Mathematical Biosciences and Engineering

MBE, 20(10): 18695–18716.
DOI: 10.3934/mbe.2023829
Received: 21 June 2023
Revised: 09 September 2023
Accepted: 11 September 2023
Published: 08 October 2023

http://www.aimspress.com/journal/MBE

Research article

Multiple prescription pattern recognition model based on Siamese network

Wangping Xiong[1], Kaiqi Wang[1], Shixiong Liu[1], Zhaoyang Liu[1], Yimin Zhu[1], Peng Liu[1], Ming Yang[2,*] and Xian Zhou[1,*]

[1] School of Computer, Jiangxi University of Chinese Medicine, Nanchang 330004, Jiangxi, China
[2] School of Pharmacy, Jiangxi University of Chinese Medicine, Nanchang 330004, Jiangxi, China

* **Correspondence:** Email: A2751947695@outlook.com, 20030731@jxutcm.edu.cn.

Abstract: Prescription data is an important focus and breakthrough in the study of clinical treatment rules, and the complex multidimensional relationships between Traditional Chinese medicine (TCM) prescription data increase the difficulty of extracting knowledge from clinical data. This paper proposes a complex prescription recognition algorithm (MTCMC) based on the classification and matching of TCM prescriptions with classical prescriptions to identify the classical prescriptions contained in the prescriptions and provide a reference for mining TCM knowledge. The MTCMC algorithm first calculates the importance level of each drug in the complex prescriptions and determines the core prescription combinations of patients through the Analytic Hierarchy Process (AHP) combined with drug dosage. Secondly, a drug attribute tagging strategy was used to quantify the functional features of each drug in the core prescriptions; finally, a Bidirectional Long Short-Term Memory Network (BiLSTM) was used to extract the relational features of the core prescriptions, and a vector representation similarity matrix was constructed in combination with the Siamese network framework to calculate the similarity between the core prescriptions and the classical prescriptions. The experimental results show that the accuracy and F1 score of the prescription matching dataset constructed based on this paper reach 94.45% and 94.34% respectively, which is a significant improvement compared with the models of existing methods.

Keywords: prescription classification matching; AHP; BiLSTM; Siamese network; traditional Chinese medicine

1. Introduction

Prescription data is a valid document for doctors to diagnose and treat patients according to their conditions [1,2] and contains a wealth of information and knowledge that can be used as a reference for the treatment of clinically relevant diseases. During most Traditional Chinese medicine (TCM) treatments, doctors choose to make adjustments and additions to or subtractions from the classical prescriptions based on the patient's specific situation [3], where multiple herbs are dispensed in reasonable proportions. The classical prescriptions in the classical ancient texts of TCM are the origin and essence of TCM and are the basis for the transmission of TCM [4,5]. From classical Chinese medical literature to practical clinical prescriptions, there exist valuable medical resources. However, due to the relatively low level of informatization in the field of TCM, extensive clinical TCM data has not been fully utilized. There is an imbalance in the ratio between experienced clinical practitioners and patients. TCM inherits its knowledge from classical texts in ancient Chinese medical literature [6,7]. Leveraging existing technologies to effectively classify and match complex TCM prescriptions with suitable classical prescriptions holds the promise of compensating for the shortcomings in experiential knowledge transfer in clinical practice. This can provide a solid basis for clinical decision-making [8].

The classification and matching of TCM prescriptions heavily rely on prescription representation and feature extraction methods [9]. Unlike standard textual formats, TCM prescription texts are characterized by complex and irregular natural language. Firstly, TCM prescriptions exhibit intricate semantic relationships among various herbal components, as opposed to isolated entities [10,11]. Moreover, conventional text mining methods often focus on data with explicit and rich semantics [12,13]. However, TCM prescription texts, being integral components of herbal combinations, frequently lack sufficient semantic information. The format of TCM prescription texts renders traditional text processing approaches ineffective in knowledge extraction from TCM prescription data.

In the field of classification and matching of TCM prescriptions, many scholars have conducted research in three main directions: rule-based matching at the herbal vocabulary level, machine learning-based models, and deep learning-based models. Firstly, in rule-based matching at the herbal vocabulary level, Zhao et al. [14] proposed a clinical prescription recognition algorithm based on template matching, identifying classical prescriptions in TCM prescriptions by comparing templates to subgraphs. Zhang et al. [15] utilized the Jaccard similarity score method to classify and match herbs in prescriptions, aiming to distinguish prescriptions with similar clinical efficacy and summarize prescription usage patterns. Additionally, Wang et al. [16] introduced an algorithm based on weighted similarity distance to match and classify TCM prescriptions with corresponding classical prescriptions. However, these methods typically handle herbal knowledge information and overlook the intrinsic semantic relationships within prescriptions. Secondly, in research based on traditional machine learning models, Ung et al. [17] used Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) methods for classifying TCM prescriptions and non-TCM prescriptions. Chen et al. [7] proposed a probabilistic model-based algorithm for classifying TCM formulae according to the China Academy of Chinese Medical Sciences [18], matching them with classical prescriptions. These methods overlook the semantic information of herbs in TCM prescriptions, and their ability to handle herbal semantics and generalization are relatively weak, making it difficult to effectively extract features from prescription texts. The emergence of deep learning models has overcome the limitations of traditional algorithms in capturing semantic relationships within text and extracting crucial feature information. This development is conducive to driving the inheritance of TCM. Hu et al. [19] employed

Convolutional Neural Networks (CNN) for classifying TCM texts. Ning et al. [20] introduced an improved deep learning model, composed of Bidirectional Long Short-Term Memory Network (BiLSTM) and CNN, for classifying the efficacy of TCM formulas. Li et al. [21] utilized a neural network with hierarchical attention layers to identify specific herbal combinations for prescription efficacy. Yang et al. [22] utilized a multi-head attention mechanism to focus on the varying impacts of herbal ingredients on symptoms. However, these methods primarily concentrate on establishing connections between prescriptions and all the herbs contained within them, overlooking the semantic relationships among different herbs within the prescription. Additionally, it is worth noting that the therapeutic effects of traditional Chinese herbs are intricate and high-dimensional. Chinese medicinal terminologies are vast, and distinguishing synonyms among these terms is quite complex. Therefore, relying on herbal effect analysis for the representation of TCM prescription data poses challenges.

This study proposes a complex prescription recognition algorithm based on Siamese networks, with the primary objective of achieving accurate matching between traditional Chinese clinical prescriptions and their corresponding classical prescriptions. To address the complexity of natural language in traditional Chinese prescriptions, we introduce a novel method for weighting herb compatibility and dosage to assess the importance of each herb within a prescription. Furthermore, we take into account the performance characteristics of traditional Chinese herbs to tackle the challenge of quantifying herbal efficacy attributes. By combining a Siamese network with a BiLSTM structure and herb properties (TCM-Ps) features, we aim to comprehensively capture the semantic relationships within traditional Chinese herbal prescriptions and sensitively classify minor differences between different prescriptions. Through this algorithm, our goal is to effectively match complex traditional Chinese clinical prescriptions with suitable classical prescriptions, thereby identifying the therapeutic patterns contained within clinical prescriptions. This not only contributes to the in-depth exploration and inheritance of diagnostic and therapeutic experiences from classical Chinese medicine but also provides robust support and reference value for clinical disease prediction, auxiliary treatment, and prescription evaluation.

2. Materials and methods

In this study, we propose a complex prescription recognition model (MTCMC), and Figure 1 shows the overall design of MTCMC. Firstly, the pre-processed TCM prescription text data are extracted from the key herbs by a combination of compatibility weights and dosage weights to obtain a core prescription consisting of a collection of key herbs, which is used as the input to a matching model for similar matching with classical prescriptions; secondly, a vector space model-based approach is combined with a TCM corpus to construct a word vector with semantic meaning; then, the core prescription attributes are obtained using the TCM performance quantification Finally, a prescription matching model based on the Siamese network architecture is designed to integrate the semantic information of the TCM performance level, calculate the similarity between the actual prescription and the classical prescription, and match the prescription with the classical prescription in a categorical manner.

