Mathematical Biosciences and Engineering

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

MBE, 20(3): 4574–4591.
DOI: 10.3934/mbe.2023212
Received: 14 October 2022
Revised: 27 November 2022
Accepted: 04 December 2022
Published: 27 December 2022

Research article

Prediction of coronary heart disease in gout patients using machine learning models

Lili Jiang[1,†], Sirong Chen[2,†], Yuanhui Wu[1], Da Zhou[3,*] and Lihua Duan[1,*]

[1] Department of Rheumatology and Clinical Immunology, Jiangxi Provincial People's Hospital, The First Affiliated Hospital of Nanchang Medical College, Nanchang, China

[2] School of Mathematical Sciences, Soochow University, Suzhou, China

[3] School of Mathematical Sciences, Xiamen University, Xiamen, China

† The authors contributed equally to this work.

* **Correspondence:** Email: zhouda@xmu.edu.cn, lh-duan@163.com.

Abstract: Growing evidence shows that there is an increased risk of cardiovascular diseases among gout patients, especially coronary heart disease (CHD). Screening for CHD in gout patients based on simple clinical factors is still challenging. Here we aim to build a diagnostic model based on machine learning so as to avoid missed diagnoses or over exaggerated examinations as much as possible. Over 300 patient samples collected from Jiangxi Provincial People's Hospital were divided into two groups (gout and gout+CHD). The prediction of CHD in gout patients has thus been modeled as a binary classification problem. A total of eight clinical indicators were selected as features for machine learning classifiers. A combined sampling technique was used to overcome the imbalanced problem in the training dataset. Eight machine learning models were used including logistic regression, decision tree, ensemble learning models (random forest, XGBoost, LightGBM, GBDT), support vector machine (SVM) and neural networks. Our results showed that stepwise logistic regression and SVM achieved more excellent AUC values, while the random forest and XGBoost models achieved more excellent performances in terms of recall and accuracy. Furthermore, several high-risk factors were found to be effective indices in predicting CHD in gout patients, which provide insights into the clinical diagnosis.

Keywords: gout; CHD; machine learning; diagnostic model; imbalance data; risk factor selection

1. Introduction

Gout is the leading cause of inflammatory arthritis, affecting 3.9% of adults in the USA [1] and 1–3% of adults in China [2], and the worldwide prevalence has shown an increasing trend year by year. Typically, it presents as acute peripheral inflammatory arthritis and is associated with a range of comorbidities such as hypertension, diabetes mellitus, obesity, metabolic syndrome and cardiovascular (CV) disease [3]. There is growing evidence that patients with gout have an increased risk of CV disease [4, 5], especially CHD. Presumably, this is partially attributed to the presence of systemic inflammation and oxidative stress in both gout and CHD. Besides, certain uric acid-lowering drugs (such as febuxostat) and anti-inflammatory drugs (such as non-steroidal anti-inflammatory drugs) may also increase the risk of CV disease [6].

Thus, it is necessary to screen for CV disease, especially CHD, in gout patients. Serum markers, electrocardiography (ECG), echocardiography, coronary computed tomography angiography (CTA), as well as coronary angiography, are commonly used methods, which remain less than satisfactory. Serum markers and ECG signals are always normal in asymptomatic patients and echocardiography cannot be used to assess coronary artery lesions, while both CTA and coronary angiography require relatively high radiation exposure and may have a negative impact on kidneys. Hence, it is of great significance to build a diagnostic model based on readily available clinical data, so as to avoid missed diagnoses or over exaggerated examinations as much as possible.

Machine learning, as a series of algorithm models trained from high-dimensional data to make predictions or classifications, has the potential to uncover high-risk clinical factors and automatically screen for disease by performing efficient data analysis [7, 8]. In previous clinical studies, machine learning has been used to identify gout flares from electronic clinical notes [9], and classifying leukemia and gout patients from their uric acid signatures [10]. Besides, machine learning has also been used to predict CHD in general populations [11–13]. However, few studies have been reported on predicting CHD in gout patients by using machine learning models. The aim of this study was to build a reliable diagnostic model for screening CHD in gout patients based on simple clinical factors, making use of statistical analysis and machine learning algorithms. A total of 38 clinical factors were collected from raw data; after the variable selection procedure, only eight clinical factors were used as the final feature inputs for eight machine learning models (logistic regression, decision tree, random forest, extreme gradient boosting (XGBoost), artificial neural networks, light gradient boosting machine (LightGBM), gradient boosting decision tree (GBDT) and support vector machine (SVM) models). Our results showed that stepwise logistic regression and SVM achieved more excellent area under the receiver operating characteristic (ROC) curve (AUC) values, while the random forest and XGBoost models achieved more excellent performances in terms of recall and accuracy. Furthermore, interpretable machine learning analysis revealed high-risk factors to facilitate the prediction of CHD in gout patients.

The rest of this paper is organized as follows. In Section 2, the clinical data used in this study are briefly described, and then data pre-processing and descriptive statistical analysis are presented, after which eight machine learning models and evaluation metrics used for classification are introduced. The main results are shown in Section 3, including model performances and risk factor selection. Conclusions and discussions are included in Section 4.

2. Materials and methods

2.1. Data sources

Clinical data of adult gout patients admitted to Jiangxi Provincial People's Hospital from January 1, 2016 to November 1, 2020 were retrospectively screened, and patients who were diagnosed with gout or gout with CHD were enrolled. The diagnosis of gout was performed according to the 2015 Gout Classification Criteria [14], and the diagnosis of CHD included at least one of the following: myocardial infarction, unstable angina, and coronary revascularization. Patients with liver and renal insufficiency for reasons other than gout, secondary hyperuricemia, secondary hypertension or a history of carcinoma were excluded. This study was approved by the ethics committee of Jiangxi Provincial People's Hospital (2022-048).

