RESEARCH ARTICLE

Optimization of Traditional Chinese Medicine concoction process based on machine learning algorithm

Jinyang Li, Tingting Hu*

Bozhou University, Bozhou, Anhui, China.

Received: July 16, 2024; **accepted:** November 20, 2024.

The traditional Chinese medicine (TCM) concoction process involves the preparation of herbal medicines and remedies based on ancient Chinese medicinal principles. Herbalists begin by choosing herbs based on their medicinal qualities and interactions and then mixing them into specific formulations catered to the needs of each patient. In this study, a novel sea lion-optimized efficient random forest (SLO-ERF) method was proposed for the concoction process of traditional Chinese medicine. The efficacy of the herbs and formulas were analyzed using the proposed method. The TCM herbs were collected from traditional Chinese medicine system pharmacology (TCMSP). The names, properties, and effectiveness features of the herbs were chosen to encode them, and then the formulas and herb vectors were created. A total of 15 formula effectiveness categories were gathered based on formulas found in TCM literature. The results showed a substantial correlation between herb efficacy and formula efficacy, suggesting that the herb efficacy-formula vector had the greatest influence on the proposed model. The proposed method achieved better performance in precision (94.38%), recall (92.62 %), and F1-score (91.85 %). When combining the SLO-RF model with formula feature representation, how well a formulation functioning could be more accurately classified, which offered novel directions for researching the compatibility of TCM formulas.

Keywords: Traditional Chinese Medicine (TCM); Sea Lion Optimized Efficient Random Forest (SLO-ERF); herbs; concoction.

***Corresponding author:** Tingting Hu, Bozhou University, Bozhou 236800, Anhui, China. Email: 1833100881@163.com.

Introduction

Chinese herbal medicine processing is a type of pharmaceutical technology that has been extensively refined over generations to improve the potency and lessen the toxicity of herbs used in traditional Chinese medicine (TCM) [1]. Considering that herbal processing modifies the chemical active components of herbal medicines and, hence, their actions, it is necessary for the safe and efficient use of TCM in healthcare. Chinese herbal treatments contain alkaloids as one of their main active components. Chinese herbal remedies rich in alkaloids can undergo a variety of processing techniques that cause intricate alterations in the alkaloids [2]. TCM typically prepares herbal medications high in alkaloids by washing, chopping, stirring, adding liquid adjuvant while stirring, and using water decoction. The manufacturing of herbal remedies high in alkaloids frequently uses river sand, wine, vinegar, brine solution, honey, and herbal juice as an adjuvant [3]. Research has been focused on standardizing TCM syndromes to clarify the scientific foundations of treating multiple illnesses with the same technique [4]. A realistic molecular analysis of the concept was performed using information extraction and advanced network technologies and found the molecular signs of TCM diseases like a lack of Qi and blood clots linked to heart disease and stroke. Zhang *et al.* developed a fine-grained entity classification corpus and established corresponding annotation guidelines for clinical records associated with Traditional Chinese Medicine (TCM) [5]. The research created a four-step process that worked well for creating TCM medical records inside the corpus. This process served as a framework for tasks related to named entity identification and the creation of a corpus within the TCM field.

