• Intelligent Ophthalmology •

HHO optimized support vector machine classifier for traditional Chinese medicine syndrome differentiation of diabetic retinopathy

Li Xiao[1], Cheng–Wu Wang[2], Ying Deng[3], Yi–Jing Yang[3], Jing Lu[4], Jun–Feng Yan[2], Qing–Hua Peng[3,4]

[1]School of Chinese Medicine, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China

[2]School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China

[3]Hunan Provincial Key Laboratory for Prevention and Treatment of Ophthalmology and Otolaryngology Diseases with Chinese Medicine, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China

[4]Hunan Provincial Engineering and Technological Research Center for Prevention and Treatment of Ophthalmology and Otolaryngology Diseases with Chinese Medicine and Protecting Visual Function, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China

Correspondence to: Jun-Feng Yan and Qing-Hua Peng. Hunan University of Chinese Medicine, 300 Xueshi Road, Yuelu District, Changsha 410208, Hunan Province, China. junfengyan@hnucm.edu.cn; pqh410007@126.com

Received: 2023-07-28 Accepted: 2024-02-04

Abstract

● **AIM:** To develop a classifier for traditional Chinese medicine (TCM) syndrome differentiation of diabetic retinopathy (DR), using optimized machine learning algorithms, which can provide the basis for TCM objective and intelligent syndrome differentiation.

● **METHODS:** Collated data on real-world DR cases were collected. A variety of machine learning methods were used to construct TCM syndrome classification model, and the best performance was selected as the basic model. Genetic Algorithm (GA) was used for feature selection to obtain the optimal feature combination. Harris Hawk Optimization (HHO) was used for parameter optimization, and a classification model based on feature selection and parameter optimization was constructed. The performance of the model was compared with other optimization algorithms. The models were evaluated with accuracy, precision, recall, and F1 score as indicators.

● **RESULTS:** Data on 970 cases that met screening requirements were collected. Support Vector Machine (SVM) was the best basic classification model. The accuracy rate of the model was 82.05%, the precision rate was 82.34%, the recall rate was 81.81%, and the F1 value was 81.76%. After GA screening, the optimal feature combination contained 37 feature values, which was consistent with TCM clinical practice. The model based on optimal combination and SVM (GA_SVM) had an accuracy improvement of 1.92% compared to the basic classifier. SVM model based on HHO and GA optimization (HHO_GA_SVM) had the best performance and convergence speed compared with other optimization algorithms. Compared with the basic classification model, the accuracy was improved by 3.51%.

● **CONCLUSION:** HHO and GA optimization can improve the model performance of SVM in TCM syndrome differentiation of DR. It provides a new method and research idea for TCM intelligent assisted syndrome differentiation.

● **KEYWORDS:** traditional Chinese medicine; diabetic retinopathy; Harris Hawk Optimization; Support Vector Machine; syndrome differentiation

DOI:10.18240/ijo.2024.06.02

Citation: Xiao L, Wang CW, Deng Y, Yang YJ, Lu J, Yan JF, Peng QH. HHO optimized support vector machine classifier for traditional Chinese medicine syndrome differentiation of diabetic retinopathy. *Int J Ophthalmol* 2024;17(6):991-1000

INTRODUCTION

Traditional Chinese medicine (TCM) has gone through thousands of years of accumulation and sedimentation, playing an irreplaceable role in treating diseases. TCM emphasizes the prevention of disease and dialectical treatment, and has advantages in disease prevention and treatment, including endocrine diseases[1-2], cardiovascular disease[3-4], tumor[5-6], plague[7-8], *etc*. In 2019, the World Health Organization officially released the 11th edition of the International Classification of Diseases (ICD-11)[9], which for the first time included TCM in the global diagnostic criteria for diseases, achieving parallel with modern medicine. Especially after the COVID-19 epidemic, TCM has attracted more attention around the world. Based on the theory of adapting to physical condition, times, and local conditions, TCM focuses on the personalized management of chronic diseases, achieving good therapeutic effects with fewer side effects and reducing economic burden, such as diabetes and its complications[10-11].

How to use TCM correctly is a big challenge. Accurate diagnosis is the key to using TCM. TCM diagnosis includes the diagnosis of TCM diseases and TCM syndrome types. TCM syndrome type is a generalization of the pathological nature of the body at a certain stage in the course of disease development, including etiology, disease location, disease nature, *etc*. In fact, TCM syndrome differentiation is based on the past clinical experience of TCM doctors. Inevitably, it takes a long time to manually understand and learn a large amount of experience, and it is prone to subjective biases. This limits the ability of TCM to better serve people around the world. How to quickly use the experience of experts to assist doctors to determine the TCM syndrome type of diseases more accurately and quickly is a very worthwhile problem to explore.

With the development of machine learning technology, there are more and more applications in the field of medical disease diagnosis. For example, Support Vector Machine (SVM)[12], Gradient Boosting[13], Adaptive Boosting (AdaBoost)[14], K-nearest Neighbor (KNN) algorithms[15], and other machine learning technologies are used to establish corresponding disease diagnosis models. TCM intelligent syndrome differentiation has become a hot spot and has made some achievements[16-18]. However, existing machine learning models are still faced with problems such as complex operation and insufficient generalization ability to varying degrees, so it is of great significance to build machine learning models with strong generalization ability. In practical applications, the key to improving the accuracy and generalization ability of the model lies in the selection of dataset features and the setting of model parameters.

In general, the appearance of redundant and uncorrelated features not only increases the dimension of the feature vector, but also reduces the performance of machine learning. Therefore, removing redundant features from disease diagnosis data sets through feature selection can make the model have lower complexity and better classification performance. Traditional feature selection methods are usually based on statistical analysis or manual selection, but these methods have certain limitations. First, these methods cannot process high-dimensional data, and when the number of features is too large, these methods will become very time-consuming. Second, these methods are difficult to find nonlinear relationships. Therefore, an appropriate feature selection algorithm can effectively improve the prediction performance of the model, reduce the amount of over fitting and calculation, which is crucial to improve the accuracy and efficiency of medical diagnosis. At the same time, the optimization of model parameters is also a key issue that needs to be faced. The setting of parameters usually directly affects the performance and accuracy of the model, and the optimal combination of parameters should be obtained as much as possible to achieve the best performance of the model. Traditional parameter adjustment methods are usually based on grid search or greedy algorithm, but these methods need to manually set the parameter range, with high computational complexity, and may not be able to find the optimal parameter combination.

