



OPEN A supported decision-making model for idiopathic pulmonary fibrosis based on feature screening and optimized neural network

Hua Ye^{1,3}, Wenjie Gong^{1,3}, Ping Yuan², Ruiqi Zhang¹, Beibei He¹ & Wei Lin¹✉

Idiopathic pulmonary fibrosis (IPF) severely impairs human respiratory function, with an increasing incidence and mortality. Treatment based on syndrome differentiation is the characteristics and essence of traditional Chinese medicine (TCM). This study proposes an interpretable intelligent classification model—the MIV-GA-LM-BP (MGLB) model—to support syndrome differentiation in TCM for IPF. Based on 956 real-world clinical cases, the mean impact value (MIV) algorithm was employed for feature screening to identify key symptom-syndrome relationships and improve model interpretability. A hybrid optimization strategy combining the Levenberg–Marquardt (LM) and genetic algorithm (GA) was applied to enhance the convergence speed and generalization performance of the BP neural network. Comparative experiments with GRA, PCA, and PSO demonstrated that the MGLB model achieved the highest classification accuracy of 81.22%, outperforming other models in both accuracy and stability. More importantly, the MIV-based feature screening enables transparent mapping between symptoms and syndromes, aligning well with TCM diagnostic logic. The proposed model not only provides a standardized reference for IPF diagnosis but also offers a new methodological framework for developing AI-driven TCM diagnostic tools. It contributes to the modernization and standardization of TCM and supports the integration of intelligent systems into clinical decision-making.

Keywords Idiopathic pulmonary fibrosis, Traditional Chinese medicine, Feature screening, Syndrome classification, Neural network

Idiopathic pulmonary fibrosis (IPF) is a class of scarring chronic lung diseases with complex etiology, and is the most common and severe form of idiopathic interstitial pneumonia¹, with unclear etiology, unpredictable course², high misdiagnosis, high mortality, high recurrence rate and poor prognosis³. At present, there is no uniform assessment method and diagnostic standard for this disease, which brings great disturbance to patients and society⁴ and has become a major respiratory disease threatening public health. Therefore, it is urgent to promote the improvement of standardized diagnosis and treatment of IPF.

Traditional Chinese medicine (TCM) is the essence of the wisdom of the Chinese nation and is widely recognized as one of the effective means of treating diseases due to its unique methodological system, rich diagnostic and therapeutic techniques, the advantages of low side effects and good efficacy⁵. TCM data is characterized by nonlinearity, ambiguity, unstructuredness and multidimensionality⁶. As computer technology advances incessantly, machine learning algorithms extract feature learning from complex data and process it, and its application to the field of TCM is increasingly rich in research. The results of a large number of studies^{7–10} indicate that back propagation (BP) neural network models are suitable for syndrome classification. However, BP neural network is not without its limitations like slow convergence, difficulty in guaranteeing network generalization ability, easy to fall into the local minima, and large dependence on the initial weights^{11,12}. These drawbacks can significantly affect the accuracy of model classification. Levenberg–Marquardt (LM) algorithm is a modified algorithm based on the BP neural network, which can effectively address the limitations of slow convergence speed and weak generalization ability associated with the BP neural network¹³. Genetic algorithm (GA) is an optimization method with parallel random search, which can achieve global search to prevent the BP

¹School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China.

²School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China.

³Hua Ye and Wenjie Gong have contributed equally to this work. ✉email: 3012557902@qq.com

neural network from getting trapped in local optimum¹⁴, and optimize the initial weights and thresholds of the network to further enhance model performance.

The incidence and prevalence rates of IPF show an increasing trend year by year¹⁵, and the study of TCM auxiliary diagnosis and treatment of IPF has important practical value. While previous studies have applied machine learning to TCM diagnosis, most have prioritized classification accuracy over model interpretability. In contrast, this study introduced the mean impact value (MIV) algorithm for syndrome-specific feature screening, achieving a transparent mapping between symptoms and syndromes. Additionally, the proposed MIV-GA-LM-BP (MGLB) model was constructed based on 956 real TCM cases of IPF patients, ensuring its clinical relevance and practicality. Therefore, this study provides valuable references in both methodological integration and practical application. As shown in Fig. 1, this experiment utilizes the effective medical case data of IPF treatment in TCM, combines various machine learning algorithms to explore the rules and connections between symptoms and syndromes of the disease, and applies the MIV algorithm to screen the key influencing factors of the symptom features, avoiding the redundant information brought by the input features. The GA-LM-BP neural network model with better fit and higher prediction accuracy provides more informative results for the diagnosis of the TCM syndrome classification of IPF and helps to exert the characteristics and advantages of TCM in preventing and treating major and difficult diseases. These contributions represent not only an application of machine learning to TCM syndrome classification but also a meaningful integration tailored for interpretability and real-world deployment, setting a foundation for future intelligent TCM diagnostic systems.

Related works

Syndrome differentiation in TCM is a diagnostic thinking process guided by the theories of TCM to clarify the nature of the disease and make a judgement based on the four diagnostic information¹⁶. In recent years, the classification methods of AI technology are well adapted to non-linear, complex and fuzzy TCM data, and the use of AI technology to assist in TCM syndrome differentiation and diagnosis has become a research theme of great significance¹⁷.

Machine learning models in TCM syndrome classification

The current research on intelligent syndrome differentiation in TCM mainly includes decision tree (DT), random forest (RF), Extreme Gradient Boosting (XGBoost), support vector machine (SVM), K-nearest neighbor (KNN), etc.^{18–29}. Each of the above algorithms has its own strengths in dealing with high-dimensional data, data of different sizes and complex non-linear relationships, and most of the research focuses on the processing of data and the construction of models, mainly through the comparison of multiple algorithms to derive a single algorithm's ability to predict the syndrome. Previous studies on intelligent syndrome differentiation in TCM have made less mention of feature screening and optimization algorithms, and fewer of them have incorporated treatment based on syndrome differentiation thinking, resulting in poor interpretability of the constructed models and limited application in the real world. Clinical application needs to follow the principle of treatment based on syndrome differentiation, play the interpretable role of feature screening, and establish a closed loop, so that the model can have the interpretability in line with the clinical reality, and truly provide clinical decision support. However, most of these works focus solely on prediction accuracy and do not consider model interpretability or alignment with TCM diagnostic logic.

BP Neural networks in TCM applications

Although the existing intelligent TCM syndrome differentiation models have achieved good results, they still need to be optimized and should continue to be researched in depth so as to ensure the rapid development of the modernisation of TCM. BP neural network is one of the most widely used neural network models³⁰, and the results of a large number of researches have shown that it is a suitable model for TCM syndrome classification^{9,10,31}. Shenghao Yang et al.³² proposed a BP neural network-based prediction model for the syndrome classification of gastrointestinal disease chronic atrophic gastritis, which adopts the correlation-based feature selection method and improves the initialization of the BP weights based on the Gaussian distribution method to achieve good results. Despite their success, standard BP networks suffer from slow convergence, weak generalization ability, and sensitivity to initial weights and thresholds, which can lead to unstable predictions and limited practical use in clinical settings.

Optimization strategies for BP neural networks

To address the limitations of the standard BP network, researchers have proposed optimization strategies. Ye Wang³³ et al. found that the method of optimizing BP neural network based on artificial bee colony (ABC) algorithm can find the mapping relationship between TCM symptoms and TCM syndromes and greatly improves the accuracy of TCM syndrome differentiation, and then the information of TCM syndrome differentiation auxiliary diagnosis and treatment was realized, indicating that it is feasible to use the ABC-BP neural network algorithm in the auxiliary diagnosis and treatment of TCM syndrome differentiation. Zhang Mingqi³⁴ found that after ensemble learning and integration of BP neural network, the evaluation indexes of the TCM syndrome prediction model for liver cirrhosis reached the most balanced state with excellent performance. All of the above methods apply the standard or improved BP neural network to the field of TCM, but ignore the principle of diagnosis and treatment and the interpretability of the model. In real-world diagnosis and treatment, the emphasis is on comprehensive judgement based on individual differences of patients, and the connection between symptoms and syndromes in the neural network modelling process is determined by the connection weights and probabilities generated, which is not the reasoning of TCM syndrome differentiation. Therefore, the modelling cannot rely solely on the automated processing of the algorithm, but must fully consider the core thinking of TCM and follow the principle of diagnosis and treatment. The weights are determined according