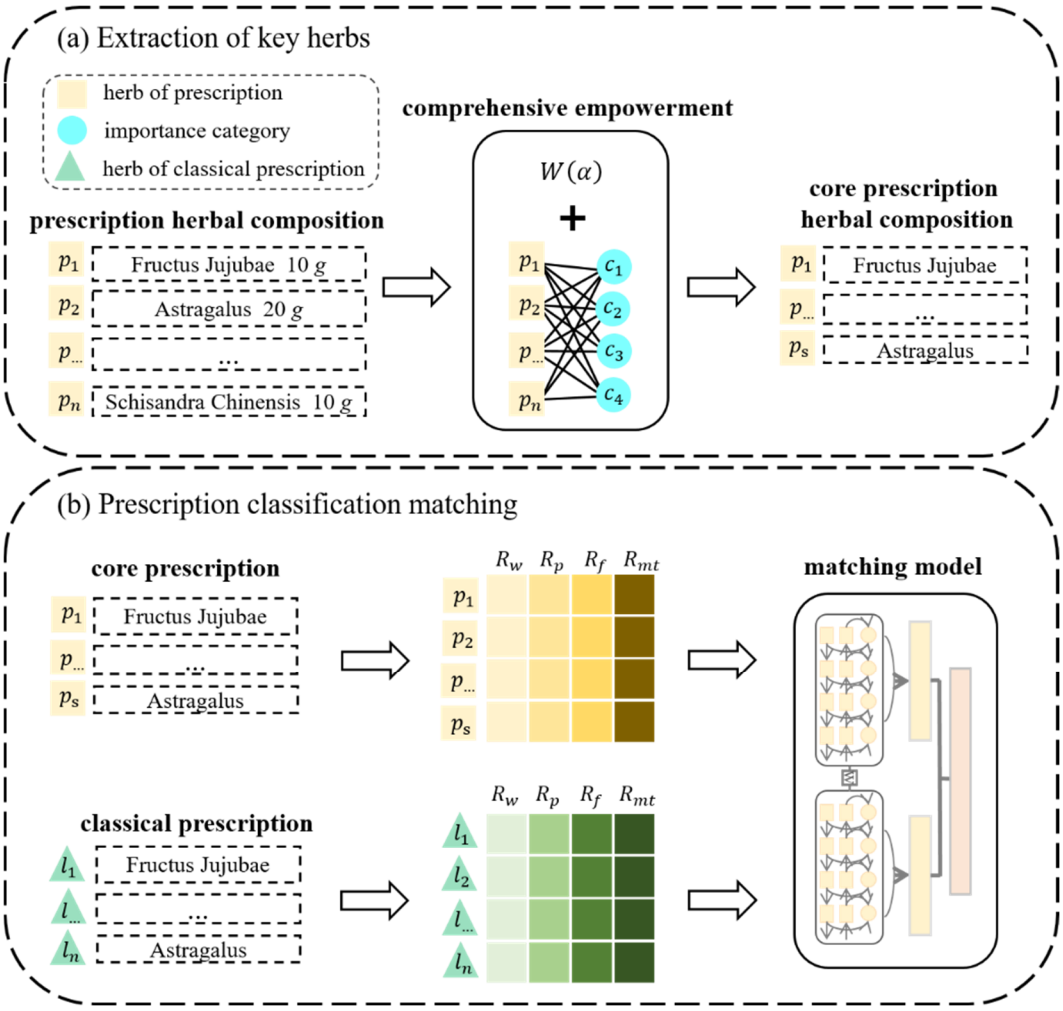

Figure 1. The overall design of MTCMC.

2.1. Preprocessing of prescription text data

There are practical problems such as complexity, mutual redundancy, and non-standardization in TCM prescription text data [23], such as the inconsistent expression of terms and the non-uniform measurement units for medication quantities in prescriptions [17,24]. The main preprocessing operations are as follows:

1) Text segmentation: First, stop words, special characters, and punctuation marks in the text are removed using the method of loading a stop word list to reduce noise and interference with the semantic meaning of the text. Then, the Jieba Chinese word segmentation library is used to segment the text. To address the difficulty of recognizing proprietary terms in the field of TCM, a dictionary of Chinese medicine vocabulary is loaded as the segmentation dictionary to improve the recognition rate of proprietary nouns [25].

2) Terminology normalization: According to the guidance of clinical experts in TCM, referring to the Coding Rules and Coding of Chinese Medicine (National Standard [GB/T 31774-2015]) and the Chinese Pharmacopoeia (2020 Edition), the preprocessed drugs are strictly normalized and annotated, and a synonym table is constructed to define the standard terms for each class of synonyms, realizing

the organization and standardization of TCM text data.

2.2. Extraction of key herbs

According to the compatibility principle of "monarch, minister, assistant, and guide" in TCM, different groups of medicines are fused into a coordinated and orderly organic whole in a prescription, achieving the effect of the whole prescription being greater than the sum of its parts. This relationship leads to significant differences in the importance of medicines [26–28]. To measure the compatibility weight of each medicinal herb in different prescriptions, the importance of each drug in the prescription was divided into four levels. C1–C4 represent the four drug weight categories. Based on the expert evaluation of TCM, the basic scaling rules of the Analytic Hierarchy Process (AHP) [29] method were used to assign importance to the prescription dispensing rules, and the basic scaling rules are shown in Figure 2. A two-by-two judgment matrix was established according to the scaling rule and the weight coefficients were calculated, at which time the weight vector of prescription dispensing $W(\beta) = (\beta_1, \beta_2, \cdots, \beta_n)$, and finally the matrix was tested for consistency.

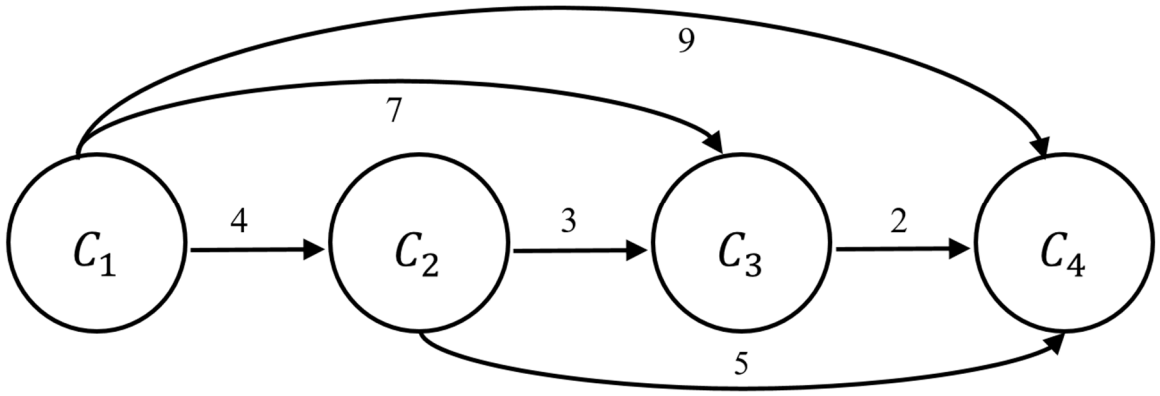

Figure 2. Basic scale rules of the AHP method in TCM data.

To measure the pharmacological effects of different doses of traditional Chinese medicines [30,31], a dosage weighting method is proposed. The dosage characteristics of Chinese medicine make the commonly used dosage range, for example, the commonly used dosage of pseudo-ginseng is 3 to 9 g. However, the dosage of Chinese medicine may exceed this range. To eliminate the influence of the disparity in various indicators, standardization is required. Equation (1) is the standardization of herbal doses.

$$Z_i^* = \frac{Z_i}{Z_{MIN} + Z_{MAX}} \tag{1}$$

Z_i^* is the standardized dose of TCM. Z_i is the dose of TCM . Z_{MIN} is the minimum value in the commonly used dose of TCM. Z_{MAX} is the maximum value in the commonly used dose of TCM. The commonly used dose of TCM is referred to in the 2020 edition of the Chinese Pharmacopoeia, the Dictionary of TCM, and Chinese Pharmacology. At the same time, the prescription dosage weight vector $W(\alpha) = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ is calculated by normalizing the data set of drug dosage after standardization.

The total weight of medicinal ingredients is obtained by combining the compatibility weight and dosage weight through linear weighting, and then the herbs are sorted in descending order. The top 70% of key herbs are selected to represent the composition of the prescription. Equation (2) calculates the comprehensive weight of medicinal ingredients in a prescription.

$$W(f) = \lambda W(\beta) + (1 - \lambda)W(\alpha) \tag{2}$$

In Eq (2), the parameter λ represents the weighting coefficient, which is used to regulate the balance between the objective weighting method and the subjective weighting method. Through comparative analysis, we find that when $\lambda < 0.5$, the smaller the value of λ, the greater the influence of objective factors, while when $\lambda > 0.5$, the greater the value of λ, the greater the influence of subjective factors. When $\lambda = 0.5$, a good balance between subjective and objective weights is achieved, which enables accurate extraction of key herbs. This effectively supports the subsequent steps of the prescription similarity detection framework.

