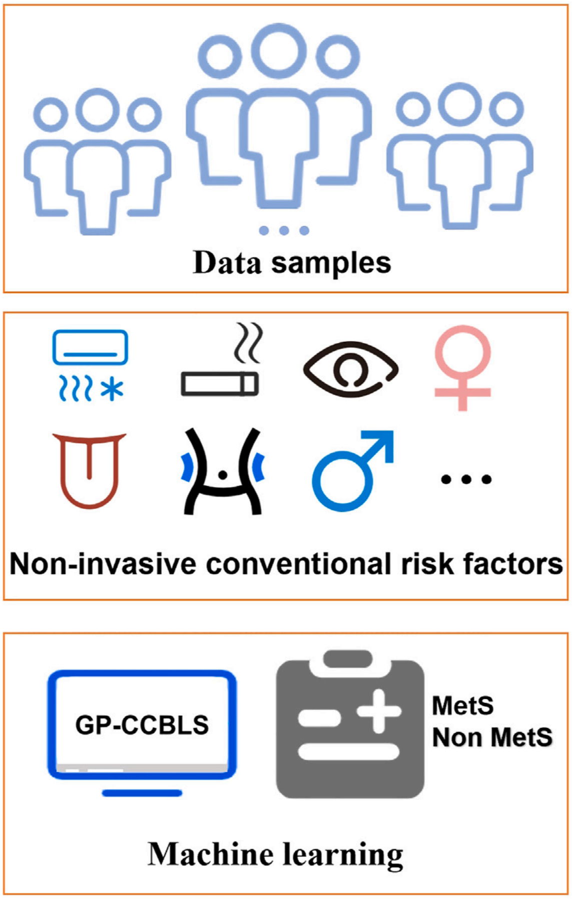
iScience

CellPress OPEN ACCESS

Article

Employing broad learning and non-invasive risk factor to improve the early diagnosis of metabolic syndrome



Junwei Duan, Yuxuan Wang, Long Chen, C. L. Philip Chen, Ronghua Zhang

jwduan@jnu.edu.cn (J.D.)
tzrh@jnu.edu.cn (R.Z.)

Highlights

15 noninvasive CRF are first utilized for early diagnosis of MetS

Broad learning is first proposed to improve the early diagnosis of MetS

The performance of our proposed GP-CCBLS model is superior to seven other models

Duan et al., iScience *27*, 108644
January 19, 2024 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.108644

iScience

CellPress
OPEN ACCESS

Article

Employing broad learning and non-invasive risk factor to improve the early diagnosis of metabolic syndrome

Junwei Duan,[1,6,7,*] Yuxuan Wang,[2] Long Chen,[3] C. L. Philip Chen,[4] and Ronghua Zhang[5,6,*]

SUMMARY

Metabolic syndrome (MetS) as a multifactorial disease is highly prevalent in countries and individuals. Monitoring the conventional risk factors (CRFs) would be a cost-effective strategy to target the increasing prevalence of MetS and the potential of noninvasive CRF for precisely detection of MetS in the early stage remains to be explored. From large-scale multicenter MetS clinical dataset, we discover 15 non-invasive CRFs which have strong relevance with MetS and first propose a broad learning-based approach named Genetic Programming Collaborative-competitive Broad Learning System (GP-CCBLS) with noninvasive CRF for early detection of MetS. The proposed GP-CCBLS model can significantly boost the detection performance and achieve the accuracy of 80.54%. This study supports the potential clinical validity of noninvasive CRF to complement general diagnostic criteria for early detecting the MetS and also illustrates possible strength of broad learning in disease diagnosis comparing with other machine learning approaches.

INTRODUCTION

Metabolic syndrome (MetS) is a group of metabolic risk factors whose main clinical manifestations include obesity, hyperglycemia (diabetes or impaired glucose regulation), dyslipidemia [high fasting triglyceride (TG) and/or low fasting high-density lipoprotein cholesterol (HDL-C)], and hypertension.[1,2] In epidemiological studies, the prevalence of MetS ranges from approximately 20% to 45% of the total population, with an average prevalence of 31%. It is estimated that the incidence rate of MetS will increase to about 53% by 2035.[3] The current diagnostic criteria for MetS were proposed by the Chinese Medical Association Diabetes Branch in 2017 which involves many important invasive test data such as triglycerides and high-density lipoprotein (HDL).[4] In addition, studies on the incidence of MetS have shown that hyperinsulinemia is also a common risk factor.[5] However, these data are highly dependent on medical equipment and resources; especially, regular monitoring of risk factors is the cornerstone of early detection and management the MetS.[6]

If patients do not have timely physical examinations and blood tests, they may miss the optimal treatment period and lead to exacerbation of the disease. However, the number of health check-ups in China in 2019 was 444 million, with a coverage rate of only 31.71%.[7] With the increasing incidence of MetS in the population, a novel early diagnostic model that does not rely on invasive risk factors may be a new idea to address the problem. In order to meet this need, we consider the noninvasive CRFs which have great value in disease diagnosis. For instance, early signs of disease in the heart, spleen, and stomach can be seen through the tongue, and visceral organ lesions can be seen from the look of the face.[8]

Moreover, many researchers have shown that machine learning can be used to build models for disease diagnosis. For example, Shimoda, Ichikawa, and Oyama[9] built machine learning models such as logistic regression (LR) based on health data collected by Japan's National Health Examination System to predict the likelihood of disease in participants and provide guidance to high-risk groups to improve their lifestyles. And some researches have focused on the diagnosis of MetS. For instance, decision tree (DT) is a popular machine learning method for the diagnosis of MetS due to the fact that it is easy to understand and its high accuracy.[10–12] In addition, a better machine learning model could be selected by comparing multiple evaluation metrics. A good case in point is that Karimi-Alavijeh, Jalili and Sadeghi[13] used support vector machine (SVM) and DT to predict the risk of MetS and the performance of SVM model was better than the DT model with accuracy, sensitivity and specificity of 75.7%, 77.4% and 74.0% respectively.

[1] College of Information Science and Technology, Jinan University, Guangzhou, Guangdong 511436, China
[2] Jinan University – University of Birmingham Joint Institute, Jinan University, Guangzhou, Guangdong 511436, China
[3] Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau 999078, China
[4] School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510006, China
[5] College of Pharmacy, Jinan University, Guangzhou, Guangdong 510006, China
[6] Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Jinan University, Guangzhou, Guangdong 511436, China
[7] Lead contact
*Correspondence: jwduan@jnu.edu.cn (J.D.), tzrh@jnu.edu.cn (R.Z.)
https://doi.org/10.1016/j.isci.2023.108644

However, current machine learning models for MetS diagnosis still rely heavily on clinical data, especially for some invasive factors such as blood glucose, which are not very practical in real-life situations. In this context, the main goal of this paper is to build a reliable and usable early diagnosis model for MetS through machine learning techniques that rely only on basic information and noninvasive risk factors.

RESULTS

Subjects

This study included 1849 participants aged 18–90 years who were admitted to and hospitalized in Guangdong Provincial Hospital of Traditional Chinese Medicine, Jiangsu Provincial Hospital of Traditional Chinese Medicine, and Xinjiang Traditional Chinese Medicine Hospital from November 2019 to November 2021, with the following conditions were excluded from the study population: (1) patients with malignancy; (2) patients with severe cardiac and renal insufficiency including blood creatinine clearance <30 mL/min, alanine aminotransferase ≥2.5 times the normal upper limit, total bilirubin ≥1.5 times the upper limit of normal, chronic cardiac insufficiency, cardiac function class III or above); (3) patients with new onset of cardiovascular disease within the last six months; (4) patients with acute infection; (5) pregnant or lactating women; (6) patients with secondary dyslipidemia, hypertension, abnormal blood glucose, (7) patients with Type 1 diabetes, (8) patients with hyperthyroidism or hypothyroidism, (9) patients who are taking corticosteroids, contraceptives, diet pills or other medications that affect their weight.

