Heliyon 9 (2023) e22727

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Metabolic syndrome prediction model using Bayesian optimization and XGBoost based on traditional Chinese medicine features☆

Jianhua Zheng [a,b], Zihao Zhang [a], Jinhe Wang [c], Ruolin Zhao [a], Shuangyin Liu [a,b], Gaolin Yang [a], Zhengjie Liu [d,e,*], Zhengyuan Deng [a,f]

[a] College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China
[b] Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Guangzhou, 510630, China
[c] Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing, 100091, China
[d] Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, 510120, China
[e] The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, 510120, China
[f] Network and Educational Technology Center, Jinan University, Guangzhou, 510630, China

ARTICLE INFO

Keywords:
XGBoost
Machine learning
Metabolic syndrome
Traditional Chinese medicine (TCM)
Bayesian optimization

ABSTRACT

Metabolic syndrome (MetS) has a high prevalence and is prone to many complications. However, current MetS diagnostic methods require blood tests that are not conducive to self-testing, so a user-friendly and accurate method for predicting MetS is needed to facilitate early detection and treatment. In this study, a MetS prediction model based on a simple, small number of Traditional Chinese Medicine (TCM) clinical indicators and biological indicators combined with machine learning algorithms is investigated. Electronic medical record data from 2040 patients who visited outpatient clinics at Guangdong Chinese medicine hospitals from 2020 to 2021 were used to investigate the fusion of Bayesian optimization (BO) and eXtreme gradient boosting (XGBoost) in order to create a BO-XGBoost model for screening nineteen key features in three categories: individual bio-information, TCM indicators, and TCM habits that influence MetS prediction. Subsequently, the predictive diagnostic model for MetS was developed. The experimental results revealed that the model proposed in this paper achieved values of 93.35 %, 90.67 %, 80.40 %, and 0.920 for the F1, sensitivity, FRS, and AUC metrics, respectively. These values outperformed those of the seven other tested machine learning models. Finally, this study developed an intelligent prediction application for MetS based on the proposed model, which can be utilized by ordinary users to perform self-diagnosis through a web-based questionnaire, thereby accomplishing the objective of early detection and intervention for MetS.

☆ Zhenjie LIU reports administrative support, article publishing charges, equipment, drugs, or supplies, statistical analysis, travel, and writing assistance were provided by National Key R&D Program of China. Jianhua ZHENG reports administrative support, article publishing charges, equipment, drugs, or supplies, statistical analysis, travel, and writing assistance were provided by Research Fund Program of Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization.

* Corresponding author. Guangdong Provincial Hospital of Chinese Medicine, Guangzhou, 510120, China.
E-mail addresses: zhengjianhua@zhku.edu.cn (J. Zheng), zhangzihaoff@163.com (Z. Zhang), 1182767797@qq.com (J. Wang), z7799517@163.com (R. Zhao), shuangyinliu@zhku.edu.cn (S. Liu), yanggaolin@zhku.edu.cn (G. Yang), gzlzjie@gzucm.edu.cn (Z. Liu), dzy2018@jnu.edu.cn (Z. Deng).

https://doi.org/10.1016/j.heliyon.2023.e22727
Received 31 January 2023; Received in revised form 16 November 2023; Accepted 17 November 2023

1. Introduction

Metabolic syndrome (MetS) is a pathological state in which the metabolism of carbohydrates, proteins, fats, and other substances in the human body is disturbed, which manifests clinically as a series of diseases, such as obesity, diabetes, hypertension, and hyperlipidemia. Worldwide, the average prevalence of MetS is 31 % [[1]], and it has become a significant public health problem. Studies have found that MetS is closely related to the occurrence of malignant tumors, polycystic ovary syndrome, cardiovascular diseases, and other diseases and is considered a high-risk factor for many diseases. Therefore, early diagnosis and timely intervention of MetS are essential to reduce the prevalence of MetS and reduce serious complications.

Current diagnostic criteria for MetS depend on physical and chemical indicators [[1,2]] and are not conducive to user self-detection and early identification. Chinese medicine mainly determines the disease through four diagnostic methods: looking, listening, asking, and cutting. Current studies have shown that traditional Chinese medicine (TCM) evidence and physical factors are associated with MetS [[3]]. For example, some literature [[4]] suggested that phlegm dampness and qi deficiency, which are wet evidence of disease in TCM, are risk factors for MetS, which provides a possibility to achieve MetS diagnosis based on TCM characteristics. Pei-Li Chien et al. utilized characteristics from individual physical examination data to establish the association between MetS and TCM constitution, guiding pharmacological and dietary care [[5]]. Their approach primarily relied on statistical methods to uncover the connection between the frequency of specific symptoms and MetS in patients with MetS. However, such results cannot yet be directly used for MetS diagnosis. Due to the rapid advancement of artificial intelligence in recent years, increasing numbers of researchers have combined artificial intelligence techniques with medical treatment and achieved better results; for example, they have used machine learning to predict diabetes [[6,7]] and artificial intelligence to assist doctors in medical imaging judgments [[8,9]]. In studies using machine learning for MetS diagnosis, the prediction methods can be divided into those that do and do not use TCM indicators, depending on the indicators chosen. Prediction methods that do not use TCM indicators use physicochemical and bioinformatic indicators as the selected features to model MetS using machine learning algorithms. Eun Kyung Choe et al. constructed a MetS prediction model in 7502 non-obese Koreans, which included 647 individuals with confirmed MetS. The model was based on a Bayesian classifier that incorporated clinical indicators and genetic information, achieving the highest AUC value of 0.69 [[10]]. Karimi-Alavijeh F et al. analyzed data from 2107 Iranian individuals who did not meet the ATP III criteria for Metabolic Syndrome (MetS) and developed an SVM-based classification model using physicochemical and bioinformatic indicators. The model achieved an accuracy of 75.7 % [[11]]. Guadalupe Obdulia Gutiérrez-Esparza et al. used the ATP III criterion as a framework to construct MetS prediction models using multiple machine learning algorithms after ranking health parameters on a dataset of 2942 participants from Mexico. They found that the Jrip model had the highest accuracy value of 86.91 % [[12]]. Prediction methods without the use of TCM indicators are not easy for the general population to perform on their own using physical and chemical indicators, despite the results achieved in MetS research and complementary diagnostic methods. In contrast, prediction methods that include TCM indicators combine TCM physical and physiological indicators, which can provide new ideas for MetS detection. For example, Tang Y et al. used physicochemical, bioinformatic, and TCM indicators to build a prediction model by TreeNet, which had an accuracy value of 73.23 %, and their experimental results indicated that the combination of TCM and physicochemical indicators could provide early warning for MetS [[13]]. Shu-Jie Xia et al. constructed a random forest MetS prediction model for 586 cases in China using 47 TCM indicators and 20 physicochemical indicators, including alanine aminotransferase (ALT), aspartate aminotransferase (AST), and glutamyl transpeptidase (γ-GT), with an accuracy of 94.2 %, however, in the data using purely TCM indicators, the model's accuracy was only 80.1 % [[14]]. The above research on machine learning prediction methods for MetS including TCM indicators shows that it is possible to achieve MetS prediction using TCM indicators, but there are three shortcomings: 1) The above models use a large number of TCM indicators [[14]], and there are complex TCM balance theory [[15,16]] indicators, which need the experience of TCM doctors to determine; 2) The above models also need to combine physical and chemical indicators; it is not clear whether physical and chemical indicators or TCM indicators contribute more to the model prediction, and the model with physical and chemical indicators is not convenient for self-testing by ordinary users; 3) There needs to be proper optimization of these models. Their performance is closely related to the choice of model hyperparameters, and different hyperparameters will have different effects on the robustness and accuracy of the model; thus, so the model needs to be optimized using hyperparameter optimization methods to maximize the model performance.