2.3. Quantification of TCM properties

Single-drug properties make it difficult to reflect the overall characteristics of TCM prescriptions, while multidimensional quantification indicators based on drug properties can provide objective and comprehensive information for studying the overall efficacy of prescriptions. The distribution of TCM properties (TCM-Ps) is based on the 2020 edition of the Pharmacopoeia of the People's Republic of China and is supplemented by the planning textbook of national TCM colleges and universities. According to Table 1, TCM is divided into three categories: property (P), flavor (F), and meridian tropism (MT). Each drug in a TCM prescription has its own unique property, flavor, and meridian tropism, represented by a label code. For example, the TCM herb Moutan bark has a slightly cold property, bitter and pungent flavor, and tropism towards the heart, liver, and kidney meridians. The TCM property vector of Moutan bark is represented as $f = (f_1, f_2, f_3, f_4, f_5, f_6, f_7) = (0,1,0,1,0,1,0,0,0)$. Different TCM herbs are stacked and fused in prescriptions, resulting in multidimensional property characteristics of property, flavor, and meridian tropism for each prescription [1,32]. Therefore, Eq (3) represents the TCM prescription for n herbs as a vector of TCM properties.

$$R_f = \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} & f_{15} & f_{16} & f_{17} \\ f_{21} & f_{22} & f_{23} & f_{24} & f_{25} & f_{26} & f_{27} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{n1} & f_{n2} & f_{n3} & f_{n4} & f_{n5} & f_{n6} & f_{n7} \end{pmatrix} \tag{3}$$

The *i*-th row represents the drug property vector of the *i*-th herb in the prescription, $i = 1, 2, 3, \ldots, n$. The final result is the matrix of prescription property vectors R_p, flavor vectors R_f and meridian vectors R_{mt}.

Table 1. TCM properties of Chinese herbal medicines.

Herb property		Herb flavor	Herb meridian tropism	
Subclass (P)	TCM-P	TCM-F	Subclass (MF)	TCM-MT
P I: cold	p1: great cold	f1: sour	MT I: Yin meridian	mt1: heart
	p2: cold	f2: bitter		mt2: liver
	p3: mildly cold	f3: sweet		mt3: spleen
		f4: pungent		mt4: lung
P II: cool	p4: cool	f5: salty		mt5: kidney
		f6: bland		mt6: pericardium
	p5: even	f7: astringent		
			MT II: Yang meridian	mt7: stomach
P III: warm	p6: mildly warm			mt8: gallbladder
	p7: warm			mt9: large
				mt10: small
P IV: hot	p8: hot			mt11: bladder
	p9: great hot			mt12: triple

2.4. The core prescription and classical prescription matching

To further help doctors identify classical TCM prescriptions contained in clinical prescriptions, this section describes in detail the Siamese network-based BiLSTM neural network architecture for core prescription and classical prescription matching, as shown in Figure 3. The model framework mainly includes the word vector representation layer, Siamese network layer, BiLSTM layer, and similarity calculation layer. The output of each layer is used as the input of the next network layer, where the concatenated vector of the word vector, herb property vector, herb flavor vector, and meridian tropism vector in the word embedding layer is used as the input of the BiLSTM network layer for feature extraction and training, and the matching results are output through the similarity calculation layer.

2.4.1. Word vector representation

A word2vec skip-gram model is used to train the Chinese medicine word vector model on a large corpus of Chinese medicine text. The model obtains continuous and low-dimensional word vectors, which effectively solve the problems of semantic isolation and dimension explosion caused by traditional word vector models [33]. For a given corpus, the word vectors are stored in a word vector query matrix $M = R^{|V| \times d_w}$, where $|V|$ is the vocabulary size of the given unlabeled Chinese medicine text corpus, and d_w is the dimension of the word vector [34]. For a sentence $S = \{s_1, s_2, \cdots, s_n\}$ of length n, the sentence matrix is represented as $X = \{x_1, x_2, \cdots, x_n\}$, where x_i is the word vector obtained from the word vector query matrix M for the word s_i in the sentence. Assuming that the corpus consists of $\omega_1, \omega_2, \cdots, \omega_M$ words, the skip-gram model aims to maximize Eq (4).

$$F = \frac{1}{M}\sum_{m=1}^{M}\sum_{-n\le j\le n, j\ne 0}\log p\left(\omega_{t+j}|\omega_t\right) \tag{4}$$

where: n is a parameter that represents the training window size. In this study, the context window size for word2vec training was set to 5, and the number of iterations was set to 8, resulting in the generation of word vectors with dimensions of 100, 200, 300, and 400. After extracting the key herbs, the set of key herbs was used to form the core prescription. The core prescription text was then used as input for the model training, which involved concatenating the word vector matrix R_w, the medicinal property vector matrix R_p, the medicinal flavor vector matrix R_f, and the meridian vector matrix. R_{mt} to form the prescription vector matrix $R_s = \left[R_w^T \oplus R_p^T \oplus R_f^T \oplus R_{mt}^T\right]$. The symbol $\oplus$ denotes the vector concatenation operation.

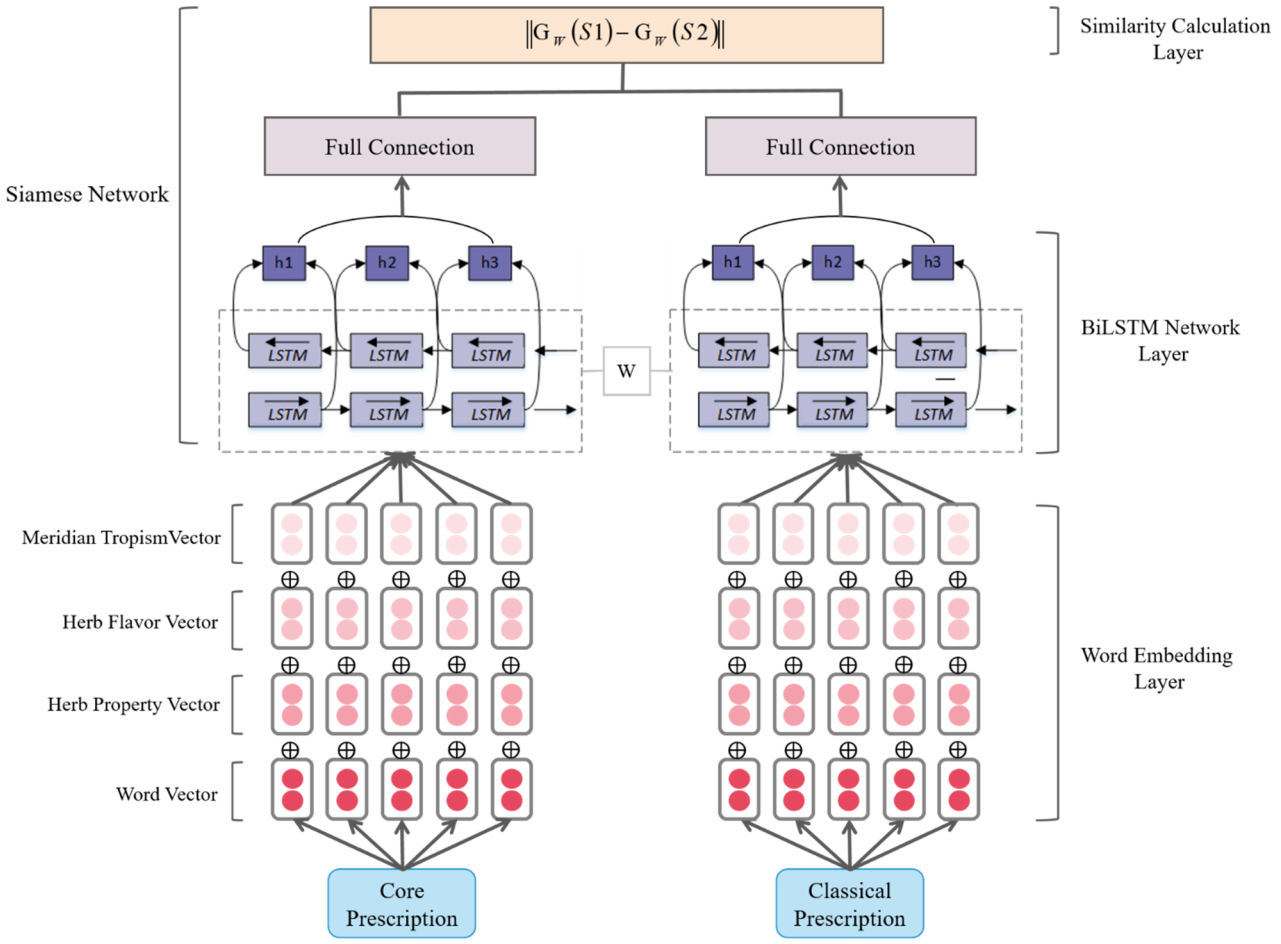

Figure 3. Prescription matching model framework.