Li *et al*. introduced a natural language processing (NLP) method based on text categorization to the field of TCM with the goal of efficiently and scientifically lengthening the TCM compound decoction time [6]. Weighing the feature vectors utilized for plant identification and medicinal components was accomplished by the improvement of Term Frequency-Inverse Word Frequency (TF-IWF) with a multi-dimensional herbal feature vector combining those weighted feature vectors, enabling a more thorough depiction. Wang *et al*. presented a network-based approach to measure the interactions between herb combinations and found shorter distances between commonly used herb pairs and random herb pairs, suggesting that a medicinal herb pair was more likely to influence nearby enzymes in the human interactome [7]. Calculating the center distance at the ingredient level improved the ability of differentiation between random and common herb pairings. Another study used Danshen-Chuanxiong (DS–CX), a most common herbal pair, to find more scientific information on possible herb-drug interactions and investigated the molecular interactions between Western cardiovascular drugs and herbal medicines that helped with blood clotting [8]. Herb-drug interactions were important to consider from a therapeutic standpoint, particularly when using many herbs simultaneously. Eigenschink *et al*. presented a study result supporting the fundamental theories of TCM including Qi, meridians, acupuncture, pulse, tongue, and herbal remedies, which addressed whether TCM-related study material corresponded to the current standards for evidence-based research as outlined in recommendations for excellent clinical practice and science. An old understanding of herbal species and preparations could indicate an inheritance as opposed to a Pandora's Box [9]. Wu *et al*. proposed an objective and adaptive neuro-fuzzy inference system (ANFIS) to diagnose sleep disturbances in TCM using four diagnostic techniques including palpation, hearing and smelling, inspection, and questioning. The results found that 8 out of 92 cases diagnosed by six TCM physicians had different diagnoses, suggesting that the proposed model was capable of unbiased reasoning [10]. The frequency of allergy disorders has risen dramatically. Three factors could be responsible for this increase, which include environmental changes, hygiene theory, and epigenetic alterations. The term "allergic diseases" refers to a collection of long-term, diverse illnesses that can be fatal or cause severe morbidity such as asthma, allergenic rhinitis, atopic eczema, and anaphylaxis. The main purpose of symptomatic therapy for these individuals is symptom management. However, there is currently no known cure [11, 12]. The treatment of TCM is intricate and varied, particularly in the use of herbal remedies. TCM generally uses mixed prescriptions of many herbs to treat various ailments. As science and technology have progressed, a multitude of techniques and tools have surfaced, offering increasingly potent and accurate ways to examine intricate systems [13]. Traditional processing philosophy and techniques have consistently endured in the clinical application of herbal medicine despite the lack of knowledge about their underlying scientific foundations [14]. The extensive collection of clinical medical records, particularly those pertaining to refined procedures collected by ancient doctors, serves as the foundation for the examination of efficaciousness and formulae. TCM theory does not have any objective norms, which is empirical. There are complex interactions between procedures, botanicals, and effectiveness [15].

There is a lack of standardized methods for TCM formulation. Variations in substances, quantities, and preparation techniques used by different practitioners for the same ailment could result in distinctions between safety and effectiveness. This study proposed a new method for the traditional Chinese medicine concoction process using Sea lion optimized efficient random forest (SLO-ERF) to streamline the TCM formulation procedure. The proposed approach analyzed the effectiveness of the procedure using the herbs gathered from the traditional Chinese medicine system pharmacology (TCMSP) with 15 formula efficacy categories based on formulae discovered in TCM literature. This study would provide a robust correlation between the effectiveness of herbs and formulas and investigate the influence of herb efficacy formula vector on the proposed model.

Materials and methods

Dataset

A total of 2,664 formulas from both ancient and contemporary TCM texts were included in this study, which were categorized into 15 effective types based on moistening dryness, refreshing, regulating blood, to handle a variety of stroke symptoms. Further, details from 1,054 TCM herbs including their characteristics and effectiveness were included in additional categorization tests, which were retrieved from TCMSP database (http://sm.nwsuaf.edu.cn/lsp/tcmsp.php). Structure files of molecules were downloaded from PubChem (https://pubchem.ncbi.nlm.nih.gov/), Compound database (https://www.chemdiv.com), ChEMBL (https://www.ebi.ac.uk/chembl/), ChemSpider (https://www.chemspider.com/), and further optimized by Sybyl 6.9 (Tripos, Inc., St. Louis, Missouri, USA) with Sybyl force field and default parameters. Python 3.10.1 (https://www.python.org/) was employed to implement the proposed method on a Windows 10 laptop with an Intel i7 core CPU and 8 GB of RAM. Tensor Flow/Keras (https://www.tensorflow.org/) or Scikit-Learn (https://scikit-learn.org/stable/) were used to train the proposed model using the training dataset. The proposed method SLO-ERF was compared with Word-to-Vector model + long short-term memory (W2V+LSTM), Word-to-Vector model + self-attention (W2V+SA), and bidirectional encoder representations from transformers (BERT) approach (Roberta-large) [16]. The performance metrics included precision, recall, and F1 score.

Sea lion optimized efficient random forest (SLO-ERF) for TCM concoction process

SLO-ERF is an innovative machine learning method created for TCM formulation process optimization, which combines effective random forest techniques with sea lion-inspired behavior tactics for foraging and decision-making. By improving the precision and efficacy of forecasting the ideal TCM concoction techniques, this hybrid method intends to improve TCM formulation safety and medicinal efficacy. SLO was applied to maximize the choice and blending of therapeutic herbs or components in TCM concoction procedures.