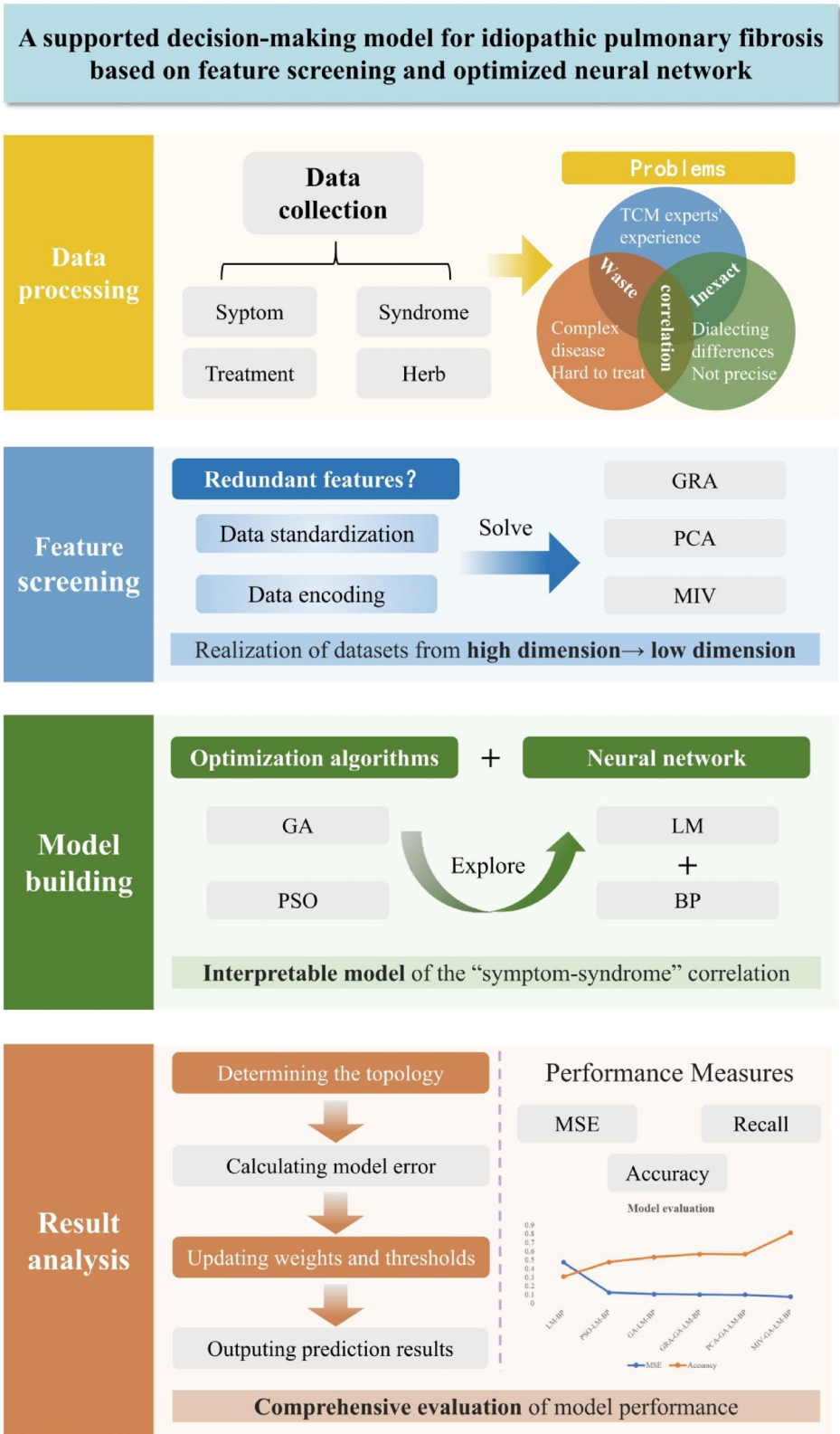


Fig. 1. Structure drawing of this paper.

to the correlation between the actual clinical symptoms and syndromes to ensure the clinical applicability and effectiveness of the model.

Current research has made some progress in the use of AI technology to assist TCM syndrome differentiation, but there is still a need for more in-depth exploration and research on model interpretability and precise

symptom-syndrome relationships. “Interpretability” is an important condition for AI technology to be trusted and used in TCM clinical diagnosis and treatment, and previous intelligent syndrome differentiation research has mainly been devoted to the study of the statistical relationship between symptoms and syndromes³⁵. Many researchers have implemented various TCM syndrome classification models, but there is a lack of classification method that highlight symptoms of key diagnostic significance for specific syndrome during the syndrome differentiation process. Since models such as neural networks are not interpretable and can not understand the relationship between symptoms and syndromes from the perspective of knowledge of TCM syndrome differentiation and diagnosis, calculating the exact value of the “symptoms-syndromes” relationship in TCM to enhance the interpretability of the model is the focus of intelligent syndrome differentiation research in TCM. The presence of too many redundant and irrelevant features in the original data will directly affect the performance of machine learning algorithms. Feature screening methods can effectively remove irrelevant and redundant features and improve classification accuracy. TCM syndrome differentiation focuses on individual differences and dynamic changes, and contains a large amount of complex information, and the selection of specific symptoms related to the diagnosis is crucial for the establishment of the syndrome classification model.

In this paper, we explore the application of various optimization algorithms in the intelligent TCM syndrome classification model to improve its prediction accuracy in response to the shortcomings of the standard BP neural network, and further modify the BP neural network by combining the global optimal solution searching ability of GA and particle swarm optimization (PSO) algorithms to overcome its typical shortcomings in practical applications. Importantly, exploring the use of grey relational analysis (GRA), principal component analysis (PCA), and MIV feature dimensionality reduction algorithms for the complex and multi-dimensional characteristics of TCM symptom data helps to enhance the trust in model prediction and improve the reliability of the model.

Data preprocessing

Data source

Data quality directly determines the upper limit of the model, and due to the limited data available, most existing work relies heavily on datasets that are not directly derived from clinical cases. The data for this study are derived from both the literature (open literature and medical monographs) and clinical data supported by National Key R&D Program of China (Grant No. 2018YFC1704104). Literature data were selected from VIP, Wanfang, and CNKI databases. Cross-searches were conducted using keywords “IPF”, “medical cases”, “experience”, “Famous senior TCM practitioners”, “National Master of TCM” and so on. The search time was from the establishment of the database to September 2022. Medical monographs data were manually consulted in national, provincial, and municipal contemporary medical monographs about IPF diagnosis and treatment by famous TCM experts. Cases of famous TCM experts’ experience in the diagnosis and treatment of IPF were collected such as *The Collection of Academic Thoughts and Clinical Experience of Du Yumao*, *Clinical Experience of Senior Physician Gao Yimin*, *Medical Theories and Cases of National Master Hong Guangxiang*, etc. Clinical data were collected from outpatient follow-up cases of famous TCM experts in Southwest China. A total of 956 datasets were screened according to the inclusion and exclusion criteria in Table 1.

According to the *Diagnostic Criteria for TCM Symptoms of Idiopathic Pulmonary Fibrosis (2019 version)*³⁶, the criteria was formulated by the Internal Medicine Branch of CACM, Lung Disease Branch of CMAM and CACM by integrating the outcomes of statistic, artificial neural network and Delphi method on the analysis of the data of the medical cases, and the final artificial standardization of the syndrome types are 8, after the discussion of experts in combination with the clinical practice. They are: 249 cases of syndrome of lung qi deficiency complicated with phlegm and stasis obstructing the collaterals, 151 cases of syndrome of lung qi deficiency complicated with accumulation and binding of phlegm and heat, 136 cases of syndrome of qi deficiency in the lung and kidney complicated with phlegm and stasis obstructing the collaterals, 132 cases of syndrome of lung dryness with yin deficiency, 115 cases of syndrome of lung qi deficiency, 108 cases of syndrome of qi deficiency in the lung and kidney, 61 cases of syndrome of lung qi deficiency complicated with turbid phlegm obstructing the lung, and 4 cases of syndrome of liver qi invading the lung.

Standardized research is the basis for achieving the accuracy of syndrome identification³⁷. In this paper, different expressions of the same symptom are unified and described. Deleting descriptions without statistical significance, such as good appetite, good sleep, normal bowel movement, normal urination, etc. We finally obtained 267 symptoms with a total frequency of 11,034, which were divided into common symptoms and four diagnostic symptoms. The common symptoms include 224 symptoms such as gasping, cough with little phlegm, white phlegm, etc., with a total frequency of 6,478 times, while the four diagnostic symptoms include

Inclusion criteria	Exclusion criteria
(1) Medical case records of nationally, provincially, and municipally renowned TCM practitioners treating IPF (≥ 3 consultations) (2) Cases meeting the diagnostic criteria of the Diagnostic Criteria for TCM Symptoms of Idiopathic Pulmonary Fibrosis (2019 version) (3) Complete information on the four diagnostic methods and prescription data for treatment (4) Treatment primarily involves oral Chinese herbal decoctions, including empirical formulas and self-formulated formulas (5) Case records are detailed and reliable	(1) The primary diagnosis is not IPF, or there are other concomitant diseases (2) The source of the medical case records is unclear (3) Information from the four diagnostic methods or data on syndrome-based prescriptions and medication is incomplete (4) Treatment measures do not primarily involve oral Chinese herbal medicine (e.g., acupuncture, acupoint injection, topical application, etc.) (5) Reviews, animal experiments, case reports, or duplicate publications

Table 1. The inclusion and exclusion criteria.

43 symptoms such as white coating, thin coating, thready pulse, etc., which include 26 types of tongue and 17 types of pulse, with a total frequency of 4,556 times.

Data encoding

The data of common symptoms and four diagnostic symptoms after screening and normalization were coded. The top 20 symptoms ranked by frequency of common symptoms and four diagnostic symptoms were counted separately, as shown in Table 2.

To facilitate computer recognition, for common symptoms, the code "0" means no such symptom and "1" means having such symptom. Degrees of symptoms light, medium and heavy were coded as 1, 2, 3, such as dry cough symptoms are divided into mild dry cough, moderate dry cough, severe dry cough, and it will be merged into the dry cough column of dataset and coded as 1, 2, 3. Symptoms of the same category were coded as 1, 2, 3, etc. in descending order of frequency, e.g., sticky phlegm, frothy phlegm, clear thin phlegm, etc. were merged into one column of phlegm nature and coded as 1, 2, 3, etc. in descending order of frequency. The four diagnostic symptoms were coded in eight columns of data: tongue color, tongue with stasis, form of the tongue, coating texture, coating color, pulse 1, pulse 2, and pulse 3, respectively. For symptoms of the same type, each column is coded starting with 1 and increasing by 1 unit in turn until all symptoms of that type have been coded. When both symptoms are presented in a column, we start coding at 1.5 and increasing by 1 unit in sequence. If the coating nature symptom column contains thin, greasy, thick, and scanty coating, it is coded 1, 2, 3, 4, etc., in that order, with 1.5 indicating the presence of both thin and greasy coating, 2.5 indicating the presence of both greasy and thick coating, and so forth.

The dataset after adopting this coding rule contains a total of 76 symptom feature dimensions. Since insufficient data can adversely impact the accuracy of training the neural network model, the first 6 types of syndromes (number of cases > 100) totaling 891 data were selected for inclusion in the experiment. The coding was accomplished by using a one-hot code, with syndrome of lung qi deficiency complicated with phlegm and stasis obstructing the collaterals as [1, 0, 0, 0, 0, 0], syndrome of lung qi deficiency complicated with accumulation and binding of phlegm and heat as [0, 1, 0, 0, 0, 0], syndrome of qi deficiency in the lung and kidney complicated with phlegm and stasis obstructing the collaterals as [0, 0, 1, 0, 0, 0], syndrome of lung dryness with yin deficiency as [0, 0, 0, 1, 0, 0], syndrome of lung qi deficiency as [0, 0, 0, 0, 1, 0], and syndrome of qi deficiency in the lung and kidney as [0, 0, 0, 0, 0, 1]. Table 3 is a table of the dataset after coding was completed.