2.4.2. Siamese network

A Siamese network is a neural network with weight sharing, which can determine whether a sample outside the training set is similar. It has significant advantages in terms of the number of classifications and generalizations [35]. The basic structure is shown in Figure 4. The Siamese Network is mainly composed of two parts: sample representation and similarity calculation. It uses two sub-

networks with the same structure and shared parameters to represent samples $S1$ and $S2$. The similarity $E_w(S1, S2)$ of the sample representation vectors $G_w(S1)$ and $G_w(S2)$ is calculated to determine whether the two samples belong to the same category [36].The loss function of a Siamese network depends on the inputs and parameters. The variable y is the label indicating whether $S1$ and $S2$ are a match, where $y \in \{0,1\}$. Let $L_+(S1, S2)$ denote the loss when $y = 1$ and let $L_-(S1, S2)$ denote the loss when $y = 0$. Equation (5) is the loss function for a single text.

$$L_w = (1 - y)L_-(S1, S2) + yL_+(S1, S2) \tag{5}$$

Equation (6) is the total loss function for N texts.

$$L_w(S) = \sum_{i=1}^{N} L_w^{(i)}(S1, S2, y) \tag{6}$$

The loss function $L_w^{(i)}$ is the cross-entropy loss function. The Adaptive Moment Estimation (ADAM) algorithm is the optimization algorithm used for the model, and the dropout technique is a regularization method.

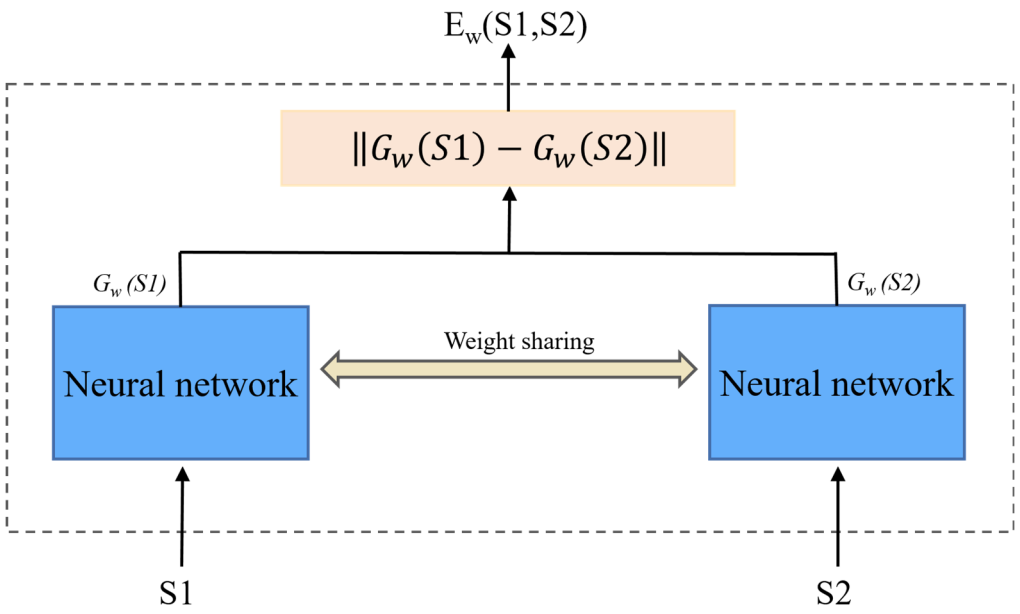

Figure 4. Siamese network framework.

2.4.3. BiLSTM

Long short-term memory (LSTM) networks can effectively preserve the historical information of long sequences while avoiding problems such as gradient vanishing and explosion that standard Recurrent Neural Networks (RNNs) often face [37]. A LSTM model is shown in Figure 5. For a prescription text sequence $S = \{x_1, x_2, \cdots, x_n\}$ consisting of n words, x_t represents the mixed vector corresponding to the word in the input sequence at time step t in the LSTM unit.

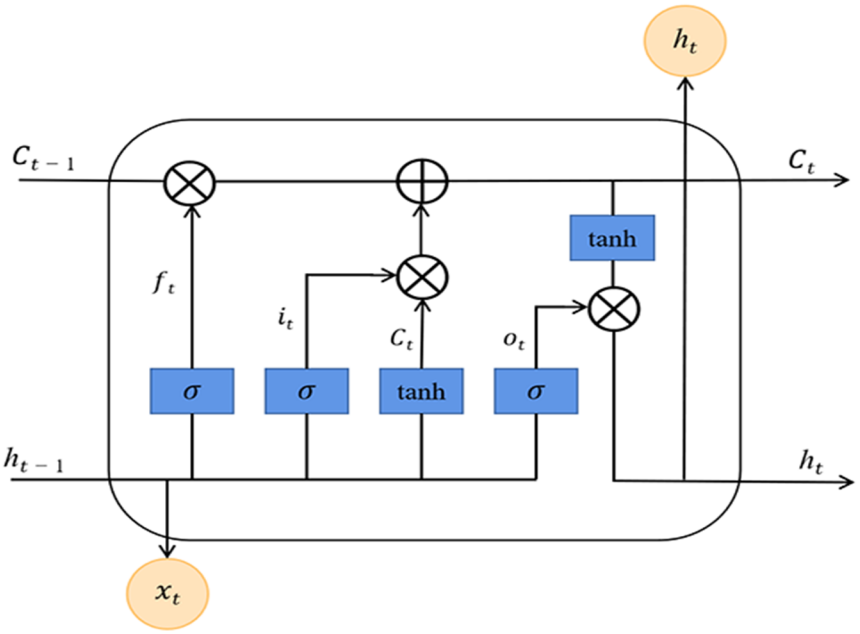

Figure 5. LSTM cell structure.

Equation (7) calculates the three gates and the memory cell in the LSTM unit.

$$\begin{cases} X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \\ f_t = \sigma(W_f \cdot X + b_f) \\ i_t = \sigma(W_i \cdot X + b_i) \\ o_t = \sigma(W_o \cdot X + b_o) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot X + b_c) \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{7}$$

In the above prescription, $W_f, W_i, W_o \in R^{m \times d}$ represent weight matrices, $b_f, b_i, b_o \in R^m$ represent biases, σ is a non-linear activation function, $m = 100$ is the number of LSTM units in the network, $\odot$ represents element-wise multiplication; x_t contains input vectors of the LSTM unit; $h_t \in m$ is the hidden layer vector. For a given prescription text $S = \{x_1, x_2, \cdots, x_n\}$, each word x_i is mapped to its corresponding word vector $v_w^i \in R^{d_w}$, each word's medicinal property is mapped to a medicinal property vector $v_p^i \in R^{d_p}$, the word's flavor vector is mapped to a flavor vector $v_f^i \in R^{d_f}$ and the word's meridian vector is mapped to a meridian vector $v_{mt}^i \in R^{d_{mt}}$ (where $d_w = 300$, $d_p = 9$, $d_f = 7$, $d_{mt} = 12$ represent the dimensions of word vector, medicinal property vector, flavor vector and meridian vector respectively). Finally, Eq (8) connects these vectors to generate the prescription vector matrix:

$$[x_1, x_2, \cdots, x_n] \to \left[R_w^T \oplus R_p^T \oplus R_f^T \oplus R_{mt}^T\right] \to \begin{bmatrix} v_w^1 \oplus v_p^1 \oplus v_f^1 \oplus v_{mt}^1 \\ v_w^2 \oplus v_p^2 \oplus v_f^2 \oplus v_{mt}^2 \\ \vdots \\ v_w^n \oplus v_p^n \oplus v_f^n \oplus v_{mt}^n \end{bmatrix} \tag{8}$$

To measure the positional information of the text in different orders equally and further explore the contextual information associated with the prescription, a BiLSTM [38] model is used to extract text features. After passing through the forward and backward LSTM networks, the network contains two hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ in both directions [39]. The vector $h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$ generated by concatenating them is the output, which is a fused result. Then, the max pooling layer takes the maximum value of the vector to obtain the most significant feature vector as the input of the fully connected layer.