After the screening, demographic characteristics, clinical characteristics and laboratory parameters were collected from medical records on admission and during hospitalization. Since the aim of this study was to build a reliable diagnostic model to assist clinical doctors in screening CHD in gout patients, the principle of variable selection was based on variables that are non-invasive, cost-effective, stable and easily available, as this would allow the model to be widely used in hospitals at all levels. Variables such as blood pressure on admission, lipid levels and blood glucose were considered unstable or susceptible to drug therapy, while other drug therapies such as antihypertensive drugs, hypoglycemic agents and antiplatelet drugs were considered to have high heterogeneity. In addition, levels of inflammatory factors, such as interleukin-1 beta, interleukin-6 and tumor necrosis factor-α were also considered unstable and not easily accessible in some primary hospitals. Based on the above considerations, only eight clinical factors were selected as the final feature inputs for machine learning models, including age, gender, body mass index (BMI), smoking history, creatinine, hypertension, diuretic use, and nonsteroidal anti-inflammatory drugs (NSAIDs) use.

2.2. Data pre-processing

Among all of the eight features, gender, hypertension, diuretic use and NSAIDs use were qualitative (categorical) variables that needed to be coded, which was done as follows: gender was coded as 1 for males and 0 for females; smoking history was coded as 1 for smokers and 0 for non-smokers; hypertension was coded as 1 for hypertensive patients and 0 for non-hypertensive patients; diuretic use was coded as 1 for diuretic and 0 for no diuretic; NSAIDs use was coded as 1 for NSAIDs and 0 for no NSAIDs. Moreover, diagnosis is the dependent variable (1 for gout complicated with CHD patients, 0 for gout patients without CHD).

We then dealt with outliers and the imbalanced problem in the dataset. In statistics, an outlier is a data point that has an abnormal distance from other samples. The distance between the upper and lower quartiles is known as the interquartile range (IQR). Starting above the upper quartile (Q3), a distance of 1.5 times the IQR was measured. Similarly, a distance of 1.5 times the IQR was measured below the lower quartile (Q1). In this way, all of the observations outside of the 1.5 times IQR range were considered to be outliers [15]. We searched all continuous variables for outliers, checked medical records and corrected data for recording errors. Finally, there were 326 patient samples left in the dataset, including 64 gout patients and 262 gout+CHD patients. In this way, an imbalance problem arose, i.e., the number of gout patients (minority class) was much less than that of gout+CHD patients (majority class).

We divided the samples into the training set and the testing set at a ratio of 75% : 25%. Due to the imbalanced problem that the two classes were not represented equally [16], both over- and under-sampling techniques were used in the training data to overcome this issue. Over-sampling is the method of random sampling with the replacement for the minority class until its sample size meets the requirement. Under-sampling randomly deletes samples from the majority class until its sample size meets the requirement. The combined sampling method resolves the imbalance problem by combining these two methods. Noticing that there were 41 gout patients and 203 gout+CHD patients in the original training set, after a combined sampling process, there were 123 gout patients and 121 gout+CHD patients in the updated training set. It should be noted that we did not use the well-known SMOTE method [17] to deal with the imbalanced problem. This is because there were four categorical variables among all eight features, so it was less suitable to use the interpolation method in our case. Besides, our experimental result shows that the combined sampling technique significantly improved the training quality (see Figure 1).

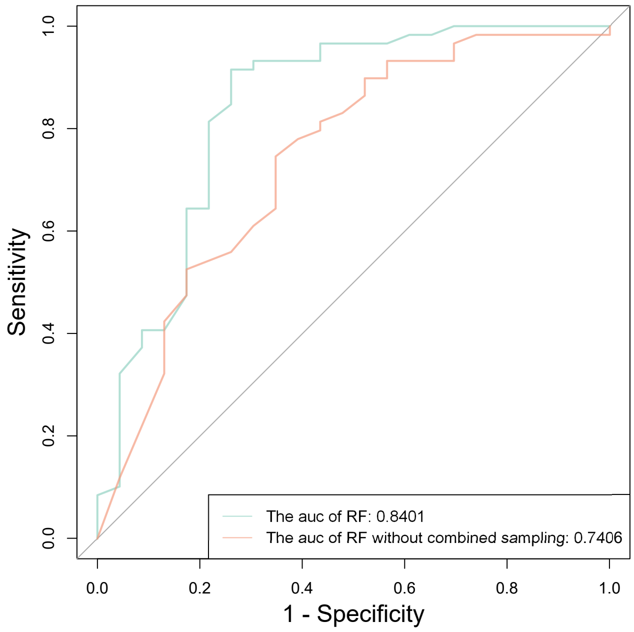


Figure 1. Comparison of random forests (RFs) with and without combined sampling.

2.3. *Descriptive statistics*

Table 1 demonstrates the mean values of all eight variables of two different groups. We used the [1]Mann-Whitney U-test to determine the statistical significance of continuous features between two groups and the [2]Pearson's chi-square test to determine the associations between the discrete features in two groups of patients. It turns out that, age, gender, smoking history, hypertension, diuretic use and NSAIDs use were significantly different between the two groups. Besides, The gout complicated with CHD patients were on average older and had a higher proportion of males, smokers, hypertensive patients, diuretic users and NSAIDs users.

Table 1. Descriptive statistics of all features.

Features	Gout (n = 64)	Gout + CHD (n = 262)	*p*-value
Age (year)	57.7813 ± 13.0000	67.2710 ± 11.1272	< 0.001[1]
Gender (%)	92.1875	93.8931	< 0.001[2]
BMI (kg/m^2)	24.1558 ± 3.7094	24.7940 ± 3.4487	0.1857[1]
Smoking history (%)	29.6875	45.8015	< 0.001[2]
Creatinine (mmol/L)	133.7812 ± 118.7193	108.5687 ± 44.8845	0.7403[1]
Hypertension (%)	37.5000	44.6565	< 0.001[2]
Diuretics use (%)	0.0000	2.6718	< 0.001[2]
NSAIDs use (%)	6.2500	27.0992	< 0.001[2]

2.4. *Classification models*

2.4.1. Logistic regression and stepwise regression

Logistic regression is a traditional statistical machine learning method for classification problems [18]. It uses the linear regression method to estimate the probability of an event occurring. In logistic regression, a logit transformation is used on the odds. Specifically, the expression for the logistic regression is given as follows:

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}, \tag{2.1}$$

where x_i represents the value of the i-th independent variable and p represents the conditional probability of a positive class outcome occurring in the presence of m independent variables.

