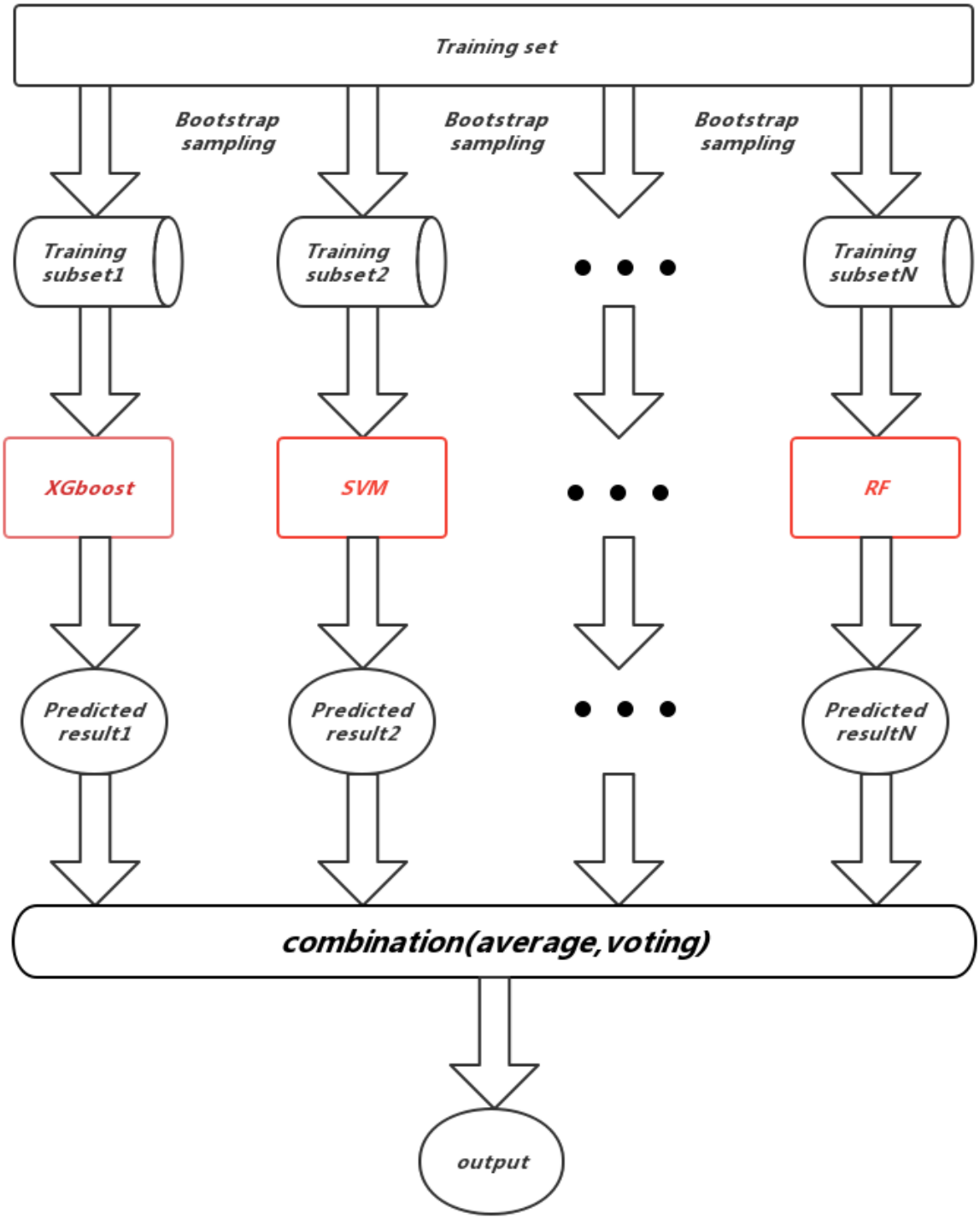

Figure 2

Ensemble learning model based on strong classifier