(1) Identifying and monitoring stage:

The form, position, and size of a target were all discerned by sea lions to identify the existence and location of the targets. The forefront sea lion alerted the other members of its subgroup to the target location, while the others adjusted their positions accordingly as described in equation 1.

$$\overrightarrow{Dist} = \left|\overrightarrow{2C}.\overrightarrow{O(s)} - \overrightarrow{SL}(s)\right| \qquad (1)$$

where $\overrightarrow{Dist}$ was the distance between the target and the sea lion. $\overrightarrow{SL}(s)$and $\overrightarrow{O(s)}$ were the vector positions of the sea lion and the target. s was the current iteration of concoction process. $\vec{C}$ was a random vector ranging from $0\ to$ 1. The sea lion approached its target in the subsequent iteration as equation 2.

$$SL\ \overrightarrow{(s+1)} = \overrightarrow{O(s)} - \overrightarrow{Dist}.\vec{D} \qquad (2)$$

where $(s+1)$ was the next repetition and decreases during the repetition period from two to zero because this decrease induced the whole group moving to the direction of the current target and encircling its contents.

(2) Vocal Stage:

Sea lions utilize various vocalizations to convey messages to one another for hunting and catching targets in both air and water. Similarly, the Chinese medicine concoction process helps in identifying the disease and diagnosed fast. This tendency was expressed as follows.

$$\overrightarrow{sp_{leader}} = \left|\left(\frac{\vec{U}_1(1+\vec{U}_2)}{\vec{U}_2}\right)\right| \quad (3)$$

$$\vec{U}_1 = sin\theta \quad (4)$$

$$\vec{U}_2 = sin\phi \quad (5)$$

where $\vec{U}_2$ and $\vec{U}_1$ were the voices' velocities in the air and water, respectively. $\overrightarrow{sp_{leader}}$ was the velocity of the sea lion leader's voice. sin ϕ was the first item. sin θ was the second item.

(3) Phase of attack (exploitation):

Once the target was located, the most skilled search agent acted as a leader by deciding on the best hunting strategy and alerting other individuals. Target was typically considered to be the greatest answer offered by the present candidates. However, improved target identification and surroundings were possible by creating a new search agent using reduction in encirclement strategy that the group leader moved to encircle the target. However, the search individual’s input location could vary from anywhere between the starting point and the position of the best agent. Therefore, the circle adjusting position could be used by pursuing and starting the search from the margins. Both methods could be conjunct as follows.

$$\overrightarrow{SL}(s+1) = \left|\vec{O}(s) - \overrightarrow{SL}(s)\right|.\cos(2\pi n) + \vec{O}(s) \quad (6)$$

where n was an arbitrary amount between -1 and 1. $\left|\vec{O}(s) - \overrightarrow{SL}(s)\right|$ was the distance between the search agent and the target in best optimum solution. The search agent circled around the target if it was situated on the outermost portion.

(4) Searching for target (exploration):

To locate the target, sea lions swim in zigzag patterns. Thus, in the given procedure, $\vec{D}$ was employed with the random quantity. Sea lions were compelled to migrate away from the target and the leading agent if $\vec{D}$ was larger than 1 or less than negative. Under these circumstances, sea lions began searching for different targets as below.

$$\overrightarrow{D\imath st} = \left|\overrightarrow{2C}.\overrightarrow{SL}_{rnd}(s) - \overrightarrow{SL}(s)\right| \quad (7)$$

$$\overrightarrow{SL}(s+1) = \overrightarrow{SL}_{rnd}(s) - \overrightarrow{D\imath st}.\vec{D}, \quad (8)$$

where the randomly selected sea lion from the current group was represented as $\overrightarrow{SL}_{rnd}(s)$.