Stepwise regression is a statistical method of finding an independent variable subset that has a significant effect on the dependent variable through step-by-step iterative selection. There are three types of stepwise regression. 1) Forward stepwise regression gradually adds independent variables while starting with no variable in the model. It tests the addition of each variable, adds the variable that improves the quality of fit the most and repeats this process until no variable significantly improves the model. 2) Backward stepwise regression gradually removes independent variables starting with all variables, in order to find a reduced model that explains the data the best. It is a reverse process to forward stepwise regression [19]. 3) A combination of forward and backward stepwise regressions is frequently used in practice. In this work, we used a combination of these two methods.

2.4.2. Decision tree

Decision trees constitute one of the most popular tools for classification [20]. The logic of decision trees is transparent and interpretable. Decision trees provide a flowchart-like structure that is logically in the form of a tree by classifying feature judgments at each node. The nodes of a decision tree are divided into root nodes, internal nodes and leaf nodes. The root node contains all samples; each internal node contains the samples that satisfy the condition from the root node to the current internal node; each leaf node represents a classification result or decision. There are three main classification criteria used in the decision tree, including the information gain, information gain rate and Gini index. In this study, we used the Gini index to train the model and the corresponding decision tree model is called

CART. There is a complexity parameter in decision trees which determines the process of pruning. We performed five-fold cross-validation to optimize the complexity parameter.

2.4.3. Random forest

Random forest is a popular ensemble learning method [21]. It is a bagging algorithm that uses a decision tree as a base classifier, and it has been used in both classification and regression problems. Unlike the original bagging algorithm, random forests use a modified feature-bagging strategy, that is, the tree learning algorithm selects a random subset of all of the candidate features in order to promote the diversity of base classifiers. Compared to decision trees, random forests normally show better predictive performance. However, random forests are not as interpretable as decision trees. Even so, the random forest provides feature importance measurements such as mean decrease accuracy and mean decrease Gini [22]. The random forest algorithm involves the selection of hyper parameters such as the number of trees to grow and the number of variables randomly sampled as candidates at each split. Five-fold cross-validation was used to achieve optimal predictive performance.

2.4.4. Gradient boosting

Gradient boosting is a powerful ensemble learning technique that is widely used in practice. In this work, we used three types of gradient-boosting-based models: XGBoost, GBDT and LightGBM.

XGBoost provides a regularizing gradient boosting framework [23]. It learns a new function by adding a tree to fit the residuals of the previous prediction. The objective function of XGBoost is given as follows [24]:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.2}$$

where l represents the loss function and Ω represents the complexity of the tree. Adding the complexity of K trees as a regular term to the objective function prevents the overfitting of the model. XGBoost has many hyper parameters that can be adjusted such as the maximum depth of the tree, the step size of each boosting step and the maximum number of iterations. We optimized the maximum depth of the tree and the maximum number of iterations using by five-fold cross-validation.

GBDT is a specific gradient-boosting method using regression decision trees as base learners [25]. A GBDT can be regarded as an additive model composed of K decision trees as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in F, \tag{2.3}$$

where f_k represents a regression decision tree and $\hat{y}_i$ represents the prediction result of x_i. The learning rate and the total number of trees to fit were adjusted to achieve good performances in our model. Five-fold cross-validation was used for parameter optimization.

LightGBM is a distributed gradient boosting framework that uses tree-based learning algorithms [26]. It provides a nonparametric method for implementing the GBDT algorithm. The LightGBM improves on the traditional GBDT by using various computational techniques such as a histogram algorithm, gradient-based one-side sampling, exclusive feature bundling and a leaf-wise algorithm with depth limitation, which make the LightGBM much faster than the conventional GBDT algorithm. We used five-fold cross-validation to optimize the maximum number of leaves in a tree.

2.4.5. Neural networks

Artificial neural networks are learning algorithms that mimic the structure and function of biological neural networks for computation [27] and have become one of the most popular methods in machine learning. There are many different types of neural network models. Here, we used a fully connected feedforward neural network model which contains an input layer, some hidden layers and an output layer. The number of hidden layers and the number of neurons in each layer are the most important hyper-parameters which essentially determine the complexity of the model. Due to the sample size and the dimension of our data, the number of hidden layers was set to one. We performed five-fold cross-validation to optimize the number of neurons in the hidden layer. The optimization algorithm we used for training the networks is the resilient backpropagation (Rprop) with and without weight backtracking [28].

2.4.6. SVM

SVMs map training data points into space so as to maximize the separation or margin between the two classes. In other words, SVMs aim to solve for the maximum-margin hyperplane. SVMs can be divided into three types: a linear SVM with a hard margin, linear SVM with a soft margin [29] and non-linear SVM with kernel tricks [30], as follows:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i K\left(x, x_i\right) + b^*\right), \tag{2.4}$$

where K represents kernel function, (x_i, y_i) represents a training sample and α_i^* and b^* are parameters to be learned from training data. We solve the binary classification problem by using a non-linear SVM with the following Gaussian radial basis kernel function:

$$K(x, y) = \exp\left(-\gamma\|x - y\|^2\right). \tag{2.5}$$

In Eq (2.5), $\gamma > 0$ is the hyperparameter that determines the smoothness of the decision boundary and the model variance. In addition, the parameter C indicates the cost of constraint violation. We performed a five-fold cross-validation to optimize these parameters.

2.5. Evaluation metrics

In order to evaluate the effectiveness of the above machine learning models, we used the following evaluation metrics: ROC, AUC, accuracy, and recall, which are common metrics in binary classification problems [31–33].

For the binary classification problem, we have the following:

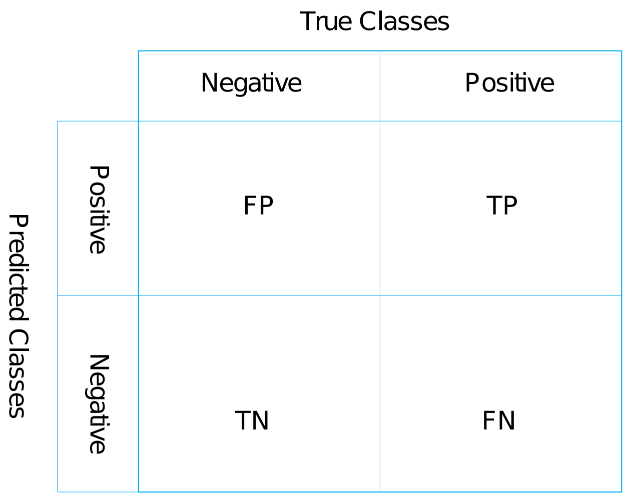
TP (True Positive): the number of samples that are actually positive and predicted to be positive.