3. Results

3.1. Datasets

The data for the experimental prescription were selected from authoritative Chinese medicine prescription books and data platforms. The included Chinese medicine prescriptions meet the following standards: 1) they were clinically used by academic experts and instructors recognized by the National Administration of TCM, including the first three batches of national TCM masters and the first batch of nationally renowned TCM practitioners; 2) they were published in authoritative TCM journals. The exclusion criteria are: 1) incomplete records of medical cases, prescriptions, etc.; 2) duplicated literature, in which only one is selected; 3) duplicated medical cases or prescriptions, in which only one is selected.

Table 2 shows some sample data from the dataset used in this study. According to the above-mentioned criteria, the prescription dataset included in this study consisted of 5034 prescriptions, each containing an average of 13 to 15 herbs, of which prescription data indicating classical prescriptions were collated as matching data, and the non-matching data were generated by combining prescription data with classical prescriptions for different main treatment conditions, involving 562 classical prescriptions, to establish a matching dataset of clinical prescriptions of national masters and national famous Chinese medicine practitioners (MTCMFE). The dataset consists of prescriptions, corresponding classical prescriptions, following grid and a label. A label of 1 indicates that the actual prescription matches the classical prescription, while a label of 0 indicates no match. The MTCMFE dataset is divided into a 7:3 training set and a test set, Table 3 provides a detailed distribution of positive and negative samples in each dataset. Here, positive samples represent matching pairs of prescriptions, while negative samples represent non-matching pairs of prescriptions.

Table 2. Some sample data from the dataset.

Prescription	Classical prescription	Similarity label
10 g Ephedra, 15 g Almonds, 50 g Gypsum, 30 g Houttuynia, 15 g Great Burdock Fruit, 15 g Sunflower Seed, 10 g Baikal Skullcap Root, 10 g Tendrilled Fritillary Bulb, 30 g Honeysuckle Flower, 10 g Balloon Flower Root, and 10 g Licorice Root.	9 g of Ephedra, 9 g of Almonds, 6 g of Licorice Root, and 18 g of Gypsum.	1
12 g Rehmannia, 8 g Dogwood Fruit, 10 g Chinese Yam, 8 g Tree Peony Bark, 8 g Fried Common Rush Rhizome, 8 g Poria, 12 g Salvia Root, 8 g Sichuan Lovage Rhizome, 10 g Goji Berry, 8 g Vitex Fruit, 8 g Dodder Seed, 10 g Angelica Root, and 6 g Tangerine Peel.	6 g of Ephedra, 6 g of Almonds, 3 g of Licorice Root, and 12 g of Coix Seed.	0

Table 3. Distributions of positive samples and negative samples in the dataset.

Dataset	Training		Test	
	Positive	Negative	Positive	Negative
Number of samples	1753	1771	757	753

3.2. Experimental environment

The specific experimental environment settings for this study are shown in Table 4.

Table 4. Experimental environment configuration.

Experimental environment	Configuration
Development tools	Pycharm
programming languages	Python3.6
Development framework	Pytorch 1.4.0
Operating system	64-bit Windows 10
Central processing unit	Intel (R) Core (TM) i5-9300H CPU @ 2.40 GHz
Hardware platform	NVIDIA GeForce GTX 1650

3.3. Evaluation indicators and model configuration

The experiment used accuracy, precision, recall, and F1 score as evaluation metrics to comprehensively measure the performance of the model. Their definitions are as shown in Eq (9). Where TP represents the number of samples that are predicted as positive in the positive class, FP represents the number of samples predicted as positive in the negative class, FN represents the number of samples predicted as negative in the positive class, and TN represents the number of samples predicted as negative in the negative class.

$$\begin{cases} A = \frac{TP+TN}{TP+FP+TN+FN} \\ P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ F1 = \frac{2\times P\times R}{P+R} \end{cases} \tag{9}$$

In the subsequent experiments, we use the grid search method to adjust the model parameters of the training dataset, and F1 score serving as the evaluation metric for the grid search. Finally, we will assess the model's performance using the test dataset.

The MTCMC model used in our experiments was established as follows: we assigned weights to herbs using the Analytic Hierarchy Process (AHP) in combination with standardization and normalization techniques. These weights were based on evaluations by domain experts. The fundamental scaling rules for herbal weight assignment are depicted in Figure 2. We introduced a Siamese-BiLSTM matching model that integrates multiple features. To determine the model's hyperparameters, including word vector feature dimensions (w_dim), the number of iterations (epochs), learning rate (lr), optimization functions, LSTM hidden layer size (LSTM_hidden), and the number of LSTM layers (LSTM_layer), we employed a grid search approach. To prevent model overfitting, we incorporated dropout regularization to control model complexity and avoid overfitting issues. The magnitude of the dropout regularization term was also determined via grid search.

For the other models used in our experiments, the procedures were as follows: We constructed SVM and KNN models using Scikit-Learn. The regularization parameter (C) and sigmoid kernel parameter (coef0) for SVM, as well as the number of nearest neighbors (n_neighbors) for KNN, were determined using grid search techniques. We implemented BiLSTM, CNN, Siamese-BiLSTM, and Siamese-CNN models using PyTorch. Hyperparameters including LSTM hidden layer size (LSTM_hidden), the number of LSTM layers (LSTM_layer), the number of convolutional filters (num_filters), the number of convolutional layers (CNN_layer), the number of iterations (epochs), learning rate (lr), regularization strength (dropout), and optimization functions (optimizer) were determined through grid searches. The parameter grids and specific values following short in effectively the grid search are presented in Tables 5 and 6.

Table 5. Model parameters in the experiment.

Model	C	coef0	n_neighbors
Parametric Grid	[0.01, 0.1, 1]	Arrange (0.1, 1, 0.1)	Arrange (1, 10, 1)
SVM	0.1		
KNN		0.2	4

Table 6. Model parameters in the experiment.

Model	w_dim	LSTM_ hidden	LSTM_ layer	num_ filters	CNN_ layer	epochs	lr	dropout	optimizer
Parametric Grid	Arrange (100, 500, 100)	Arrange (50, 500, 1)	Arrange (1, 3, 1)	Arrange (16, 512, 1)	Arrange (1, 4, 1)	Arrange (1, 100, 1)	[0.001,0.01]	Arrange (0.1, 0.8, 0.1)	{“SGD”, “Adam”, “RMSprop”}
CNN	300			256	2	30	0.001	0.3	Adam
BiLSTM	300	256	2			56	0.001	0.5	Adam
Siamese-CNN	300			256	2	35	0.001	0.3	Adam
Siamese-BiLSTM	300	300	2			45	0.001	0.5	Adam
MTCMC	300	300	2			40	0.001	0.5	Adam

3.4. Comparison experiment with the baseline method

Table 7. Results of comparative experiments on classical similarity matching.

Models	Dataset of MTCMFE		
	Precision %	Recall %	F1 score %
SVM	58.12	60.76	60.23
KNN	57.32	58.46	59.76
CNN	71.62	70.32	71.47
BiLSTM	61.72	60.58	61.01
Siamese-CNN	74.72	71.45	74.36
Siamese-BiLSTM	68.52	70.13	72.15
MTCMC	89.24	85.63	89.19

To compare our proposed method with a range of models in the field of TCM data mining, experiments were conducted on the MTCMFE dataset. To ensure a fair comparison of the matching performance of each method, only pre-trained word vectors with a dimension of 300 were used as input for the comparative methods. The results are shown in Table 7, it can be seen that the proposed MTCMC model performs the best in the task of matching TCM prescriptions in terms of precision, recall, and F1 score, and outperforms other baseline methods. The specific analysis is as follows:

1) Compared to traditional machine learning models SVM and KNN, MTCMC shows significant improvements in precision, with an average increase of 31.52%, recall with an average increase of 26.02%, and F1 score with an average increase of 29.19%. SVM is suitable for handling linearly separable problems, while KNN's simple distance metric struggles to capture the nonlinear relationships present in traditional Chinese medicine (TCM) prescription texts. Both SVM and KNN fall short of effectively dealing with the nonlinear semantic relationships and multidimensional attribute features found in TCM prescriptions. MTCMC leverages the nonlinear feature extraction capabilities of the Siamese-BiLSTM deep learning model, allowing it to better capture complex features and deep semantic relationships in TCM prescription text data, while also exhibiting strong generalization capabilities.