In recent years, intelligent optimization algorithm has been widely proposed. It possesses global search ability, enabling to locate the global optimal solution or an approximate optimal solution within the search space. Moreover, it can cope with non-smooth, non-continuous, non-linear, and other complex objective functions, and has better robustness. Additionally, it can speed up the search process, improve the efficiency of optimization, and bring new ideas and methods for solving optimization problems. Consequently, intelligent optimization algorithms emerge as a promising methodology for tackling challenges related to finding the optimal feature subset and parameter optimization problems.

Diabetic retinopathy (DR) is one of the most common microvascular complications of diabetes and the main cause of adult acquired blindness. With increasing DR trends worldwide, the prevention, treatment and associated costs have become global public health concerns. The pathogenesis of this disease is very complex. It has been proved that TCM has good clinical effect on the prevention and treatment of DR[19-22].

In other words, there are many effective methods for treating DR in TCM[23-25], including tonifying Qi and nourishing Yin, clearing heat and breeding body fluids, activating blood and removing blood stasis. So, how to correctly diagnose TCM syndrome type is a problem that must be solved. In this study, we used machine learning algorithms to establish an intelligent syndrome differentiation model for TCM, and utilized optimization algorithms to improve the classification performance of the model, which was applied to the classification of TCM syndrome types in DR.

SUBJECTS AND METHODS

Ethical Approval This study collected DR patients who were hospitalized in Bailing Diabetes Hospital of Guizhou Province and the First Affiliated Hospital of Hunan University of Chinese Medicine from March 2015 to January 2019. This study was reviewed and approved by the hospital ethics committee (approval number: HN-LL-YJSLW-2020-77).

Diagnostic Criteria Refer to Guidelines on Diabetic Eye Care[26] for the diagnosis of DR: the classic retinal lesions of DR are well described and include microaneurysms, intraretinal hemorrhages, venous beading (venous caliber changes consisting of alternating areas of venous dilation and constriction), intraretinal microvascular abnormalities, hard exudates (lipid deposits), and retinal neovascularization.

Inclusion Criteria Meet the DR diagnostic criteria mentioned above and no other diseases affecting the presentation of symptoms at the time of consultation with the doctor.

Exclusion Criteria Medical record information is missing and there are obvious errors in medical record information.

Information Collection By summarizing the description of DR in *Ophthalmology of Traditional Chinese Medicine*[27] and *Ophthalmology of Integrated Chinese and Western Medicine*[28], a unified clinical medical record data collection table was developed, which included general information (name, gender, age, diabetes course, *etc*.) and disease information (TCM four diagnostic information, past history, family history, personal history, auxiliary examination, *etc*.). For hospitalized patients, personal information, admission records, course records and discharge records of electronic medical records were exported, and then features were extracted using Python language according to clinical observation records.

Desensitization In order to prevent the leakage of sensitive information of patients, desensitization processing is carried out on the aggregated medical record database, which involves hiding personal privacy fields such as the patient's name and hospitalization number, and using the number as the identifier of the medical record.

Data Screening and Cleaning The case data were screened according to inclusion criteria and exclusion criteria, and errors, duplication, and missing values (data loss does not exceed 10%) in the data were corrected, redundancy was removed.

Standardized Processing According to *Ophthalmology of Traditional Chinese Medicine*[27], *Syndrome Element Differentiation*[29], *Terminology of Clinical Diagnosis and Treatment of Traditional Chinese Medicine (Revised edition)*[30], and *ICD-11*[9], TCM syndrome types, disease names, and TCM symptoms are standardized.

Data Quality Control No one shall alter, disclose or misappropriate the research data at will. The data were entered independently by three researchers, and cross-checked twice until the data results were consistent. The survey physician shall supervise and review, and if there is disagreement, the general responsible person shall rule.

Data Preprocessing Excel was used to establish a database, TCM symptoms and TCM syndrome types were sorted into binary categorical variables, with "yes" assigned as "1" and "no" assigned as "0". Sort according to the frequency of TCM symptoms and delete symptoms with a frequency of less than 8 (occasional symptoms lack diagnostic significance).

Classifier Construction The data is randomly divided into training set and verification set at a ratio of 7:3. KNN model, Random Forest (RF) model, AdaBoost model, SVM model are established respectively, and the accuracy, precision, recall and F1-score were calculated to evaluate.

Feature Selection The existence of redundant features in the data set will reduce the accuracy of model classification and increase the training time of the model. Genetic Algorithm (GA), a heuristic search method that is inspired by natural genetic mechanisms and biological evolution and can search for the optimal solution to a problem in a given spatial range. GA consists of five main parts, which are population initialization, fitness function design, selection operator, crossover operator and mutation operator. In this study it was used to select features in the disease dataset.

The population is composed of chromosomes, and different chromosomes are generated by random values of each gene position. In each iteration process, the chromosomes in the population will be evaluated by a fitness function, which represents the adaptability of individuals to the environment, so as to judge the merits of individuals in the population. The individuals with high fitness function value represent the dominant individuals in the population. Individuals with low fitness function value represent the inferior individuals in the population. According to the survival of the fittest criterion, the superior individuals are more likely to be selected in the subsequent operation, while the inferior individuals will be gradually eliminated in the iterative process. The specific implementation process is shown in Figure 1.

Parameter Optimization The setting of parameters has a great impact on the performance and effect of the model. In order to obtain better classification results, it is necessary to optimize the parameters. Harris Hawk Optimization (HHO) is a new intelligent optimization algorithm proposed by Heidari *et al*[31]. It has the characteristics of fewer parameters, easy implementation and wide application, and is suitable for a variety of optimization problems.

The whole search process can be divided into three stages: exploration, conversion of exploration and development, and development. According to it, the mathematical model is to take the Harris Hawk with the optimal fitness value as the prey or expected prey in order to determine the next position update strategy during each iteration.