FP (False Positive): the number of samples actually negative and predicted to be positive.

TN (True Negative): the number of samples that are actually negative and predicted to be negative.

FN (False Negative): the number of samples that are actually positive and predicted to be negative.

To show the prediction results for the dichotomous problem, we present the above indicators as a confusion matrix in Figure 2:



		True Classes	
		Negative	Positive
Predicted Classes	Positive	FP	TP
	Negative	TN	FN

Figure 2. Form of the confusion matrix.

In addition, we have:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.6}$$

$$recall = \frac{TP}{TP + FN}, \tag{2.7}$$

$$TPR = \frac{TP}{TP + FN}, \tag{2.8}$$

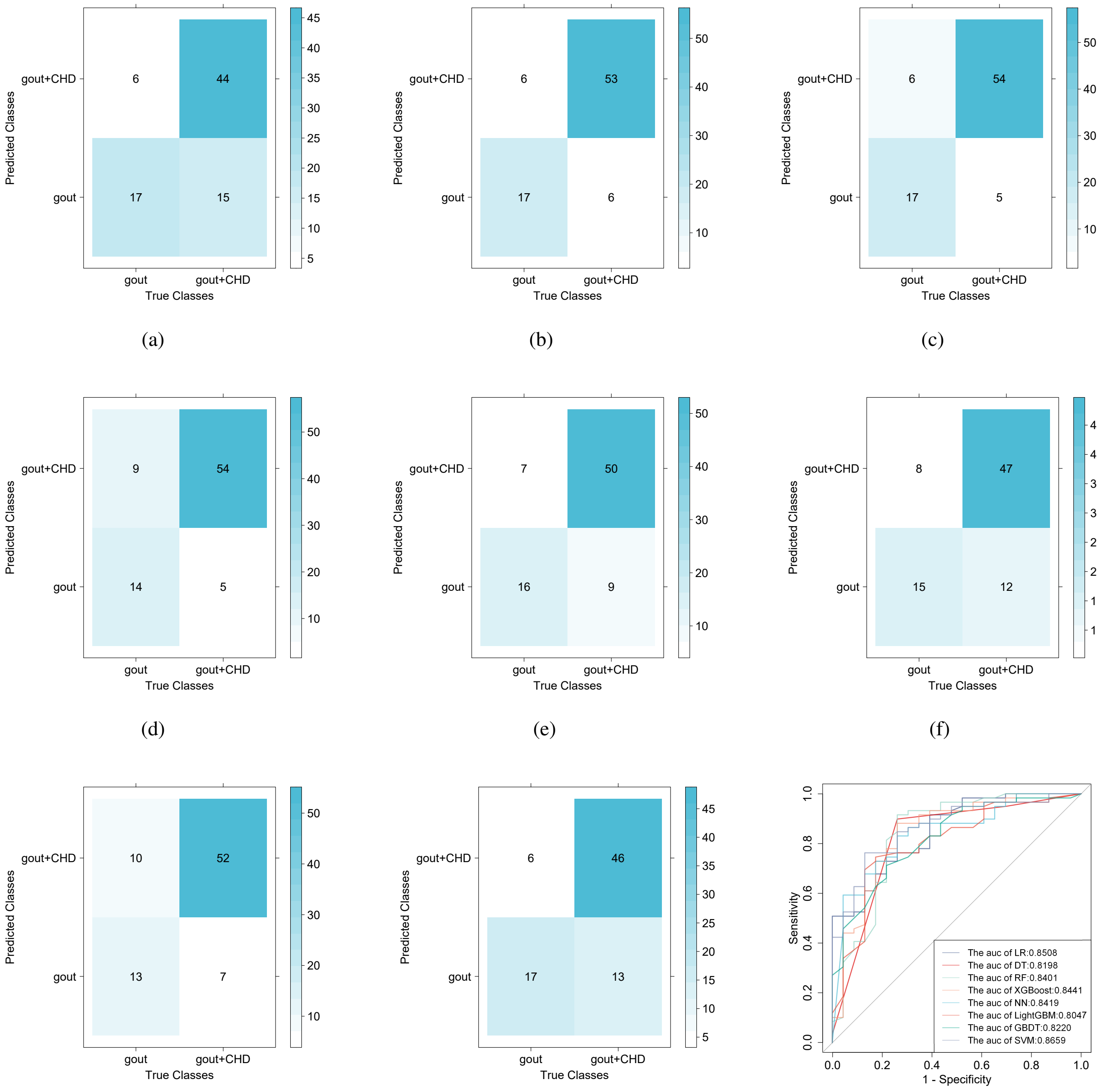
$$FPR = \frac{FP}{TN + FP}, \tag{2.9}$$

where accuracy denotes the probability of correct predictions and recall denotes the probability that a sample that is actually a positive sample is correctly predicted. In a binary classification problem, both recall and accuracy are within [0,1]. The ROC is a curve with FPR on the x axis and TPR on the y axis, and the area under the curve is called the AUC. Both the ROC curve and AUC reflect the reliability of the model. They are widely used in medical research. A higher AUC indicates more excellent model performance. By the rule of thumb, AUC values above 0.8 are considered excellent [34].