2) Compared to deep neural networks CNN and BiLSTM, MTCMC demonstrates improvements in precision, with an average increase of 22.57%, recall with an average increase of 20.18%, and F1 score with an average increase of 22.95%. CNN primarily focuses on local information, making it challenging to capture the overall semantic relationships within traditional Chinese herbal medicine prescriptions. Additionally, CNN requires text to be padded to a fixed length, potentially leading to information loss. BiLSTM typically used for modeling single sequences, lacks a contrastive learning mechanism and may perform weakly in handling complex text. MTCMC combines the global semantic relationship extraction capabilities of BiLSTM with the contrastive learning mechanism of Siamese, resulting in the extraction of richer feature information, facilitating a more precise differentiation of semantic relationships between different prescriptions.

3) Compared to Siamese-CNN and Siamese-BiLSTM, MTCMC exhibits improvements in precision, with an average increase of 17.48%, recall with an average increase of 14.84%, and F1 score with an average increase of 15.94%. The Siamese network improves the accuracy of the matching task by sharing weights to better capture the similarities and differences between prescriptions and sensitively classify small differences between prescriptions. However, these models may not fully consider the attributes and compatibility of herbal medicines, limiting their performance in traditional Chinese herbal medicine prescription matching. MTCMC, building upon Siamese-BiLSTM, integrates a weighted calculation method for herbal compatibility and dosage. This further enhances the model's understanding and modeling capabilities for prescription texts, enabling it to better capture complex semantic relationships within prescriptions. MTCMC comprehensively considers multidimensional attributes, compatibility relationships, and global semantic relationships, resulting in more accurate modeling of prescription text characteristics.

To visualize the performance of the MTCMC model, we created bar charts reflecting precision, recall, and F1 score, as depicted in Figure 6. Taking into account the overall performance of the MTCMC model, it consistently outperforms the other six models. This superiority can be attributed to several key factors. Firstly, the MTCMC model leverages Siamese-BiLSTM to extract nonlinear features from traditional Chinese medicine prescription data, effectively capturing the intricate global semantic relationships within prescriptions. Additionally, the model introduces a weighted calculation method for herbal compatibility and dosage, further enhancing its ability to comprehend and model

prescription text. MTCMC takes a holistic approach by considering multiple dimensions of attributes, nonlinear relationships, and global semantic relationships. Combining attribute weights with text features, achieves a more precise modeling of the unique characteristics of prescription text, thereby efficiently capturing the complex semantic relationships within traditional Chinese medicine prescription texts.

To examine the learning process of the MTCMC model, Figure 7 illustrates the descent of the loss during the training of the MTCMC model. The horizontal axis represents training epochs, while the vertical axis represents the train loss. As observed from the graph, the train loss continually decreases and gradually converges to a lower loss value, reaching a state of convergence.

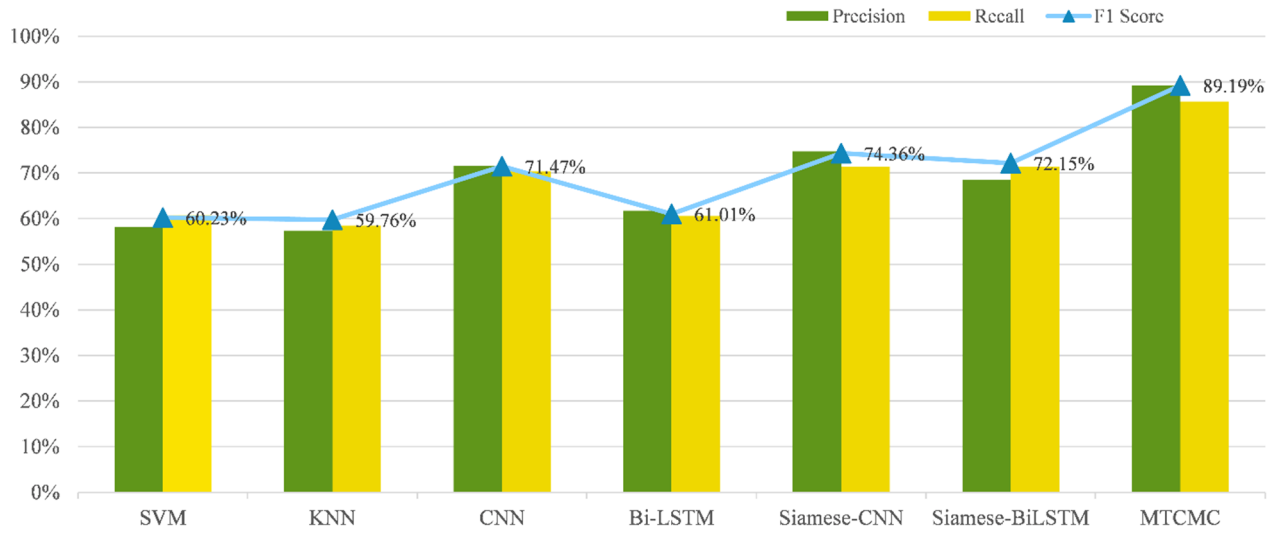

Figure 6. Model evaluation visualization results.

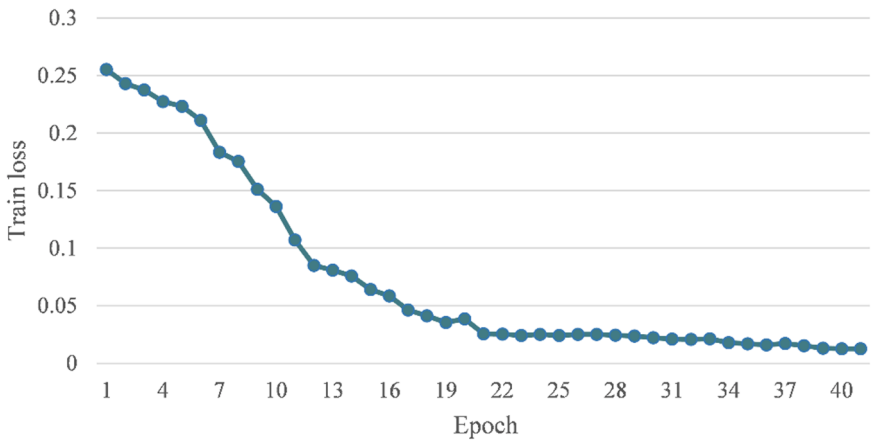

Figure 7. Descent of Train loss in MTCMC learning process.

3.5. The impact of multiple features on prescription matching

To verify the effectiveness of the drug property, flavor, and meridian features on the matching of

Chinese herbal prescriptions, different feature combinations were used as model inputs in this experiment. The F_p, F_f and F_{mt} were added to the concatenated mixed vector $F = F_p \oplus F_f \oplus F_{mt} \oplus F_w$ generated by the word vector F_w and used as the model input for training. The results of the MTCMC multi-feature comparison experiment are shown in Figure 8. The decrease dosage MTCMC model proposed in this paper, which uses a fusion of word vectors, drug property vectors, flavor vectors, and meridian vectors, has the best performance. Specifically:

Training the model using only word vectors as input resulted in a matching score of 89.19%, indicating that using only word vectors for feature extraction is not sufficient to fully capture the feature information of prescription texts. Adding drug property vectors, flavor vectors, and meridian vectors on top of word vectors improved accuracy, with matching values increasing by 0.13%, 0.24%, and 2.24%, respectively. These vectors contributed differently to the semantic representation of prescription texts, with meridian vectors making a larger contribution.

Concatenating word vectors, drug property vectors, and flavor vectors to generate a vector $F_w \oplus F_p \oplus F_f$ as input for the model increased the matching score by 1.19%. It performed better than using $F_w \oplus F_p$ or $F_w \oplus F_f$ feature vectors, but its expressive power is still limited compared to using a single comprehensive vector $F_w \oplus F_{mt}$, especially in terms of $F_w \oplus F_p \oplus F_f$.