Methods

Based on extensive literature and expert consultation, a questionnaire was designed to cover five main categories, including basic information, habits and customs, syndrome types, main observed symptoms, all relevant test indicators of the subjects were collected through the questionnaires and medical records. Table 1 shows the characteristics of the study sample.

Diagnostic criteria

According to the 2017 unified criteria of the Diabetes Branch of the Chinese Medical Association, the diagnosis of MetS can be made when three or more of the following five items are met.

- Abdominal obesity (i.e., central obesity): male waist circumference ≥90 cm, female waist circumference ≥85 cm;
- Hyperglycemia: fasting blood glucose ≥6.1 mmol/L or 2 h postprandial blood glucose ≥7.8 mmol/L and/or diagnosed and treated for diabetes mellitus.
- Hypertension: blood pressure ≥130/85 mmHg and/or diagnosed with hypertension and receiving treatment;
- TG $\geq$ 1.70 mmol/L;
- HDL-C < 1.04 mmol/L.

Statistical processing

We use Statistical Product and Service Solutions (SPSS, version23.0) statistical software for the analysis. For different kinds of data, we used different statistical and testing methods. The data that conform to a normal distribution were represented by their mean ($\bar{x}$) and standard deviation (s) $\bar{x} \pm s$, and tested by ANOVA. The data that did not follow a normal distribution were examined by the Kruskal-Wallis test and shown with interquartile ranges M (P_{25}, P_{75}). The chi-squared test was applied for qualitative variables that were reported as percentages n (%). Statistically significant indicators (P-value <0.05) were screened out. Indicator characteristics were complex and large in number, so feature selection was required. Taking MetS as the dependent variable and all the above-mentioned statistically significant indicators were used as independent variables for correlation analysis. Spearman correlation analysis[14] was used to test the correlation between these statistically significant indicators and MetS, and the indicators with stronger correlations were selected as risk factors. In the experiment, the level of statistical significance was fixed at an alpha error of less than 5%. Detailed information about the noninvasive CRF can be seen in Table 2, and the results of the correlation analysis are displayed in Table 3.

Tongue color, fur color, and nature of pee belong to the main observation symptoms.

- The six indexes of tongue color are expressed respectively 1: light red; 2: red; 3: dark red; 4: dark; 5: magenta; 6: purple.
- The five indicators of fur color, referring to the color of the tongue coating, are expressed as follows, 1: white moss; 2: slightly yellow moss; 3: yellow moss; 4: gray moss; 5: black moss.
- The meanings of the four indicators of the nature of pee are 0: normal; 1: clear and long urine; 2: yellow urine; 3: frothy urine.

Table 3 shows the Spearman correlation coefficient between noninvasive CRF and MetS and the significance level of the correlation. The P-value represents whether the Spearman correlation coefficient has statistical significance. The statistical significance level is fixed at an alpha error of less than 5%, and P-value in Table 3 are all less than 0.05, indicating significant correlation between the noninvasive CRF and MetS. With regard to Spearman column, it represents the values of the correlation coefficients and the range of the correlation coefficient should be in [-1, 1]. If the correlation coefficient is positive, it indicates there is a positive correlation between MetS and this indicator; if the correlation coefficient is negative, it means a negative correlation between MetS and that indicator. Nonparametric statistical inference. New York: M. Dekker. Retrieved from EBSCO Publishing: e-book Collection on 3, 2016.]. For example, the age correlation coefficient with MetS is 0.185, which means that the likelihood of developing MetS will increase with age growth. In addition, the gender correlation coefficient with

Table 1. The characteristics of study data

	Female	Male
Anthropometric measurements		
n=1849	986 (53.33%)	863 (46.67%)
Age	58.73	55.46
Body fat rate	32.10	26.37
BMI (kg/m2)	24.93	25.86
Waist-hip ratio	0.93	0.97
Habits and customs		
Average daily air conditioning Duration in summer (hours/day)	6.98	8.67
Fatty diet	46 (4.67%)	99 (11.47%)
Smoking	23 (2.33%)	408 (47.28%)
Syndrome types		
Qi-deficiency	403 (40.87%)	326 (37.78%)
Yin-deficiency	259 (26.27%)	195 (22.60%)
Yang-deficiency	154 (15.62%)	82 (9.50%)
Blood-deficiency	42 (4.26%)	9 (1.04%)
Spleen-deficiency	410 (41.58%)	297 (34.41%)
Kidney-deficiency	124 (12.57%)	97 (11.24%)
Liver-depression	88 (0.89%)	32 (3.71%)
Air-stagnation	44 (0.45%)	30 (3.48%)
Congestion	486 (49.29%)	436 (50.52%)
Phlegm	99 (10.04%)	135 (15.64%)
Wet	148 (15.01%)	169 (19.58%)
Damp-heat	129 (13.08%)	204 (23.64%)
Main observation symptoms		
Nocturia frequency(times/night)	1.37	1.32
Nature of pee	392 (39.76%)	210 (24.33%)
Analytical variables		
Visceral Fat Index	8.89	12.01

(*Continued on next page*)

Table 1. *Continued*

	Female	Male
ALT (U/L)	18.39	25.78
AST (U/L)	18.97	20.66
Cr (μ mol/L)	64.90	82.71

Mean [$\bar{x}$] for continuous variables; n% for categorical variables.

BMI: body mass index; Waist-hip ratio: waist divided by hip; Fatty diet: dietary preference for fat; Syndrome types: a classification of disease states in Chinese medicine; ALT: alanine aminotransferase; AST: aspartate aminotransferase; Cr: creatinine.

Available for 1,849 persons.

MetS is −0.138, i.e., compared to females, males maybe more likely to develop MetS. The closer the absolute value of the correlation coefficient is to 1, the stronger the correlation is.

Broad learning system

The broad learning system (BLS) is a new type of flattened and incremental learning neural network proposed by Chen, and Liu.[15] As shown in Figure 1, BLS is an efficient learning system without deep architecture, which is designed based on the idea of linking neural networks with random vector functions.[16–18]

Suppose the training dataset $\{(A_k, B_k) \middle| A_k \epsilon R^a, B_k \epsilon R^b, k = 1, \cdots, N\}$, where N is the number of samples in the training set, a and b represent the dimension of input and output data respectively. It is assumed that there are n_1 groups of feature nodes, and each group contains p nodes. Therefore, the i th group of feature nodes can be expressed as

$$P_i = f\left(AW_{c_i} + \beta_{c_i}\right), i = 1, \cdots, n_1, \tag{Equation 1}$$

where $f(\cdot)$ is a mapping function. All feature nodes can be expressed as $P^{n_1} \triangleq [P_1 \cdot \cdot \cdot P_{n_1}], W_{c_i}, \beta_{c_i}$ are the weight matrix and bias term, which are randomly generated by the network.

Then, suppose there are n_2 groups of enhancement nodes, and each group contains q nodes. Thus, the j th group of enhancement nodes can be expressed as,

$$Q_j = g\left(P^{n_1} W_{d_j} + \beta_{dj}\right), j = 1, \cdots, n_2, \tag{Equation 2}$$

where $g(\cdot)$ is an activation function, W_{h_j}, β_{h_j} are randomly generated by the network, and all the enhancement nodes can be represented as $Q^{n_2} \triangleq [Q_1 \cdot \cdot \cdot Q_{n_2}]$.

Hence, the final broad learning network output can be expressed as,

$$B = HW, \tag{Equation 3}$$

where $H = [P^{n_1} | Q^{n_2}]$, and W is the output weight connecting the feature nodes and the enhancement nodes to the output layer.