Considering the previous analysis, this study suggests a BO-XGBoost algorithm to screen out 19 TCM-related indicators and construct a MetS prediction model. The algorithm screens simple indicators to facilitate self-testing by ordinary users. The model can provide a simple and accurate method for diagnosing MetS. The main contributions of this paper are listed as follows.

1. A BO-XGBoost prediction model for MetS based on TCM characteristics is proposed. Using this model, users don't need to draw blood for testing, but only need a simple questionnaire survey, and the model can predict the diagnosis results.
2. Nineteen indicators were screened based on the BO_XGBoost algorithm, which belongs to three categories of individual bioinformation, TCM indicators, and TCM lifestyle habits, excluding any physical and chemical indicators, facilitating self-testing by the general population.
3. In this study, 2040 examples were experimentally investigated, and the prediction model based on BO XGBoost exhibited excellent performance, outperforming all seven comparison models. Furthermore, an intelligent MetS prediction application was developed based on the proposed model, which is very convenient for user to test.

The second section introduces the data sources and preprocessing methods as well as the method suggested in this paper; the third

section of the paper discusses the feature screening and experimental results; and the last section of the paper is the conclusion.

2. Materials and methods

2.1. Experimental data sources

The experimental data were collected from the electronic medical records of outpatient clinics of Guangdong Provincial Hospital of Chinese Medicine from 2020 to 2021. A total of 2040 cases, aged between 18 and 90 years old, included 268 cases in the ordinary population and 1772 cases in the population with confirmed MetS. To achieve early identification of MetS, the case labeling method uses the MetS diagnostic criteria of the IDF [[2]], and people who meet two or more of these diagnostic criteria are labeled as confirmed cases of MetS. Otherwise, they are labeled healthy people. The following three types of TCM information are collected for all people: TCM indicators, individual biological information, and TCM habits, totaling 400 pieces of information. The specific individual biological information includes age, height, weight, etc. TCM indicators include urination, sweating conditions, abdominal and stomach distention, taste and diet, etc., and this information must be easily self-measured by ordinary people. Information on TCM habits includes frequency of diet, frequency of exercise, frequency of raw or cold food, etc. Its TCM indicators were classified into four levels of severity using a standard quantitative scale: no symptoms 0, mild symptoms 1, moderate symptoms 2, and severe symptoms 3, the details of which can be found in Table 3.

2.2. Data preprocessing

This study performs preprocessing techniques on the dataset, including feature transformation, missing value filling, and feature binning [[17]], to inhance the accuracy of the machine learning model.

(1) Feature Transformation. The string type features and labels in the dataset are converted to a numeric type. For example, for the string type feature of sleep time, 0 represents before 23:00, and 1 represents after 23:00. The gender "male" is represented as 0, and the gender "female" is represented as 1.
(2) Missing value padding. In the process of data sampling, there are generally missing values, and to ensure the integrity of the data, the missing values need to be filled. For features with missing values more than half of the sample size, this paper discards the feature; for features with missing values less than 5 % of the sample size, the median is used to fill the missing features; considering that some symptoms may differ in gender, such as whether menstruation is abnormal (female), this paper splits the data sample by gender using the plural to fill the missing features.
(3) Feature discretization. In machine learning, continuous-type features are usually discretized operations [[18]], and then one-hot coding or dummy variable coding is performed on the discrete features. For example, waist circumference is binary classified according to the criteria of abdominal obesity: $\geq$90 cm for men and $\geq$85 cm for women. The disease duration was classified into three intervals of mild, moderate, and severe by time duration, and then one-hot coding was performed on the discretized data. This practice reduced the influence of feature value perturbation and improved the model stability and robustness.
(4) Hybrid sampling. Due to the class imbalance between positive and negative samples in the original dataset, the model tends to classify samples into the majority class. To address this issue, this study proposes a hybrid sampling strategy. Firstly, the data set is divided into a training set and a test set. Subsequently, the SMOTE-ENN algorithm [[19]] is applied to the training set, augmenting 268 positive samples to match the quantity of negative samples. Then, the ENN algorithm is used to perform downsampling on both positive and negative samples. After the hybrid sampling process, there are 1030 positive instances and 963 negative instances, ensuring a balanced representation of positive and negative cases in the training samples and mitigating the impact of imbalanced data on model predictions. Furthermore, the adoption of hybrid sampling aims to avoid excessive noise introduced by oversampling.

2.3. BO-XGBoost model design

2.3.1. XGBoost algorithm

Extreme gradient boosting, also known as XGBoost (eXtreme Gradient Boosting)[[20]], is a machine learning algorithm that integrates multiple regression trees using the boosting method. It has recently gained popularity in classification and regression tasks due to its high generalization and efficiency [[21,22]]. An enhanced variant of the GBDT (Gradient Boosting Decision Tree) algorithm is the XGBoost algorithm. It differs from the GBDT algorithm in that it uses a combination of first- and second-order derivatives to optimize the objective function while also including the complexity of the tree as a regular term to prevent overfitting. The fundamental idea is to continuously construct new trees to rectify the errors made by the initial classifier and to weigh the total of each tree to determine the final prediction result. The specific XGBoost algorithm is implemented as follows for a dataset with m features and n samples.

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

$$F = \left\{ f(x) = w_{q(x)} \left(q : R^m \to T, w \in R^T \right) \right\} \tag{2}$$

In Eqs. (1) and (2), K is the number of additive models utilized, F is the regression tree space, $f(x)$ is one of the regression trees, and $q(x)$ denotes the mapping relationship between samples and leaf nodes. T stands for the quantity of leaves in each regression tree, and w stands for the percentage of distinct leaf nodes. The XGBoost objective function L is as follows:

$$L = \sum_{i=1}^{n} l(\widehat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda w_j^2 \tag{4}$$

Ω denotes the complexity of a regression tree, l is the loss function, and γ and λ are coefficients that regulate the model to prevent overfitting in Equation (3). New trees must be created to fit the residuals produced by the previous forecast, because XGBoost is an additive model. The prediction score can be stated as follows after the tth round of optimization and the generation of the nth regression tree:

$$\widehat{y}_i^{(t)} = \widehat{y}_i^{t-1} + f_t(x_i) \tag{5}$$

where $\widehat{y}_i^{t-1}$ is the cumulative model prediction score for the first t-1 rounds, at which point the objective function L can be rewritten as:

$$L(t) = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i) \right) + \Omega(f_t) \tag{6}$$

Assuming $g_i = \frac{\partial l\left(y_i, \widehat{y}_i^{t-1}\right)}{\partial \widehat{y}_i^{t-1}}$, $h_i = \frac{\partial^2 l\left(y_i, \widehat{y}_i^{t-1}\right)}{\partial \widehat{y}_i^{t-1}}$ is true, the objective function is replaced and Taylor's formula is used to increase the objective function by the second order to obtain the following equation:

$$L(t) = \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y}_i^{(t-1)} \right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_j^2 \tag{7}$$

The final objective function formula L is created as follows: by increasing the loss function of the samples, recombining the samples, and then using the vertex formula to determine the ideal w:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{8}$$

$$L = -\frac{1}{2} \sum_{i=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{9}$$

$$G_i = \sum_{i \in I_j} g_i \tag{10}$$

$$H_i = \sum_{i \in I_j} h_i \tag{11}$$

To determine the best segmentation point, XGBoost employs the greedy algorithm and the approximation algorithm, lists several potential candidates using the percentile method, and then determines the best segmentation point using (8). (9).