Feature screening

The selection of training sample influencing factor features will determine the algorithmic model³⁸. Multi-dimensional features can provide more comprehensive information. However, irrelevant or unrepresentative features will affect the model prediction performance³⁹. So before building the algorithmic model, it is necessary to extract feature of the symptoms of the dataset to screen out the key influencing factors in the sample features. Currently, the main feature selection methods include filter, embedded and wrapper, and the common feature extraction algorithms are Pearson correlation coefficient (PCC)⁴⁰, kernel principal component analysis (KPCA)⁴¹, latent dirichlet allocation (LDA)⁴², independent component analysis (ICA)⁴³, GRA⁴⁴, PCA⁴⁵, and MIV⁴⁶. Methods widely employed in the field of TCM are GRA algorithm and PCA algorithm^{47–49}. This study found that the MIV algorithm applied to the classification of TCM syndrome has a strong interpretability through pre-experimentation. The experiment compares the performance of the above three algorithms to reduce the dimensionality of the symptom features of the dataset.

GRA

The GRA algorithm is a multi-factor correlation evaluation method that shows unique advantages in solving grey system and incomplete information problems for determining the relationship and degree of influence between different factors. If a particular comparison sequence has a high correlation with the reference sequence, it can indicate that the factor has a high impact on the target. The so-called degree of correlation essentially refers to the correlation between different sequences by comparing the degree of similarity between them⁵⁰, which can be expressed by the correlation coefficient $\varphi_i(k)$ defined in Eq. (1).

Common symptoms				Four diagnostic symptoms			
Symptom	Frequency	Symptom	Frequency	Symptom	Frequency	Symptom	Frequency
Gasping	395	Dry mouth	174	White coating	534	Deep pulse	210
Cough with little phlegm	387	Rapid breathing	168	Thin coating	403	Dusky tongue	147
White phlegm	321	Poor sleep	144	Thready pulse	387	Light red tongue	139
Mild cough	310	Dyspnea 2	143	Yellow coating	307	Moderate pulse	126
Moderate cough	283	Dry and itchy throat	129	Red tongue	301	Thick coating	104
Shortness of breath	283	Cough with moderate phlegm	123	Wiry pulse	299	Floating pulse	97
Chest oppression	273	Soreness and weakness of waist and knees	109	Crimson tongue	290	Scanty coating	63
Fatigued spirit and lack of strength	270	Spontaneous sweating	86	Slippery pulse	268	Pale tongue	59
Sticky phlegm	212	Dry stool	83	Greasy tongue	245	Weak pulse	55
Poor appetite	200	Aggravated by movement	82	Rapid pulse	243	Enlarged tongue	51

Table 2. Symptom frequency statistics (top 20).

NO	Gasp	Aggravated by movement	Dry cough	...	Pulse 1	Pulse 2	Pulse 3	Syndrome						
1	0	1	0	...	2	0	0	1	1	0	0	0	0	0
2	0	0	0	...	2	0	0	1	1	0	0	0	0	0
3	0	0	0	...	2	0	0	1	1	0	0	0	0	0
4	0	0	3	...	1	2	0	1	1	0	0	0	0	0
5	0	1	2	...	1	2	0	1	1	0	0	0	0	0
6	0	0	1	...	2	0	0	1	1	0	0	0	0	0
7	0	0	0	...	2	6	0	1	1	0	0	0	0	0
8	0	0	0	...	1	6	0	1	1	0	0	0	0	0
9	0	0	0	...	1	2	0	1	1	0	0	0	0	0
...
883	1	0	0	...	1	5	0	6	0	0	0	0	0	1
884	0	0	0	...	6	0	0	6	0	0	0	0	0	1
885	0	0	0	...	6	0	0	6	0	0	0	0	0	1
886	0	0	0	...	5	6	0	6	0	0	0	0	0	1
887	0	0	0	...	5	6	0	6	0	0	0	0	0	1
888	1	0	0	...	2	7	12	6	0	0	0	0	0	1
889	1	0	0	...	1	5	6	6	0	0	0	0	0	1
890	0	1	0	...	3	5	9	6	0	0	0	0	0	1
891	0	0	0	...	3	5	9	6	0	0	0	0	0	1

Table 3. Encoded data sets.

$$\varphi_i(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|} \tag{1}$$

where $\varphi_i(k)$ is the correlation coefficient between the *i*th subsequence X_i and the reference sequence X_0 , which essentially characterizes the degree of correlation between the reference sequence and the comparison sequence, $x_0(k)$ is the reference sequence (dependent variable), i.e., the value of the syndrome column data, and $x_i(k)$ is the comparison sequence (independent variable), i.e., the value of all the symptom column data. In the GRA method, the reference sequence X_0 and the comparison sequence X_i play a crucial role, min and max denote the minimum and maximum values respectively, and ρ is the distinguishing coefficient between 0 and 1, and the correlation coefficient obtained is generally more accurate when 0.5 is taken.

The correlation value corresponding to each data in each data column is too scattered. In order to facilitate the overall comparison, it is proposed to average the correlation coefficients of each data column as the comparison value of the degree of correlation between the comparison series and the reference series⁵¹. Symptom correlations are calculated as shown in Eq. (2).

$$r_i = \frac{\sum_{k=1}^n \varphi_i k}{n} \tag{2}$$

where r_i is the degree of correlation and N is the total number of samples.

In this paper, the GRA method was applied to assess the correlation of symptom-influencing factors in the classification of TCM syndromes by processing and analyzing the relevant data. We constructed the correlation sequence and used GRA method to rank the influencing factors, and statistically counted the data whose symptom correlation value is greater than 0.967, as shown in Table 4.

The correlation degree, between 0 and 1, indicates the degree of similarity and correlation between each symptom and the syndrome. A higher value indicates a stronger correlation and a closer relationship between symptoms and syndromes, and thus the correlation degree is higher. Combined the correlation values for all the symptoms to get the ranking of each symptom. For the current 76 symptoms, tongue color had the highest correlation (correlation: 0.985), followed by coating color (correlation: 0.983). To ensure that we find the number of influencing factors that make the prediction accuracy optimal, symptom factors with a correlation of less than 0.963 are excluded, and the remaining 55 symptom factors are left as inputs to the neural network model.

PCA

The PCA algorithm, which maps the original data onto a new set of linearly independent composite variables by linear transformation, is a commonly used multivariate data analysis method. It aims to reduce the number of features and retain the maximum information value of the input data integrally⁵², which helps to simplify the processing of complex problems and improve the efficiency of algorithms and model building. Used for data dimensionality reduction, the PCA algorithm maps high-dimensional data into a low-dimensional space by selecting the most informative principal components, which helps to reduce redundant features and improve the prediction accuracy of the model, solving the problem of dimensionality catastrophe while making it easier to analyze, visualize and process the data. The steps for screening the symptom factors using PCA are as follows.

Symptom	Correlation	Rank	Symptom	Correlation	Rank
Tongue color	0.985	1	Suffocation	0.968	14
Coating color	0.983	2	Feverish sensations in palms and soles	0.967	15
Pulse 1	0.983	3	Shallow breathing and less speech	0.967	16
Coating texture	0.981	4	Hoarseness	0.967	17
Pulse 2	0.98	5	Mental exhaustion	0.967	18
Cough	0.98	6	Stomach gastric upset	0.967	19
Expectoration	0.977	7	Sallow facial complexion	0.967	20
Phlegm color	0.972	8	Yellow urine	0.967	21
Gasping	0.972	9	Phlegm nature	0.967	22
Distending headache	0.968	10	Nasal congestion	0.967	23
Dry skin and Pruritus	0.968	11	Hard to cough out	0.967	24
Pain in the subcostal region	0.968	12	Cough with blood in the phlegm	0.967	25
A sticky, greasy taste	0.968	13			

Table 4. Symptom correlation value (correlation value ≥ 0.967).

NO	Eigenvalue	Corresponding contribution rate (%)	Cumulative contribution rate (%)	NO	Eigenvalue	Corresponding contribution rate (%)	Cumulative contribution rate (%)
1	3.3335	4.38	4.38	11	1.6214	2.13	31.67
2	2.9917	3.93	8.31	12	1.6035	2.11	33.78
3	2.3982	3.15	11.46	13	1.5352	2.02	35.79
4	2.3146	3.04	14.50	14	1.5098	1.98	37.78
5	2.1194	2.78	17.29	15	1.4547	1.91	39.69
6	2.0561	2.70	19.99	16	1.3804	1.81	41.50
7	1.9682	2.59	22.58	17	1.3419	1.76	43.27
8	1.8464	2.43	25.00	18	1.3329	1.75	45.02
9	1.7511	2.30	27.30	19	1.3032	1.71	46.73
10	1.7021	2.24	29.54	20	1.2507	1.64	48.37

Table 5. Symptom descending eigenvalues and contribution values (top 20).

First of all, symptom-influencing factors are standardized and normalized to eliminate the impact of different attributes between different indicators on the comparison of variables, as shown in Eq. (3). Secondly, the covariance matrix of the sample was calculated by using Eq. (4), and then its eigenvalues and the corresponding eigenvectors are deduced. The corresponding contribution rates (eigenvalues as a percentage of the sum of the eigenvalues) and cumulative contribution rates of all the principal elements are calculated. The top 20 cases ranked by the eigenvalues are shown in Table 4. The top 20 statistical symptom descending eigenvalues, corresponding contribution rate and cumulative contribution rate are shown in Table 5.

$$x_i^* = \frac{x_i - \text{mean}(X_i)}{\sigma(X_i)} \tag{3}$$

Let the variable type matrix of the original sample point X be $X = (x_{ij})_{n \times m} = (X_1, X_2, \dots, X_m)$, where $X_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$, $i = 1, 2, \dots, m$. Where X_i is the i -th influencing factor of each sample, $\text{mean}(X_i)$ is the mean of the influencing factor, and $\sigma(X_i)$ is the standard deviation of the influencing factor.

$$C_x = \begin{bmatrix} \text{cov}(x_1^*, x_1^*) & \text{cov}(x_1^*, x_2^*) & \cdots & \text{cov}(x_1^*, x_m^*) \\ \text{cov}(x_2^*, x_1^*) & \text{cov}(x_2^*, x_2^*) & \cdots & \text{cov}(x_2^*, x_m^*) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_n^*, x_1^*) & \text{cov}(x_n^*, x_2^*) & \cdots & \text{cov}(x_n^*, x_m^*) \end{bmatrix} \tag{4}$$

$$\text{cov}(x_i^*, x_j^*) = E \{ [x_i^* - E(X_i^*)] [x_j^* - E(X_j^*)] \}$$

where $i = 1, 2, \dots, n$ $j = 1, 2, \dots, m$, The $E(x)$ function represents the mathematical expectation, and X_i^* denotes the result after standardization and normalization for the i -th influence factor.