The MTCMC model proposed in this paper uses drug property, flavor, and meridian vectors in addition to word vectors for training, demonstrating the specialization of TCM prescription data and enriching the semantic representation of prescription texts. It strengthens the connection between internal herbs in prescription texts and improves the matching F1 score by 5.15%, further enhancing the experimental effect of TCM prescription text matching.

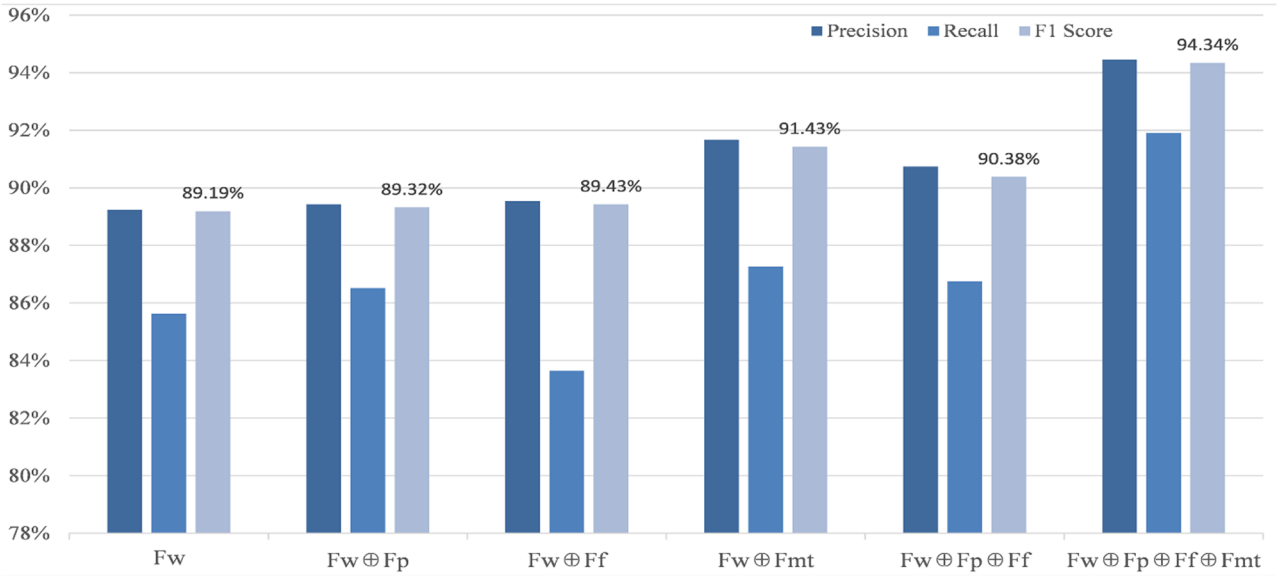

Figure 8. Multi-feature comparison experiment.

4. Discussion

Although the model achieved high performance, improvements are still needed to achieve the overall optimization of the model.

1) The quality and quantity of the dataset have a significant impact on the accuracy of the model. Therefore, more high-quality data on the matching of TCM prescriptions and prescriptions will be collected, and TCM experts will be invited to review the data to optimize and learn models with higher performance.

2) In the prescription process of TCM physicians, the description of herbs may vary. Many physicians use the aliases of Chinese medicine. This semantic diversity of Chinese medicine may interfere with the results of prescription recognition, so further research is needed on the discernment of Chinese medicine semantics.

3) Due to the differences in the pathological symptoms and personal constitution of patients, physicians may appropriately add or remove drugs, increase or decrease the dosage, or combine several prescriptions based on actual situations and personal experience on top of classical prescriptions. Additionally, physicians may be influenced by their preferences and replace drugs in classical prescriptions with other drugs of the same efficacy. Therefore, clinical prescriptions may differ greatly from classical prescriptions, which is difficult to handle.

4) This study lacked the inclusion of underlying diseases, signs and symptoms, living areas, to optimize and other patient information; therefore, the findings are somewhat limited. In the future, the parameters of the included information and validation data will be further expanded to optimize the model.

5. Conclusions

The complex prescription recognition algorithm based on Siamese networks, denoted as MTCMC, offers several key advantages when applied to the task of TCM prescription matching. Three primary strengths of the MTCMC model can be highlighted:

1) Robust Semantic Modeling Capability: The MTCMC model effectively captures the intricate semantic relationships within TCM herbal prescriptions by introducing a weighted calculation method for herbal compatibility and usage. In comparison to traditional machine learning models, the MTCMC model excels at handling non-linear semantic relationships in TCM herbal prescriptions, showcasing superior performance in addressing complex natural language tasks.

2) Multi-level Feature Fusion: The MTCMC model integrates deep learning with herbal performance features, combining them with a Siamese-BiLSTM structure. This comprehensive approach allows for a more thorough consideration of herb attributes and compatibility relationships. Consequently, it enhances the model's ability to accurately capture the overall semantic relationships among herbs in prescriptions, thereby elevating its performance.

3) Cross-Modal Learning and Similarity Measurement: Through the introduction of the Siamese network structure, the MTCMC model adeptly captures the similarity between different prescriptions. It exhibits sensitivity to subtle differences, further boosting the accuracy of the matching task.

Based on real clinical data and classical Chinese medicine knowledge, our proposed complex prescription recognition model (MTCMC) encompasses the multidimensional attributes, nonlinear relationships, and global semantic context found within TCM prescriptions. By leveraging deep learning and feature fusion strategies, MTCMC achieves a higher level of accuracy in handling complex and irregular TCM prescription data, resulting in an average improvement of 15% in F1 score. The significance of this method is not only in the ability to extract complex semantic features of Chinese medicine texts, but also in its ability to provide powerful support for the in-depth excavation and inheritance of the knowledge of Chinese medicine diagnosis and treatment experience.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82274680, 82160955) and the university-level research team of Jiangxi University of Traditional Chinese Medicine for the innovation team of Chinese medicine preparation technology and equipment (CXTD22006).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. C. Klonoff, The new FDA real-world evidence program to support development of drugs and biologics, *J. Diabetes Sci. Technol.,* **14** (2020), 345–349. https://doi.org/10.1177/1932296819832661
2. G. Li, X. Zuo, B. Liu, Scientific computation of big data in real-world clinical research, *Front. Med.*, **8** (2014), 310–315. https://doi.org/10.1007/s11684-014-0358-7
3. Q. Zhang, C. Yang, S. Ma, S. Guo, X. Hu, Z. Zhou, Shiwei Qingwen decoction regulates TLR4/NF-κB signaling pathway and NLRP3 inflammasome to reduce inflammatory response in lipopolysaccharide-induced acute lung injury, *J. Ethnopharmacol*., **313** (2023). https://doi.org/10.1016/j.jep.2023.116615
4. G. Song, Y. Wang, R. Zhang, B. Liu, X. Zhou, X. Zhou, et al., Experience inheritance from famous specialists based on real-world clinical research paradigm of traditional Chinese medicine, *Front. Med.*, **8** (2014), 300–309. https://doi.org/10.1007/s11684-014-0357-8
5. X. Xiong, C. Huang, F. Wang, J. Dong, D. Zhang, J. Jiang, et al., Qingxue jiedu formulation ameliorated DNFB-induced atopic dermatitis by inhibiting STAT3/MAPK/NF-κB signaling pathways, *J. Ethnopharmacol.*, **270** (2021). https://doi.org/10.1016/j.jep.2020.113773
6. J. Su, X. M. Chen, Y. L. Xie, M. Q. Li, Q. Shang, D. K. Zhang, et al., Clinical efficacy, pharmacodynamic components, and molecular mechanisms of antiviral granules in the treatment of influenza: A systematic review, *J. Ethnopharmacol*., **318** (2023). https://doi.org/10.1016/j.jep.2023.117011
7. X. Chen, C. Ruan, Y. Zhang, H. Chen, Heterogeneous information network based clustering for precision traditional Chinese medicine, *BMC Med. Inf. Decis. Making*, **19** (2019), 264. https://doi.org/ 10.1186/s12911-019-0963-0
8. X. Ren, X. X. Shao, X. X. Li, X. H. Jia, T. Song, W. Y. Zhou, et al., Identifying potential treatments of COVID-19 from Traditional Chinese Medicine (TCM) by using a data-driven approach, *J. Ethnopharmacol.*, **258** (2020). https://doi.org/10.1016/j.jep.2020.112932
9. K. W. DeGregory, P. Kuiper, T. DeSilvio, J. D. Pleuss, R. Miller, J. W. Roginski, et al., A review of machine learning in obesity, *Obes. Rev*., **19** (2018), 668–685. https://doi.org/10.1111/obr.12667