2.3.2. Bayesian optimization algorithm

From Eqs. (8)–(11), we find that the XGBoost model involves a more significant number of hyperparameters, and these hyperparameters have a greater impact on the model performance, so it is necessary to optimize the hyperparameters of XGBoost. Grid search, random search, and other standard hyperparameter optimization techniques are listed below. The best parameters can be discovered using grid search, which explores all possible parameter combinations. However, XGBoost has a sizable number of hyperparameters, and grid search is time-consuming. The goal of the random search is to repeatedly sample the parameter domain at random to find the set of parameters that produces the best results. Random search has the disadvantage of inconsistent results while increasing randomness and speeding up optimization.

The Bayesian optimization algorithm is a global optimization method that rapidly identifies a set of globally optimal solutions by updating the prior probability model while taking into account previous parameter information. The following is an expression for the Bayesian optimization hyperparameters:

$$a^* = \underset{a \in A}{\operatorname{argmin}} L(a) \tag{12}$$

where A is the best combination of parameters discovered and a^* is the set of hyperparameters. The choice of an appropriate probabilistic agent model and acquisition function is the foundation of applying Bayesian optimization to solve real-world issues.

There are several different forms of proxy functions, including the Gaussian process (GP) and Tree Parzen Estimator (TPE)[[23]]. This paper uses TPE, which does not define a predictive distribution for the objective function but creates l(a) and g(a), indicating probability distribution modeling from existing historical data, and uses both as generative models for hyperparametric domain variables [[24]]. This is because TPE outperforms Gaussian processes in terms of accuracy and efficiency.

TPE is defined as follows:

$$p(a|h) = \begin{cases} l(a), & \text{if} h < u^* \\ g(a), & \text{if} h \geq u^* \end{cases} \tag{13}$$

where h is the response value of the objective function's evaluation metric and a is the hyperparameter combination. The weighted values of AUC and accuracy are the evaluation metrics used in this paper. u^* is the highest possible value for the objective function's evaluation metric response value. The probability distribution is $l(a)$ when the value of the parameter $a^{(i)}$'s h is less than u^*, and it is $g(a)$ when the value of the parameter $a^{(i)}$'s h is greater than u^*.

The two most frequently used acquisition functions are PI (probability of improvement) and EI (expected improvement). EI is chosen as the collection function because it integrates the probability of improvement and reflects the various lift amounts. In contrast, PI only reflects the probability of improvement and does not account for the size of improvement because the PI strategy considers all improvements equal [[25]]. The EI acquisition function is:

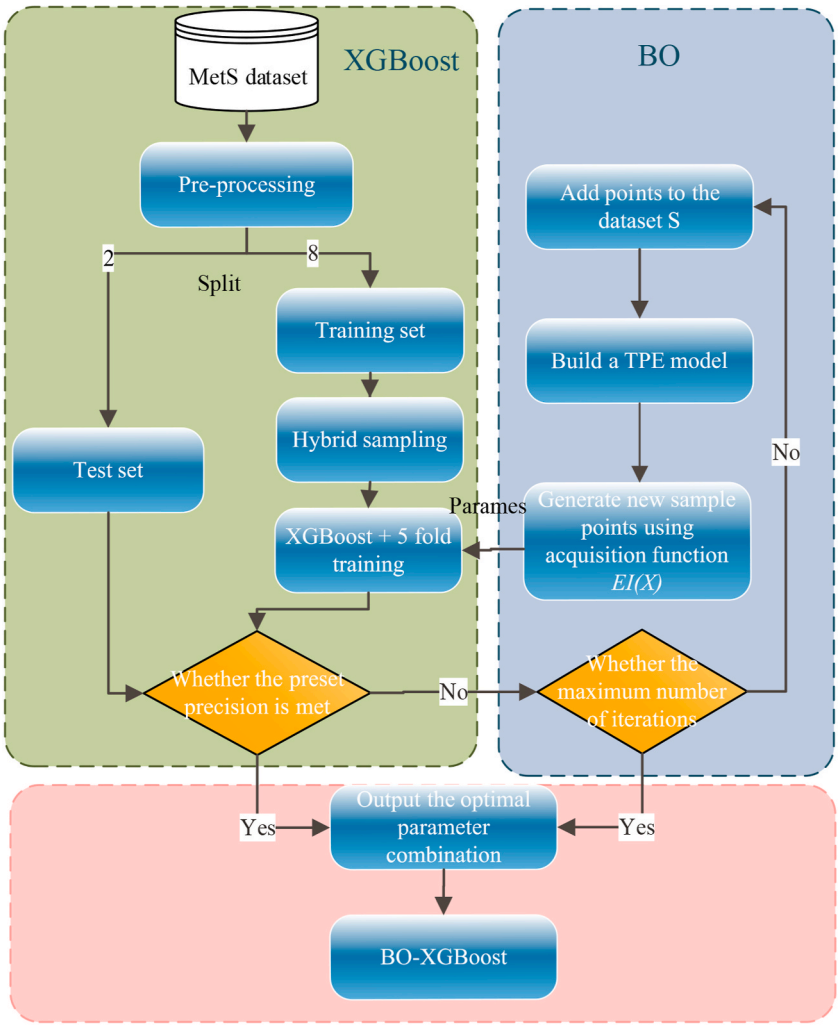

Fig. 1. Flow chart of the BO-XGBoost algorithm.

$$EI_{u^*}(a) = \frac{\gamma u^* l(a) - l(a)\int_{-\infty}^{u^*} p(h)\mathrm{d}h}{\gamma l(a) + (1-\gamma)g(a)} \propto \left(\gamma + \frac{g(a)}{l(a)}(1-\gamma)\right)^{-1} \tag{14}$$

where $\gamma = p(h < u^*)$. Updating the probability density functions $l(a)$ and $g(a)$, the new x^* can be determined according to the minimum of $g(a)/l(a)$, and the EI is larger when $l(a)$ is larger and $g(a)$ is smaller. Iterating is continued by replacing (15) with (16) until the accuracy is met or the allotted number of iterations has been used.

2.3.3. MetS BO-XGBoost prediction model design

In this study, XGBoost and a Bayesian optimization algorithm were combined to create a BO-XGBoost-based algorithm that automatically looks for the best parameters using heuristic techniques to optimize the XGBoost algorithm's parameters. The BO-XGBoost prediction model based on MetS is created using the MetS dataset as input, and Fig. 1 depicts the step-by-step procedure of the BO-XGBoost algorithm implementation.

Step 1. Electronic case information of MetS patients is collected, operations such as data preprocessing are performed, and according to the 8:2 rule, the processed dataset is divided into a training set and a test set. Subsequently, the training set undergoes hybrid sampling.

Step 2. Based on the XGBoost hyperparameters to be optimized, n initial points are generated, $X_n = [X_1, X_2, X_3 \ldots X_n]$, as the initial hyperparameters of the model.

Step 3. To determine the goal value of X_n for the XGBoost objective function $Y_n = [Y_1, Y_2, Y_3 \ldots Y_n]$, the parameters are substituted into the XGBoost model and a 5-fold cross-training on the training set is run. The dataset $S = [(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)]$ is created.

Step 4. : Whether the model meets the predetermined accuracy threshold on the MetS test set is determined. If so, the algorithm ends and outputs the corresponding optimal parameter combination X_n. If not, whether the maximum number of iterations has been achieved is determined. If so, the corresponding optimal parameter combination X_n is output. Otherwise, a cycle of the Bayesian optimization process is carried out.