The selection of p principal components is based on their cumulative contribution, with the aim of achieving it up to 85%, and corresponding components can retain the original information well and simplify the problem. The eigenvectors $a = (a_1, a_2, \dots, a_p)^T$ of the first p principal components are calculated to obtain the principal component expression as $F_i = a_{i1} X_1^* + a_{i2} X_2^* + \dots + a_{im} X_m^*$, where $a_i = (a_{i1}, a_{i2}, \dots, a_{im})$, $i = 1, 2, \dots, p$. Therefore,

the first 51 components are selected as principal components, and these 51 principal components are used to describe the relationship between these 76 symptom-influencing factors and can basically summarize these 76 symptom variables. There is a linear relationship between the screened principal components and the original symptom features. Symptom features are selected by the PCA algorithm, which consists of 76 original symptom features reduced dimension to 51 principal components as inputs to the neural network model.

MIV

The MIV algorithm, one of the optimal algorithms for evaluating the correlation of indicators in a neural network⁵³, determines the importance of the influence of input variables on the output results. Variables with a high degree of influence are selected after screening by the MIV algorithm, and the input variables for constructing the BP neural network model are reduced. This method plays a significant role in increasing the training accuracy of the model and reducing the error. The steps for screening symptom-influencing factors using the MIV algorithm are as follows.

A matrix of symptom-influencing factors $X_{m \times n}$ is constructed as the original input data, with the number of rows m denoting the number of patient cases and the number of columns n denoting the number of symptom characteristics. The sample data values of the symptom-influencing factors are increased or decreased by δ (δ is taken as 10% in this paper) to obtain two new input matrices X_1 and X_2 respectively, as shown in Eq. (5). Two new input matrices X_1 and X_2 are fed into the constructed BP neural network model and two outputs Y_1 and Y_2 are obtained. The MIV values of the symptom-influencing factors corresponding to each syndrome type are calculated as shown in Eq. (6). The absolute value of the MIV value reflects the relative importance of the input indicators to the output results, which can be ranked according to the absolute value, and the contribution and cumulative contribution rate of each symptom-influencing factor of each syndrome can be calculated to obtain the ranking of the relative importance of the influence of each symptom on each type of syndrome. The formula for calculating the contribution of the symptom influencing factor C_i is shown in Eq. (7).

$$X_{m \times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad X_1, X_2 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k}(1 \pm \delta) & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2k}(1 \pm \delta) & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & \cdots & \cdots & x_{kn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mk}(1 \pm \delta) & \cdots & x_{mn} \end{bmatrix} \quad (5)$$

$$MIV_i = \frac{1}{m} \sum_{j=1}^m (Y_1 - Y_2) \quad (6)$$

$$C_i = \frac{|MIV_i|}{\sum_{i=1}^n |MIV_i|} \times 100\% \quad (7)$$

After processing the dataset with the MIV algorithm, the MIV values corresponding to the symptoms for each syndrome and their contribution are obtained and shown in Table 6.

Each type of syndrome whose cumulative contribution is less than 95 percent are coded, and other symptoms are coded as 0. If the symptom code in a column is all 0, then the column of this symptom is deleted. Also remove symptom columns with a frequency of less than 10. Symptoms are statistically analyzed and after processing

Syndrome of lung qi deficiency complicated with phlegm and stasis obstructing the collaterals	MIV	Contribution rate (%)	Cumulative contribution rate (%)	Syndrome of lung qi deficiency complicated with accumulation and binding of phlegm and heat	MIV	Contribution rate (%)	Cumulative contribution rate	...
cough	0.006402	11.1884	11.1884	Tongue color	0.01044	14.4498	14.4498	
gasping	0.006377	11.1455	22.3339	Coating color	0.0057	7.8891	22.3389	
pulse 1	0.005291	9.2473	31.5812	Coating nature	0.0057	7.8883	30.2272	
tongue color	0.00503	8.7901	40.3714	Pulse 1	0.00516	7.1454	37.3725	
shortness of breath	0.004291	7.4992	47.8705	Gasping	0.00516	7.1358	44.5083%	
expectoration	0.003073	5.3711%	53.2416%	Cough	0.0046	6.3653%	50.8736%	
mental fatigue and lassitude	0.002547	4.4505%	57.6921%	Pulse 2	0.00441	6.1037%	56.9773%	
coating texture	0.00211	3.6868%	61.3788%	Expectoration	0.00379	5.2382%	62.2155%	...
...	
thirst	2.07E-05	0.0362%	99.9121%	Clear runny nose	2.4E-05	0.0332%	99.9586%	
blood in phlegm	1.81E-05	0.0317%	99.9438%	Yellow urine	1.3E-05	0.0181%	99.9767%	
a stick, greasy taste	1.55E-05	0.0271	99.9708	Cold hands and feet	7.8E-06	0.0108	99.9876	
edema of lower limbs	9.2E-06	0.0161	99.9869	Aggravated at night	5.4E-06	0.0075	99.9950	
gastric discomfort	7.48E-06	0.0131	100.0000	Tinnitus	3.6E-06	0.0050	100.0000	

Table 6. MIV values and their contribution to the symptoms corresponding to each syndrome.

by the MIV algorithm the symptom feature dimension is 55, and finally, the neural network model takes the 55-dimensional symptom dataset as input.

GA-LM-BP model construction

BP

BP neural network is a popular multilayer feed-forward network⁵⁴ and one of the most commonly used neural network models⁵⁵. The main idea is to input data samples, and through continuous learning and feedback along the direction of gradient descent. The weights and thresholds of the network are repeatedly adjusted and trained to minimize the sum of squared errors, ensuring that the output value gradually approaches the desired value. The BP neural network model can be regarded as a supervised approach self-learning training model. Its training process has three main layers: input layer, hidden layer and output layer. The schematic structure of a typical three-layer BP neural network is shown in Fig. 2.

The number of neurons in the input layer is generally equal to the number of features in the sample. The number of neurons in the output layer is generally equal to the number of output result categories. The number of neurons in the hidden layer is determined by an empirical formula $\sqrt{m+n+a}$. Where, m is equal to the number of neurons in the input layer; n is equal to the number of neurons in the output layer; a is a constant between 1 and 10. The BP neural network uses gradient descent, which is prone to local minima and slow convergence. A drawback will be improved in this paper by the optimization algorithms introduced subsequently. The different layers are connected to each other by weight matrixes, and the initialization weights and thresholds of the network are generated randomly, leading to uncertainty in the evaluation results. In this paper, the initialization weights and thresholds will be determined through subsequent introduction of optimization algorithms.

LM

The LM algorithm is an optimization algorithm for nonlinear least squares problems⁵⁶. Mathematically, the algorithm is known as an improved version of the Gauss–Newton method, which provides more stable and efficient solutions to nonlinear least squares problems. The main idea is to continuously adjust the model parameters through an iterative process to minimize the residuals between the fitted function and the observed data.

The LM algorithm incorporates the damping factor into Newton's method, combining the strengths of both Newton's method and gradient descent method⁵⁷, which show the advantages and properties of Newton's method and gradient descent method, respectively, with different damping factors. The iterative formula in the optimization process is shown in Eq. (8).

$$u_{k+1} = u_k - [J^T(u_k)J(u_k) + \mu I]^{-1} J^T(u_k) e(u_k) \quad (8)$$

where u_k is the k th iteration control input sequence; J denotes the Jacobi matrix; I is the unit matrix; μ denotes the damping factor, which is a constant greater than zero; When μ gradually increases, the algorithm is similar to the gradient descent method with slow learning speed and global convergence. When μ gradually decreases, the algorithm is similar to Newton's method with fast learning speed and local convergence; $e(u_k)$ is the resulting error.

The LM algorithm can locally converge and converge quickly, with good stability⁵⁸. It is one of the most recommended basic optimization algorithms in MATLAB training attributes, and its introduction as an optimization algorithm into BP neural networks can effectively improve the model convergence efficiency and

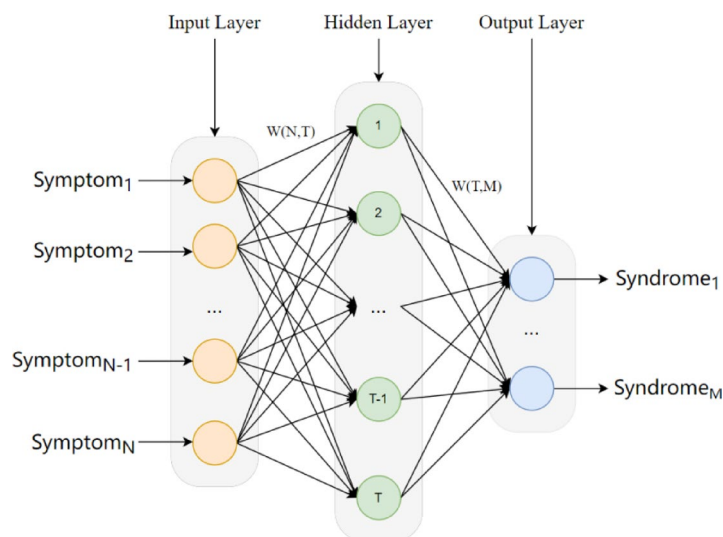


Fig. 2. Schematic structure of BP neural network.

learning rate while significantly mitigating the risk of sticking in local minima⁵⁹. The adjustment formula for the weights and thresholds of the network is shown in Eq. (9).

$$W_{(k+1)} = W_{(k)} - [J^T(X) J(X) + \mu I]^{-1} J^T(X) e(X) \tag{9}$$

where X denotes the input sample of the BP neural network; $W_{(k)}$ denotes the weight vector at the kth iteration of the BP neural network; the network output is obtained by learning and training the sample dataset through the model; the weights and thresholds are updated based on the error between the output and the desired output, resulting in the acquisition of a new vector $W_{(k+1)}$.