(5) Efficient random forest (ERF):

Random forest classification has varied accuracy, which is recommended to pick the decision tree. The most suitable feature for dividing the nodes will be selected to build a new splitting rule that corresponds to the selection and separation of node characteristics. The distinct properties were observed in various decision trees when alternative node-splitting strategies were chosen for an identical data set. The node splitting technique was separated into a linear composition as described in equations 9 and 10.

$$Gain\ (C,b) = Ent(C) - \sum_{u=1}^{U}\frac{|C^u|}{|C|}Ent(C^u) \quad (9)$$

$$Gini(C,b) = \sum_{u=1}^{U}\frac{|C^u|}{C}Gini(C^u) \quad (10)$$

As the random forest approach of Spark mllib was connected to $ID3$ and CART, the optimization of node split was considered in both methods. The node splitting algorithm showed statistics like *Gain* and $Gini$ indices after dividing the sample set C according to characteristic b in equations 11 and 12.

$$Ent(C) = -\sum_{l=1}^{|z|} o_l log_2 o_l \tag{11}$$

$$Gini(C) = \sum_{l=1}^{|z|} \sum_{l' \neq l} pkpk' = 1 - \sum_{l=1}^{|z|} pk^2 \tag{12}$$

where samples in C with a value of u on the attribute b^u were contained in the u branch node as indicated by the value C^u in equation 13.

$$G = \min_{\propto,\beta \in Q} E\{C,b\} = \alpha\, Gini(C,b) - \beta Gain(C,b) \begin{cases} \alpha + \beta = 1 \\ 0 \leq \alpha, \beta \leq 1 \end{cases} \tag{13}$$

The combination of node dividing formula and adaptive parameter selection procedure were based on the idea that node splitting should strive for a greater integrity of the data collection after division, where α and β were the attribute splitting weight coefficients. G has a negligible value in the interim period. To find the best combination of parameters, the adaptive parameter selection method was used, which indicated that, to maximize the classification impact, $ID3$ and CART were the best options for node partition criteria. To gauge the experiment's performance, the accuracy and classification error rates were determined using equation 14 for sample C and equation 15 for the reliability rate.

$$F(e; C) = \frac{1}{n} \sum_{j=1}^{n} JJ\left(e\left(w_j\right) \neq z_j\right) \tag{14}$$

$$acc(e; C) \frac{1}{n} \sum_{j=1}^{n} JJ\left(e\left(w_j\right) = z_j\right) = 1 - F(e; C) \tag{15}$$

SLO-ERF strategy

The SLO-ERF strategy optimized the TCM formulation process by utilizing machine learning techniques. In traditional Chinese medicine, concoction refers to the methodical blending and preparation of many herbs to provide the intended therapeutic outcomes, while guaranteeing efficacy and safety. SLO-ERF combined the effectiveness and forecasting capability of RF algorithm with SLO concepts. SLO is a population-based search method that emulates the motion and decision-making of sea lions in the wild to optimize complicated problems. The technique was motivated by the foraging behavior of sea lions and was used to improve the Random Forest algorithm, a powerful machine learning method that is well-known for its capacity to manage high-dimensional data while preserving forecast reliability. TCM practitioners could be able to streamline and maximize the selection and mixing of herbal combinations by utilizing SLO-ERF, which would improve the efficacy and consistency of TCM procedures, while maintaining the conventional values of herbal medicine. The algorithm of SLO-ERF was shown in Figure 1.

Results and discussion

Precision

When discussing the Chinese medicine preparation process, precision refers to the accuracy and effectiveness of measuring, mixing, and preparing the medicinal ingredients according to the principles of Chinese medicine. Every stage of the procedure requires careful attention to details from choosing the appropriate herbs and components to figuring out their amounts and preparation techniques. The precision performance of this study showed that the proposed SLO-ERF method achieved 94.38% precision rate compared to the W2V+LSTM, W2V+SA, and Roberta-large models, which achieved 78.74%, 82.83%, and 92.22%, respectively (Figure 2). The results confirmed that, when compared to the current existing TCM concoction process methods, the proposed method worked better.

Recall

When discussing the concoction methods used in TCM, the term recall usually refers to the capacity to easily retrieve or replicate a particular herbal composition or prescription, which details the exact classification and measurements of the components utilized, together with information on their ratios, preparation techniques, and any other pertinent guidelines for the mixing

```
function SLO_ERF(data, parameters):
    selected_features = SeaLionOptimization(data)
    data_selected = preprocess_data(data, selected_features)
    train_set, test_set = split_data(data_selected)
    forest = RandomForestClassifier(parameters)
    forest.train(train_set)
    predictions = forest.predict(test_set)
    accuracy = calculate_accuracy(predictions, test_set.labels)
    return accuracy
function SeaLionOptimization(data):
    initialize sea lions randomly
    while not terminate_condition:
        for each sea lion:
            update_position_based_on_fitness_function
            update_position_based_on_social_behavior
            update_position_based_on_crowding_distance
    return selected_features
function preprocess_data(data, selected_features):
    processed_data = data[selected_features]
    processed_data = normalize(processed_data)
    return processed_data
function split_data(data):
    train_set, test_set = split(data)
    return train_set, test_set
function calculate_accuracy(predictions, true_labels):
    accuracy = correct_predictions / total_predictions
    return accuracy
```

Figure 1. The algorithm of SLO-ERF.