Step 5. A TPE model is built using dataset S.

Step 6. Based on the TPE, the acquisition function $EI(X)$ is used to calculate X_{n+1}.

Step 7. The calculated new sampling points X_{n+1} are used as hyperparameters of the XGBoost model for 5-fold cross-validation training and validation to obtain Y_{n+1}.

Step 8. Whether the model meets the preset accuracy requirements is determined. If it does not, steps 3–8 are repeated, and if it does, the best hyperparameter is X_{n+1} , and no steps need to be repeated.

Step 9. The optimized best hyperparameter combination is added to the XGBoost model to form the BO-XGBoost model.

The BO-XGBoost model obtained in step 9 is the MetS prediction model, and the BO-XGBoost model can also be used for the feature selection of the MetS dataset.

2.3.4. MetS feature selection based on BO-XGBoost

The raw data for MetS contain several features, some of which are redundant and useless. Feature selection of the raw features is necessary to increase the model's accuracy while lowering the feature dimension and running time. Filtering, packing, and embedding approaches are common categories for feature selection techniques. The filtering approach, which is popular due to its simplicity and effectiveness [[26]], chooses features based on the statistical characteristics of each feature. Common filtering techniques include the Pearson correlation coefficient, distance correlation coefficient, and mutual information criteria, which often operate on the dataset in a noniterative manner and can remove redundant features but not irrelevant ones. Wrapper and embedding methods of feature selection are model-oriented implementation schemes with typical methods, such as recursive feature elimination (RFE), particle swarm optimization (PSO), genetic algorithms, and random forest feature selection [[27,28]]. These methods typically require multiple iterations on the dataset to calculate the best performance of each combination of features and feature subsets on their models with better outcomes. To more accurately predict MetS, this paper suggests the forward-selective wrapper method BO-XGBoost feature selection. This method selects the features that have the highest contribution to the model, eliminates the redundant features that remain, and then classifies the selected features to produce the optimal selection of MetS features. The steps for implementing BO-XGBoost feature selection are as follows:

Step 1: The MetS dataset containing all the original features after preprocessing is trained using the BO-XGBoost model.

Step 2: The importance of each feature is calculated by the BO-XGBoost model feature_importances and subsequently sorted in ascending order of feature importance.

Step 3: The features are added to the BO-XGBoost model one by one, and a 5-fold validation is performed to calculate the AUC values.

Step 4: The AUC value of each feature is saved and recorded to form an AUC-feature number curve.

Step 5: The curve inflection point is identified, the optimal number of features, V, is determined, and the subset of V features is considered the reduced-dimensional MetS dataset.

3. Experimental results and analysis

3.1. Test environment and configuration

The experimental environment in this paper is: AMD Ryzen 9 5900HX CPU @3.30 GHz, Ubuntu18 OS, python3.8.

3.2. Evaluation indicators and methods

The main goal of this study was to identify MetS predictively, which effectively changes the challenge of predictive identification into a classification problem and uses machine learning to evaluate whether the user is healthy or ill. The most popular evaluation metric for classification problems, *Accuracy*, is used in this paper to assess the performance of the BO-XGBoost prediction model. However, *Accuracy* is not always the best indicator of the model's validity, so this paper also includes *Precision*, *Recall*, the sum of specificity and sensitivity mean (*F*1_*score*), and AUC values.

As shown in Table 1, assuming that sickness is the positive category and health is the negative category, this thesis tends to focus on the performance indicators of sickness prediction, and the precision, specificity, sensitivity (recall) and accuracy of sickness prediction can be expressed as:

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

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

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

On the basis of *Precision* and *Recall*, *F*1_*score* can be calculated as:

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall}$$

On the basis of *Recall* and *Specificity*, We propose a new evaluation index named as *FRS_score*,which considers both sensitivity and specificity.

$$FRS\_score = \frac{2 \times \text{specificity} \times recall}{\text{specificity} + recall}$$

The following formula was used to calculate the AUC values:

$$AUC = \int_0^1 \frac{TP}{TP + FN} d\frac{FP}{FP + TN}$$

3.3. Model parameter optimization results

To ensure that the model achieves accurate identification of MetS, this paper uses the Bayesian optimization algorithm for intelligent tuning of XGBoost. Five hundred rounds is the defined limit for the number of Bayesian optimization iterations. In this paper, ten core parameters of the XGBoost model are chosen for optimization: learning_rate, importance_type, max_depth, n_estimators, reg_alpha, reg_lambda, subsample, colsample_bytree, max_delta_step and min_child_weight. To find the optimal parameters, the parameter ranges need to be set reasonably, and the parameter fields in Table 2 of this paper are shown.

The Bayesian optimization parameter distribution is shown in Fig. 2(a), and the random search optimization parameter distribution is shown in Fig. 2(b). Fig. 2 compares the two different parameter search methodologies. Fig. 2 demonstrates that Bayesian optimization, as opposed to random search, can use the data from the previously studied samples as an a priori to plot the subsequent sample. Bayesian optimization can exactly locate the point that fits the posterior maximum as the number of iterations rises, effectively reducing the time and computational effort required for the parameter search. In accordance with Table 2, the optimal XGBoost parameters are indicated in Fig. 2 with asterisks. From Table 2, we can see that colsample_bytree is optimized from the default of 1 to

Table 1
Confusion Matrix.

	Sickness Positive (Pre)	Healthy Negative (Pre)
Sickness Positive (True)	TP	FN
Healthy Negative (True)	FP	TN

Table 2
XGBoost parameter settings and optimal parameters.

Parameter Name	Parameter Space	Default Parameters	BO optimal parameters
learning_rate	Loguniform (0.1,0.2)	0.1	0.0462
importance_type	['weight','gain']	gain	weight
max_depth	Uniform (1,100,1)	6	65
n_estimators	Uniform (50,5000,1)	100	115
reg_alpha	Uniform (0,1)	0	0.7120
reg_lambda	Uniform (0,1)	1	0.2514
subsample	Uniform (0,0.99)	1	0.9501
min_child_weight	Uniform (0,20,1)	1	6
colsample_bytree	Uniform (0.1,0.99)	1	0.3640
max_delta_step	Uniform (0,2)	0	1.684

Table 3
Features after dimensionality reduction by BO-XGBoost feature selection for 19 features.

Number	Category	Name	Description of the values
X1	Individual biological information	Age	Unit Age (int)
X2		BMI	Unit: kg/m^2 Formula: BMI = Weight (kg)/Height (m)
X3		Waistline	Unit cm (float)
X4		Career	1: teachers, 2: caregivers and nannies, 3: catering and food industry, 4: commercial services, 5: medical personnel, 6: workers, 7: civilian workers, 8: farmers, 9: pastoralists, 10: fishermen, 11: cadres and employees, 12: retired persons, 13: domestic and waiting for work, 14: other
X5	TCM indicators	Bitter taste in the mouth	0: none, 1: rarely, 2: sometimes, 3: often, 4: always (int)
X6		Abdominal and stomach distention	0: Normal, no abdominal distension. 1: Yes, mild. Occasional episodes, 1–2 times a week, obvious after eating, sometimes stopping, relieved within half an hour, not affecting daily life 2: Yes, moderate. 2–3 days, one episode, obvious after eating, frequent episodes, relieved within 0.5–1 h, partially affecting daily life 3: Yes, severe. Daily seizures, obvious after eating, relieved only in 1 h, or even not relieved all day, affecting work and life.
X7		Feel short of breath	0: none, 1: rarely, 2: sometimes, 3: often, 4: always (int)
X8		Easy to panic	0: none, 1: rarely, 2: sometimes, 3: often, 4: always (int)
X9		Abdominal fat and flabby	0: none, 1: rarely, 2: sometimes, 3: often, 4: always (int)
X10		Character of urine	0: normal, 1: clear and long urine, 2: yellow urine, 3: foamy urine (int)
X11		Character of sweat and spontaneous sweating	0: No, 1: Yes (int)
X12		Eye discomfort	0: No, 1: Yes (int)
X13	TCM habits	Frequency of raw and cold foods	Unit day/week (int)
X14		Work-stress score	0: retired or not working, 1: 0 points, 2: 1 point, 3: 2 points, 4: 3 points, 5: 4 points, 6: 5 points (int)
X15		Average daily hours of air conditioning use in summer	Unit: hour/day, retain one decimal
X16		Frequency of fruit consumption	Unit day/week (int)
X17		Frequency of nut consumption	Unit day/week (int)
X18		Frequency of tea consumption	Unit day/week (int)
X19		Frequency of dairy product consumption	Unit day/week (int)