Finally, the output weight W is solved by the following formula, while this optimization problem can also be an alternative to solve the pseudo-inverse directly in order to reduce the amount of computation.

$$\underset{W}{\operatorname{argmin}} \|B - HW\|_2^2 + \lambda \|W\|_2^2, \tag{Equation 4}$$

Equation 4 is the objective function of BLS, by setting the derivative of W to 0, the solution of the output weight can be obtained as

$$W = \left(\lambda I + H^T H\right)^{-1} H^T B, \tag{Equation 5}$$

where λ is the regularization parameter and I is the identity matrix.

Sparse autoencoder and collaborative-competitive representation

In BLS, a sparse autoencoder (SAE) is adopted to fine-tune the mapped features which are randomly generated at first. As mentioned above, the random feature P is obtained from $P = AW$, and W is randomly initialized. SAE adds L1 regularization on the basis of autoencoder, constraining most of the nodes in each layer to be zero, and only a few are not zero. In order to get sparse features, we need to solve the minimization problem in (6),

Table 2. Noninvasive CRF grouped by Mets [$\overline{x} \pm s$, M (P25, P75), n (%)]

Group	Quantity	Gender-Male	Eye anomaly	Smoking	Fatty diet	Congestion	Thirsty					
							(%)					
		(%)	(%)	(%)	(%)	(%)	Normal	Drink more	Not drinking much			
Patient	1107	579	621	309	112	623	511	558	38			
		(52.30)	(56.10)	(27.91)	(10.12)	(56.28)	(46.16)	(50.41)	(3.43)			
Non-patient	742	284	302	112	33	299	470	251	21			
		(38.27)	(40.70)	(15.09)	(4.45)	(40.30)	(63.34)	(33.83)	(2.83)			
P-value		<0.01	<0.01	<0.01	<0.01	<0.01	<0.01					
Group	Quantity	Tongue color						Fur color				
		(%)						(%)				
		1	2	3	4	5	6	1	2	3	4	5
Patient	1107	325	216	329	199	4	34	700	217	188	1	1
		(29.36)	(19.51)	(29.72)	(17.98)	(0.36)	(3.07)	(63.23)	(19.60)	(16.98)	(0.09)	(0.09)
Non-patient	742	363	126	141	103	5	4	543	133	66	0	0
		(48.92)	(16.98)	(19.00)	(13.88)	(0.67)	(0.54)	(73.18)	(17.92)	(8.89)	(0.00)	(0.00)
P-value		<0.01						<0.01				
Group	Quantity	BMI	Body fat rate	Waist-Hip Ratio	Age	Daily air conditioner usage time		Nocturia frequency	Nature of pee			
									(%)			
									0	1	2	3
Patient		26.06	1	29.8	61	8		1	587	69	190	261
	1107	(24.22,28.31)	(0.9,1)	(26.2,34.7)	(51,69)	(4,10)		(1,2)	(53.03)	(6.23)	(17.16)	(23.58)
Non-patient	742	23.24	0.9	28	56	6		1	514	41	97	90
		(21.37,25.49)	(0.9,1)	(23.3,32.9)	(43,64)	(1,10)		(0,2)	(69.27)	(5.53)	(13.07)	(12.13)
P-value		<0.01	<0.01	<0.01	<0.01	<0.01		<0.01	<0.01			

Median *x*, lower quartile P_{25} and upper quartile P_{75} for continuous variables not conforming to a normal distribution.
n% for categorical variables.
Available for 1,849 persons.

Table 3. Noninvasive CRF and MetS correlation analysis results

Variable	Spearman	P-value
Nocturia frequency (times/night)	0.183	<0.05
Nature of pee	0.173	<0.05
Tongue color	0.190	<0.05
Fur color	0.116	<0.05
Eye anomaly	0.151	<0.05
Thirsty	0.164	<0.05
Fatty diet	0.103	<0.05
Daily air conditioner usage time (hours/day)	0.135	<0.05
Smoking	0.133	<0.05
Congestion	0.164	<0.05
Age	0.185	<0.05
Gender	−0.138	<0.05
Body fat rate	0.362	<0.05
Waist-Hip Ratio	0.173	<0.05
BMI (kg/m^2)	0.404	<0.05

$$\underset{W}{\text{argmin}}\|A - P\tilde{W}\|_2^2 + \lambda\|\tilde{W}\|_1, \qquad \text{(Equation 6)}$$

where $\tilde{W}$ is the solution of SAE, λ is a regularization parameter greater than 0. There are many ways to find the solution of the above optimization problem, such as K-SVD,[19] Alternating Direction Multiplier Method (ADMM),[20] etc.

A collaborative-competitive representation based on classification model (CCRC) was first proposed by Yuan.[21] The model adds the competition term to the formula of collaborative representation-based classification (CRC),[22] and the objective function formula is as follows:

$$\underset{\beta}{\text{argmin}}\ \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_2^2 + \lambda_2\sum_{i=1}^{C}\|y - X_i\beta_i\|_2^2, \qquad \text{(Equation 7)}$$

where y is the label of data, X is the data used for training and c indicates how many categories there are. The first term is the collaborative term while the third term is the competition term, and the parameters λ_1, λ_2 are used to balance them. collaborative-competitive representation (CCR) uses all training data to cooperatively represent test samples while encouraging competing representations of distinct classes at the same time.[23]

Collaborative-competitive representation based broad learning system

In 2022, Wu and Duan replaced SAE in BLS with collaborative-competitive representation based autoencoder (CCRAE) in order to improve the feature representation, and proposed collaborative-competitive representation based broad learning system (CCBLS), which was motivated by the collaborative-competitive representation (CCR) methodology.[24] The objective function of the replaced autoencoder is:

$$arg\,\underset{\hat{W}}{\min}\left(\|A - G\widehat{W}\|_2^2 + \lambda_1\|\widehat{W}\|_2^2\right) + \lambda_2\sum\nolimits_{i=1}^{k_1}\left\|A - \overline{G^{(i)}}\widehat{W}\right\|_2^2, \qquad \text{(Equation 8)}$$

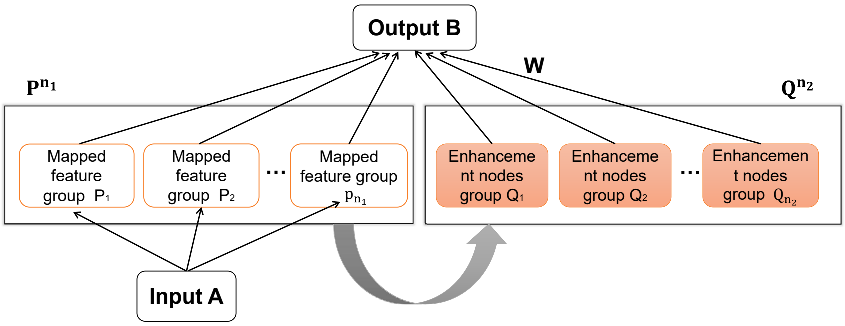

Output B
W
P^{n_1}
Q^{n_2}
Mapped feature group P_1
Mapped feature group P_2
...
Mapped feature group P_{n_1}
Enhancement nodes group Q_1
Enhancement nodes group Q_2
...
Enhancement nodes group Q_{n_2}
Input A