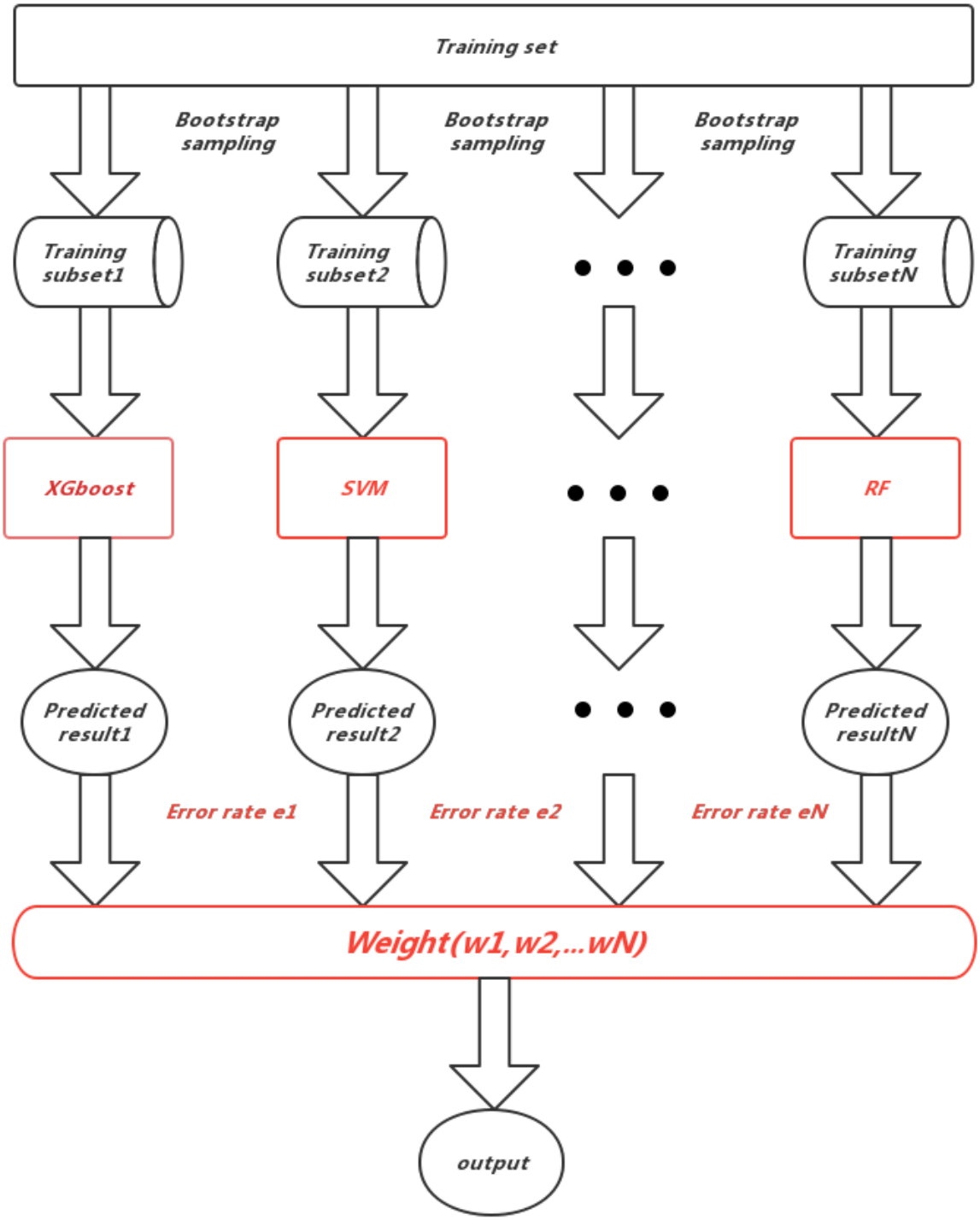

Figure 3

Weighted-Ensemble model

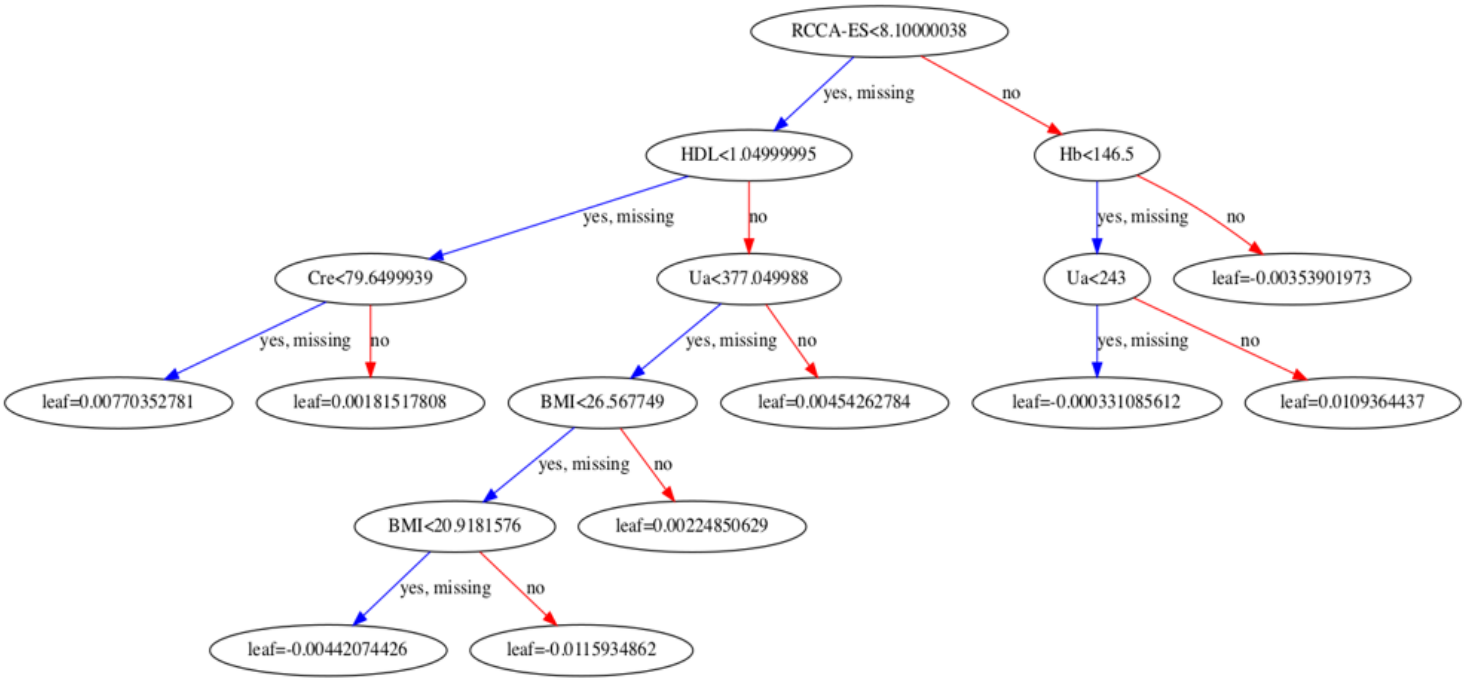