During the exploration phase, Harris Hawks fly to random locations and explore prey based on two strategies. If each flight strategy has an equal chance, the randomly generated size. When $P \geq 0.5$, it will roost randomly in large trees within the range of the eagles. When $P < 0.5$, the Harris Hawk perches according to the position of other members and prey. The specific mathematical model is as follows.

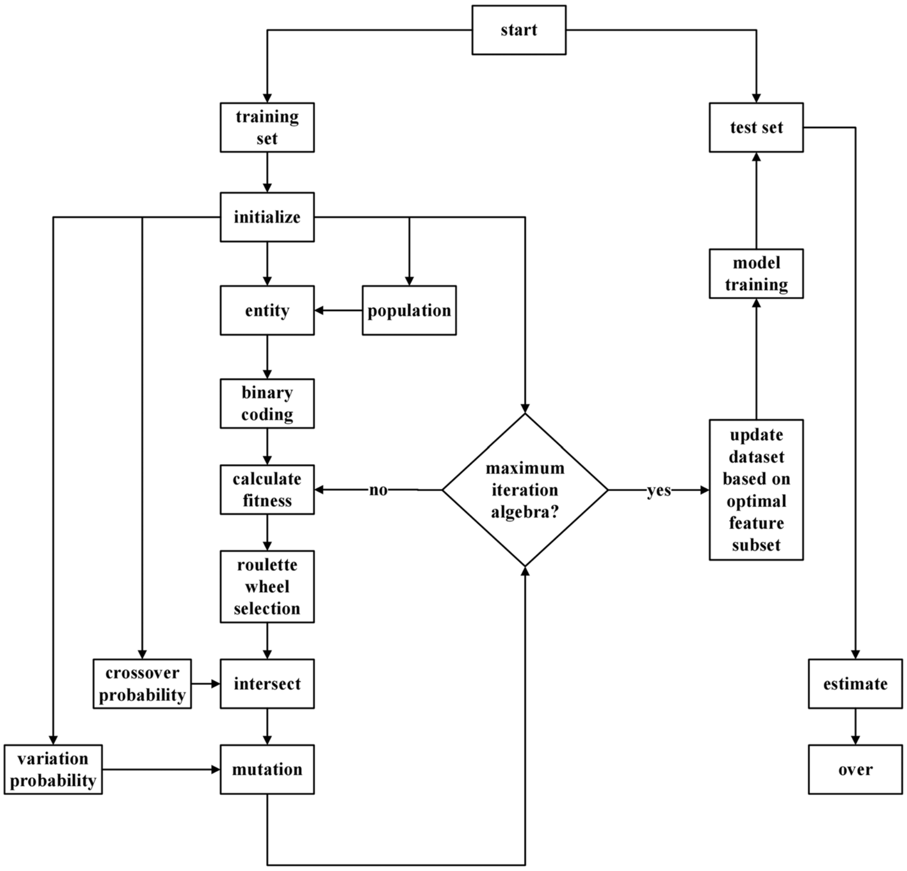

Figure 1 Flow chart of feature selection.

$$X(t+1)=\begin{cases} X_{rand}(t)\text{-}r_1|X_{rand}(t)\text{-}2r_2X(t)| & P\geq 0.5 \\ X_{rabbit}(t)\text{-}X_m(t)\text{-}r_3(LB+r_4(UB\text{-}LB)) & P<0.5 \end{cases} \quad (1)$$

$$X_m(t)=\frac{1}{N}\sum_{i=1}^{N}X_i(t) \quad (2)$$

where *X(t+1)* is the position vector of hawks in the next iteration *t*, $X_{rabbit}(t)$ is the position of rabbit, *X(t)* is the current position vector of hawks, *r1*, *r2*, *r3*, *r4*, and *P* are random numbers inside (0, 1), which are updated in each iteration, *LB* and *UB* show the upper and lower bounds of variables, $X_{rand}(t)$ is a randomly selected hawk from the current population, and X_m is the average position of the current population of hawks, where $X_i(t)$ indicates the location of each hawk in iteration *t* and *N* denotes the total number of hawks.

During the search and exploitation transition phase, the escape energy of the prey is used to determine whether to enter the exploration phase or the exploitation phase.

$$\mathrm{E}=2\mathrm{E}_0(1\text{-}\frac{\mathrm{t}}{\mathrm{T}}) \quad (3)$$

Where E_0 is the initial energy of the prey, a random number between -1 and 1 that is updated with each round of computation. Harris Hawk was in the search phase when $|E|\geq 1$ and in the development phase when $|E|<1$.

During the development phase, *r* is a random number between 0 and 1 that determines which development strategy to execute. When $0.5\leq|E|<1$ and $r\geq 0.5$, the position update method is soft siege. This behavior can be modeled as follows.

$$X(t+1)=\Delta X(t)\text{-}E|JX_{rabbit}(t)\text{-}X(t)| \quad (4)$$

$$\Delta X(t)=X_{rabbit}(t)\text{-}X(t) \quad (5)$$

Where *ΔX(t)* represents the difference between the position vector of the rabbit and the current location in iteration *t*, *J* is the random number between 0 and 2.

When $|E|<0.5$ and $r\geq 0.5$, the position update mode is hard siege. This behavior can be modeled as follows.

$$X(t+1)=X_{rabbit}(t)\text{-}E|\Delta X(t)| \quad (6)$$

When $0.5\leq|E|<1$ and $r<0.5$, the position update mode is soft encircle with asymptotic fast dive. This behavior can be modeled as follows.

$$Y=X_{rabbit}(t)\text{-}E|JX_{rabbit}(t)\text{-}X(t)| \quad (7)$$

$$Z=Y+S\times LF\,(D) \quad (8)$$

$$X(t+1)=\begin{cases} Y, f(Y)<f(X(t)) \\ Z, f(Z)<f(X(t)) \end{cases} \quad (9)$$

Where, *f()* is the fitness function, *D* is the problem dimension, *S* is a D-dimensional random vector, *LF* is a Levy function.

When $|E|<0.5$ and $r<0.5$, the position update mode is hard encircling with asymptotic fast dive. This behavior can be modeled as follows.

$$X(t+1)=\begin{cases} Y, f(Y)<f(X(t)) \\ Z, f(Z)<f(X(t)) \end{cases} \quad (10)$$

$$Y=X_{rabbit}(t)-E|JX_{rabbit}(t)-X_m(t)| \tag{11}$$

$$Z=Y+S\times LF(D) \tag{12}$$

The specific experimental steps are shown in Figure 2.