Figure 1. Structure of BLS

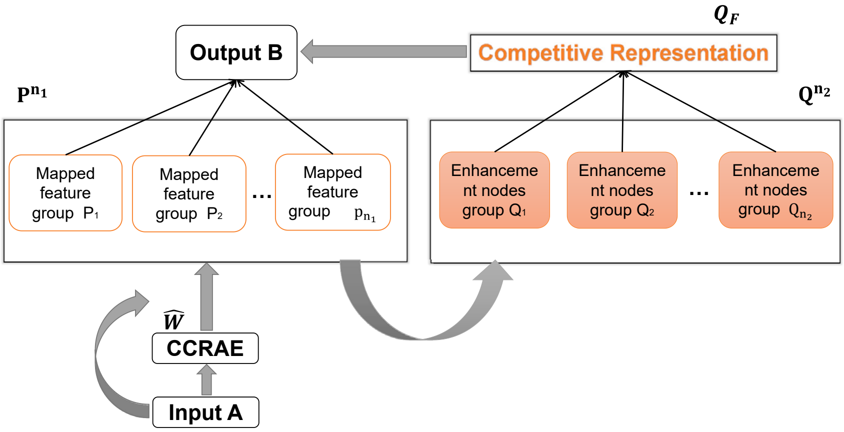

Q_F
Output B
Competitive Representation
P^{n_1}
Q^{n_2}
Mapped feature group P_1
Mapped feature group P_2
…
Mapped feature group p_{n_1}
Enhancement nodes group Q_1
Enhancement nodes group Q_2
…
Enhancement nodes group Q_{n_2}
$\widehat{W}$
CCRAE
Input A

Figure 2. Structure of CCBLS

where G is the randomly generated feature node, $\widehat{W}$ is the solution of CCRAE and A is the input. In addition, k_1 is the number of columns in every mapped feature group, $\overline{G^{(i)}} = [0,\cdots,0,G^{(i)},0,\cdots 0]$, and $G^{(i)}$ is the i th column of G. The first term expresses collaborative and the second term expresses competition, which is similar with Equation 7.

On the basis of the competitive representation, CCRAE and BLS, we can obtain the new network structure, namely CCBLS, as shown in the Figure 2.

It can be noted that the overall framework of the CCBLS is very similar to the BLS, with changes in the mapped features,

$$G = f(A\widehat{W}), \tag{Equation 9}$$

where $f(\cdot)$ is a mapping function, and $\widehat{W}$ is the weight generated by the CCRAE. In addition, as the mapped feature changes, the enhancement features will also change to the following equation:

$$Q = g(PW_q), \tag{Equation 10}$$

where $g(\cdot)$ is an activation function, and W_q is randomly generated by the network. However, this also brings about the problem that the randomly generated weights W_q do not allow us to ensure the performance of the enhancement features. Hence, we need to finetune the random weight W_q by introducing the competitive representation again. Assuming that the finetuned weight and finetuned enhancement features are W_F and Q_F respectively, which can be obtained through the following equations:

$$Q_F = QW_F, \tag{Equation 11}$$

$$\arg\min_{W_F} \sum_{j=1}^{k_2} \left\| \overline{Q_F{}^{(j)}} W_F - Q \right\|_2^2, \tag{Equation 12}$$

where k_2 is the number of columns in the enhancement layer, $\overline{Q_F^{(j)}} = [0,\cdots,0,Q_F^{(j)},0,\cdots 0]$, and $Q_F^{(j)}$ is the j th column of Q_F.

Genetic programming

Genetic programming (GP) is a kind of evolutionary algorithm, which can dynamically construct trees like mathematical formulas, as well as being suitable for feature construction due to its flexible representation.[25–27] It has been tested that GP can extract factors with incremental information from a limited amount of data and has been applied to stock selection.

Like many other evolutionary algorithms, GP starts with a randomly generated set of formulas, and the genetic operations in GP are suitable for individuals selected based on fitness probability, that is, better individuals are more likely to have more children than poor ones.[28] According to the different goals, it is possible to choose to use different fitness. The one with the highest fitness is selected as the parent, do genetic operations to produce new generation for the next iteration and repeat the process. Figure 3 shows the structure of genetic programming.

By simulating the process of genetic evolution in nature, we can gradually generate formula sets that satisfy a specific goal. The genetic evolution of organisms involves inheritance, mutation, and adaptation to the ecological environment. This is also true in the genetic programming algorithm, where there will also be operators such as copy, mutation, and crossover. As shown in Algorithm 1 and Figure 2, the algorithm has two termination conditions: (1) reaching the maximum fitness; (2) reaching the maximum number of generations. In this study, the maximum number of generations is used as the termination condition. As a supervised learning method, the advantage of GP is that it

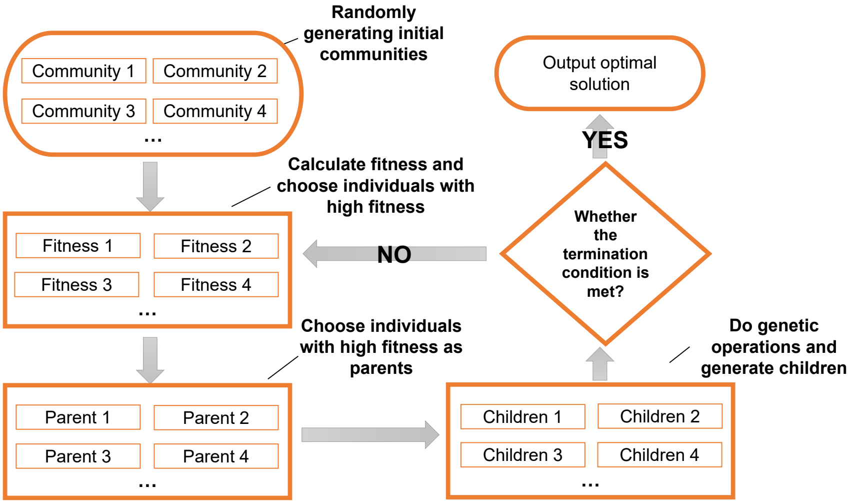

Randomly generating initial communities
Community 1
Community 2
Community 3
Community 4
...
Calculate fitness and choose individuals with high fitness
Fitness 1
Fitness 2
Fitness 3
Fitness 4
...
Choose individuals with high fitness as parents
Parent 1
Parent 2
Parent 3
Parent 4
...
Children 1
Children 2
Children 3
Children 4
...
Do genetic operations and generate children
Whether the termination condition is met?
NO
YES
Output optimal solution

Figure 3. Structure of genetic programming

can make full use of the powerful computing capability of computers for heuristic search while breaking through the limitations of human thinking to uncover some hidden factors.

Genetic programming collaborative-competitive broad learning system

As mentioned above, GP, as a suitable method for feature engineering, can use the strong computing power of computers to dig more information about factors that are hard to discover directly. Likewise, CCBLS has the advantage of introducing a competitive collaboration mechanism among features in the network. Considering the strength of both, we decided to introduce GP to CCBLS and propose a completely new model Genetic Programming Collaborative-Competitive Broad Learning System (GP-CCBLS). Figure 4 illustrates the overall framework of this model.

In GP-CCBLS, firstly, GP needs to be deployed to dig out more information from the original input data and to generate new factors through iterations. After that, the factors obtained from the iterations are used as input to the CCBLS and the final predictions are output.

Training process

The proposed method can judge whether a patient has MetS according to his/her external performance and living habits. 1849 samples were available in the dataset, including 1107 patients and 742 non-patients. We randomly selected 80% of the data as the training set and the remaining 20% as the test set. During the division process, we tried to maintain a balance between the number of sample categories in the training and test sets. By using Numpy and gplearn in Python 3.9, we implemented the above method.

In the process of the training experiment, we decided the values of the essential parameters by grid search. Firstly, we set a wide search scope with a large step. After experiments, we found that it performs better in some certain ranges, and then we further adjusted the search scope and reduced the step. For the parameters of CCBLS, n_1 is searched within the range of [1,200], and n_2 is searched within the range of [100,400]. As for the parameter of genetic programming, we finally chose the number of generations of evolution, generations = 18, which was also the end condition of our experiment. We set the number of individuals generated in each generation to 3000, that is, population_size = 3000, and we chose n_components = 20, which means that eventually 20 optimal children would be selected as the newly generated features.