3. Results

3.1. Model performances

There were 82 samples in the testing set, including 59 gout+CHD patient samples and 23 gout patient samples. The predictions from the stepwise logistic regression, decision tree, random forest, XGBoost, neural network, LightGBM, GBDT and SVM models are shown in Figure 3. The confusion matrices and ROC curves of the eight machine learning models are shown. Here we designated the gout+CHD group as the positive class of the binary classification problem.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

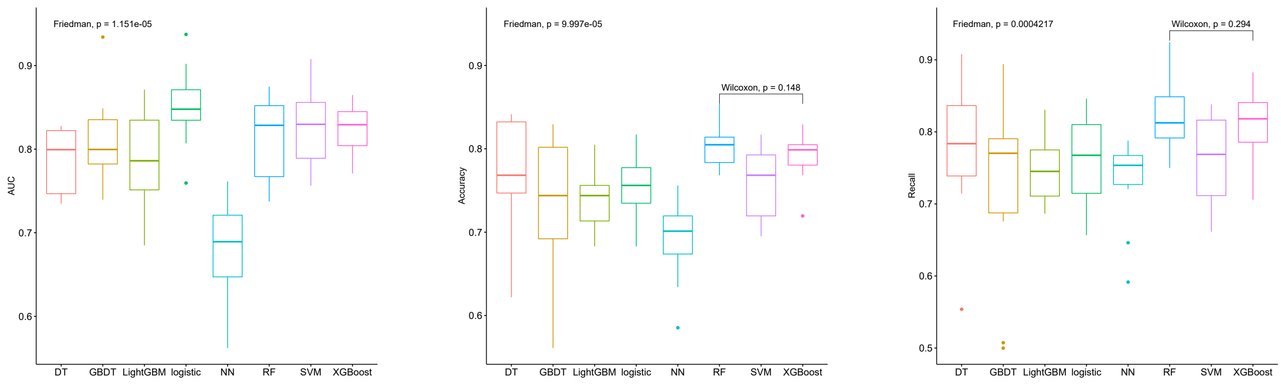
Figure 3. Confusion matrices and ROC curves of the eight models. (a) Logistic regression with the stepwise method. (b) Decision tree: the complexity parameter is 0.046. (c) Random forest: the number of trees to grow is 65, and the number of variables randomly sampled as candidates at each split is 7. (d) XGBoost: the maximum depth of the tree is 50, the step size of each boosting step is 0.25 and the max number of iterations is 17. (e) Artificial neural networks: the number of hidden layers is 1, and the number of neurons in the hidden layer is 40. (f) LightGBM: the maximum number of leaves in one tree is 7, the learning rate is 1.0, and the number of training rounds is 2. (g) GBDT: the learning rate is 0.008, and the total number of trees to fit is 60. (h) SVM: the cost of constraints violation is 1, and γ is 0.1. (i) ROC curves of the eight models.

Figure 3(i) shows the ROC curves and AUC values of different models. The AUC values for the stepwise logistic regression, decision tree, random forest, XGBoost, neural network, LightGBM, GBDT and SVM models are 0.8508, 0.8198, 0.8401, 0.8441, 0.8419, 0.8047, 0.8220 and 0.8659, respectively. Their AUC values all exceed 0.8. The SVM and stepwise logistic regression models achieved more excellent performances (0.8659 and 0.8508) among all eight models.

Table 2 compares the eight models in terms of the AUC, recall and accuracy metrics. Multiple experiments were performed by randomly splitting the dataset into the training set and testing set. Both the average value and the standard deviation are shown in the table. Among all of the models, stepwise logistic regression achieved the highest average value of AUC (0.8505), which was followed by the SVM (0.8278); the random forest had the highest accuracy (0.8049), which was followed by XGBoost (0.7902); random forest had the highest recall (0.8261), which was also followed by XGBoost (0.8124). Figure 4 demonstrates the significant test results for comparing the eight machine learning models. The Friedman test showed that there were significant differences between the eight models in terms of the AUC, accuracy and recall. Furthermore, the Wilcoxon test showed that the random forest and XGBoost are comparable in terms of accuracy and recall, i.e., there was no statistically significant difference.

Table 2. Comparison of eight machine learning models.

Model	AUC	Accuracy	Recall
LR	0.8505 ± 0.0490	0.7573 ± 0.0392	0.7593 ± 0.0637
DT	0.7867 ± 0.0387	0.7707 ± 0.0685	0.7774 ± 0.1001
RF	0.8113 ± 0.0525	0.8049 ± 0.0270	0.8261 ± 0.0556
XGBoost	0.8242 ± 0.0333	0.7902 ± 0.0309	0.8124 ± 0.0505
NN	0.6780 ± 0.0605	0.6915 ± 0.0514	0.7314 ± 0.0636
LightGBM	0.7890 ± 0.0587	0.7402 ± 0.0377	0.7493 ± 0.0505
GBDT	0.8119 ± 0.0541	0.7256 ± 0.0929	0.7288 ± 0.1340
SVM	0.8278 ± 0.0492	0.7573 ± 0.0465	0.7603 ± 0.0664



(a) (b) (c)

Figure 4. Statistical significance test results for model comparison.

Recall that we used a combined sampling method to solve the imbalanced problem. Figure 1 illustrates the effectiveness of the combined sampling method in the random forest model. Before using the combined sampling, the AUC value of the random forest model in the testing set is 0.7406, whereas, after combined sampling, the AUC value significantly increased to 0.8401.

3.2. Risk factor analysis

The above experimental results have shown that the predictive models based on machine learning achieved excellent performances. Furthermore, an even more important question in clinical practice is to find high-risk factors associated with CHD in gout patients.

We first checked the risk factors by using logistic regression (see Table 3) and stepwise logistic regression (see Table 4). It is shown that age, smoking history, creatinine, NSAIDs use were considered to have significant effects on gout+CHD prediction. Hypertension could also have an impact. In other words, compared to the gout group, patients in the gout+CHD group were generally older, more likely to smoke and use NSAIDs and tended to have hypertension and lower creatinine levels.

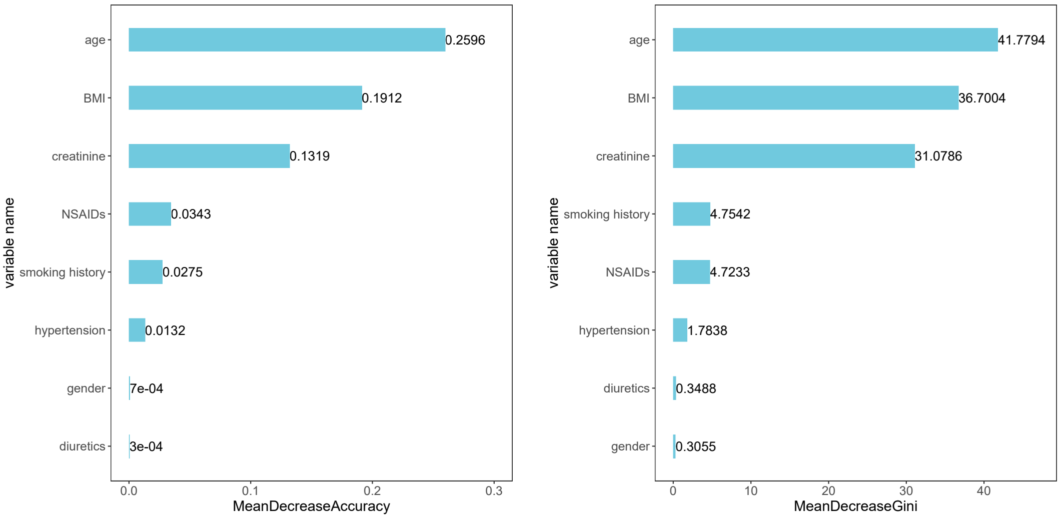
Table 3. Prediction of gout+CHD using logistic regression.