GA

GA is a heuristic algorithm that simulates the process of biological heredity and evolution, with good global search ability⁶⁰. The algorithm understands the data individual problem as a “chromosome”, and evaluates the advantages and disadvantages of the “chromosome” using the fitness function. In the process of GA optimizing BP neural network, real number coding is used to generate the initialization population, and the chromosome length is defined as $m \times s + s + n \times s + n$. Where m denotes the number of input vector dimensions; s denotes the number of neurons in the hidden layer; n denotes the number of output vector dimensions. Individual fitness is calculated based on the test error results. The selection process favors individuals with higher fitness values, indicating a greater chance of being chosen. The fitness function, represented by Eq. (10), is the inverse of the absolute value of the absolute error. Through operations such as selection, crossover and mutation, individuals with better adaptation are retained. After repeated iterations, the optimized weights and thresholds are incorporated into the BP neural network for training and subsequent prediction.

$$F = \frac{1}{\sum_{i=1}^m |y_i - o_i|} \tag{10}$$

where F is the fitness function; m is the number of training samples; y_i and o_i are the predicted output and true output of the ith training sample, respectively.

In this paper, the GA-LM-BP neural network model is trained through the programming implementation of MATLAB simulation platform, and the flow chart of model construction is shown in Fig. 3. The parameters of the model are selected as follows: The number of neurons in the input layer is 76 for the number of symptom features. The number of neurons in the hidden layer is 15 for the number of nodes in the optimal hidden layer according to the empirical formula $s = \sqrt{m \times n} + a$ where $m = 55$ (input feature dimensions), $n = 6$ (output classes), $a \in [1, 10]$. The training set consists of 80% of the samples, with 20% reserved for testing. This ratio has been widely used in similar studies to ensure sufficient training while maintaining reliable validation.

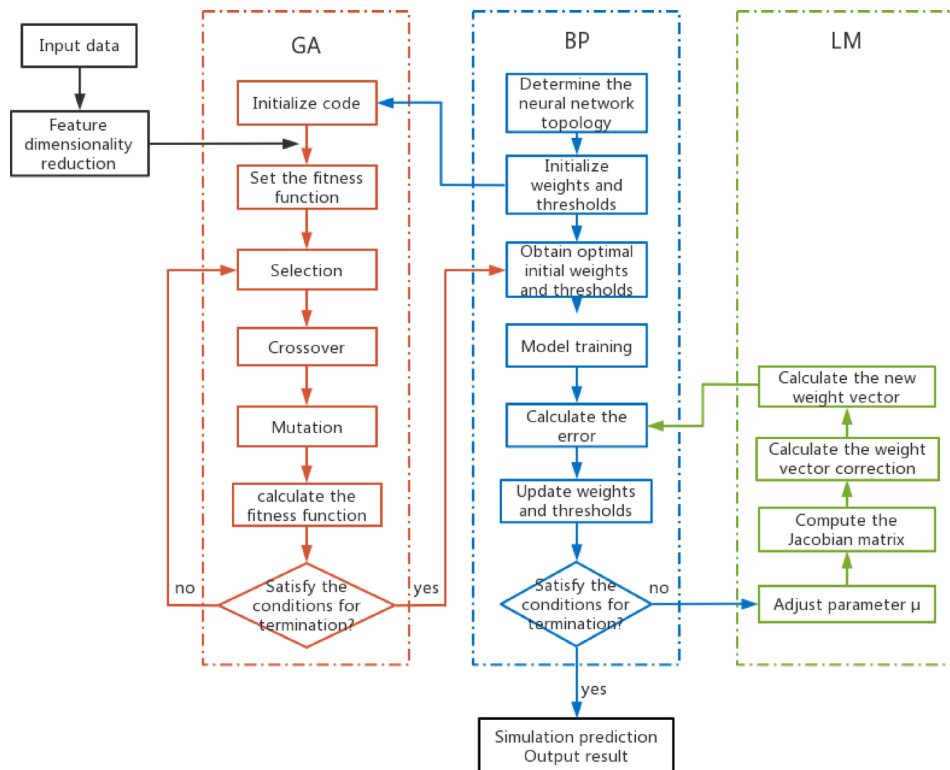


Fig. 3. Flow chart of GA-LM-BP neural network model construction.

Other hyperparameters, such as the maximum number of training epochs (1000), learning rate (0.01), and training goal (1e-5), were selected based on default recommendations in MATLAB's Neural Network Toolbox and validated through early-stage experiments. GA parameters including population size (100), crossover rate (0.6), and mutation rate (0.08) were also based on standard practices in evolutionary computation and verified through multiple trials.

Results and discussion

Optimization algorithm control group

Checking the literature related to algorithm optimization in the field of machine learning^{61–63}, it is found that the PSO algorithm is often used to do controlled experiments with GA. The PSO algorithm possesses superior global search characteristics, and it can be used to optimize the weight and threshold parameters of the BP neural network to improve the convergence speed and prediction accuracy of the neural network⁶⁴. Based on the results of a large number of literature studies, this experiment constructs three control groups of LM-BP, GA-LM-BP, and PSO-LM-BP, and sets other parameters, such as the number of iterations and the number of neurons in the hidden layer, in order to ensure the validity of the prediction results to remain consistent. Confusion matrix is a widely employed tool for evaluating the performance of a classification model. The numbers on the diagonal of the matrix indicate the number of correctly predicted cases. A denser concentration of predicted values along the diagonal signifies superior model performance. Recall is the proportion of all outcomes where the true value is in the positive category that the model predicts correctly, with a higher recall indicating a higher probability that the syndrome type will be predicted. Accuracy is the proportion of all correctly predicted results of a classification model to the total sample values, calculating how close the predicted values of the model are to the true values, with higher accuracy indicating better model classification. In statistics, the mean squared error (MSE) are used to measure the average of the squared errors between estimated and actual values⁶⁵. In this study, the confusion matrix is used to represent the difference between the prediction results and the actual results of the model, and the three indicators of recall, accuracy and mean squared error are used to judge the classification advantages and disadvantages. The calculation method of the three assessment indicators is shown in Eqs. (11–13), and the prediction results of each model are shown in Table 7.

$$recall = \frac{TP}{TP + FN} \quad (11)$$

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (12)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2 \quad (13)$$

where TP is the number of true positives; TN is the number of true negatives; FP is the number of false positives and FN is the number of false negatives. y_i and y'_i are the true value and predicted value of sample i , respectively.

Through comparative analysis, it is evident from the aforementioned table that the utilization of the optimization algorithm has led to an enhancement in model prediction accuracy compared to the non-optimized performance of 30.94%. The GA-LM-BP neural network model possesses the highest prediction accuracy of 53.35%, implying its superior predictive performance. As a result, the GA-LM-BP neural network model is chosen for conducting subsequent control experiments involving the dimensionality reduction algorithms.

Control group for dimensionality reduction algorithms

In “Feature screening” section, three feature extraction algorithms have been employed to reduce the dimensionality of the dataset, with the GRA algorithm screening for correlations of 0.963 and above, the PCA algorithm selecting results until the cumulative contribution reaches 85% and the MIV algorithm selecting results until the cumulative contribution reaches 95%. Based on the feature screening of the above parameters, the GRA algorithm optimizes the topology of the GA-LM-BP neural network model as “55-15-6”; the PCA algorithm optimizes the topology of the GA-LM-BP neural network model as “51-15-6”; the MIV algorithm optimizes the topology of the GA-LM-BP neural network model as “55-15-6”. In order to clarify the impact of the above feature extraction methods on the prediction results, the samples after screening using the above three features respectively are re-trained and predicted for the model, still dividing the training set and the test set with a ratio of 8:2, and the prediction results after utilizing the three feature screening methods are shown in Table 8.

Through comparative analysis, it is evident from the aforementioned table that after feature screening, the model prediction accuracy is improved compared to 53.35% before dimensionality reduction. It also suggests that feature screening plays a great importance in model optimization. The highest prediction accuracy of the model after the dimensionality reduction process of the MIV algorithm is 81.22%. The GRA algorithm calculates and arranges the overall descending order of the degree of correlation of symptom by using the symptom data and the syndrome data as the comparison sequence and the reference sequence, respectively, and it cannot screen out the symptoms that have high correlation among various types of syndromes. The meanings of each feature dimension of the principal components processed through the PCA algorithm have a certain degree of ambiguity and are not as interpretive as the original sample features. The MIV algorithm has the best interpretability for multi-class problems, and can better interpret the symptoms that have the top ranking of contribution to each type of syndrome through feature extraction, and further screening as the distinctive symptom features of that type of syndrome to help the neural network model to better categorize. The structure of MGLB model is shown in Fig. 4, and its topological network structure is “55-15-6”. The mean squared error

Model	MSE	Recall						Accuracy (%)
		Syndrome of lung qi deficiency complicated with phlegm and stasis obstructing the collaterals (249) (%)	Syndrome of lung qi deficiency complicated with accumulation and binding of phlegm and heat (151) (%)	Syndrome of qi deficiency in the lung and kidney complicated with phlegm and stasis obstructing the collaterals (136) (%)	Syndrome of lung dryness with yin deficiency (132) (%)	Syndrome of lung qi deficiency (115) (%)	Syndrome of qi deficiency in the lung and kidney (108) (%)	
LM-BP	0.47351	30.00	70.00	11.11	42.31	22.73	9.52	30.94
PSO-LM-BP	0.112721	70.00	80.65	50.00	62.96	13.04	9.09	47.62
GA-LM-BP	0.110784	68.00	90.32	64.29	66.67	21.74	9.09	53.35

Table 7. Model prediction results statistics.

is shown in Fig. 5; the prediction result is shown in Fig. 6; the prediction model confusion matrix is shown in Fig. 7; the network correlation graph is shown in Fig. 8.