Table 4 displays the training process of the GP, where the average length is growing as the number of generations increases. This means that the genetic operations performed become more and more complicated and at the same time, the training time may be extended. In

Algorithm 1. Genetic programming

```
1: An initial population created at random.
    2: repeat.
    3: Run every program and determine fitness.
    4: Choose one or two program(s) with fitness-based probabilities from the population to engage in genetic operations, and then apply the chosen genetic operations to produce new individuals.
    5: Until a solution is identified or other conditions are satisfied that would stop it.
    6: Return the optimal solution so far.
```

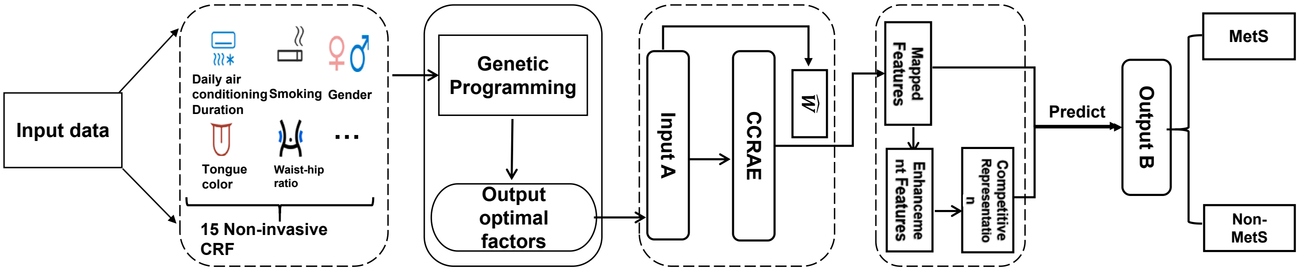


Figure 4. The overall framework of GP-CCBLS

addition, the algorithm will select the best individuals based on fitness for reproduction and gradually optimizes their fitness through iteration. In the early stages of iteration, the value of the fitness will increase as the number of iterations rises. As the number of iterations continues to increase, the individuals in the population gradually converge and the fitness tends to be stable. According to Figure 5, when the evolution reached to the sixth generation, the average fitness achieved a high level and did not improve significantly in the following generations. However, the fitness of the best individuals stayed at a high level.

Evaluation metrics

Receiver operator characteristic (ROC) curves were conducted and the area under the curve (AUC) was calculated to determine which models best displayed the risk of MetS. To conduct diagnostic test accuracy study, accuracy, sensitivity, specificity, F1-score, and precision were analyzed and defined as follows:

$$Accuracy = \frac{T\ P+T\ N}{T\ P+FP+FN+T\ N}, \tag{Equation 13}$$

$$Sensitivity = \frac{TP}{TP+FN}, \tag{Equation 14}$$

$$Specificity\ \ = \frac{TN}{TN+FP}, \tag{Equation 15}$$

$$Precision = \frac{TP}{TP+FP}, \tag{Equation 16}$$

$$F1\ -score = 2 \times \frac{Precision \times Sensicivity}{Precision+Sensicivity}, \tag{Equation 17}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Classification experiment results

Table 5 shows the result of the evaluation metrics (accuracy, sensitivity, specificity, F1-score, precision and AUC) for each model, where we also show the classification results on the same dataset using several traditional supervised learning methods. In addition, we further use 5-fold cross-validation to obtain average evaluation metrics in Table 6.

In previous studies, some traditional machine learning methods such as DT, SVM, KNN, and LR have all been used for the classification of MetS and achieved good experimental results. In addition, these traditional machine learning models have their own characteristics. LR is a probability-based model with strong interpretability, performs well in handling large-scale data and it is widely used in the medical field due to its simple model and fast training speed. KNN is a classification model based on distance measurement. As a non-parametric classification algorithm, KNN can adaptively learn the distribution of data and is easy to implement. DT is a classification model based on the tree structure, which can automatically select the most important features, making it very useful when processing high-dimensional data. Similarly, SVM is a model based on maximum margin classification and it also performs well in handling high-dimensional data. In order to compare with the BLS, we also included a Deep Neural Network (DNN) model. DNN is an artificial neural network composed of multiple hidden layers, used to solve complex nonlinear problems. In recent years, broad learning system has been widely applied to disease diagnosis and has shown better classification ability than other traditional machine learning models. For instance, Han, Liu, Chen, Xu, and Peng[29] predicted mortality in COVID-19 patients by BLS. This model achieved 94.50% sensitivity and 94.80% specificity in blood samples from 375 patients, performing better than other models and providing a reliable method for mortality prediction in COVID-19 patients. Therefore, we chose these five models and two BLS series models as the baseline models for comparison.

Table 4. GP training process of the dataset

Generation	Average length	Average fitness	Best individual length	Best individual fitness
1	8.91	0.123053	8	0.448338
2	5.54	0.270810	11	0.465023
3	5.17	0.337414	9	0.484392
4	7.35	0.375631	8	0.491079
5	8.33	0.393073	12	0.492437
6	9.22	0.399159	21	0.501052
7	9.01	0.401346	18	0.498767
8	9.11	0.401295	21	0.505983
9	10.24	0.404796	12	0.516635
10	13.31	0.402767	28	0.514059
11	18.30	0.414306	33	0.519095
12	21.37	0.421058	14	0.518305
13	22.71	0.417728	23	0.523636
14	22.22	0.409068	26	0.525935
15	21.05	0.407933	27	0.528289
16	21.37	0.407680	27	0.528182
17	20.60	0.403239	36	0.526666
18	19.88	0.399382	33	0.527850

Average length: the average program length of the generation.
Average fitness: the average program fitness of the generation.
Best individual length: the length of the best program in the generation.
Best individual fitness: the fitness of the best program in the generation.

Karimi-Alavijeh, Jalili, and Sadeghi have employed DT and SVM to predict the 7-year incidence of MetS and the accuracy were 0.757 (0.739) in SVM (DT) method. Similarly, SVM also outperformed DT in terms of accuracy in our study. What's more, K-Nearest Neighbors (KNN) and LR have also been used for the risk prediction in MetS, with external validation AUC of 0.780 (KNN) and 0.782 (LR), respectively.[30] In a recent study published in 2023,[31] DNN was used to develop a MetS classification and prediction model, which showed high accuracy and reliability. They constructed the DNN model, which consists of hidden layer with 16, 8 and 4 nodes and the output layer with one node, and the developed the DNN model of deep learning shown the improved accuracy compared with traditional models such as LR.

With regards to the experimental results in Table 5, BLS and CCBLS are better than the other five models in accuracy and AUC, but the performance in other evaluation metrics is not very ideal. In terms of the proposed model, the accuracy of GP-CCBLS improves about 7% compared to BLS and CCBLS. Meanwhile, all the remaining evaluation metrics have obvious increases. For accuracy, precision, F1-score and AUC, the GP-CCBLS all performs the best In Table 6, we can see that the similar results of performance can be