10. S. K. Poon, J. Poon, M. McGrane, X. Zhou, P. Kwan, R. Zhang, A novel approach in discovering significant interactions from TCM patient prescription data, *Int. J. Data Min. Bioinf.*, **5** (2011), 353–368. https://doi.org/10.1504/ijdmb.2011.041553
11. S. Li, B. Zhang, D. Jiang, Y. Wei, Y. Zhang, Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae, *BMC Bioinf.*, **11** (2010), 1–12. https://doi.org/10.1186/1471-2105-11-S11-S6
12. J. Liu, C. Li, Y. Huang, J. Han, An intelligent medical guidance and recommendation model driven by patient-physician communication data, *Front. Public Health*, **11** (2023). https://doi.org/10.3389/fpubh.2023.1098206
13. M. Arbane, R. Benlamri, Y. Brik, A. D. Alahmar, Social media-based COVID-19 sentiment classification model using Bi-LSTM, *Expert Syst. Appl.*, **212** (2023). https://doi.org/10.1016/j.eswa.2022.118710
14. H. Zhao, J. Luo, C. Liu, X. Lei, Z. Wang, Method of TCM prescription recognition based on template matching algorithm, in *2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*, (2018), 336–339. https://doi.org/10.2991/cmsa-18.2018.77
15. Y. Zhang, X. Li, Y. Shi, T. Chen, Z. Xu, P. Wang, et al., ETCM v2.0: An update with comprehensive resource and rich annotations for traditional Chinese medicine, *Acta Pharm. Sin. B*, **13** (2023), 2559–2571. https://doi.org/ 10.1016/j.apsb.2023.03.012
16. Y. Wang, J. Xu, J. Zhang, H. Xu, Y. Sun, Y. Miao, et al., SIAP: an intelligent algorithm for multiple prescription pattern recognition based on weighted similarity distances, *BMC Med. Inf. Decis. Making*, **23** (2023), 79. https://doi.org/10.1186/s12911-023-02141-3
17. C. Y. Ung, H. Li, C. Y. Kong, J. F. Wang, Y. Z. Chen, Usefulness of traditionally defined herbal properties for distinguishing prescriptions of traditional Chinese medicine from non-prescription recipes, *J. Ethnopharmacol.*, **109** (2007), 21–28. https://doi.org/10.1016/j.jep.2006.06.007
18. Prescription identification assistant, The institute of information on traditional Chinese medicine of China academy of Chinese medical sciences, 2021. Available from: http://tcmks.cintcm.com:81/.
19. Q. Hu, T. Yu, J. Li, Q. Yu, L. Zhu, Y. Gu, End-to-End syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine, *Comput. Methods Programs Biomed.*, **174** (2019), 9–15. https://doi.org/10.1016/j.cmpb.2018.10.011
20. N. Cheng, Y. Chen, W. Gao, J. Liu, Q. Huang, C. Yan, et al., An improved deep learning model: S-TextBLCNN for traditional Chinese medicine formula classification, *Front. Genet.*, **12** (2021). https://doi.org/10.3389/fgene.2021.807825
21. L. Chen, X. Liu, S. Zhang, H. Yi, Y. Lu, P. Yao, Efficacy-specific herbal group detection from traditional Chinese medicine prescriptions via hierarchical attentive neural network model, *BMC Med. Inf. Decis. Making*, **21** (2021), 66. https://doi.org/10.1186/s12911-021-01411-2
22. X. Yang, C. Ding, SMRGAT: A traditional Chinese herb recommendation model based on a multi-graph residual attention network and semantic knowledge fusion, *J. Ethnopharmacol.*, **315** (2023). https://doi.org/10.1016/j.jep.2023.116693
23. Q. Jia, D. Zhang, S. Yang, C. Xia, Y. Shi, H. Tao, et al., Traditional Chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism, *J. Biomed. Inf.*, **116** (2021). https://doi.org/10.1016/j.jbi.2021.103718

24. W. T. Huang, H. H. Hung, Y. W. Kao, S. C. Ou, Y. C. Lin, W. Z. Cheng, et al., Application of neural network and cluster analyses to differentiate TCM patterns in patients with breast cancer, *Front. Pharmacol.*, **11** (2020). https://doi.org/10.3389/fphar.2020.00670
25. G. Xu, H. Jin, Using artificial intelligence technology to solve the electronic health service by processing the online case information, *J. Healthcare Eng.*, **26** (2021). https://doi.org/10.1155/2021/9637018
26. R. Yuan, Y. Lin, Traditional Chinese medicine: an approach to scientific proof and clinical validation, *Pharmacol. Ther.*, **86** (2000), 191–198. https://doi.org/10.1016/s0163-7258(00)00039-5
27. H. Zhang, H. Huang, H. Song, B. Chin, H. Zheng, J. Ruan, et al., Serum metabolomics reveals the intervention mechanism and compatible regularity of Chaihu Shu Gan San on chronic unpredictable mild stress-induced depression rat model, *J. Pharm. Pharmacol.*, **72** (2020), 1133–1143. https://doi.org/10.1111/jphp.13286
28. C. Zhang, Y. He, T. Li, X. Miao, M. Song, C. Qian, et al., Analysis of chemical components in herbal formula Qi Bai Granule by UPLC-ESI-Q-TOF-MS, *Nat. Prod. Res.*, **33** (2019), 2271–2275. https://doi.org/10.1080/14786419.2018.1495641
29. J. Cancela, G. Fico, M. T. Arredondo Waldmeyer, Using the Analytic Hierarchy Process (AHP) to understand the most important factors to design and evaluate a telehealth system for Parkinson's disease, *BMC Med. Inf. Decis. Making*, **15** (2015), 1–11. https://doi.org/10.1186/1472-6947-15-S3-S7
30. X. Zhu, Discovery of traditional Chinese medicine prescription patterns containing herbal dosage based on multilevel Top-K weighted association rules, *Comput. Intell. Neurosci.*, **2022** (2022). https://doi.org/10.1155/2022/5466011
31. D. Wang, R. Zhao, H. X. Duan, M. M. Zhang, L. He, X. Ye, et al., Research progress regarding potential effects of traditional Chinese medicine on postoperative intestinal obstruction, *J. Pharm. Pharmacol.*, **73** (2021), 1007–1022. https://doi.org/10.1093/jpp/rgaa054
32. W. Dan, J. Liu, X. Guo, B. Zhang, Y. Qu, Q. He, Study on medication rules of traditional Chinese medicine against antineoplastic drug-induced cardiotoxicity based on network pharmacology and data mining, *Evidence-Based Complementary Altern. Med.*, **2020** (2020). https://doi.org/10.1155/2020/7498525
33. B. Jang, I. Kim, J. W. Kim, Word2vec convolutional neural networks for classification of news articles and tweets, *PLOS ONE*, **14** (2019). https://doi.org/10.1371/journal.pone.0220976
34. S. Akbar, M. Hayat, M. Tahir, S. Khan, F. K. Alarfaj, cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model, *Artif. Intell. Med.*, **131** (2022). https://doi.org/10.1016/j.artmed.2022.102349
35. Z. Ma, L. Zhao, J. Li, X. Xu, J. Li, SiBERT: A Siamese-based BERT network for Chinese medical entities alignment, *Methods*, **205** (2022), 133–139. https://doi.org/10.1016/j.ymeth.2022.07.003
36. H. Zhao, B. Tao, L. Huang, B. Chen, A siamese network-based approach for vehicle pose estimation, *Front. Bioeng. Biotechnol.*, **10** (2022). https://doi.org/10.3389/fbioe.2022.948726
37. Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, *Neural Comput.*, **31** (2019), 1235–1270. https://doi.org/10.1162/neco_a_01199
38. M. Arbane, R. Benlamri, Y. Brik, A. D. Alahmar, Social media-based COVID-19 sentiment classification model using Bi-LSTM, *Expert Syst. Appl.*, **212** (2022). https://doi.org/10.1016/j.eswa.2022.118710

39. K. Liu, W. Gao, Q. Huang, Automatic modulation recognition based on a DCN-BiLSTM network, *Sensors*, **21** (2021). https://doi.org/10.3390/s21051577

AIMS Press