The above experiments used different optimization algorithms and dimensionality reduction algorithms to achieve the classification of TCM syndromes, and LM-BP, PSO-LM-BP, GA-LM-BP, GRA-GA-LM-BP, PCA-GA-LM-BP and MGLB models are used for predicting the test samples, respectively. To visualize which model in the test set predicted the correct number of cases for the syndrome type more, a comparison of the recall rate of the models is shown in Fig. 9. The broken line representing the MGLB neural network model stands out at the highest position, exhibiting superior prediction accuracy and stability.

The TCM syndrome differentiation accuracy rate of the MGLB model for IPF is 81.22%, which is superior to the baseline model without feature screening and algorithm optimization. In Western medicine, two literatures based on HRCT imaging for diagnosing IPF have been found, and the accuracy rates are 78.9%⁶⁶ and a slightly higher accuracy of 85.7%⁶⁷. However, it is worth noting that TCM syndrome differentiation involves high-dimensional, nonlinear and subjective symptom data, which brings greater challenges compared with the structured biomedical indicators used in Western diagnosis. Our model is based on real-world clinical cases from multiple sources and reflects the real practice of TCM diagnosis. This enhances its applicability and interpretability in the real world, making it a valuable tool for intelligent TCM diagnosis.

Conclusion

IPF patients' scar tissue proliferates over time and it is difficult for oxygen to enter from the alveoli to the bloodstream after the onset of illness, which makes patients feel abnormally short of breath, with an average survival expectancy of only 3–5 years⁶⁸. Disease conditions of some patients may continue to deteriorate, leading to worsening of symptoms. Therefore, it is important to accurately predict the type of syndrome and prescribe the right medicine for the diagnosis and treatment of IPF. In this paper, a MGLB model is constructed to predict the TCM syndromes of IPF. The main conclusions are as follows:

1. The MIV algorithm can achieve a clear mapping between symptoms and syndrome, which is in line with the logic of syndrome differentiation, screen more characteristic variables and separately represent the symptoms with a higher contribution rate of each syndrome, offering greater interpretability and suitability for multi-class TCM syndrome classification. Compared with GRA and PCA, MIV achieves higher prediction accuracy and retains the original symptom features, ensuring both input and output of the model remain clinically interpretable within the context of TCM syndrome differentiation, though at the cost of increased computational load and greater manual processing complexity.
2. Two optimization algorithms, LM and GA, were used to optimize and update the convergence speed, weights and thresholds of the BP neural network, respectively. This approach addresses the issues commonly associated with traditional BP neural networks, including sluggish convergence speed, easy falling into local minima, and weak generalization ability. By comparing the prediction results of GA and PSO algorithm optimization model, it is found that the GA-LM-BP model based on feature screening has better recall, accuracy and stability, which is feasible to be used for the prediction of TCM syndrome in IPF.
3. The MGLB classification model proposed in this paper provides a more effective auxiliary means of TCM syndrome differentiation for unexperienced TCM doctors and improves the shortcomings of traditional TCM diagnosis and treatment with strong subjectivity. It can also reduce errors in the process of diagnosis and treatment and enable TCM diagnosis and treatment schemes to be formulated more quickly and accurately, which provides a certain reference value for the diagnosis of IPF. The model can be extended to be used for the prediction of syndromes in other chronic lung diseases, which is of great practical significance.

The main contribution of this work lies in their effective integration into a unified framework specifically designed for TCM syndrome classification. This includes: (1) The first systematic and interpretable application of the MIV algorithm in TCM syndrome differentiation, enabling transparent symptom-syndrome mapping that aligns well with the diagnostic logic of TCM; (2) The development of a novel integrated intelligent diagnostic thinking process framework; (3) The validation of the model's generalization potential through preliminary experiments on other chronic diseases; (4) The use of a real-world TCM clinical dataset comprising 956 IPF cases, ensuring practical relevance and applicability in actual diagnostic settings. In summary, this study presents a novel and interpretable machine learning-based framework for TCM syndrome classification, integrating feature screening, hybrid optimization, and real-world data validation to support accurate and interpretable diagnosis of IPF.

Limitations and future works

This paper demonstrates a novel application of machine learning technology to the research of the classification of TCM syndromes of IPF. The findings offer valuable insights for the intelligent TCM syndrome differentiation and contribute to enhancing the accuracy and scientific rigor of TCM syndrome classification. However, the research still has shortcomings. For neural network algorithms, more data is needed to put into training. The sample size of 956 medical cases included in the experiment of this study is slightly insufficient, and the collected dataset is not balanced. Furthermore, most of the clinical data are derived from expert outpatient clinics in the southwest region, which may introduce regional biases. In addition, as the data were retrospectively collated, some degree of subjectivity cannot be ruled out. To enhance the model's generalization ability, future work will expand the data collection to include more regions and hospitals, thereby increasing sample diversity and representativeness. Clustering analysis methods will also be explored to assist in identifying and correcting potential data inconsistencies. Moreover, statistical validation methods will be introduced to rigorously assess the significance of feature screening using MIV and evaluate the overall classification performance of the model.

Model	MSE	Recall						Accuracy (%)
		Syndrome of lung qi deficiency complicated with phlegm and stasis obstructing the collaterals (249) (%)	Syndrome of lung qi deficiency complicated with accumulation and binding of phlegm and heat (151) (%)	Syndrome of qi deficiency in the lung and kidney complicated with phlegm and stasis obstructing the collaterals (136) (%)	Syndrome of lung dryness with yin deficiency (132) (%)	Syndrome of lung qi deficiency (115) (%)	Syndrome of qi deficiency in the lung and kidney (108) (%)	
GRA-GA-LM-BP	0.101160	74.00	83.87	25.00	70.37	69.57	18.18	56.83
PCA-GA-LM-BP	0.100680	78.00	67.74	78.57	70.37	26.09	18.18	56.49
MIV-GA-LM-BP	0.075336	78.00	93.55	78.57	81.48	73.91	81.82	81.22

Table 8. Feature screening prediction results statistics.

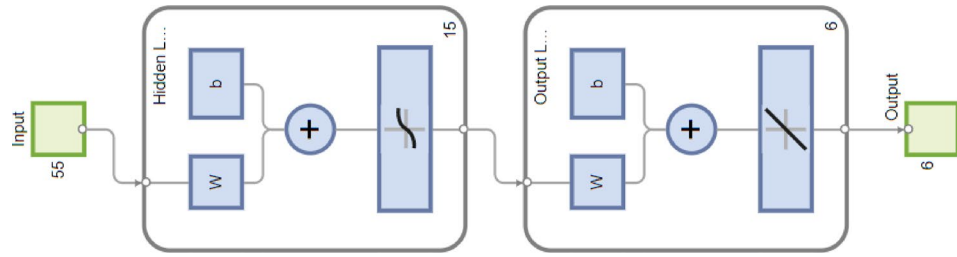


Fig. 4. Training interface.

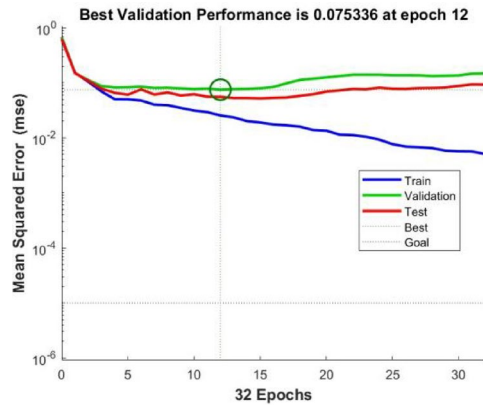


Fig. 5. Mean squared error.

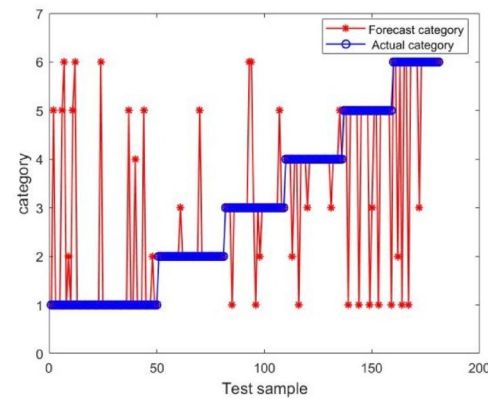


Fig. 6. Prediction result.

In addition, the MGLB model was applied to a dataset of 1205 insomnia cases and 1403 COPD cases, achieving an accuracy of 82.99% and 82.92%, showing its application potential for expansion to other chronic diseases. This research is mainly modeled and evaluated through the MATLAB platform, which can be embedded in the auxiliary diagnosis and treatment system for practical use in the medical field, increasing the depth of research and the value of use. The current research on machine learning algorithms has not yet involved the recommendation of treatments and herbs, and subsequent research on multi-label classification algorithms can be carried out for the part of treatment prediction and herb recommendation to achieve a closed loop.

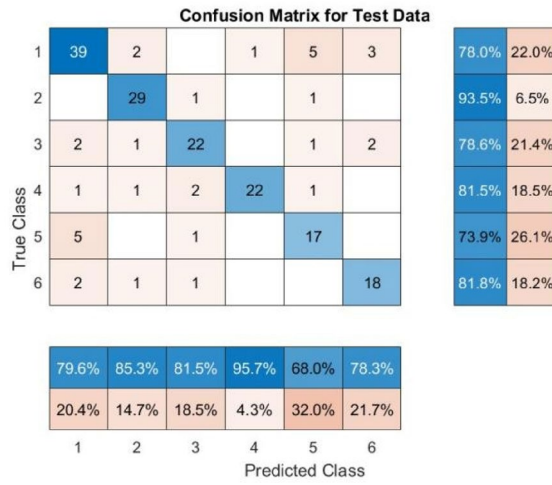


Fig. 7. Confusion matrix.