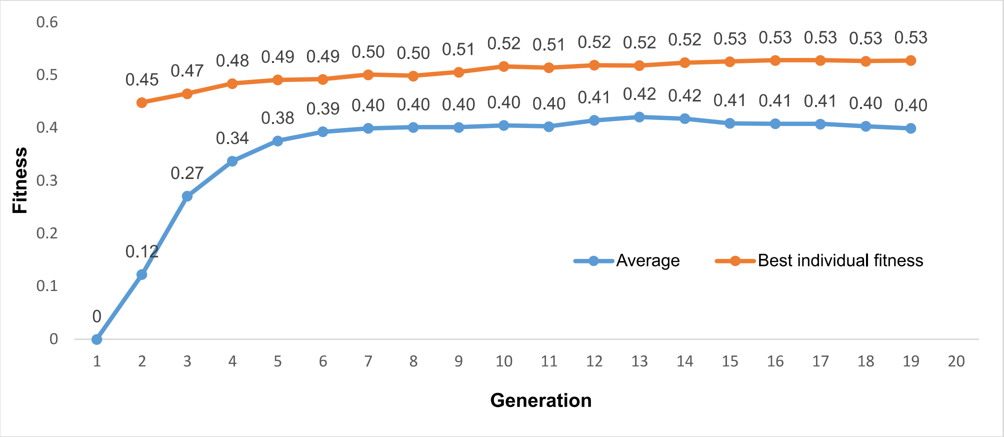


Figure 5. Learning curve of fitness

Table 5. Comparison of other machine learning methods

Method								
Metrics	DT	SVM	KNN	LR	DNN	BLS	CCBLS	GP-CCBLS
Accuracy	0.6703	0.7108	0.6892	0.7216	0.7000	0.7297	0.7324	**0.8054**
Sensitivity	0.6802	0.9549	0.7703	0.8514	0.9550	0.8784	0.8829	**0.9189**
Specificity	0.6554	0.3446	0.5676	0.5270	0.3176	0.5068	0.5068	**0.6351**
Precision	0.7475	0.6861	0.7277	0.7297	0.6773	0.7276	0.7286	**0.7907**
F1-score	0.7123	0.7985	0.7484	0.7859	0.7925	0.7959	0.7984	**0.8500**
AUC	0.6678	0.6498	0.6689	0.6892	0.6363	0.6926	0.6948	**0.7770**

DT: decision tree; SVM: support vector machine; KNN: k-nearest neighbors; LR: logistic regression; BLS: broad learning system; DNN: Deep Neural Network; CCBLS: collaborative-competitive representation based BLS; GP-CCBLS: genetic programming collaborative-competitive broad learning system.

obtained by 5-fold cross-validation and our proposed GP-CCBLS method achieves the best performance compared to seven other methods.

Validation of the new proposed model

To examine the validity of the proposed model, we collect some new samples for the experiment. This new dataset has 153 new participants, 79 of whom are with MetS (51.63%). In terms of validating our proposed model for the early detection of MetS, and 106 of the 153 participants were diagnosed with MetS. The proposed new method identified 69 of the 79 subjects as with MetS and 37 of the 74 as not having MetS, thus achieving a sensitivity of 87.34% and specificity of 50.00%. The validation results can be found in Table 7.

However, the results of the evaluation metrics are not very high because of the limited newly collected validation dataset.

Web app for early diagnosis of MetS

To better validate, explain and bring the model into real-life scenarios, we have created a demo that allows the patient to assess the risk of MetS as long as the patient inputs the 15 noninvasive CRFs, which can further support the development of early diagnosis of MetS. Using Gradio, a visual interface may be created for our model, and we can also perform input operations, interactions, and output conclusions with this open-source Python module.[32] Here, we provide a permanent public uniform resource locator (URL) address for this model (https://huggingface.co/spaces/WangYX/WYX_DEMOforMetS) as well as an illustration of the interactive user interface in Figure 6. On the left side of the interface, there are the 15 non-intrusive traditional risk factors that we need to input. After entering them, click the "Submit" button and the results will be displayed in the output section.

DISCUSSION

In this study, we analyze the clinical validity of noninvasive CRF to improve the diagnosis of future MetS using machine learning. We demonstrate that noninvasive CRF can be used as variables for the model and have good diagnostic performance and, more importantly, the advantage of not requiring blood tests (invasive and analytical variables) further enhances the general usefulness of the model and expands its application scenarios and scope. To the best of our knowledge, it is the first time for BLS to be used for the diagnosis of diseases based on noninvasive CRF, and it performs well compared to other more traditional popular machine learning methods. Based on the BLS and CCBLS models, we apply GP to extract more features, which provides a new approach for primary diagnosis of MetS. In addition, we compare the proposed model with other machine learning algorithms such as DT, SVM, LR and KNN that have been used to the diagnosis of MetS and further illustrate the advantages of GP-CCBLS. With these 15 noninvasive variables, the result shows that the proposed model presents 80.54% accuracy, 91.89% sensitivity, 79.09% precision, and 85% F1-score on the test data.

Table 6. Results of 5-fold cross-validation

	DT	SVM	KNN	LR	DNN	BLS	CCBLS	GP-CCBLS
Accuracy	0.6587	0.7236	0.6993	0.7161	0.6341	0.7231	0.7204	**0.7773**
Sensitivity	0.7173	0.9449	0.7814	0.8384	0.5930	0.8555	0.8726	**0.8991**
Specificity	0.5715	0.3936	0.5768	0.5337	0.6948	0.5256	0.4932	**0.5956**
Precision	0.7145	0.6993	0.7340	0.7286	0.7478	0.7304	0.7203	**0.7693**
F1-score	0.7155	0.8037	0.7568	0.7792	0.6370	0.7872	0.7889	**0.8288**
AUC	0.6444	0.6692	0.6791	0.6860	0.6439	0.6906	0.6829	**0.7474**

DT: decision tree; SVM: support vector machine; KNN: k-nearest neighbors; LR: logistic regression; BLS: broad learning system; DNN: Deep Neural Network; CCBLS: collaborative-competitive representation based BLS; GP-CCBLS: genetic programming collaborative-competitive broad learning system.

Table 7. Validation of the proposed model

	DT	SVM	KNN	LR	DNN	GP-CCBLS
Accuracy	0.6993	0.6144	0.6667	0.5686	0.6275	**0.6928**
Sensitivity	0.8354	0.9241	0.8101	0.6582	0.8987	**0.8734**
Specificity	0.5541	0.2838	0.5135	0.4730	0.3378	**0.5000**
Precision	0.6667	0.5794	0.6400	0.5714	0.5917	**0.6509**
F1-score	0.7416	0.7122	0.7151	0.6118	0.7136	**0.7459**
AUC	0.6947	0.6039	0.6618	0.5656	0.6183	**0.6867**

DT: decision tree; SVM: support vector machine; KNN: k-nearest neighbors; LR: logistic regression; DNN: Deep Neural Network; GP-CCBLS: genetic programming collaborative-competitive broad learning system.

Limitations of the study

There are also some limitations to this study. This study was conducted in some hospitals in different regions of China and represents only a portion of the Chinese population. It is possible to conduct model validation studies in more regions, and even prospective cohort studies are necessary to verify the accuracy of our proposed model. However, this study demonstrates that BLS has a wide range of application prospects in the field of disease diagnosis.

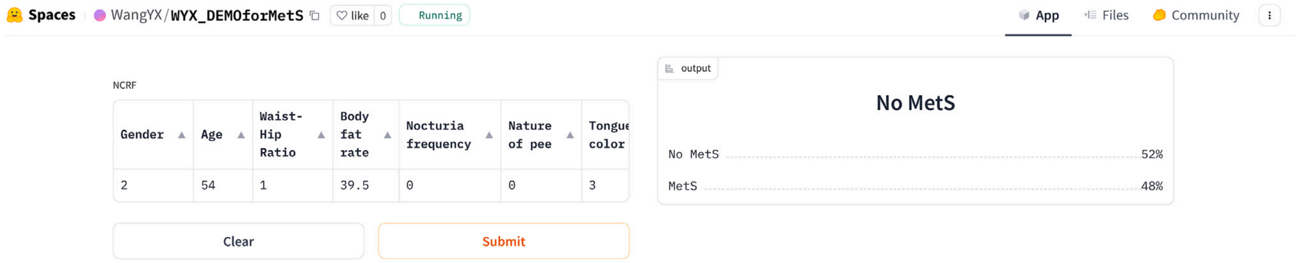
STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Study population
  - ○ Data collection
- METHOD DETAILS
  - ○ Disease case definitions
  - ○ Exclusion criteria
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Data processing
  - ○ General framework of modelling
  - ○ Model development
  - ○ Model evaluation
- ADDITIONAL RESOURCES

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC2002500, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011999 and in part by Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization under Grant 2021B1212040007, and in part by the Fundamental Research Funds for the Central Universities under Grant 11623201.