0.3640, learning_rate is optimized from the default of 0.1 to 0.0462, and reg_lambda is optimized from the default of 1 to 0.2517.

The parameters of the final optimized XGBoost are shown in Table 3, and the results show that the optimized parameters were changed significantly from the default parameters.

3.4. Results of feature selection based on BO_XGBoost

For the 400-dimensional features obtained after preprocessing, this paper combines the BO optimization algorithm with ExtraTrees (ET), Decision tree (DT), RandomForest (RF), and XGBoost algorithms, trains the optimized algorithm on the original MetS dataset, and takes a 5-fold validation average. The experimental results show that the AUC value of XGBoost is significantly higher than that of the other three models starting from the 11th feature and tends to be stable, as shown in Fig. 3. Fig. 3 demonstrates that the AUC value of the BO-XGBoost algorithm reaches its highest value when the number of features is 19, and the addition of subsequent features has little effect on the model performance. Therefore, we screened the first 19 features as the features used in this paper by the BO-XGBoost feature selection method and determined the importance of each feature. The filtered features are shown in Table 3, and the visualization of feature importance is shown in Fig. 4. Fig. 4 shows that the 10 features of greatest importance are waist circumference, nature of urination, occupation, whether the eyes are abnormal, age, spontaneous sweating, epigastric distention, work stress score, ease of panic, and frequency of fruits. Chen Shujiao et al. [[29]] concluded that the symptoms with higher frequency in MetS

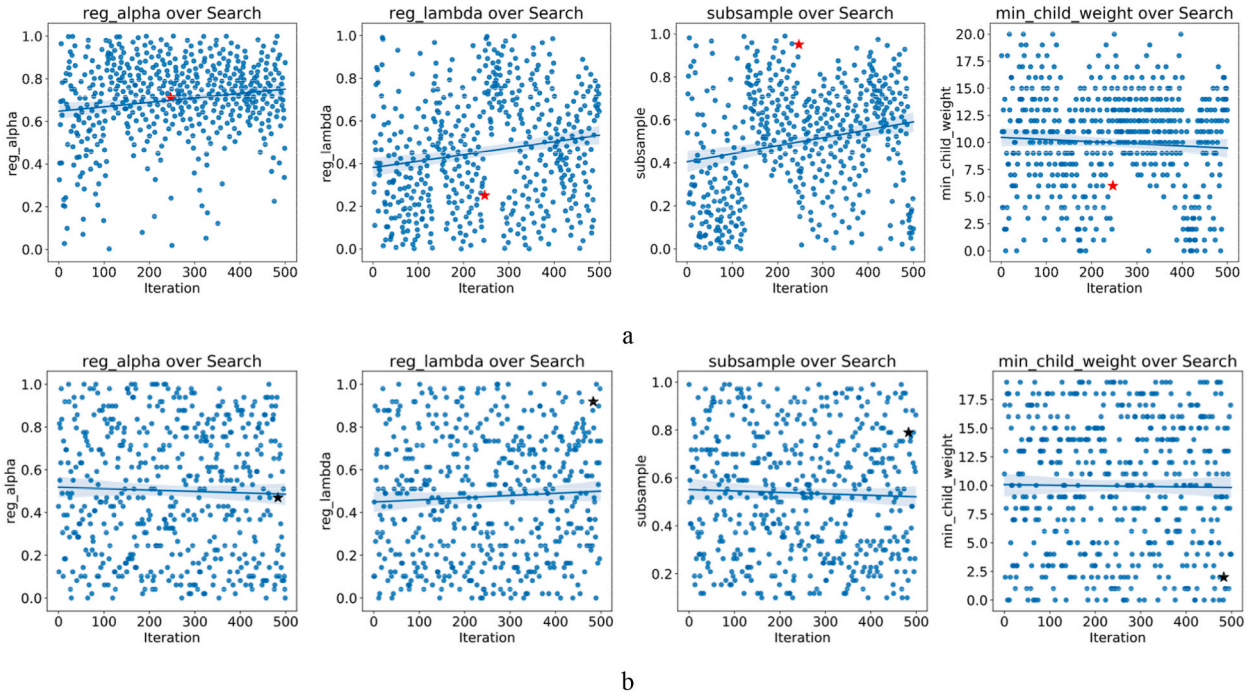

Fig. 2. Parameter distributions of two different parameter optimization methods. a: Bayesian optimization parameter distribution graph, b: random search optimization parameter distribution graph.

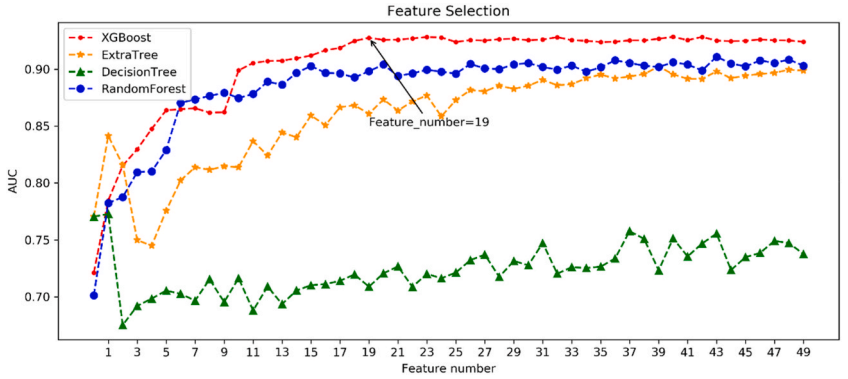

Fig. 3. Comparison effect of different feature selection methods.

symptoms were physical obesity, thirst, epigastric distention, and thirst for hot drinks, and this information has a good match with the features screened in this paper.

3.5. Feature visualization

To better interpret the 19 features screened, this paper introduces the SHAP-value pair proposed by Lundberg for analysis [[30]].