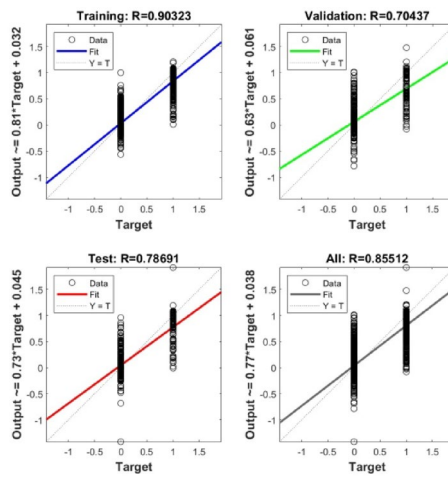


Fig. 8. Network correlation.

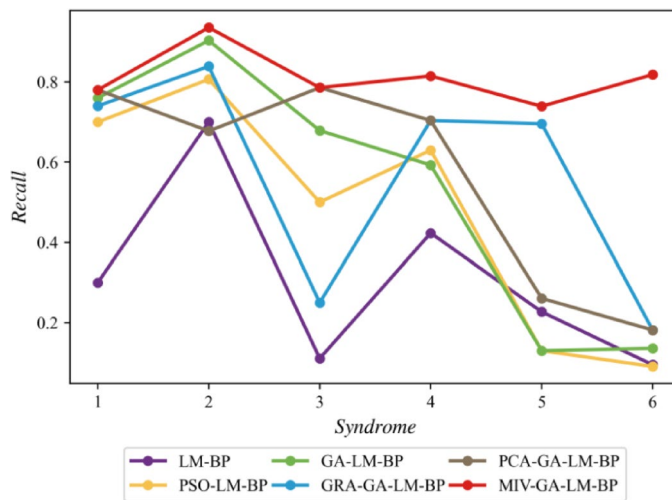


Fig. 9. Model comparison.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Received: 12 November 2024; Accepted: 1 July 2025

Published online: 28 November 2025

References

- Honda, K., Saraya, T. & Ishii, H. A real-world prognosis in idiopathic pulmonary fibrosis: A special reference to the role of antifibrotic agents for the elderly. *J. Clin. Med.* **12**(10), 3564. <https://doi.org/10.3390/jcm12103564> (2023).
- Mai, T. et al. Idiopathic pulmonary fibrosis therapy development: a clinical pharmacology perspective. *Ther. Adv. Respir. Dis.* **17**, 17534666231181536. <https://doi.org/10.1177/17534666231181537> (2023).
- Li, Y. et al. S100A12 as biomarker of disease severity and prognosis in patients with idiopathic pulmonary fibrosis. *Front. Immunol.* <https://doi.org/10.3389/fimmu.2022.810338> (2022).
- Yu, N. et al. Research progress on risk factors for acute exacerbation of Idiopathic pulmonary fibrosis and thoughts on prevention and treatment of traditional Chinese medicine. *Chin. Arch. Tradition. Chin. Med.* **41**(5), 119–123. <https://doi.org/10.13193/j.issn.1673-7717.2023.05.028> (2023).
- Li, N. et al. The research and development thinking on the status of artificial intelligence in traditional Chinese medicine. *Evid. Based Complement. Altern. Med.* **2022**, e7644524. <https://doi.org/10.1155/2022/7644524> (2022).
- Tao, Z. & Chen, H. Advances in application of data mining in TCM syndrome research. *Shanghai J. Tradition. Chin. Med.* **55**(6), 91–95. <https://doi.org/10.16305/j.1007-1334.2021.1910162> (2021).
- Fan, S. et al. Machine learning algorithms in classifying TCM tongue features in diabetes mellitus and symptoms of gastric disease. *Eur. J. Integr. Med.* **43**, 101288 (2021).
- Tang, Y. et al. Research of insomnia on traditional Chinese medicine diagnosis and treatment based on machine learning. *Chin. Med.* **16**(1), 2. <https://doi.org/10.1186/s13020-020-00409-8> (2021).
- Liu, C. et al. Study on TCM syndrome differentiation and diagnosis model based on BP neural network for syndrome elements and their common combinations in patients with borderline coronary lesion. *Chin. J. Inf. Tradition. Chin. Med.* **28**(3), 104–110. <https://doi.org/10.19879/j.cnki.1005-5304.202003248> (2021).
- Li, N. et al. Prediction of therapeutic effect on TCM syndromes of Qi stagnation and phlegm obstruction syndrome in dyslipidemia based on BP neural network optimized by particle swarm. *Modern. Tradition. Chin. Med. Mater. Med. World Sci. Technol.* **24**(10), 4089–4097 (2022).
- Yang, J. et al. An improved evolution algorithm using population competition genetic algorithm and self-correction BP neural network based on fitness landscape. *Soft. Comput.* **25**(3), 1751–1776. <https://doi.org/10.1007/s00500-020-05250-7> (2021).
- Cheng, P., Chen, D. & Wang, J. Research on prediction model of thermal and moisture comfort of underwear based on principal component analysis and genetic algorithm-back propagation neural network. *Int. J. Nonlinear Sci. Numer. Simul.* **22**(6), 607–619. <https://doi.org/10.1515/ijnsns-2020-0068> (2021).
- Guo, X. et al. A novel fast solving method for targeted drug-delivery capsules in the gastrointestinal tract. *Technol. Health Care* **27**(3), 335–341. <https://doi.org/10.3233/thc-181484> (2019).
- Cheng, P., Chen, D. & Wang, J. Clustering of the body shape of the adult male by using principal component analysis and genetic algorithm-BP neural network. *Soft. Comput.* **24**(17), 13219–13237. <https://doi.org/10.1007/s00500-020-04735-9> (2020).
- Mei, Q. et al. Idiopathic pulmonary fibrosis: An update on pathogenesis. *Front. Pharmacol.* **12**, 797292. <https://doi.org/10.3389/fphar.2021.797292> (2021).
- She, K. et al. The development status, problems, and solutions of machine LearningDriven intelligence in traditional Chinese medicine diagnosis. *J. Basic Chin. Med.* **30**(03), 398–406. <https://doi.org/10.19945/j.cnki.issn.1006-3250.2024.03.006> (2024).
- Jiang, Q., Sun, X., Xie, B. et al. Exploration of intelligent inference model for syndrome differentiations of traditional Chinese medicine. In *Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology*, 1–14 [2024-03-30]. <https://link.cnki.net/urlid/11.5699.r.20240228.1709.004>
- Wu, S., Wang, P., Kong, Z. et al. A multi-model fusion model for traditional Chinese medicine syndrome classification of omicron patients. In *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*. (IEEE, 2023). <https://doi.org/10.1109/ICCCBDA56900.2023.10154654>
- Yao, S. et al. Research on intelligent diagnosis model of traditional Chinese medicine based on ensemble learning: Taking perimenopausal syndrome as an example. *J. Tradition. Chin. Med.* **64**(06), 572–580. <https://doi.org/10.13288/j.11-2166/r.2023.06.007> (2023).
- Xie, J. et al. Feature selection and syndrome classification for rheumatoid arthritis patients with Traditional Chinese Medicine treatment. *Eur. J. Integr. Med.* **34**, 101059. <https://doi.org/10.13288/j.11-2166/r.2023.06.007> (2020).
- Zhang, Y. et al. Study on TCM influenza syndrome differentiation model based on machine learning. *Chin. J. Inf. Tradition. Chin. Med.* **31**(09), 48–57. <https://doi.org/10.19879/j.cnki.1005-5304.202403074> (2024).
- Li, R. et al. Exploring the identification pattern of Professor JiangXiaomin's treatment of bone paralysis based on CART decision tree and BP neural network algorithm. *Modern. Tradition. Chin. Med. Mater. Med. World Sci. Technol.* **25**(01), 401–412. <https://doi.org/10.13288/j.11-2166/r.2024.17.007> (2023).
- Guo, X. et al. Establishment of a traditional Chinese medicine syndrome diagnostic model based on stacking ensemble learning: Take lung cancer as an example. *J. Tradit. Chin. Med.* **65**(17), 1775–1783 (2024).
- Gong, W. et al. Research on traditional Chinese medicine syndrome diagnosis model of pediatric pneumonia with accumulation of phlegm-heat syndrome based on random forest and partial correlation analysis. *China J. Tradition. Chin. Med. Pharm.* **38**(09), 4497–4501 (2023).
- Yao, S. et al. Study on construction of TCM intelligent syndrome differentiation model of perimenopausal syndrome based on machine learning. *Chin. J. Inf. Tradition. Chin. Med.* **30**(06), 68–75. <https://doi.org/10.19879/j.cnki.1005-5304.202207558> (2023).
- Cao, Y. et al. Establishment of TCM intelligent pattern identification mode of gastroesophageal reflux disease based on machine learning. *J. Beijing Univ. Tradition. Chin. Med.* **42**(10), 869–874 (2019).
- Sun, Z. et al. Construction and application of stroke TCM pattern differentiation model based on machine learning. *J. Hunan Univ. Chin. Med.* **43**(04), 694–699 (2023).
- Jiang, C. et al. A comparative study on the application of three machine learning methods in the classification of cerebral haemorrhage in traditional Chinese medicine diagnosis. *Chin. J. Health Stat.* **40**(06), 921–928 (2023).
- He, P. et al. Construction and internal validation of an intelligent model for the identification of Chinese medicine in chronic glomerulonephritis using machine learning. *J. Liaoning Univ. Tradition. Chin. Med.* **26**(06), 73–78. <https://doi.org/10.13194/j.issn.1673-842x.2024.06.014> (2024).
- Chen, D. & Cheng, P. A perceptual image prediction model of professional dress style based on PSO-BP neural network. *J. Eng. Fabrics* <https://doi.org/10.1177/15589250231189816> (2023).
- Zhu, Z. et al. Research on classification of traditional Chinese medicine dysmenorrhea syndrome based on BP neural network with adaptive moment estimation. *Modern. Tradition. Chin. Med. Mater. Med. World Sci. Technol.* **23**(12), 4560–4568 (2021).