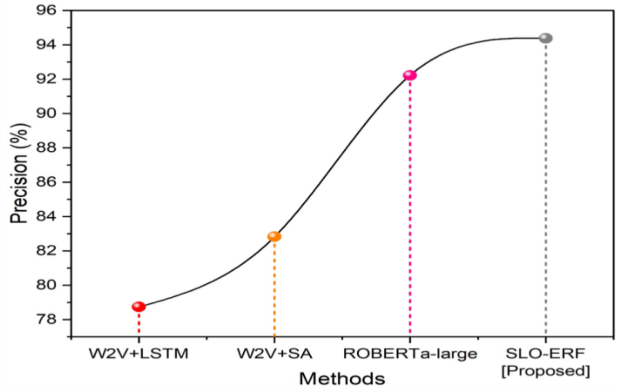

Figure 2. Precision performance.

procedure. In the TCM concoction process, the recall performance showed that the proposed SLO-ERF approach achieved 92.62% of recall compared to 66.14%, 66.86%, and 86.71% in W2V+LSTM, W2V+SA, and Roberta-large methods, respectively (Figure 3). The results confirmed that the proposed approach outperformed the existing methods.

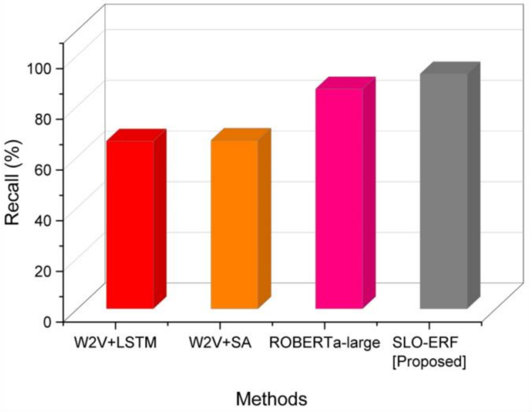

Figure 3. Recall performance.

F1-Score

F1 score integrates recall and precision into a single metric that is frequently utilized in the assessment for the TCM formulation process. Recall evaluates the coverage of real positive events, whereas precision evaluates the accuracy of an optimistic forecast. The F1 score denotes a well-balanced achievement in discovering efficient combinations of TCM components since it reflects both high and strong recall. The results of F1 score performance were displayed in Figure 4. The proposed SLO-ERF approach obtained 91.85% F1 score, while W2V+LSTM, W2V+SA, and Roberta-large methods achieved 71.89%, 73.99%, and 89.38% F1 scores, respectively. As compared to the existing approaches in the TCM concoction process, the proposed method performed better.

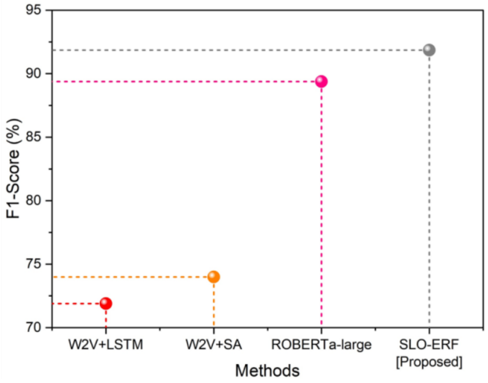

Figure 4. F1 score performance.

The technique of making herbal medications and treatments using TCM were based on TCM theories. In this research, a new SLO-ERF method for the TCM concoction process was proposed to evaluate the effectiveness of the formula and herbs. The TCM system of pharmacology gathered the TCM herbs. Formulas and herb vectors were created by encoding the herbs through the selection of their names, characteristics, and efficacy factors. Based on formulae discovered from TCM literatures, 15 formula efficacy categories were constructed in this study. The results indicated that there was a strong relationship between the effectiveness of herbs and formulas, indicating that the herb efficacy formula vector had the most impact on the proposed model. When compared with existing approaches, the proposed approach performed better in recall (92.62%), precision (94.38%), and F1 score (91.85%) than that of other existing models. Challenges with data integration, standardization of input variables, and interpretation of outcomes arose when traditional expertise was combined with contemporary computational approaches. These areas are critical to practitioners and regulatory organizations. Further investigation into hybrid models that fuse domain-specific knowledge with machine learning may improve the safety and effectiveness of TCM formulation procedures, opening the door to a wider range of applications in clinical practice.