NCRF

Gender	Age	Waist-Hip Ratio	Body fat rate	Nocturia frequency	Nature of pee	Tongue color
2	54	1	39.5	0	0	3

output

No MetS

```
No MetS ........................ 52%
MetS ........................... 48%
```

Figure 6. Interactive user interface for MetS

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, study design, and funding acquisition, J.D.; Data analysis and interpretation, Y.W.; Validation, visualization and investigation, L.C.; Project administration and supervision, C.L.P.C. All authors reviewed the manuscript and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 24, 2023
Revised: September 12, 2023
Accepted: December 1, 2023
Published: December 5, 2023

REFERENCES

1. Nilsson, P.M., Tuomilehto, J., and Rydén, L. (2019). The metabolic syndrome – What is it and how should it be managed? Eur. J. Prev. Cardiol. *26*, 33–46.
2. Bishehsari, F., Voigt, R.M., and Keshavarzian, A. (2020). Circadian Rhythms and the Gut Microbiota: From the Metabolic Syndrome to Cancer. Nat. Rev. Endocrinol. *16*, 731–739.
3. Engin, A. (2017). The Definition and Prevalence of Obesity and Metabolic Syndrome. In Obesity and Lipotoxicity Advances in Experimental Medicine and Biology, A.B. Engin and A. Engin, eds. (Springer International Publishing), pp. 1–17.
4. Society, C.D. (2018). Guidelines for the prevention and control of type 2 diabetes in China (2017 Edition). Chinese Journal of Practical Internal Medicine *38*, 292–344.
5. Han, T.S., Williams, K., Sattar, N., Hunt, K.J., Lean, M.E.J., and Haffner, S.M. (2002). Analysis of obesity and hyperinsulinemia in the development of metabolic syndrome: San Antonio Heart Study. Obes. Res. *10*, 923–931.
6. Hsieh, S.D., and Muto, T. (2004). [A simple and practical index for assessing the risk of metabolic syndrome during routine health checkups]. Nihon Rinsho. *62*, 1143–1149.
7. China Statistical Yearbook (2020. http://www.stats.gov.cn/tjsj/ndsj/2020/indexeh.htm.
8. Xin-lu, W., Si-seng, T., and Yong-hong, Z. (2002). Disease Prediction of Traditional Chinese Medicine (China Medical Science and Technology Press).
9. Shimoda, A., Ichikawa, D., and Oyama, H. (2018). Prediction models to identify individuals at risk of metabolic syndrome who are unlikely to participate in a health intervention program. Int. J. Med. Inform. *111*, 90–99.
10. Worachartcheewan, A., Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2013). Quantitative population-health relationship (QPHR) for assessing metabolic syndrome. EXCLI Journal *12*, 569–583.
11. Miller, B., Fridline, M., Liu, P.Y., and Marino, D. (2014). Use of CHAID Decision Trees to Formulate Pathways for the Early Detection of Metabolic Syndrome in Young Adults. Comput. Math. Methods Med. *2014*, 242717.
12. Romero-Saldana, M., Fuentes-Jimenez, F.J., Vaquero-Abellan, M., Alvarez-Fernandez, C., Molina-Recio, G., and Lopez-Miranda, J. (2016). New non-invasive method for early detection of metabolic syndrome in the working population. Eur. J. Cardiovasc. Nurs. *15*, 549–558.
13. Karimi-Alavijeh, F., Jalili, S., and Sadeghi, M. (2016). Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler. *12*, 146–152.
14. Kendall, M., and Gibbons, J.D. (1990). Rank Correlation Methods (Charles Griffin Book Series (Oxford University Press).
15. Chen, C.L.P., and Liu, Z. (2018). Broad learning system: An effective and efficient incremental learning system without the need for deep architecture. IEEE Trans. Neural Netw. Learn. Syst. *29*, 10–24.
16. Pao, Y.-H., Park, G.-H., and Sobajic, D.J. (1994). Learning and generalization characteristics of the random vector functional-link net. Neurocomputing *6*, 163–180.
17. Tyukin, I.Y., and Prokhorov, D.V. (2009). Feasibility of Random Basis Function Approximators for Modeling and Control (IEEE Control Applications, (CCA) & Intelligent Control, (ISIC) (IEEE)), pp. 1391–1396.
18. Pao, Y.-H., and Takefuji, Y. (1992). Functional-link net computing: theory, system architecture, and functionalities. Computer *25*, 76–79.
19. Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. *54*, 4311–4322.
20. Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. FNT. in Machine Learning *3*, 1–122.
21. Yuan, H., Li, X., Xu, F., Wang, Y., Lai, L.L., and Tang, Y.Y. (2018). A collaborative-competitive representation based classifier model. Neurocomputing *275*, 627–635.
22. Zhang, L., Yang, M., and Feng, X. (2011). International Conference on Computer visionIEEE, 2011 (IEEE), pp. 471–478.
23. Li, Z.-Q., Sun, J., Wu, X.-J., and Yin, H.-F. (2020). Multiplication fusion of sparse and collaborative-competitive representation for image classification. Int. J. Mach. Learn. Cybern. *11*, 2357–2369.
24. Wu, G., and Duan, J. (2022). BLCov: A novel collaborative–competitive broad learning system for COVID-19 detection from radiology images. Eng. Appl. Artif. Intell. *115*, 105323.
25. Koza, J.R. (1992). Genetic Programming, on the Programming of Computers by Means of Natural Selection. A Bradford Book (MIT Press).
26. Banzhaf, W., Nordin, P., Keller, R.E., and Francone, F.D. (1998). Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and its Applications (Morgan Kaufmann Publishers Inc.).
27. Tran, B., Xue, B., and Zhang, M. (2016). Genetic programming for feature construction and selection in classification on high-dimensional data. Memet. Comput. *8*, 3–15.
28. Poli, R., Langdon, W.B., McPhee, N.F., and Koza, J.R. (2008). A Field Guide to Genetic Programming. Lulu. Com. With Contributions by JR Koza.
29. Han, R., Liu, Z., Chen, C.L., Xu, L., and Peng, G. (2020). Mortality Prediction for COVID-19 Patients via Broad Learning System. 2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), pp. 837–842.
30. Zhang, H., Chen, D., Shao, J., Tang, L., Wu, J., Xue, E., and Ye, Z. (2023). Construction, Validation, and Comparison of a Metabolic Syndrome Risk Prediction Model Based on KNN Algorithm [J/OL]. Chongqing Medicine. https://kns.cnki.net/kcms/detail/50.1097.R.20230403.0959.002.html.
31. Kim, H., Heo, J.H., Lim, D.H., and Kim, Y. (2023). Development of a Metabolic Syndrome Classification and Prediction Model for Koreans Using Deep Learning Technology: The Korea National Health and Nutrition Examination Survey (KNHANES) (2013-2018). Clin. Nutr. Res. *12*, 138–153.
32. Abid, A., Ali, A., Ali, A., Khan, D., Alfozan, A., and Zou, J. (2019). Gradio: hassle-free sharing and testing of ML models in the wild. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.02569.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Validation data	This study	https://github.com/WYX-ID/Data-sample
Training and testing data	This study	https://github.com/WYX-ID/GPCCBLS
Software and algorithms		
BLS	Gong, Xinrong et al. "Research review for broad learning system: Algorithms, theory, and applications." IEEE Transactions on Cybernetics 52.9 (2021): 8922-8950.	https://broadlearning.ai/
GPCCBLS	This study	https://github.com/WYX-ID/GPCCBLS
DT	Pedregosa Fabian et al. "Scikit-learn: Machine learning in Python.", Journal of machine Learning research 12 (2011): 2825–2830.	https://scikit-learn.org/stable/modules/tree.html
SVM	Pedregosa Fabian et al. "Scikit-learn: Machine learning in Python.", Journal of machine Learning research 12 (2011): 2825–2830.	https://scikit-learn.org/stable/modules/svm.html
KNN	Pedregosa Fabian et al. "Scikit-learn: Machine learning in Python.", Journal of machine Learning research 12 (2011): 2825–2830.	https://scikit-learn.org/stable/modules/neighbors.html#regression
LR	Pedregosa Fabian et al. "Scikit-learn: Machine learning in Python.", Journal of machine Learning research 12 (2011): 2825–2830.	https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
DNN	Hyerim Kim et al. "Development of a Metabolic Syndrome Classification and Prediction Model for Koreans Using Deep Learning Technology: The Korea National Health and Nutrition Examination Survey (KNHANES) (2013–2018), Clinical Nutrition Research, (2023) 12(2): 138–153.	https://github.com/WYX-ID/DNN
Other		
Visual interface	This study	https://huggingface.co/spaces/WangYX/WYX_DEMOforMetS