variables	Coefficient	Standard Error	p-value	Odds Ratio
Age	0.0730	0.0142	< 0.0001	1.0757
Smoking history	0.7564	0.3112	0.0151	2.1305
Hypertension	0.4545	0.3229	0.1592	1.5755
Creatinine	−0.0071	0.0036	0.0497	0.9929
NSAIDs use	1.2583	0.4168	0.0025	3.5193
Gender	0.1217	0.6910	0.8602	1.1294
BMI	0.0614	0.0450	0.1723	1.0633
Diuretics use	13.4063	1029.1216	0.9896	664164.4427

Table 4. Prediction of gout+CHD using stepwise logistic regression.

variables	Coefficient	Standard Error	p-value	Odds Ratio
Age	0.0653	0.0125	< 0.0001	1.0674
Smoking history	0.7329	0.3018	0.0152	2.0810
Hypertension	0.5948	0.3041	0.0504	1.8127
Creatinine	−0.0071	0.0035	0.0463	0.9930
NSAIDs use	1.2532	0.4076	0.0021	3.5017

We then used the feature importance of the random forest to rank different clinical factors (see Figure 5). Two types of feature importance measurements were adopted: mean decrease accuracy and mean decrease Gini, where features were presented from descending importance. In terms of the mean decrease accuracy, the eight features were ranked as follows: age, BMI, creatinine, NSAIDs, smoking history, hypertension, gender, and diuretics. In terms of the mean decrease Gini values, they were ranked as follows: age, BMI, creatinine, smoking history, NSAIDs, hypertension, diuretics and gender. Combining these two results, we found that age, BMI and creatinine were considered to be the most important risk factors in predicting CHD in gout patients.



(a)

(b)

Figure 5. Feature importance plot. Features with higher values are considered to be more important in predicting CHD in gout patients. The results by mean decrease accuracy and mean decrease Gini are presented in (a) and (b) respectively.

It should be noted that the BMI was regarded as the second important feature in the feature importance plot (Figure 5), whereas the BMI was not considered as the significant feature in stepwise logistic regression (Table 4). To explain this inconsistency, we found that BMI is significantly correlated with age (see Figure 6). Therefore, only one of age and BMI could be selected in the process of stepwise logistic regression; otherwise, a collinearity problem would arise. However, feature importance methods do not take variable correlation into consideration, so both age and BMI were thus selected as important variables.

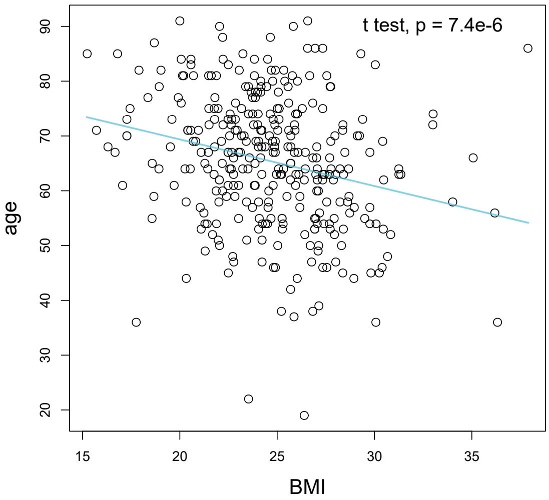


Figure 6. Correlation between age and BMI.

4. Conclusions and discussion

The association between gout and CHD has been attracting considerable attention [35] in recent years. In this study, we have established machine learning models based on simple clinical factors for predicting CHD in gout patients. The prediction task has been treated as a binary classification problem. There are three major steps in the modeling process. The first is selecting reliable clinical features. Even though there were a total of 38 candidates in the original dataset, only eight features were chosen. The principle of variable selection was based on the variables being non-invasive, cost-effective, stable and easily available so that the factors can be widely used in hospitals at all levels. The second is solving the imbalanced problem when the sample sizes of two groups of patients are quite different. A combined sampling method was used to balance the training set. The third is training the machine learning models. Extensive experimental results have shown that the stepwise logistic regression and SVM models achieved more excellent AUC values, while the random forest and XGBoost models achieved more excellent performances in terms of recall and accuracy. Besides, we have analyzed risk clinical factors by using interpretable machine learning methods such as logistic regression and feature importance plots.

From the medical point of view, CHD is a complex, multifactorial disease that involves several etiopathogenic mechanisms. To date, more than 200 risk factors of CHD have been reported, among which age [36], hypertension [37], hyperlipidemia [38], hyperglycemia [39], smoking [40], and obesity [41] are well-determined. However, how can one identify high-risk groups of CHD in the specific group of people with gout? Clinicians are still at a loss, for it is not practical to screen every gout patient with those risk factors for CHD. Our aim is to build a practical mathematical model that uses easily accessible variables to help clinicians identify CHD high-risk groups in gout patients, so as to explore the potential of machine learning in clinical problems. In other words, we are more concerned about the weight of existing risk factors in patients with gout and optimizing our recognition effect by establishing mathematical models with these existing risk factors.