RESULTS

Data Set A total of 970 cases meeting the screening criteria were included in this study. These patients were divided into 4 TCM syndrome types, including 592 cases of Qi and Yin deficiency syndrome, 113 cases of liver and stomach depression heat syndrome, 103 cases of gastrointestinal damp-heat syndrome, 162 cases of Yin and Yang deficiency syndrome, which are consistent with clinical distribution patterns. The frequency of standardized TCM symptoms was analyzed, and there were 75 symptoms with a frequency higher than 8. Twenty-nine typical TCM symptoms with a frequency higher than 10% are shown in Table 1. The most frequent TCM symptom is blurred vision.

Basic Classifier The system platform used in this experiment is Windows 11 (64 bit) operation working system, programming language is Python, and programming software is PyCharm. To solve the problem of unbalanced sample size, undersampling method is adopted in this study. The model was trained by randomly generating 70% as a training set and 30% as a test set, and the performance of the model is evaluated. The comparison of evaluation indicators of DR TCM syndrome classifier models is shown in Figure 3. The experimental results show that the classification based on SVM is the best.

Optimal Results

Classification based on feature selection In this study, the crossover probability of GA was set to 0.25, the mutation probability was set to 0.35, the number of populations was set to 10, and the number of iterations was set to 100. The fitness function is shown in formula 13.

$$f=\alpha\times acc\ (SVM)+\beta\times(1-\tfrac{n}{N}) \tag{13}$$

In the formula, *acc(SVM)* represents the ten-fold cross-validation accuracy of the SVM training model; α and β are two weight parameters, which are used to represent the weight of the accuracy of the SVM classifier and the weight of the length of the feature subset selected by the algorithm, respectively. α was set as 0.92. β was set to 0.08. n represents the length of the selected feature subset, and N is the total number of feature attributes in the original dataset.

By randomly generating 70% of the training set and running 30% of the test set independently for 10 times, the average result of 10 operations of the test set was taken as the final experimental result. The change curve of the optimal individual fitness value with the number of iterations is shown in Figure 4.

It can be seen from Figure 4 that the rate of convergence is fast at the initial stage of the algorithm evolution, the fitness value has reached about 0.7 within 40 generations, and then the fitness value gradually increases and converges around 100 generations, and the optimal fitness value is 0.78. The 37 optimal features were obtained. The feature sequence were selected, corresponding to the following features: "frequent micturition, polydipsia, sunken pulse, sleep disorders, pale red tongue, feeble, greasy and yellow fur on the tongue, clouded heavy head, slippery pulse, be languid, stringy pulse, white and greasy tongue coating, white and thick tongue coating, red tongue nature, lumbago, polyphagia, yellow and thin tongue coating, white and thin tongue coating, fluid deficiency, feeble pulse, pale tongue nature, panting, stabbing pain, palpitation, yellow tongue coating, shortness of breath, tinnitus, hyperhidrosis, white foam sputum, gain weight, thin coated tongue, dislike heat, easy to starve, dysuria, irritability, less coated tongue, bitter taste". Based on the optimal feature subset and SVM (GA_SVM), the accuracy of the classifier is 83.97%, and the precision is 85.06%, the recall is 83.92%, the

Table 1 Distribution of typical TCM symptoms in DR patients

Number	TCM symptoms	Frequency (*n*)	Proportion (%)
1	Blurred vision	849	87.53
2	Limb numbness	753	77.63
3	Frequent micturition	712	73.40
4	Polydipsia	701	72.27
5	Thready pulse	655	67.53
6	Sunken pulse	616	63.51
7	Pale red tongue	564	58.14
8	Sleep disorders	458	47.22
9	Thirst	446	45.98
10	Feeble	415	42.78
11	Dizzy	364	37.53
12	Greasy and yellow fur on the tongue	307	31.65
13	Weight loss	308	31.75
14	Chest tightness	304	31.34
15	Clouded heavy head	283	29.18
16	Palpitate	281	28.97
17	Slippery pulse	251	25.88
18	Be languid	246	25.36
19	Anorexia	222	22.89
20	Stringy pulse	211	21.75
21	White and greasy tongue coating	195	20.10
22	White and thick tongue coating	171	17.63
23	Red tongue nature	163	16.80
24	Hypomnesis	134	13.81
25	Pruritus	113	11.65
26	Headache	112	11.55
27	Lumbago	109	11.24
28	Light purple tongue nature	101	10.41
29	Polyphagia	97	10.00

TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

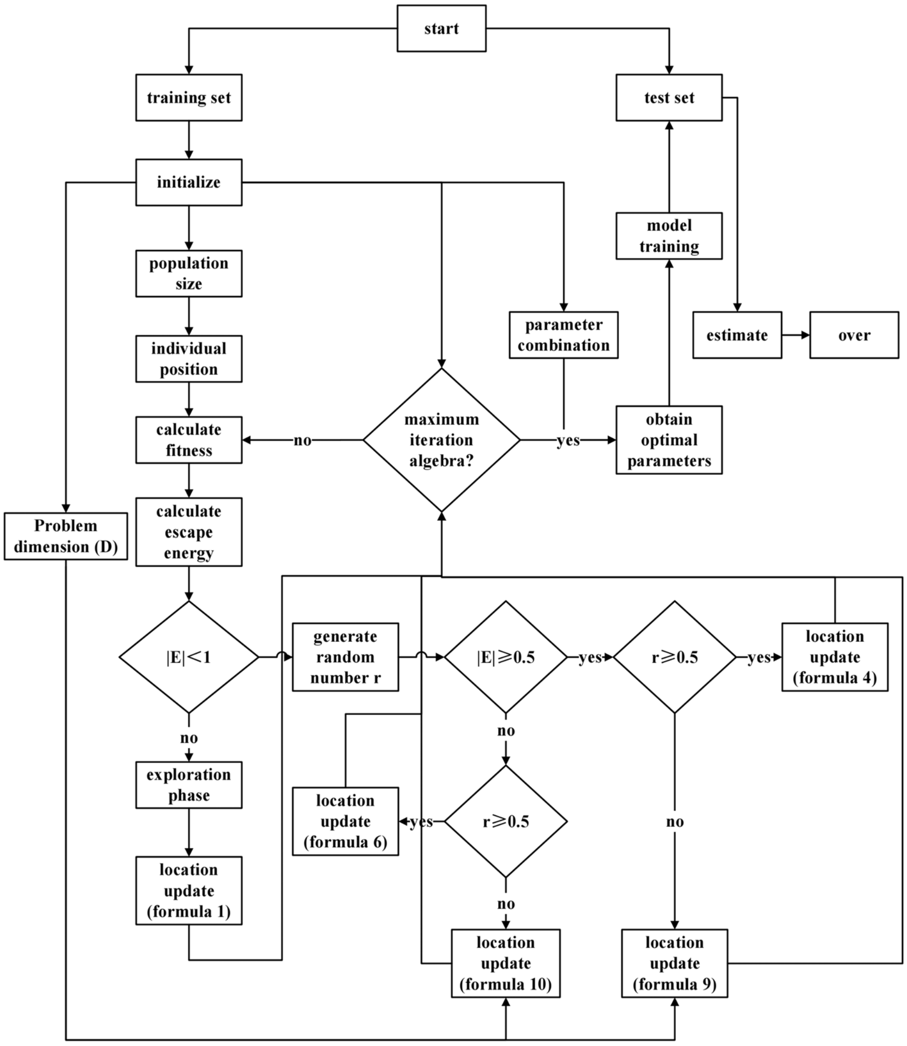

Figure 2 Flow chart of parameter optimization.

F1 is 83.94%. In other words, compared with the basic SVM, the accuracy of the model is increased by 1.92%, the precision rate is increased by 2.72%, the recall rate is increased by 2.11%, and the F1 value is increased by 2.18%.

Optimizing classifiers In this study, the number of populations was set to 30, the number of iterations was set to 100, and the parameter optimization range of SVM was set to (0.001, 100). By randomly generating 70% of the training set and running 30% of the test set independently for 10 times, the average result of 10 operations of the test set was taken as the final experimental result. SVM classification accuracy was set as a fitness function. The experimental results of 10 times are shown in Table 2. We can see that the average accuracy of the SVM model optimized based on HHO and GA (HHO_GA_SVM) is 85.56%, which is 1.59% higher than that of GA_SVM; the average precision rate is 86.45%, which is 1.39% higher; the average recall rate is 85.68%, which is 1.76% higher; the average F1 value is 85.73%, which is 1.79% higher. It can also be seen in Figure 5A.

In order to verify the effectiveness of this method, a control group experiment was set up in this chapter. On the one hand, the longitudinal comparison was performed, that is, the SVM model optimized based on HHO and GA (HHO_GA_SVM) was compared with the SVM model optimized by GA (GA_

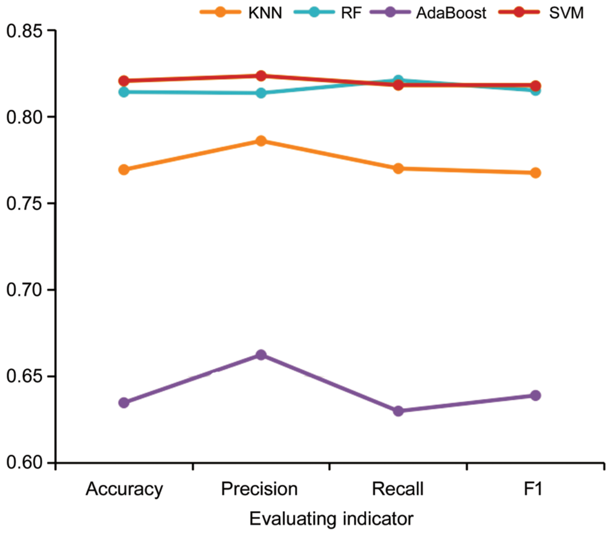

Figure 3 Evaluation indicators of the models KNN: K-nearest Neighbor algorithms; RF: Random Forest model; AdaBoost: Adaptive Boosting; SVM: Support Vector Machine.

Table 2 Ten times of optimization experiment

Number	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	85.90	86.81	86.02	86.06
2	84.62	85.02	85.11	84.94
3	85.90	86.93	86.02	86.13
4	85.90	86.81	86.02	86.06
5	85.90	86.81	86.02	86.06
6	85.68	85.68	85.48	85.45
7	85.90	86.93	86.02	86.13
8	85.26	85.69	85.58	85.54
9	85.90	86.93	86.02	86.13
10	84.62	86.92	84.49	84.83
Avg	85.56	86.45	85.68	85.73

SVM) and the basic SVM model. On the other hand, the performance of the HHO_GA_SVM model was compared horizontally with Particle Swarm Optimization (PSO) optimized, Whale Optimization Algorithm (WOA) optimized, and Grey Wolf Optimizer (GWO) optimized. It is shown in Figure 5. At the same time, the convergence curves of these models for horizontal comparison can be seen in Figure 6. In Figure 5A, we can see that the accuracy rate, precision rate, recall rate and F1 value of GA_SVM are improved compared with SVM, while HHO_GA_SVM is further improved compared with GA_SVM. In Figure 5B, we can see that HHO_GA_SVM and GWO_GA_SVM have relatively high accuracy rate, precision rate, recall rate and F1 value, followed by PSO_GA_SVM and WOA_GA_SVM have the lowest.

In order to more directly observe the convergence ability of HHO and comparison algorithms, the convergence curve of the model is plotted in Figure 6, where the horizontal axis represents the number of iterations and the vertical axis represents the average value obtained by 30 independent runs. Through the convergence curve, we can see the speed of the algorithm convergence, the accuracy of the algorithm convergence and whether the algorithm can jump out of the local optimal value. It can be seen from Figure 6 that HHO_GA_SVM and GWO_GA_SVM have the best convergence accuracy, followed by PSO_GA_SVM, and WOA_GA_SVM has the lowest. HHO_GA_SVM converges faster than GWO_GA_SVM and can jump out of local optimal values better. In general, HHO_GA_SVM has better convergence accuracy and faster convergence speed, showing strong optimization ability.