To examine the impact of their attributes on the MetS classification outcomes, two e-case datasets were randomly chosen from the MetS dataset. As Fig. 5(a) illustrates, the red area shows that a feature has a positive contribution to the goal value, while the blue area shows a negative contribution. The case will be classified as belonging to the sick group when the red response value exceeds the base value and as belonging to the normal group when the blue response value is less than the base value. From Fig. 5(a), it can be seen that the response value f(x) corresponding to the current case characteristics is 4.57, which is much larger than the base value, so the sample is judged as 1 (patient), where X10 Character of urine = 3 (foamy urine), X3 Whether abdominal obesity = 1 (yes), X15 Daily air conditioning hours in summer = 10 (hours), X16 Frequency of fruit = 2 (day/week), X18 Frequency of tea = 7 (day/week), X11 Whether spontaneous sweating = 1 (yes), X17 Frequency of nut = 0 (day/week), X2BMI = 24.79, and X12 Whether eye discomfort = 1 (yes) played a positive role in determining the person as sick. Similarly, Fig. 5(b) was judged as a normal group. Among the

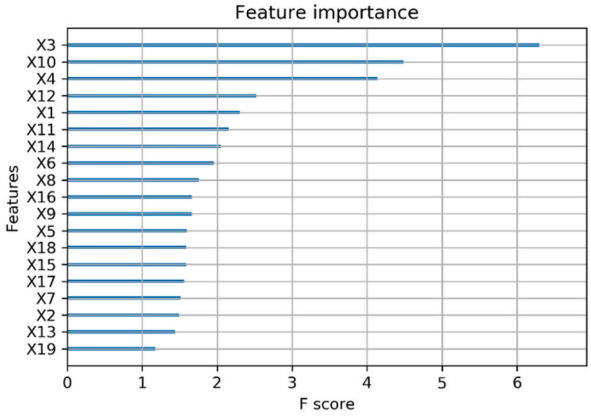

Fig. 4. MetS feature gain ranking.

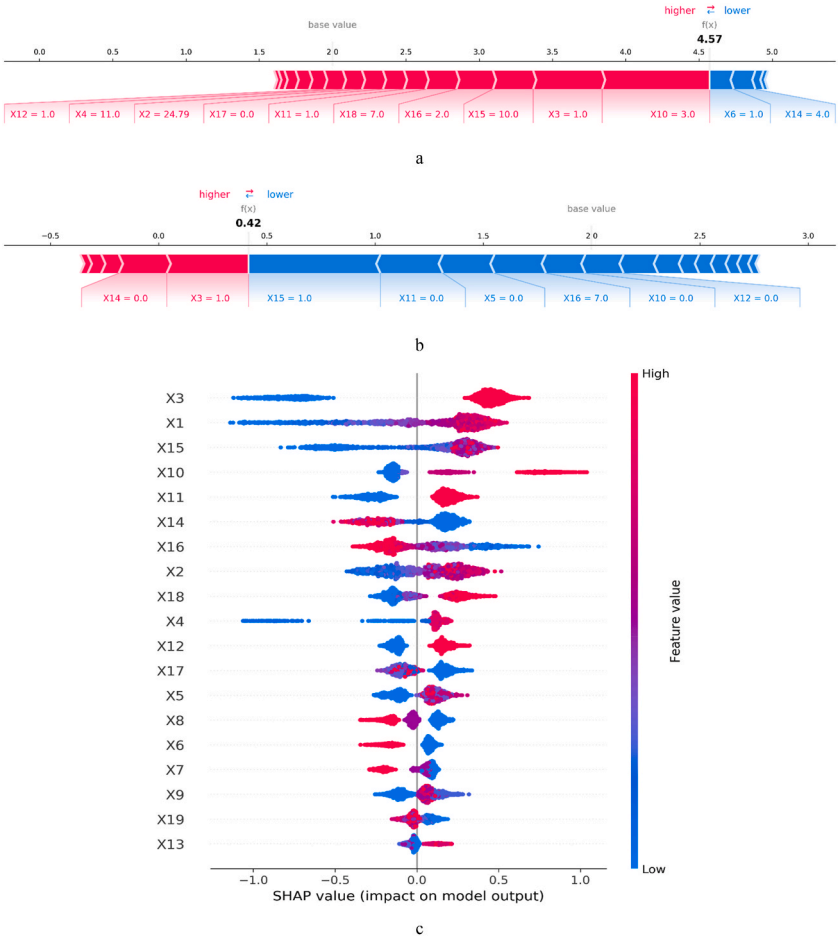

Fig. 5. SHAP feature plots. Note: a and b are the two patient predictor contribution plots. c is the SHAP-value plot for all features.

characteristics used in this case, X15 Average daily air conditioning hours in summer = 1 (hour), Whether spontaneous sweating = 0 (no), X5 Bitter taste in the mouth = 0 (no), X16 Frequency of fruit = 7 (daily weekly), X10 Character of urine = 0 (normal), and X12 Whether eye discomfort = 0 (no) played a positive role in judging the person as normal.

The global feature importance of the BO-XGBoost model with the SHAP-value for 19 features is shown in Fig. 5(c), and this SHAP plot shows approximate results in Fig. 4. The highest ranked feature in Fig. 4 is waist circumference, which is the same in Fig. 5(c), and 7 of the top 10 features selected by BO-XGBoost are again selected by SHAP global feature importance. This result indicates that the

robustness of BO-XGBoost is supported by the SHAP technique.

3.6. Comparison of evaluation value indicators between different models

To confirm the accuracy of the BO_XGBoost prediction model for MetS based on TCM features, this paper compares it with seven algorithms, including random forest (RF), K-nearest neighbor (KNN), multilayer perceptron MLP, logistic regression (LoR), support vector machine (SVM), LightGBM, and XGBoost.

The experimental dataset was divided, with 20 % serving as the validation set and 80 % serving as the training set. The trials were repeated 100 times to confirm the objectivity of the findings, and the average of these repetitions served as the basis for the final comparative findings, as shown in Table 4.

The first seven machine learning algorithms all use default parameters. Table 4 shows that the BO_XGBoost method proposed in this paper outperforms the other machine learning algorithms in all three metrics, including a recall (sensitivity) of 90.67 %, an F1 value of 93.35 %, an accuracy of 88.23 %, and a *FRS_score* of 80.40 %, which are 0.84 %, 0.91 %, 0.92 % and 1.49 % higher than the unoptimized XGBoost model, respectively. The other models have better single metrics, but the combined metrics are rather low.

Therefore, the performance of the above classifiers on the test set can be combined to conclude that the BO_XGBoost method is the best model for MetS recognition.

3.7. Comparison of ROC_AUC curves of different classifiers

The positive and negative samples of the MetS dataset are unbalanced, and for such unbalanced data, we use the ROC curve as a supplement to judge the goodness of the model. The AUC value is the area under the ROC curve, and the closer its value is to 1, the better the model's ability to classify data. The better the model performs, the closer the ROC curve of the model is to the upper left corner. The ROC curves of the experimental results are shown in Fig. 6, from which we can see that the ROC curve of the KNN model is at the bottom, and the AUC value of this model is 0.811, which is the worst result among all models. The AUC value of the BO_XGBoost prediction model is 0.920, which is the best among all models and 1.8 % higher than the next-best RF. In summary, the BO_XGBoost prediction model performs better in the MetS dataset and can achieve the MetS prediction effect.