The present study suggested that older age, a higher BMI and smoking history remained to be the main risk factors for CHD in gout patients. Hypertension turned out to be a less important risk factor for CHD according to the feature importance of the random forest, and the difference barely met our significance level in stepwise logistic regression; it seemed that hypertension is not a good predictor of CHD in gout patients in the present study. NSAIDs are widely used to relieve acute gout attacks and prevent recurrent gout flare-ups. However, the use of NSAIDs is widely recognized as a risk factor for CHD. Due to the high heterogeneity in the composition, dose and duration, it was impossible for us to perform subgroup analysis. Hence, "NSAIDs" was included as a dichotomous variable in further analysis. Consistent with previous studies [42], the present study shows that the use of NSAIDs is associated with an increased risk of CHD; thus, NSAIDs should be prescribed with caution when a patient has other risk factors of CHD. Renal insufficiency has long been considered an important risk factor for CV events in patients with CHD [43, 44]; however, we drew the opposite conclusion that lower creatinine is associated with an increased risk of CHD. We suspect the reason to be that the estimated glomerular filtration rate, rather than serum creatinine, was used in previous studies to evaluate kidney function, where the former was calculated by using the Modification of Diet in Renal Disease (MDRD) formula [45] or Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula [46] based on the value of serum creatinine and adjusted by according to patient age and sex.

Therefore, this particular finding should be interpreted with caution.

There are insufficiencies in this study, of which the leading limitation is that the total number of enrolled subjects is relatively small, which may have increased the risk of sampling errors. Future studies with larger sample sizes are needed to further test this model.

In conclusion, we built a diagnostic model for screening CHD in gout patients based on simple clinical factors in this study and found that age, BMI, smoking history and NSAIDs use showed significant effect on predicting CHD, which indicate that gout patients with these risk factors need further cardiac examinations to reveal potential CHD. In the future, it is worthwhile to improve the predictive ability of the diagnostic model by collecting more clinical data and searching for more stable, reliable, and easily available features.

Acknowledgments

We thank the reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (81960296, 81871286, and 11971405), Jiangxi Provincial Natural Science Foundation of China (20192ACB21006), Interdisciplinary Innovation Team, Frontier Science Key Research Project of Jiangxi Provincial People's Hospital (19-008), Long-term (Youth) Project for Leading Innovative Talents in Jiangxi Province (Lihua Duan), Jiangxi Provincial Clinical Research Center for Rheumatic and Immunologic Diseases (20192BCD42005), Jiangxi Province Medical Leading Discipline Construction Project (Rheumatology) and Provincial and Municipal Joint Construction Projects of Medical Disciplines in Jiangxi Province (Rheumatology).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. J. D. Fitzgerald, N. Dalbeth, T. Mikuls, R. Brignardello-Petersen, G. Guyatt, A. M. Abeles, et al., 2020 American College of Rheumatology guideline for the management of gout, *Arthritis Care Res.*, **72** (2020), 744–760. https://doi.org/10.1002/acr.24180

2. R. Liu, C. Han, D. Wu, X. Xia, J. Gu, H. Guan, et al., Prevalence of hyperuricemia and gout in mainland China from 2000 to 2014: A systematic review and meta-analysis, *Biomed Res. Int.*, **2015** (2015), 762820. https://doi.org/10.1155/2015/762820

3. Y. Zhu, B. J. Pandya, H. K. Choi, Comorbidities of gout and hyperuricemia in the US general population: NHANES 2007–2008, *Am. J. Med.*, **125** (2012), 679–687. https://doi.org/10.1016/j.amjmed.2011.09.033

4. M. A. De Vera, M. M. Rahman, V. Bhole, J. A. Kopec, H. K. Choi, Independent impact of gout on the risk of acute myocardial infarction among elderly women: a population-based study, *Ann. Rheum. Dis.*, **69** (2010), 1162–1164. https://doi.org/10.1136/ard.2009.122770

5. O. O. Seminog, M. J. Goldacre, Gout as a risk factor for myocardial infarction and stroke in England: evidence from record linkage studies, *Rheumatology*, **52** (2013), 2251–2259. https://doi.org/10.1093/rheumatology/ket293

6. W. B. White, K. G. Saag, M. A. Becker, J. S. Borer, P. B. Gorelick, A. Whelton, et al., Cardiovascular safety of febuxostat or allopurinol in patients with gout, *N. Engl. J. Med.*, **378** (2018), 1200–1210. https://doi.org/10.1056/NEJMoa1710895

7. J. Wang, Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques, *Math. Biosci. Eng.*, **19** (2022), 10407–10423. https://doi.org/10.3934/mbe.2022487

8. Z. Chen, M. Yang, Y. Wen, S. Jiang, W. Liu, H. Huang, Prediction of atherosclerosis using machine learning based on operations research, *Math. Biosci. Eng.*, **19** (2022), 4892–4910. https://doi.org/10.3934/mbe.2022229

9. C. Zheng, N. Rashid, Y. L. Wu, R. Koblick, A. T. Lin, G. D. Levy, et al., Using natural language processing and machine learning to identify gout flares from electronic clinical notes, *Arthritis Care Res.*, **66** (2014), 1740–1748. https://doi.org/10.1002/acr.22324

10. G. Bahra, L. Wiese, Parameterizing neural networks for disease classification, *Expert Syst.*, **37** (2019), e12465. https://doi.org/10.1111/exsy.12465

11. J. J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, et al., Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease), *J. Biomed. Inform.*, **97** (2019), 103257. https://doi.org/10.1016/j.jbi.2019.103257

12. K. H. Miao, J. H. Miao, G. J. Miao, Diagnosing coronary heart disease using ensemble machine learning, *Int. J. Adv. Comput. Sci. Appl.*, **7** (2016). https://doi.org/10.14569/ijacsa.2016.071004

13. A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, G. Singh, Prediction of coronary heart disease using machine learning: an experimental analysis, in *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, (2019), 51–56. https://doi.org/10.1145/3342999.3343015

14. T. Neogi, T. L. Jansen, N. Dalbeth, J. Fransen, H. R. Schumacher, D. Berendsen, et al., 2015 gout classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative, *Arthritis Rheumatol.*, **67** (2015), 2557–2568. https://doi.org/10.1002/art.39254

15. F. I. Mowbray, S. M. Fox-Wasylyshyn, M. M. El-Masri, Univariate outliers: a conceptual overview for the nurse researcher, *Can. J. Nurs. Res.*, **51** (2019), 31–37. https://doi.org/10.1177/0844562118786647

16. H. He, E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.*, **21** (2009), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