DISCUSSION

With the rapid development in recent years, TCM has been widely used in different area. After a large number of clinical studies and basic research, TCM has shown a strong charm in disease prevention and treatment, and has gained wide international attention. In this process, many people have realized that how to use it correctly is the

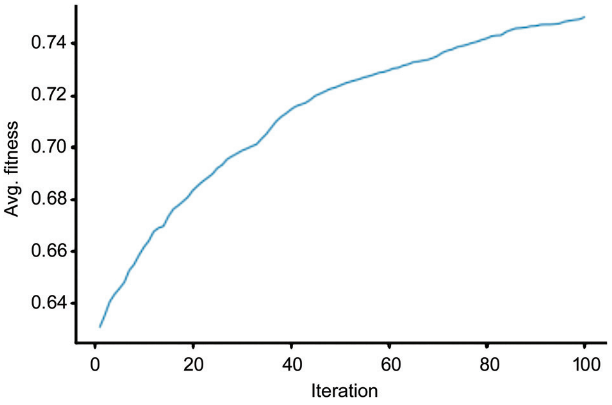

Figure 4 Convergence curve of fitness over iterations.

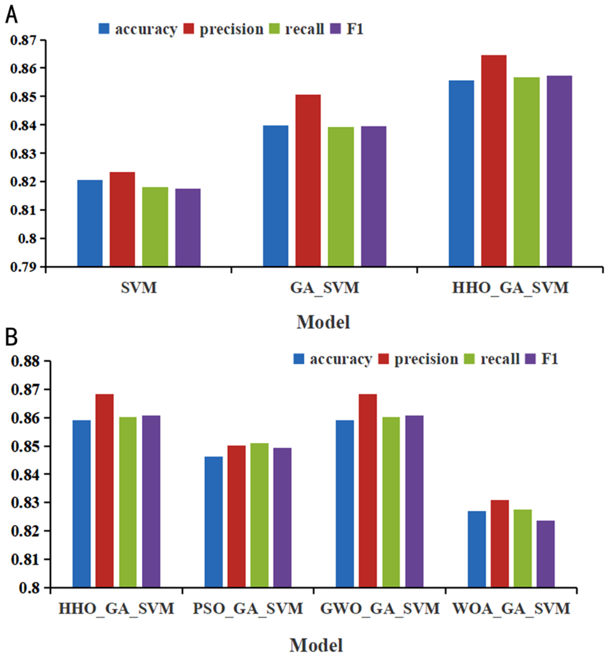

Figure 5 Comparison of Model Performance A: Vertical comparison; B: Horizontal comparison. SVM: Support Vector Machine; GA: Genetic Algorithm; HHO: Harris Hawk Optimization.

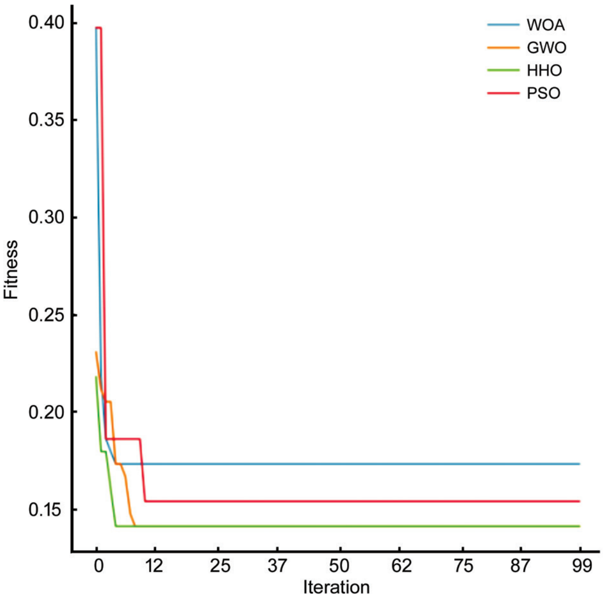

Figure 6 Convergence curve of the model WOA: Whale Optimization Algorithm; GWO: Grey Wolf Optimizer; HHO: Harris Hawk Optimization; PSO: Particle Swarm Optimization.

key, not just dissecting it. With the rapid development of artificial intelligence, it has become possible to replicate TCM experts and improve the reliability and accuracy of TCM diagnosis. TCM intelligent assisted diagnosis mainly includes the collection of TCM four diagnostic information and TCM syndrome differentiation. However, the current TCM intelligent assisted diagnosis research mainly focuses on modern electronic instruments, which have become mature. There is a lack of research on TCM intelligent syndrome differentiation, but it is worth exploring.

Accordingly, we launched a model study of TCM intelligent syndrome differentiation, and took DR as an example. First of all, mainly through the analysis of the real-world case, the proportion of TCM syndromes was consistent with the clinical distribution law. Based on the syndrome differentiation and treatment in TCM, DR is mainly characterized by the deficiencies of Qi and Yin[32], which accounts for the largest proportion. In addition, it was found that the prescriptions of TCM or herbs have great advantages in the treatment of DR, with the primary functions of clearing heat, promoting blood circulation and removing blood stasis, and replenishing Qi[21]. In other words, liver and stomach depression heat syndrome, and gastrointestinal damp-heat syndrome are also the main classification of TCM syndromes. Yin and Yang deficiency are the main TCM syndromes of severe DR, and the corresponding skin advanced glycation end products level is also higher[33]. Blurred vision is the most frequent symptom, consistent with the main symptom of DR.

Subsequently, we used a variety of machine learning methods to classify TCM syndromes, and the SVM model achieved good performance. Further, we optimize the model by feature selection and parameter optimization. Based on the optimal feature combination, which was screened by GA, the accuracy of the model has been improved by 1.92%. The optimal combination of selected features accords with the clinical manifestation of TCM syndrome differentiation. For example, feeble and shortness of breath are typical symptoms of Qi deficiency, which means that if a patient experiences symptoms such as fatigue and shortness of breath, the high probability belongs to Qi deficiency syndrome[34]. Polydipsia, irritability, dislike heat, bitter taste, red tongue nature, and yellow tongue coating are the main clinical manifestations of heat syndrome, including liver and stomach depression heat syndrome and gastrointestinal damp-heat syndrome[35]. Feeble pulse belongs to deficiency syndrome, including Qi deficiency, Yin deficiency, and Yang deficiency[36].

