Computer Fraud and Security
ISSN (online): 1873-7056

Standardizing the Translation of Traditional Chinese Medicine Terminology: A Framework for Consistency

Feng Hongli[1], Hoe Foo Terng[2], Leow Min Hui[2], and Goh Chin Shuang[2]
[1]Foreign Language Teaching School, Ningxia Medical University, Ningxia, Yinchuan, China
120130121@nxmu.edu.cn
[2]Academy of Language Studies, University Teknologi, MARA, Malaysia

Abstract

Traditional Chinese Medicine (TCM) records contain many synonymous terms with different definitions, which do not allow it to be precisely translated and applied, across languages or systems. The absence of a standardized framework hinders cross-linguistic understanding and limits the integration of TCM into global healthcare. No standardized framework could exist to transcend linguistic barriers and integrate traditional Chinese practice into global medicine. The consistent framework for TCM terminology in translation is proposed in this research. The proposed Conversion of Synonymous Terms (CST) model normalizes TCM terminology is to improve the accuracy and consistency of translation. A Dynamic Backtracking Search Optimized Siamese Long Short-Term Memory (DBSO-Siamese LSTM) algorithm is used in text classification and for translation purposes of TCM terms. A large dataset was created from TCM medical records, literature, and expert-verified term dictionaries. The dataset experienced text preprocessing, including tokenization and stop-word removal, to improve model efficiency. To select the most critical terms in the TCM description, Term Frequency-Inverse Document Frequency (TF-IDF) is used. The DBSO algorithm improves the Siamese LSTM performance over TCM terminology classification by optimizing search parameters, and dealing with multimodal data in an efficient technique. Experimental results proved that the DBSO-Siamese LSTM model reached a 92.23% accuracy, 91.78% recall, and 92.15% F1-score. The CST model, being classification-based, improves translation consistency and synonym recognition, thus ensuring the global integration of TCM into modern healthcare systems.

Keywords: Traditional Chinese Medicine (TCM), Translation, Consistency, Dynamic Backtracking Search Optimized Siamese Long Short-Term Memory (DBSO-Siamese LSTM), Text Classification.

1. Introduction

Chinese medicine has been a staple in the formative Eastern health care for thousands of years. Entrenched holistic, the mechanism on balance in the body by using herbal medicines, acupuncture, and mind-body practices. In a short time, it spread globally and became of great interest to people who wanted to integrate into modern medical systems [1]. Traditional Chinese Medicine (TCM) is a whole medical system with distinctive diagnostic techniques and treatment philosophies. Contrasting with the western approach to medicine, TCM perceives health in terms of perceptions, such as Qi, Yin-Yang balance, and the five elements theory. Such an ancient and intricate history is a reason, why TCM has become so widespread and is under continued research. Heat and cold, wet and dry, light and dark, and other characteristics that correspond with the seasons, compass directions, and the human cycle of birth, development, and death were all used to structure each of these ancient systems [2].

TCM terminology comprises complex words, which indicate symptoms, syndromes, and treatment methodologies. The terminology used is with various synonymous words and differing descriptions, so the terms tend to be hard to standardize. Standardization of these terms help to communicate effectively in the modern health system [3]. The translation of TCM terms presents a large challenge because it differs in other languages and cultural settings. TCM terms could not have counterparts in other languages, which poses challenges in achieving uniformity among medical records and research [4].Efforts that contributed to the standardization of TCM terminology include work on lexicons, expert consultations, and computational linguistics. Although these terms have generated improved exactness, inconsistencies exist because of different historical texts and regional practices [5]. Developing a structured framework for TCM terminology translation

can improve consistency and accessibility. Standardized translations can enhance the reliability of medical records, support cross-cultural research, and facilitate TCM's integration into modern healthcare [6].

The lack of standardization in translating TCM terms leads to discrepancies in medical files and research; hence, translation across languages is not properly communicated. Standardization efforts of TCM terms have proved futile because of variations in languages, cultures, and histories. This research looks to generate an all-encompassing system to normalize synonymous terms of TCM, ensuring translations are accurate and seamlessly assimilated into health systems across the world.

1.1 Research Objective

The research is to develop a standardized translation framework for the TCM terminology. The terminology is transferred without any deviations or inaccuracies across languages. Using the Conversion of Synonymous Terms (CST) models and the Dynamic Backtracking Search Optimized Siamese Long Short-Term Memory (DBSO-Siamese LSTM) algorithm used to improve the accuracy of translation in this research. The research aim includes cross-linguistic understanding for better incorporation of TCM into modern healthcare systems and ensuring accurate medical documentation.