4. Discussion

This study aims to develop a model for early identification of Mets based on TCM characteristic information, enabling non-invasive detection of patients and improving the convenience of testing. In this research, the BO-XGBoost feature selection algorithm is utilized to extract 19 features from the original medical records dataset containing 400 features. These selected features include personal biological characteristics such as age, BMI, waist circumference, as well as TCM characteristics like bitter taste in the mouth, shortness of breath, and spontaneous sweating. Additionally, TCM lifestyle habits like tea consumption frequency, consumption of cold and raw food, and daily air-conditioning usage are considered. Previous studies [29] have reported that these features are closely related to the occurrence and development of Mets.Since the original dataset exhibits an imbalance with a higher number of positive samples, this study adopts a hybrid sampling algorithm to balance the dataset. SMOTE is applied to synthesize minority class samples in the training set, while ENN is used to eliminate majority class samples, mitigating the impact of data imbalance on model predictions.Based on the aforementioned dataset and 19 selected features, the BO-XGBoost model for early identification of metabolic syndrome is trained. Compared to other machine learning models such as LoR, k-NN, RF, MLP, lightGBM, and XGBoost, the BO-XGBoost model demonstrates superior performance in terms of AUC, F1-score, and sensitivity. Specifically, the AUC is 0.920, F1-score is 93.35 %, and sensitivity is 90.67 %. High sensitivity indicates that BO-XGBoost can effectively identify individuals with metabolic syndrome, reducing the risk of missed diagnoses. Simultaneously, high specificity indicates its ability to distinguish individuals without metabolic syndrome, minimizing the risk of misdiagnoses.To comprehensively evaluate the performance, this study introduces a new evaluation metric, FRS-score, which takes both sensitivity and specificity into account. Among all the machine learning models considered, BO-XGBoost exhibits the highest performance with an impressive FRS-score of 80.40 %. Consequently, the results support the early detection of cases or disease risks in a non-invasive and convenient manner, enabling timely intervention measures.

RandomForest and Artificial Neural Network (ANN) are algorithms that can be used for nonlinear statistical modeling [[14,31]]. In contrast, LoR is a simple technique that uses linear combinations of variables, meaning it cannot effectively capture complex interactions with nonlinearities. BO-XGBoost holds theoretical advantages over LoR in capturing nonlinearity between factors and outcomes, and BO provides a better approach for optimizing numerous parameters in XGBoost. The BO-XGBoost model demonstrates excellent performance in MetS prediction, but it falls under the category of black-box algorithms, where the internal decision-making process is challenging to interpret directly.

Numerous scholars have conducted research on the identification of MetS. CHEN Shu-jiao et al. analyzed 160 MetS patients and found that symptoms such as abdominal and stomach distention, bitter taste in the mouth, and others occur frequently [[29]]. Shu-Jie Xia et al. constructed a random forest MetS prediction model for 586 cases in China using 47 TCM indicators and 20 physicochemical indicators. They discovered that TCM characteristics, including body fat, chest tightness, and spontaneous sweating, play a crucial role in predicting MetS. Additionally, the physicochemical indicator fasting blood glucose (FBG) also proves to be significant in the prediction process [[14]]. This demonstrates the importance of TCM characteristic parameters such as bitter taste in the mouth, feeling short of breath, and spontaneous sweating in predicting MetS. However, these studies require invasive procedures to collect blood samples for measuring biochemical or biophysical parameters (e.g., TG, HDL-C, FBG), which is not conducive to convenient testing.

Table 4
Classification performance of different algorithms.

Classification Algorithm	Evaluation Indicators					
	P (%)	R (%)	F1 (%)	Accuracy (%)	Specificity	*FRS_score*
RF	94.18 %	87.09 %	90.45 %	84.06 %	64.81 %	74.28 %
MLP	94.71 %	75.98 %	84.32 %	75.49 %	72.22 %	74.05 %
KNN	95.12 %	66.10 %	78.00 %	67.64 %	77.77 %	71.46 %
LoR	95.80 %	70.90 %	81.49 %	72.05 %	79.62 %	75.01 %
SVM	98.68 %	21.18 %	34.88 %	31.37 %	98.14 %	34.85 %
Lightgbm	95.12 %	90.54 %	92.77 %	87.20 %	66.66 %	77.04 %
XGBoost	95.20 %	89.83 %	92.44 %	87.25 %	70.37 %	78.91 %
BO-XGB	95.53 %	**90.67 %**	**93.35 %**	**88.23 %**	**72.22 %**	**80.40 %**

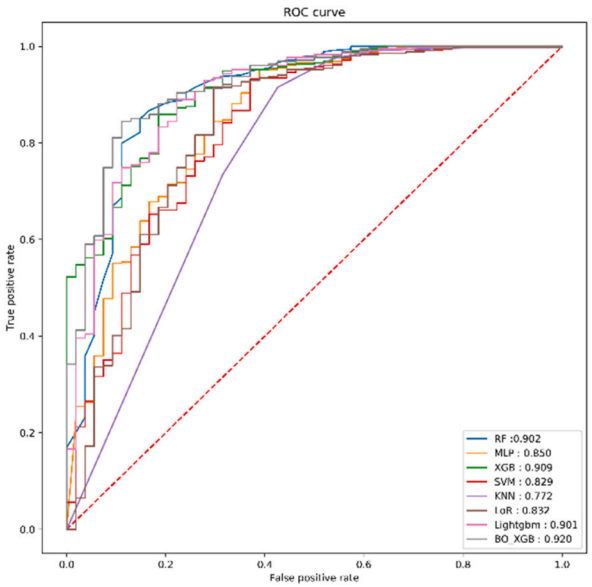

Fig. 6. ROC graphs of various machine learning algorithms, where the AUC values are in the lower right corner.

Wang, Feng-Hsu et al. investigated the diagnostic accuracy of using an ANN for MetS prediction based on socioeconomic status and lifestyle factors [[31]]. Their findings emphasize the feasibility of using non-invasive features for MetS prediction. In comparison to previous studies, this research integrates three categories of non-invasive indicators: TCM indicators, individual biological information, and individual lifestyle habits, to construct the BO-XGBoost model. The AUC value of this model allows for more convenient and accurate early identification of MetS. Additionally, this study addresses the interpretability of machine learning and black-box models. Through the analysis of SHAP values, the research provides insights into the impact of different features on the model's predictions, identifying the key features that play a crucial role in MetS prediction.

Our study also has some limitations. Firstly, all the data were collected from the Guangdong Provincial Hospital of Traditional Chinese Medicine, which means that the model is currently mainly applicable to the population in the South China region. We have not yet tested its applicability to other regions or continents, which will be further explored in future work. Secondly, our study is a cross-sectional study, which prevents us from making causal inferences. Further follow-up research is needed to evaluate causality.

5. Conclusion

In this study, a new model for the prediction of MetS was developed that includes the advantages of TCM indices, using 19 data points, including lifestyle habits, such as usual dietary habits, exercise frequency, and easily neglected symptoms, such as the nature of urination, abdominal bloating and mouth bitterness, as reference indicators. Based on the BO-XGBoost model, the 400-dimensional features of the original dataset were reduced to 19-dimensional features, which effectively eliminated the useless features and reduced the redundancy among the features. From the experimental results, the BO-XGBoost prediction model proposed in this paper achieved values of 93.35 %, 90.67 %, 80.40 % and 0.920 for the F1, sensitivity, FRS, and AUC metrics, respectively. These values are the highest compared to other model algorithms for the same indicators.

By including bad habits and TCM symptoms and ranking their degree of correlation with MetS, we have derived information on the bad habits most likely to lead to MetS and the characteristics that predict the most likely development of MetS. At the same time, a

MetS prediction model was constructed based on the BO-XGBoost model, which allows all users to quickly self-test through a questionnaire (the questionnaire is currently deployed on the cloud server and available to users at http://175.178.194.244:82/, username: usr, password:123), thus achieving the goal of early detection and early treatment to avoid serious MetS complications.

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Jianhua Zheng: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Funding acquisition. **Zihao Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation. **Jinhe Wang:** Supervision, Methodology, Investigation, Conceptualization. **Ruolin Zhao:** Visualization, Validation, Investigation. **Shuangyin Liu:** Validation, Investigation. **Gaolin Yang:** Validation, Investigation. **Zhengjie Liu:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Data curation. **Zhengyuan Deng:** Validation, Software, Project administration, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Acknowledgments

This work was supported in part by the Research Fund Program of Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization under Grant 2021B1212040007 and Grant 2021503; in part by the National Key Research and Development Program of China under Grant 2018YFC2002500.