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Junwei Duan (jwduan@jnu.edu.cn).

Materials availability

This paper did not result in the generation of novel reagents.

Data and code availability

Data have been deposited at https://github.com/WYX-ID/GPCCBLS. They are publicly available as of the date of publication. All original code has been deposited at https://github.com/WYX-ID/GPCCBLS and is publicly available as of the date of publication. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study population

This study included 1849 participants aged 18-90 years who were admitted to and hospitalized in three specific geographical areas in China, namely Guangdong Provincial Hospital of Traditional Chinese Medicine, Jiangsu Provincial Hospital of Traditional Chinese Medicine, and Xinjiang Traditional Chinese Medicine Hospital from November 2019 to November 2021. Among them, 863 were male and 986 were female.

Data collection

Based on extensive literature and expert consultation, a questionnaire was designed to cover five main categories, including basic information, habits and customs, syndrome types, main observed symptoms, all relevant test indicators of the subjects were collected through the questionnaires and medical records. All participants are Chinese and signed an informed consent. The study protocol was approved by the Ethics Committee of Guangdong Provincial Hospital of Chinese Medicine (Approval No. BF2020-177-01).

METHOD DETAILS

Disease case definitions

According to the 2017 unified criteria of the Diabetes Branch of the Chinese Medical Association, the diagnosis of MetS can be made when three or more of the following five items are met:

- Abdominal obesity (i.e. central obesity): male waist circumference $\geq$ 90 cm, female waist circumference $\geq$ 85 cm;
- Hyperglycemia: fasting blood glucose $\geq$ 6.1 mmol/L or 2 hours postprandial blood glucose $\geq$ 7.8 mmol/L and/or diagnosed and treated for diabetes mellitus.
- Hypertension: blood pressure $\geq$ 130/85 mmHg and/or diagnosed with hypertension and receiving treatment;
- TG $\geq$ 1.70 mmol / L;
- HDL-C < 1.04 mmol / L.

Exclusion criteria

The following conditions were excluded from the study population: (1) patients with malignancy; (2) patients with severe cardiac and renal insufficiency including blood creatinine clearance <30 ml/min, alanine aminotransferase $\geq$2. 5 times the normal upper limit, total bilirubin $\geq$1. 5 times the upper limit of normal. chronic cardiac insufficiency, cardiac function class III or above); (3) patients with new onset of cardiovascular disease within the last six months; (4) patients with acute infection; (5) pregnant or lactating women; (6) patients with secondary dyslipidemia, hypertension, abnormal blood glucose, (7) patients with Type 1 Diabetes, (8) patients with hyperthyroidism or hypothyroidism, (9) patients who are taking corticosteroids, contraceptives, diet pills or other medications that affect their weight.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing

1849 samples were available in the dataset, including 1107 patients and 742 non-patients. We randomly selected 80% of the data as the training set and the remaining 20% as the test set. During the division process, we tried to maintain a balance between the number of sample categories in the training and test sets. By using Numpy and gplearn in Python 3.9, we implemented the above method. In addition, for different kinds of data, we used different statistical and testing methods. The data that conform to a normal distribution were represented by their mean ($\bar{x}$) and standard deviation (s) $\bar{x} \pm s$, and tested by analysis of variance (ANOVA). The data that did not follow a normal distribution were examined by the Kruskal-Wallis test and shown with interquartile ranges M (P_{25}, P_{75}), where P_{25} is the lower quartile and P_{75} is the upper quartile. The chi-square test was applied for qualitative variables that were reported as percentages n (%), n is the number of samples for each class. Furthermore, we analyzed the relationship between statistically significant indicators and Metabolic Syndrome disease using Spearman correlation. In the experiment, the statistics were performed using the IBM SPSS Statistics 23 program, and the level of statistical significance was fixed at an alpha error of less than 5%.

General framework of modelling

Firstly, 15 non-invasive conventional risk factors were selected as input data by correlation analysis. For detailed information about the non-invasive CRF and the results of the correlation analysis, see the statistical processing section of the paper. We evaluated models on the test set and further created a random split of the internal training set and test set five times in an 8:2 ratio (five-fold cross-validation). The evaluation results are presented in the article, and it can be seen in the classification experiment results part. Moreover, we evaluated on a new dataset that was not part of the previous training and testing sets. The detailed procedures are explained in the rest of this section.

Model development

BLS extracts features from the input data and generates mapped feature layer and enhancement layer. The mapped feature layer and enhancement layer are connected to the output layer to generate labels. CCBLS is a collaborative-competitive representation based broad learning system model, SAE in BLS is replaced by CCRAE in the mapped feature layer of CCBLS, and further fine-tuning the enhancement layer using a competitive representation mechanism. Genetic programming is a suitable method for feature engineering, and it can use the strong computing power of computers to dig more information about factors that are hard to discover directly. We decided to introduce GP to CCBLS and propose a completely new model Genetic Programming Collaborative-Competitive Broad Learning System. In the process of the training experiment, we decided the values of the essential parameters by grid search. Firstly, we set a wide search scope with a large step. After experiments, we found that it performs better in some certain ranges, and then we further adjusted the search scope and reduced the step. For the parameters of CCBLS, n_1 is searched within the range of [1,200], and n_2 is searched within the range of [100,400]. As for the parameter of genetic programming, we finally chose the number of generations of evolution, generations=18, which was also the end condition of our experiment. We set the number of individuals generated in each generation to 3000, that is, population_size=3000, and we chose n_components=20, which means that eventually 20 optimal children would be selected as the newly generated features.

Model evaluation

The performance of all models was evaluated in the test set (20% sample), the five-fold cross-validation and new validation dataset. In addition, receiver operator characteristic (ROC) curves were conducted and the area under the curve (AUC) was calculated to determine which models best displayed the risk of MetS. To conduct diagnostic test accuracy study, accuracy, sensitivity, specificity, F1-score, and precision were analyzed in the paper.

Moreover, we compare and analyze our proposed model with other popular models. For example, some traditional machine learning methods such as DT, SVM, KNN, and LR have all been used for the classification of metabolic syndrome and achieved good experimental results in previous study. Additionally, in order to compare with the BLS, we also included a Deep Neural Network (DNN) model. According to the results, our proposed GP-CCBLS method achieves the best performance compared to other methods.

ADDITIONAL RESOURCES

To better validate, explain and bring the model into real-life scenarios, we have created a demo that allows the patient to assess the risk of MetS as long as the patient inputs the 15 non-invasive conventional risk factors, which can further support the development of early diagnosis of MetS. We provide a permanent public uniform resource locator (URL) address for this: https://huggingface.co/spaces/WangYX/WYX_DEMOforMetS.