1.2 Research Organization

The subsequent sections of the research are categorized into: Section 2 gives provides a literature review and establishes research gaps. Section 3 presents the methodology with which DBSO integrates with Resilient RF to achieve dynamic optimization of resources. Section 4 defines results, including experiment setup performance analysis and then comparative analysis. The discussion section is presented in Section 5. The conclusion of the research with future avenues of research is offered in Section 6.

2. Related Works

To address the disparity in TCM terminology translation and medical text interpretation, a Comprehensive Medical Benchmark (CMB) was developed [7]. The benchmark was based on the native Chinese linguistic and cultural framework, though TCM formed a key constituent. The research was used to compare and contrast numerous Large Language Models (LLMs), including Chinese-specific LLMs and domain-specific medical-domain LLMs. However, its evaluation was limited to the scope of the benchmark selected, which could not capture regional and historical variations in the terms used for TCM.

A one-stage domain adaptation protocol was developed to unify heterogeneous data from pre-training and supervised learning into an instruction-output pair format for efficient knowledge TCM injection [8]. A data priority sampling strategy was implemented, where the data mixture was dynamically adjusted during training. By utilizing the concept, the specialized Chinese medical LLM Huatuo-II was developed with competitive performance against other models, which resulted in multiple Chinese medical benchmarks. However, such simplicity could hamper the capacity of the protocol to capture significant medical domain variations.

To aid the advancement in the medical field, a large multilingual model was developed to support autoregressive domain adaptation for general-purpose LLMs [9]. A multilingual medical multi-choice question-answering benchmark (MMedBench) was also developed to track the progress of the multilingual development of medical LLMs. After testing several open-source models on it, the model with only 8B parameters performed well. The research was limited by the restricted number of languages and medical domains.

BiomedGPT was the first open-source and light mass vision-language-based framework that could be developed as a generalist for an extensive assortment of biomedical responsibilities [10]. It has achieved enhanced performance in several experiments. Human evaluations have been carried out to determine its performance in the tasks of radioscopy graphic inquiry responding, statement generation, and summarization. The model had a better forecast capability with minimal error rates in question answering. Despite that, the performance of the model was limited because of its specificity in biomedical domains and data diversity.

A model was developed with precision medicine to revolutionize translational research through the use of predictive biomarkers to direct tailored treatments, avoiding ineffective therapies and adverse effects [11].

Multi-omics analysis and large-scale collection of clinical, behavioral, and environmental data allow for the creation of digital health profiles. Real-world data applications offer an exciting alternative to traditional evidence-based medicine, especially in precision oncology. The model's drawback is that the integration of data is complex.

The LLMs and Chat models were tested on the translation, question answering, and summarization of medical text for tasks like clinical practice, research, and education [12]. The possibilities for democratizing medical knowledge and healthcare access were evaluated. The focus of the research was to check the interpretability of the medical content and whether the models could support the decision-making process. However, the research has a limitation because of the absence of real-world validation and issues related to accountability and transparency.

A MedChatZH, transformer decoder-based dialogue model was introduced and fine-tuned with the LLM architecture for optimizing performance on Chinese medical QA [13]. The model continued pre-training with curated Chinese medical texts and then fine-tuned using a domain-specific instruction dataset. Performance was evaluated by comparing MedChatZH with several Chinese dialogue baselines on real-world medical QA tasks. However, its results depend on the availability of a wide range of high-quality training data.

An open-source LLM called the BioMistral model was specifically designed for the biomedical domain and utilized the mistral foundation model [14]. A thorough evaluation was conducted on a medical QA task. Furthermore, lightweight models were discussed by using techniques from quantization and model merging. The results showed that BioMistral's performance was improved and also competitive. However, the model had the limited non-English availability of medical datasets.

2.1 Research Gaps

TCM terminology translation faces limitations, such as narrow linguistic and domain coverage, reliance on specific biomedical data, and a lack of comprehensive real-world validation. For instance, models that showed competitive performance are constrained by data diversity and limited language coverage. Furthermore, LLMs often struggle with the complexity of TCM terminology and medical domain variations. Existing TCM terminology translation works lack sufficient linguistic and domain coverage, dependence on narrow biomedical data, and limited comprehensive real-world testing. The proposed DBSO-Siamese LSTM model work addresses these limitations by introducing a structured framework for standardizing TCM terminology, integrating multilingual capabilities, and enhancing real-world applicability, ensuring more reliable, consistent, and cross-linguistic integration of TCM into healthcare systems. The model introduces a structured framework to standardize TCM terminology, integrate multilingual capabilities, and enhance real-world applicability, bringing TCM in health more reliably, consistently, and cross-linguistically.

3. Methodology

The section of methodology details the processes involved in the standardization of TCM terminology translation, specifically data collection, preprocessing, and feature extraction. The enhancement of classification and translation of synonymic TCM terms is applied using the CST model and DBSO-Siamese LSTM algorithm. Figure 1 depicts the workflow of the proposed methodology for TCM terminology normalization and translation consistency.