Acknowledgements

This research was supported by Key Project of Anhui University Natural Science Research Project [Grant No. KJ2021A1144], The University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2023-073), 2023 Anhui Province Gaofeng Cultivation Discipline Construction-Traditional Chinese pharmacy (Anhui Education Secretary Department) (Grant No. 13,2023), and 5 Kinds of Traditional Chinese Medicine (include Radix *Achyranthis Bidentatae*) Standardized Construction Project (Grant No. BYH202442).

References

1. Wang WY, Zhou H, Wang YF, Sang BS, Liu L. 2021. Current policies and measures on the development of traditional Chinese medicine in China. Pharmacol Res. 163:105-187.
2. Luo L, Jiang J, Wang C, Fitzgerald M, Hu W, Zhou Y, *et al*. 2020. Analysis of herbal medicines utilized for treatment of COVID-19. Acta Pharm Sin B. 10(7):1192-1204.
3. Jiang W, Tang M, Yang L, Zhao X, Gao J, Jiao Y, *et al*. 2022. Analgesic alkaloids derived from traditional Chinese medicine in pain management. Front Pharmacol. 13:851508.
4. Zhai X, Wang X, Wang L, Xiu L, Wang W, Pang X. 2020. Treating different diseases with the same method—a traditional Chinese medicine concept analyzed for its biological basis. Front Pharmacol. 11:946.
5. Zhang T, Wang Y, Wang X, Yang Y, Ye Y. 2020. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. BMC Medical Inform Decis Mak. 20:1-17.
6. Li Z, Zhao H, Zhu G, Du J, Wu Z, Jiang Z, *et al*. 2024. Classification method of traditional Chinese medicine compound decoction duration based on multi-dimensional feature weighted fusion. Comput Methods Biomech Biomed Engin. 2024(1):1-15.
7. Wang Y, Yang H, Chen L, Jafari M, Tang J. 2021. Network-based modeling of herb combinations in traditional Chinese medicine. Brief Bioinformatics. 22(5):106.
8. Zuo HL, Linghu KG, Wang YL, Liu KM, Gao Y, Yu H, *et al*. 2020. Interactions of antithrombotic herbal medicines with Western cardiovascular drugs. Pharmacol Res. 159:104963.
9. Eigenschink M, Dearing L, Dablander TE, Maier J, Sitte HH. 2020. A critical examination of the main premises of Traditional Chinese Medicine. Wien Klin Wochenschr. 132:260-273.
10. Wu J, Hu R, Li M, Liu S, Zhang X, He J, *et al*. 2021. Diagnosis of sleep disorders in traditional Chinese medicine based on adaptive neuro-fuzzy inference system. Biomed Signal Process Control. 70:102942.
11. Wang Z, Wang ZZ, Geliebter J, Tiwari R, Li XM. 2021. Traditional Chinese medicine for food allergy and eczema. Ann Allergy Asthma Immunol. 126(6):639-654.
12. Zhang Y, Zhang L. 2022. A brief analysis of the correlation between allergic rhinitis and traditional Chinese medicine constitution. Journal of Clinical and Nursing Research. 6(5):1-8.
13. Cheung H, Doughty H, Hinsley A, Hsu E, Lee TM, Milner-Gulland EJ, *et al*. 2021. Understanding traditional Chinese medicine to strengthen conservation outcomes. People and Nature. 3(1):115-128.
14. Rizvi SA, Einstein GP, Tulp OL, Sainvil F, Branly R. 2022. Introduction to traditional medicine and its role in prevention and treatment of emerging and re-emerging diseases. Biomolecules. 12(10):1442.
15. Song S, She Z. 2022. Quantum theory-based physical model of the human body in TCM. Digital Chinese Medicine. 5(4):354-359.
16. Shi QY, Tan LZ, Seng LL, Wang HJ. 2021. Intelligent prescription-generating models of traditional Chinese medicine based on deep learning. World Journal of Traditional Chinese Medicine. 7(3):361-369.