32. Yang, S. et al. Classification and prediction of Tibetan medical syndrome based on the improved BP neural network. *IEEE Access* **8**, 31114–31125. <https://doi.org/10.1109/ACCESS.2020.2973304> (2020).
33. Wang, Y., Wang, L., Song, J. et al. TCM syndrome differentiation based on artificial bee colony optimization BP neural network algorithm. In *2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. (IEEE, 2021). <https://doi.org/10.1109/YAC53711.2021.9486472>
34. Mingqi, Z. & Xin, D. Discussion on the construction of an intelligent TCM syndrome differentiation model for compensatory stage of liver cirrhosis after ensemble learning and fusion of neural network algorithms. *J. Guangzhou Univ. Chin. Med.* **40**(10), 2650–2660 (2023).
35. Zhang, T. et al. Challenges and ideas in constructing an interpretative system for syndrome differentiation in traditional Chinese medicine. *J. Tradition. Chin. Med.* **65**(05), 445–448+454. <https://doi.org/10.13288/j.11-2166/r.2024.05.001> (2024).
36. Li, J. et al. Syndrome diagnostic criteria of idiopathic pulmonary fibrosis in traditional Chinese medicine. *J. Tradition. Chin. Med.* **61**(18), 1653–1656. <https://doi.org/10.13288/j.11-2166/r.2020.18.021> (2020).
37. Shen, S. et al. Exploration of the problems and direction on syndrome differentiation based on the characteristics of traditional Chinese medicine thinking methods. *China J. Tradition. Chin. Med. Pharm.* **38**(2), 459–462 (2023).
38. Tong, M. et al. Predicting the natural-gas demand based on feature selection and BP neural network. *Nat. Gas Technol. Econ.* **16**(3), 59–65 (2022).
39. Elwali, A. & Moussavi, Z. A feature reduction and selection algorithm for improved obstructive sleep apnea classification process. *Med. Biol. Eng. Comput.* **59**(10), 2063–2072. <https://doi.org/10.1007/s11517-021-02421-y> (2021).
40. Tian, F., Huang, L. & Zhou, C. Photovoltaic power generation and charging load prediction research of integrated photovoltaic storage and charging station. *Energy Rep.* **9**, 861–871. <https://doi.org/10.1016/j.egy.2023.04.250> (2023).
41. Zhang, K., Zhang, K. & Bao, R. Prediction of gas explosion pressures: A machine learning algorithm based on KPCA and an optimized LSSVM. *J. Loss Prev. Process Ind.* **83**, 105082. <https://doi.org/10.1016/j.jlp.2023.105082> (2023).
42. Liu, W. et al. Research on electronic nose for compound malodor recognition combined with artificial neural network and linear discriminant analysis. *J. Intell. Fuzzy Syst.* **44**(4), 6991–7008. <https://doi.org/10.3233/JIFS-222539> (2023).
43. Ben Hassen, D. et al. A novel intelligent reasoning method to estimate the cutting system energy consumption for a sustainable manufacturing. *J. Chin. Inst. Eng.* **46**(1), 74–80. <https://doi.org/10.1080/02533839.2022.2141337> (2023).
44. Qiu, Q. et al. Study on the fingerprint spectrum and the spectrum-effect relationship of analgesic and anti-inflammatory effects of the aqueous extract from *Dalbergia Hancai* Benth. *J. Anal. Methods Chem.* **2023**, 1242756. <https://doi.org/10.1155/2023/1242756> (2023).
45. Liu, Z., Cheng, S. & Liu, P. Prediction model of BOF end-point P and O contents based on PCA–GA–BP neural network. *High Temp. Mater. Process.* **41**(1), 505–513. <https://doi.org/10.1515/htmp-2022-0050> (2022).
46. Chen, S. et al. Analysis of factors influencing wave overtopping discharge from breakwater based on an MIV–BP estimation model. *Water Multidiscip.* **14**(19), 2967. <https://doi.org/10.3390/w14192967> (2022).
47. Gu, Z. et al. Study on medication rules of TCM prescription in treatment of Idiopathic pulmonary fibrosis based on gray-fuzzy system. *Chin. J. Inf. Tradition. Chin. Med.* **25**(12), 92–96. <https://doi.org/10.1155/2020/7498525> (2018).
48. Zhan, X. Application of PCA algorithm in psoriatic BP neural network modeling. *Mod. Comput.* **4**, 25–28 (2021).
49. Tan, N. et al. Exploration of TCM symptoms and syndromes of chronic heart failure based on principal component analysis. *China J. Tradition. Chin. Med. Pharm.* **36**(7), 4265–4267 (2021).
50. Wu, P. et al. Optimization of static flocculation settlement parameters of full tailings based on grey relational analysis. *Min. Res. Dev.* **42**(9), 116–121. <https://doi.org/10.13827/j.cnki.kyyk.2022.09.019> (2022).
51. Zhao, K. & Mu, K. Evaluation of shale reservoirs based on grey relation analysis and principal component analysis. *Geol. Explor.* **59**(2), 443–450 (2023).
52. Tripathi, M. & Singal, S. Use of principal component analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecol. Ind.* **96**, 430–436. <https://doi.org/10.1016/j.ecolind.2018.09.025> (2019).
53. Liu, L. Assessment of water resource security in karst area of Guizhou Province, China. *Sci. Rep.* **11**(1), 7641. <https://doi.org/10.1038/s41598-021-87066-5> (2021).
54. Zhang, H., Li, Y. & Yan, L. Prediction model of car ownership based on back propagation neural network optimized by particle swarm optimization. *Sustain. Multidiscip.* **15**(4), 2908. <https://doi.org/10.3390/su15042908> (2023).
55. Cheng, P., Chen, D. & Wang, J. Research on underwear pressure prediction based on improved GA–BP algorithm. *Int. J. Cloth. Sci. Technol.* **33**(4), 619–642. <https://doi.org/10.1108/IJCST-05-2020-0078> (2020).
56. Cai, G. et al. Inverse kinematics solution of mining robot via APPO–LM–BP neural network. *Mech. Sci. Technol. Aerosp. Eng.* **39**(5), 706–713. <https://doi.org/10.13433/j.cnki.1003-8728.20190200> (2020).
57. Quan, L. et al. Fault diagnosis of electro hydraulic servo valves based on GA+LM algorithm optimized BP neural networks. *China Mech. Eng.* **29**(5), 505–510. <https://doi.org/10.3969/j.issn.1004-132X.2018.05.001> (2018).
58. Wu, C. & Jiang, R. Temperature compensation method of pressure sensor based on PSO–LM–BP neural network. *Instrum. Tech. Sens.* **2**, 129–133 (2018).
59. Liu, Q. & Zhao, Z. Relationship model between human resource management activities and performance based on LMBP algorithm. *Secur. Commun. Netw.* **2022**, e1125084. <https://doi.org/10.1155/2022/1125084> (2022).
60. Jin, G., Feng, W. & Meng, Q. Prediction of waterway cargo transportation volume to support maritime transportation systems based on GA–BP neural network optimization. *Sustain. Multidiscip.* **14**(21), 13872. <https://doi.org/10.3390/su142113872> (2022).
61. Zhu, C. et al. Comparison of GA–BP and PSO–BP neural network models with initial BP model for rainfall-induced landslides risk assessment in regional scale: A case study in Sichuan, China. *Nat. Hazards* **100**(1), 173–204. <https://doi.org/10.1007/s11069-019-03806-x> (2020).
62. Feng, X. et al. RSM, ANN–GA and ANN–PSO modeling of SDBS removal from greywater in rural areas via Fe₂O₃-coated volcanic rocks. *RSC Adv. Roy. Soc. Chem.* **12**(10), 6265–6278 (2022).
63. Zhang, Z. et al. Intelligent geometry compensation for additive manufactured oral maxillary stent by genetic algorithm and backpropagation network. *Comput. Biol. Med.* **157**, 106716. <https://doi.org/10.1016/j.combiomed.2023.106716> (2023).
64. Wan, Q., Chen, J., Zhu, R. & Chen, W. Tooth root stress prediction of helical gear pair with misalignment errors based on PSO–BP neural network. *IEEE Sens. J.* <https://doi.org/10.1109/JSEN.2024.3376567> (2024).
65. Nanekaran, Y. A. et al. The predictive model for COVID-19 pandemic plastic pollution by using deep learning method. *Sci. Rep.* **13**(1), 4126. <https://doi.org/10.1038/s41598-023-31416-y> (2023).
66. Zhen, T., Hu, D. & Yao, W. Predictive value of deep-learning radiomics model of chest high-resolution CT for GAP stage of idiopathic pulmonary fibrosis. *Zhejiang Med. J.* **45**(18), 1921–1926+1931 (2023).
67. Soffer, S. et al. Artificial intelligence for interstitial lung disease analysis on chest computed tomography: A systematic review. *Acad. Radiol.* **29**, S226–S235. <https://doi.org/10.1016/j.acra.2021.05.014> (2022).
68. Alsomali, H. et al. Early diagnosis and treatment of idiopathic pulmonary fibrosis: A narrative review. *Pulmon. Ther.* **9**(2), 177–193. <https://doi.org/10.1007/s41030-023-00216-0> (2023).

Author contributions

H. Y.: Data curation, Methodology, Writing—original draft, Project administration. W. G.: Data analysis, Algorithm design, Visualization, Writing—original draft. P. Y.: Data collection, writing—review& editing. R. Z.: Validation, writing—review & editing. B. H.: Data curation, Methodology. W. L.: Conceptualization, Supervision, Methodology.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 82105054), the “Xinglin Scholar” Scientific Research Promotion Plan of Chengdu University of TCM (Grant No. CCCX2023014), the National Key R&D Program of China (Grant No. 2018YFC1704104), the College Students Innovation and Entrepreneurship Training Program of Sichuan (S202310633026, S202310633033), National Natural Science Foundation of China (Grant No. 82575261).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The data for this retrospective study were sourced from existing collections of data. This article does not contain any studies with human participants or animals performed by any of the authors. The study was conducted in accordance with the 1964 Declaration of Helsinki and its later amendments or similar ethical standards.

Additional information

Correspondence and requests for materials should be addressed to W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026