Figure 4

XGboost

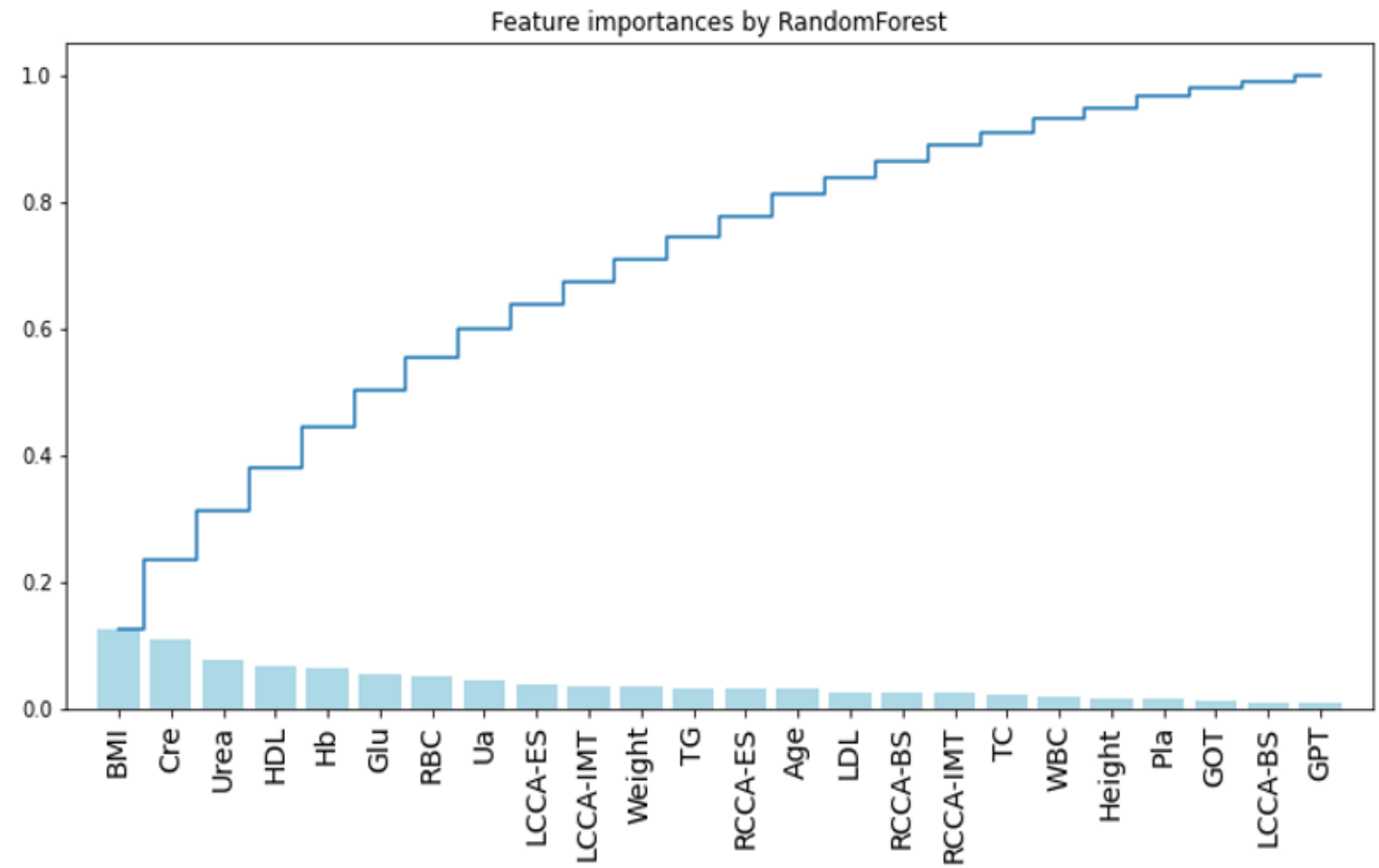

Figure 5

Features selection using RF