References

[1] A.B. Engin, A. Engin, Obesity and lipotoxicity, M/OL]. Cham: Springer International Publishing 960 (2017), 2022-09-15], http://link.springer.com/10.1007/978-3-319-48382-5.
[2] K.G.M.M. Alberti, R.H. Eckel, S.M. Grundy, et al., Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity[J/OL], Circulation 120 (16) (2009) 1640–1645, https://doi.org/10.1161/CIRCULATIONAHA.109.192644.
[3] Xinyu Dong, Guoliang Zou, Yubo Han, et al., Relationship between traditional Chinese medicine syndrome types and risk factors of metabolic syndrome, China Medicine 17 (9) (2022) 1390–1394.
[4] Xiao-Ping Cheng, L.I. Xiu-Ming, W.E.I. Hua, Investigation on the traditional Chinese medicine constitution of DampType of metabolic Syndrome：An analysis of 147 patients [J/OL], Journal of Guangzhou University of Traditional Chinese Medicine 38 (12) (2021) 2547–2551, https://doi.org/10.13359/j.cnki.gzxbtcm.2021.12.001.
[5] P.L. Chien, C.F. Liu, H.T. Huang, et al., Application of artificial intelligence in the establishment of an association model between metabolic syndrome, TCM constitution, and the guidance of medicated diet care[J/OL], Evid. base Compl. Alternative Med. 2021 (2021) 1–9, https://doi.org/10.1155/2021/5530717.
[6] M. Gollapalli, A. Alansari, H. Alkhorasani, et al., A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM[J/OL], Comput. Biol. Med. 147 (2022), 105757, https://doi.org/10.1016/j.compbiomed.2022.105757.
[7] S.K. Kalagotla, S.V. Gangashetty, K. Giridhar, A novel stacking technique for prediction of diabetes[J/OL], Comput. Biol. Med. 135 (2021), 104554, https://doi.org/10.1016/j.compbiomed.2021.104554.
[8] J. Shi, Z. Zhao, T. Jiang, et al., A deep learning approach with subregion partition in MRI image analysis for metastatic brain tumor[J/OL], Front. Neuroinf. 16 (2022), 973698, https://doi.org/10.3389/fninf.2022.973698.
[9] N. Shakhovska, P. Pukach, Comparative analysis of backbone networks for deep knee MRI classification models[J/OL], Big Data and Cognitive Computing 6 (3) (2022) 69, https://doi.org/10.3390/bdcc6030069.
[10] E.K. Choe, H. Rhee, S. Lee, et al., Metabolic syndrome prediction using machine learning models with genetic and clinical information from a nonobese healthy population[J/OL], Genomics & Informatics 16 (4) (2018) e31, https://doi.org/10.5808/GI.2018.16.4.e31.
[11] F. Karimi-Alavijeh, S. Jalili, M. Sadeghi, Predicting metabolic syndrome using decision tree and support vector, ARYA atherosclerosis 12 (3) (2016) 146.
[12] G.O. Gutiérrez-Esparza, O. Infante Vázquez, M. Vallejo, et al., Prediction of metabolic syndrome in a Mexican population applying machine learning algorithms [J/OL], Symmetry 12 (4) (2020) 581, https://doi.org/10.3390/sym12040581.
[13] Y. Tang, T. Zhao, N. Huang, et al., Identification of traditional Chinese medicine constitutions and physiological indexes risk factors in metabolic syndrome: a data mining approach[J/OL], Evid. base Compl. Alternative Med. 2019 (2019) 1–10, https://doi.org/10.1155/2019/1686205.
[14] S.J. Xia, B.Z. Gao, S.H. Wang, et al., Modeling of diagnosis for metabolic syndrome by integrating symptoms into physiochemical indexes[J/OL], Biomed. Pharmacother. 137 (2021), 111367, https://doi.org/10.1016/j.biopha.2021.111367.
[15] H.K. Wu, Y.S. Ko, Y.S. Lin, et al., The correlation between pulse diagnosis and constitution identification in traditional Chinese medicine[J/OL], Compl. Ther. Med. 30 (2017) 107–112, https://doi.org/10.1016/j.ctim.2016.12.005.
[16] S.H. Kuo, H.L. Wang, T.C. Lee, et al., Traditional Chinese medicine perspective on constitution transformations in perinatal women: a prospective longitudinal study[J/OL], Women Birth 28 (2) (2015) 106–111, https://doi.org/10.1016/j.wombi.2015.01.002.
[17] Jiang Huaiyan, Tan Lang, Li Shijie, et al. Hypertension Predicting Scheme by Analyzing Nutritional Ingredients Based on XGBoost Model [J]. Journal of Chongqing University: 1-15.
[18] Jianhua Zheng, Rong Zhu, Shuangyin Liu, et al., Research on college Students'Poverty identification model based on fusion of dual perturbation and kernel ELM (DP_KELM), Journal of Chongqing University of Technology(Natural Science) 35 (5) (2021) 243–252.
[19] G.E. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (1) (2004) 20–29.
[20] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System[C/OL]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vols. 785–794, ACM, San Francisco California USA, 2016 [2022-09-16], https://dl.acm.org/doi/10.1145/2939672.2939785.

[21] A. Prabha, J. Yadav, A. Rani, et al., Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier[J/OL], Comput. Biol. Med. 136 (2021), 104664, https://doi.org/10.1016/j.compbiomed.2021.104664.
[22] C. Dong, Y. Qiao, C. Shang, et al., Non-contact screening system based for COVID-19 on XGBoost and logistic regression[J/OL], Comput. Biol. Med. 141 (2022), 105003, https://doi.org/10.1016/j.compbiomed.2021.105003.
[23] J. Bergstra, R. Bardenet, Y. Bengio, et al., Algorithms for Hyper-Parameter Optimization[C/OL]//Advances in Neural Information Processing Systems:, vol. 24, Curran Associates, Inc., 2011, 2022-09-19], https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.
[24] I. Dewancker, M. Mccourt, S. Clark, Bayesian Optimization Primer-Sigopt, 2015 [M].
[25] J.X. Cui, B. Yang, Survey on bayesian optimization methodology and applications, Journal of Software 29 (10) (2018) 3068–3090 (in Chinese), http://www.jos.org.cn/1000-9825/5607.htm.
[26] W. Binsaeedan, S. Alramlawi, CS-BPSO: hybrid feature selection based on chi-square and binary PSO algorithm for Arabic email authorship analysis, Knowl. Base Syst. 227 (2021), 107224.
[27] M.A. Awadallah, M.A. AL-Betar, M.S. Braik, et al., An enhanced binary Rat Swarm Optimizer based on local-best concepts of PSO and collaborative crossover operators for feature selection, Comput. Biol. Med. (2022), 105675.
[28] K.N. Rajesh, R. Dhuli, T.S. Kumar, Obstructive sleep apnea detection using discrete wavelet transform-based statistical features, Comput. Biol. Med. 130 (2021), 104199.
[29] Shu-jiao Chen, L.I. Can-dong, L.A.I. Xin-mei, et al., Study on characteristics of traditional Chinese medicine syndrome of 160 patients with metabolism syndrome [D]. China Journal of Traditional Chinese Medicine and Pharmacy, March 30 (3) (2015).
[30] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 30.
[31] F.H. Wang, C.M. Lin, The utility of artificial neural networks for the non-invasive prediction of metabolic syndrome based on personal characteristics, Int. J. Environ. Res. Publ. Health 17 (24) (2020) 9288.