Then, different optimization algorithms are used for parameter optimization, and the results show that the optimization effect of HHO is the best. The accuracy of HHO_GA_SVM is better than GA_SVM improved by 1.59%, which is 3.51% higher than the basic SVM. Syndrome differentiation is a methodology unique to TCM. The syndrome is a summary of the body's reaction state according to the signs and symptoms collected by the physician using TCM four diagnostic methods and other diagnostic techniques. In other words, the syndrome type is the essence of the disease reflected by TCM symptoms. It can be seen that the dialectical method is consistent with the principle of HHO, the diagnosis of the syndrome type is equivalent to the prey caught by Harris Hawk, and the symptom group is equivalent to the Harris Hawk group. The experimental results also verify that HHO_GA_SVM can simulate the process of TCM syndrome differentiation of DR and quantitatively depict this nonlinear relationship.

In conclusion, this study uses real-world medical records as research objects, which is the basis for the conformity of research results with clinical practice. Combining mathematical algorithm and TCM syndrome differentiation, this paper puts forward a set of methods to transform the relatively fuzzy nonlinear TCM syndrome differentiation theory into objective and quantitative mathematical model. In other words, this pertains to clinical ophthalmic artificial intelligence research. According to the guidelines on clinical research evaluation of artificial intelligence in ophthalmology[37], there are certain limitations within this study that warrant attention. First of all, in terms of developing data set evaluation, the case data in this study are small sample data with uneven distribution, and the ability of the model to approach the real situation of TCM syndrome differentiation is limited. Moreover, the characteristics of the data only include the TCM four diagnostic information, neglecting to incorporate modern medical examination indices and ophthalmic images. The following work can not only increase the sample size, but also integrate the TCM four diagnostic information, examination indicators and ophthalmic images to build a comprehensive, multi-dimensional and multi-modal diagnostic model that merges Chinese and Western medicine. Standardized and reasonable clinical trial design, appropriate data collection, management, and effective data analysis methods are the key to solving these problems[37]. In addition, concerning the evaluation of model interpretability, this study primarily relies on expert experience to interpret experimental outcomes, gauging their alignment with clinical practice. The interpretability of the model needs to be improved. Providing a more systematic description of three important aspects of model building and evaluation, including normalization, hyperparameter optimization and cross-validation, which can improve the interpretability of SVM research and, in fine, its clinical adoption[38].

ACKNOWLEDGEMENTS

Foundations: Supported by Hunan Province Traditional Chinese Medicine Research Project (No.B2023043); Hunan

Provincial Department of Education Scientific Research Project (No.22B0386); Hunan University of Traditional Chinese Medicine Campus level Research Fund Project (No.2022XJZKC004).

Conflicts of Interest: Xiao L, None; **Wang CW,** None; **Deng Y,** None; **Yang YJ,** None; **Lu J,** None; **Yan JF,** None; **Peng QH,** None.

REFERENCES

1 Wang J, Ma Q, Li Y, Li P, Wang M, Wang T, Wang C, Wang T, Zhao B. Research progress on Traditional Chinese Medicine syndromes of diabetes mellitus. *Biomed Pharmacother* 2020;121:109565.

2 Tian J, Jin D, Bao Q, Ding Q, Zhang H, Gao Z, Song J, Lian F, Tong X. Evidence and potential mechanisms of traditional Chinese medicine for the treatment of type 2 diabetes: a systematic review and meta-analysis. *Diabetes Obes Metab* 2019;21(8):1801-1816.

3 Yang X, He T, Han S, Zhang X, Sun Y, Xing Y, Shang H. The role of traditional Chinese medicine in the regulation of oxidative stress in treating coronary heart disease. *Oxid Med Cell Longev* 2019;2019:3231424.

4 Shao-Mei W, Li-Fang Y, Li-Hong W. Traditional Chinese medicine enhances myocardial metabolism during heart failure. *Biomed Pharmacother* 2022;146:112538.

5 Xiao L, Ma N, Lai H, Yan J, Peng Q. Analysis of the hotspots and trends in traditional Chinese medicine immunomodulation research based on bibliometrics. *Digital Chinese Medicine* 2021;4(2):81-91.

6 Wang Y, Zhang Q, Chen Y, Liang C, Liu H, Qiu F, Dai Z. Antitumor effects of immunity-enhancing traditional Chinese medicine. *Biomed Pharmacother* 2020;121:109570.

7 Du HZ, Hou XY, Miao YH, Huang BS, Liu DH. Traditional Chinese Medicine: an effective treatment for 2019 novel coronavirus pneumonia (NCP). *Chin J Nat Med* 2020;18(3):206-210.

8 Li S, Liu C, Guo F, Taleb SJ, Tong M, Shang D. Traditional Chinese medicine as potential therapy for COVID-19. *Am J Chin Med* 2020;48(6):1263-1277.

9 Harrison JE, Weber S, Jakob R, Chute CG. ICD-11:an international classification of diseases for the twenty-first century. *BMC Med Inform Decis Mak* 2021;21(Suppl 6):206.

10 Meng X, Liu X, Tan J, Sheng Q, Zhang D, Li B, Zhang J, Zhang F, Chen H, Cui T, Li M, Zhang S. From Xiaoke to diabetes mellitus: a review of the research progress in traditional Chinese medicine for diabetes mellitus treatment. *Chin Med* 2023;18(1):75.

11 Zhang L, Miao R, Yu T, Wei R, Tian F, Huang Y, Tong X, Zhao L. Comparative effectiveness of traditional Chinese medicine and angiotensin converting enzyme inhibitors, angiotensin receptor blockers, and sodium glucose cotransporter inhibitors in patients with diabetic kidney disease: a systematic review and network meta-analysis. *Pharmacol Res* 2022;177:106111.

12 Zhang F, Petersen M, Johnson L, Hall J, O'Bryant SE. Recursive support vector machine biomarker selection for Alzheimer's disease. *J Alzheimers Dis* 2021;79(4):1691-1700.