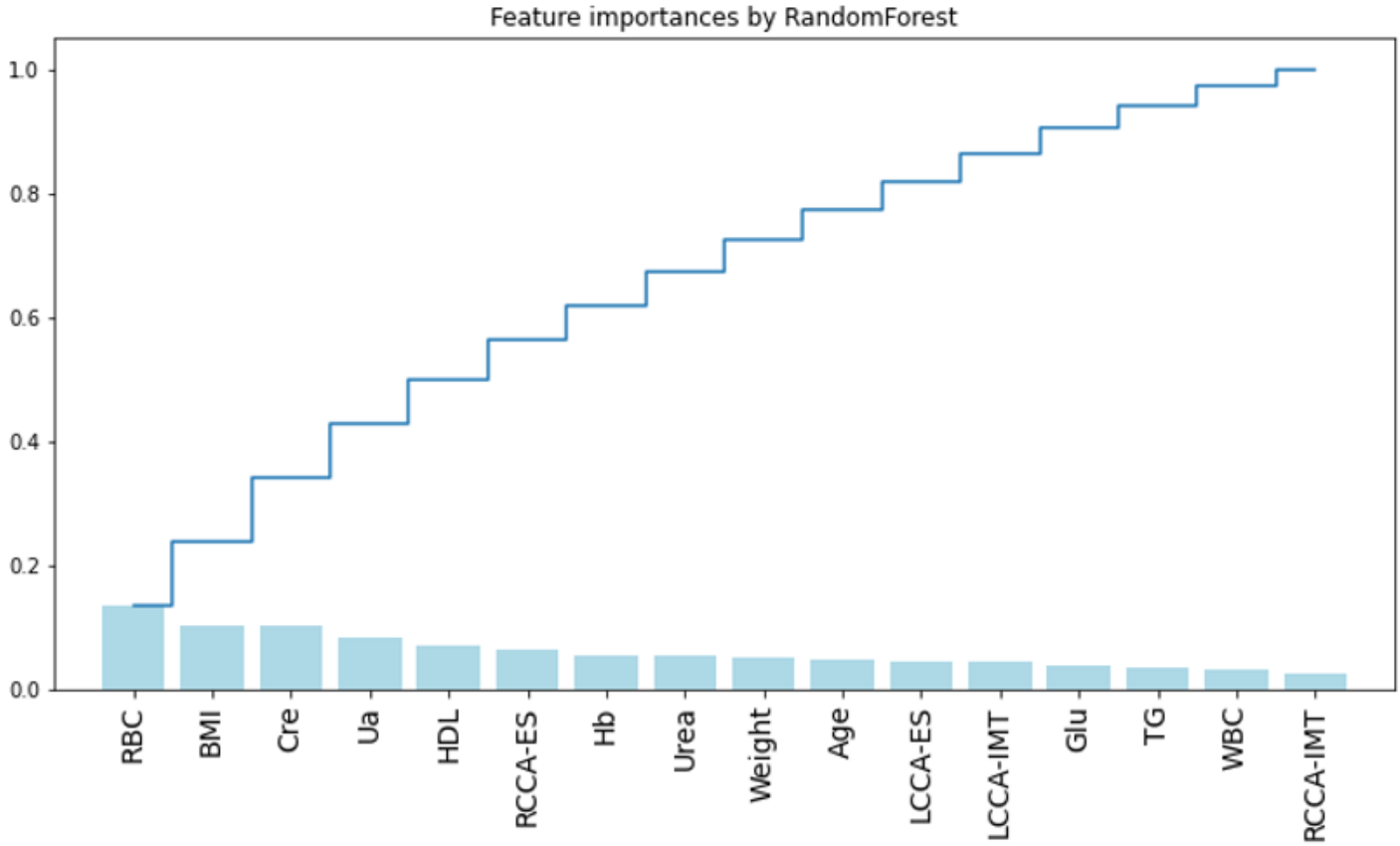

Figure 6

Features selection using ensemble feature selection algorithms