17. A. Fernandez, S. Garcia, F. Herrera, N. V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, *J. Artif. Int. Res.*, **61** (2018), 863–905. https://doi.org/10.1613/jair.1.11192

18. T. Jiang, J. L. Gradus, A. J. Rosellini, Supervised machine learning: a brief primer, *Behav. Ther.*, **51** (2020), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

19. R. R. Hocking, A Biometrics invited paper. The analysis and selection of variables in linear regression, *Biometrics*, **32** (1976), 1–49. https://doi.org/10.2307/2529336

20. L. Breiman, *Classification and Regression Trees*, 1^{st} edition, Routledge, New York, 1984. https://doi.org/10.1201/9781315139470

21. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324

22. H. Hong, G. Xiaoling, Y. Hua, Variable selection using mean decrease accuracy and mean decrease gini based on random forest, in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, (2016), 219–224. https://doi.org/10.1109/ICSESS.2016.7883053

23. P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, H. Zheng, Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer, *IEEE Trans. Biomed. Eng.*, **68** (2020), 148–160. https://doi.org/10.1109/TBME.2020.2993278

24. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785

25. J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, **29** (2001), 1189–1232. https://doi.org/10.1214/aos/1013203451

26. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al. Lightgbm: A highly efficient gradient boosting decision tree, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 3149–3157.

27. S. Agatonovic-Kustrin, R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, *J. Pharm. Biomed. Anal.*, **22** (2000), 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1

28. M. Riedmiller, Advanced supervised learning in multi-layer perceptrons-From backpropagation to adaptive learning algorithms, *Comput. Stand. Interfaces*, **16** (1994), 265–278. https://doi.org/10.1016/0920-5489(94)90017-5

29. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297. https://doi.org/10.1007/BF00994018

30. B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the fifth annual workshop on Computational learning theory*, (1992), 144–152. https://doi.org/10.1145/130385.130401

31. T. N. K. Hung, N. Q. K. Le, N. H. Le, L. Van Tuan, T. P. Nguyen, C. Thi, et al., An AI-based prediction model for drug-drug interactions in osteoporosis and Paget's diseases from SMILES, *Mol. Inform.*, **41** (2022), e2100264. https://doi.org/10.1002/minf.202100264

32. L. H. T. Lam, N. H. Le, L. Van Tuan, H. T. Ban, T. N. K. Hung, N. T. K. Nguyen, et al., Machine learning model for identifying antioxidant proteins using features calculated from primary sequences, *Biology*, **9** (2020), 325. https://doi.org/10.3390/biology9100325

33. N. Le, Y. Ou, Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins, *BMC Bioinformatics*, **17** (2016), 501. https://doi.org/10.1186/s12859-016-1369-y

34. A. E. Hendricks, S. M. Adlof, C. N. Alonzo, A. B. Fox, T. P. Hogan, Identifying children at risk for developmental language disorder using a brief, whole-classroom screen, *J. Speech Lang. Hear. Res.*, **62** (2019), 896–908. https://doi.org/10.1044/2018_jslhr-l-18-0093

35. K. H. Huang, C. J. Tai, Y. F. Tsai, Y. H. Kuan, C. Y. Lee, Correlation between gout and coronary heart disease in Taiwan: a nationwide population-based cohort study, *Acta Cardiol. Sin.*, **35** (2019), 634–640. https://doi.org/10.6515/ACS.201911_35(6).20190403B

36. M. B. Mittelmark, B. M. Psaty, P. M. Rautaharju, L. P. Fried, N. O. Borhani, R. P. Tracy, et al., Prevalence of cardiovascular diseases among older adults: the cardiovascular health study, *Am. J. Epidemiol.*, **137** (1993), 311–317. https://doi.org/10.1093/oxfordjournals.aje.a116678

37. B. B. Agbor-Etang, J. F. Setaro, Management of hypertension in patients with ischemic heart disease, *Curr. Cardiol. Rep.*, **17** (2015), 119. https://doi.org/10.1007/s11886-015-0662-0

38. D. Hu, J. Li, X. Li, Investigation of blood lipid levels and statin interventions in outpatients with coronary heart disease in China: the China Cholesterol Education Program (CCEP), *Circ. J.*, **72** (2008), 2040–2045. https://doi.org/10.1253/circj.cj-08-0417

39. L. E. Eberly, J. D. Cohen, R. Prineas, L. Yang, Impact of incident diabetes and incident nonfatal cardiovascular disease on 18-year mortality: the multiple risk factor intervention trial experience, *Diabetes Care*, **26** (2003), 848–854. https://doi.org/10.2337/diacare.26.3.848

40. U. Mons, A. Müezzinler, C. Gellert, B. Schöttker, C. C. Abnet, M. Bobak, et al., Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium, *BMJ*, **350** (2015), h1551. https://doi.org/10.1136/bmj.h1551

41. C. M. Hales, M. D. Carroll, C. D. Fryar, C. L. Ogden, Prevalence of obesity among adults and youth: United States, 2015-2016, *NCHS Data Brief*, **288** (2017).

42. I. Atukorala, D. J. Hunter, Valdecoxib: the rise and fall of a COX-2 inhibitor, *Expert Opin. Pharmacother.*, **14** (2013), 1077–1086. https://doi.org/10.1517/14656566.2013.783568

43. M. J. Sarnak, A. S. Levey, A. C. Schoolwerth, J. Coresh, B. Culleton, L. L. Hamm, et al., Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention, *Hypertension*, **42** (2003), 1050–1065. https://doi.org/10.1161/01.HYP.0000102971.85504.7c

44. E. L. Schiffrin, M. L. Lipman, J. F. Mann, Chronic kidney disease: effects on the cardiovascular system, *Circulation*, **116** (2007), 85–97. https://doi.org/10.1161/CIRCULATIONAHA.106.678342

45. A. S. Levey, J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, D. Roth, A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation, *Ann. Intern. Med.*, **130** (1999), 461–470. https://doi.org/10.7326/0003-4819-130-6-199903160-00002

46. A. S. Levey, L. A. Stevens, C. H. Schmid, Y. Zhang, A. F. Castro Iii, H. I. Feldman, et al., A new equation to estimate glomerular filtration rate, *Ann. Intern. Med.*, **150** (2009), 604–612. https://doi.org/10.7326/0003-4819-150-9-200905050-00006