13 Fernandez Escamez CS, Martin Giral E, Perucho Martinez S, Toledano Fernandez N. High interpretable machine learning classifier for early glaucoma diagnosis. *Int J Ophthalmol* 2021;14(3):393-398.

14 Fu W, Xu L, Yu Q, Fang J, Zhao G, Li Y, Pan C, Dong H, Wang D, Ren H, Guo Y, Liu QJ, Liu J, Chen X. Artificial intelligent olfactory system for the diagnosis of Parkinson's disease. *ACS Omega* 2022;7(5):4001-4010.

15 Elgammal YM, Zahran MA, Abdelsalam MM. A new strategy for the early detection of Alzheimer disease stages using multifractal geometry analysis based on K-Nearest Neighbor algorithm. *Sci Rep* 2022;12(1):22381.

16 Hu Q, Yu T, Li J, Yu Q, Zhu L, Gu Y. End-to-End syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine. *Comput Methods Programs Biomed* 2019;174:9-15.

17 Lu PH, Chiang CC, Yu WH, Yu MC, Hwang FN. Machine learning-based technique for the severity classification of sublingual varices according to traditional Chinese medicine. *Comput Math Methods Med* 2022;2022:3545712.

18 Kang H, Zhao Y, Li C, Chen YJ, Tang KL, Yang LL, Ma C, Peng JH, Zhu RX, Liu Q, Hu YY, Cao ZW. Integrating clinical indexes into four-diagnostic information contributes to the Traditional Chinese Medicine (TCM) syndrome diagnosis of chronic hepatitis B. *Sci Rep* 2015;5:9395.

19 Xu ZH, Gao YY, Zhang HT, Ruan KF, Feng Y. Progress in experimental and clinical research of the diabetic retinopathy treatment using traditional Chinese medicine. *Am J Chin Med* 2018:1-27.

20 Pang B, Li QW, Qin YL, Dong GT, Feng S, Wang J, Tong XL, Ni Q. Traditional Chinese medicine for diabetic retinopathy: a systematic review and meta-analysis. *Medicine (Baltimore)* 2020;99(7):e19102.

21 Ai X, Yu P, Hou Y, Song X, Luo J, Li N, Lai X, Wang X, Meng X. A review of traditional Chinese medicine on treatment of diabetic retinopathy and involved mechanisms. *Biomed Pharmacother* 2020;132:110852.

22 Xiao L, Yang YJ, Liu Q, Peng J, Yan JF, Peng QH. Visualizing the intellectual structure and recent research trends of diabetic retinopathy. *Int J Ophthalmol* 2021;14(8):1248-1259.

23 Xie M, Deng L, Yu Y, Xie X, Zhang M. The effects of Bushen Yiqi Huoxue prescription and its disassembled prescriptions on a diabetic retinopathy model in Sprague Dawley rats. *Biomed Pharmacother* 2021;133:110920.

24 Wang L, Li S, Wang L, Lin K, Du J, Miao W, Zhang L. Uncovering the protective mechanism of Taohong Siwu Decoction against diabetic retinopathy via HIF-1 signaling pathway based on network analysis and experimental validation. *BMC Complement Med Ther* 2020;20(1):298.

25 Liu QP, Chen YY, Yu YY, An P, Xing YZ, Yang HX, Zhang YJ, Rahman K, Zhang L, Luan X, Zhang H. Bie-Jia-Ruan-Mai-Tang, a Chinese medicine formula, inhibits retinal neovascularization in diabetic mice through inducing the apoptosis of retinal vascular endothelial cells. *Front Cardiovasc Med* 2022;9:959298.

26 Wong TY, Sun J, Kawasaki R, *et al*. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology* 2018;125(10):1608-1622.

27 Peng Q. *Ophthalmology of Traditional Chinese Medicine* (Textbook of the 14th Five Year Plan for Higher Education in the National Traditional Chinese Medicine Industry). Beijing: China Press of Traditional Chinese Medicine 2021.

28 Peng Q. *Ophthalmofundology of Integrated Chinese and Western Medicine*. Beijing: People's Military Medical Press 2011.

29 Zhu W. *Syndrome Element Differentiation*. Beijing: People's Medical Publishing House 2008.

30 Medicine SAoTC. *Traditional Chinese Medicine Clinical Diagnosis and Treatment Terminology* (Revised Edition). Beijing: National Standards Publishing House, 2020.

31 Heidari AA, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H. Harris Hawks optimization: algorithm and applications. *Future Generation Computer Systems* 2019;97:849-872.

32 Ou C, Yang YJ, Peng QH. Yiqi Yangyin Huoxue Method in Treating Diabetic Retinopathy: A Systematic Review and Meta-Analysis. *Evid Based Complement Alternat Med*. 2019;2019:6020846.

33 Zhang S, Ma P, Chen Q. The correlation between the level of skin advanced glycation end products in type 2 diabetes mellitus and the stages of diabetic retinopathy and the types of traditional Chinese medicine syndrome. *Evid Based Complement Alternat Med* 2022;2022:5193944.

34 Dou Z, Xia Y, Zhang J, Li Y, Zhang Y, Zhao L, Huang Z, Sun H, Wu L, Han D, Liu Y. Syndrome differentiation and treatment regularity in traditional Chinese medicine for type 2 diabetes: a text mining analysis. *Front Endocrinol* 2021;12:728032.

35 Lee TC, Lo LC, Wu FC. Traditional Chinese medicine for metabolic syndrome via TCM pattern differentiation: tongue diagnosis for predictor. *Evid Based Complement Alternat Med* 2016;2016:1971295.

36 Mei MF. A systematic analysis of the theory and practice of syndrome differentiation. *Chin J Integr Med* 2011;17(11):803-810.

37 Yang WH, Shao Y, Xu YW, Expert Workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023), Ophthalmic Imaging and Intelligent Medicine Branch of Chinese Medicine Education Association, Intelligent Medicine Committee of Chinese Medicine Education Association. Guidelines on clinical research evaluation of artificial intelligence in ophthalmology (2023). *Int J Ophthalmol* 2023;16(9):1361-1372.

38 Joucla C, Gabriel D, Ortega JP, Haffen E. Three simple steps to improve the interpretability of EEG-SVM studies. *J Neurophysiol* 2022;128(6):1375-1382.