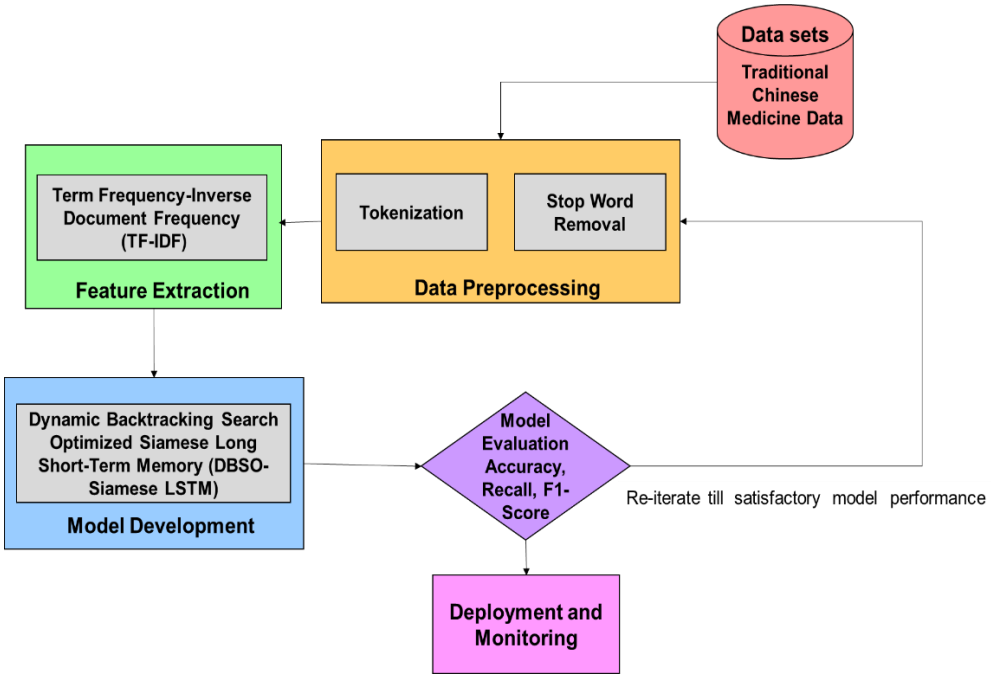

Figure 1 Workflow of the proposed methodology for TCM terminology standardization

3.1 Data Collection

The TCM data [15] is a publicly accessible Kaggle dataset that was used in the research. The dataset contains TCM herbal compositions and their indications, which is a good candidate for training and fine-tuning the DBSO-Siamese LSTM model. The dataset has a structured collection of synonymous TCM terms that would help with terminology normalization.

3.2 Data Preprocessing

Data preprocessing is the most significant phase in preparing raw text data to extract the meaningful components only. Here, tokenization and stop-word removal helped to progress the efficiency and accuracy of the model.

- **Tokenization**

Tokenization is the procedure of splitting a text into meaningful elements, referred to as tokens. The split of sentences into words based on whitespaces facilitates further analysis of the text. Irrelevant items such as tags, scripts, special characters, and punctuation marks are removed in the TCM dataset by tokenization because they do not help to improve the classification and translation of TCM terminology. Removal of these aspects reduces the features and therefore makes the classification more efficient. During this step, only meaningful components of the text are retained for further processing.

- **Stop-word Removal**

Stop-word removal is the method of eliminating stop-words, which are words that do not carry much meaning, such as "the," "this," and "but." In this methodology, this technique is employed on the TCM dataset. These words are repeated many times in the text and do not affect the classification, so removing them reduces the complexity of the computation. The process of removing stop-words compares each token in the text to the list of stop-words and removes the matching words, thus reducing the set of tokens and improving accuracy in terminology normalization.

3.3 Feature Extraction

In this methodology, the Term Frequency-Inverse Document Frequency (TF-IDF) technique is employed to extract features for key terms in TCM data. It measures the score of the importance of a term in a specific document relative to all other documents in the entire corpus. Equation (1) represents the TF-IDF.

$$T_F -= T_F \times IDF \quad (1)$$

Where T_F denotes the frequency of a term occurring, divided by the overall number of terms in the file, using the equation (2).

$$T_F = \frac{D_{ffad}}{D_{tntd}} \quad (2)$$

Here, $ffad$ is the frequency of a feature appearing in a document, and $tntd$ is the total number of terms in the document. IDF evaluates the ability of a term to distinguish between different categories. It is calculated using equation (3).

$$IDF = log\left(\frac{NDF}{T_D}\right) \quad (3)$$

Where NDF is the count of documents in which the term occurs. The total number of documents is T_D, and the TF-IDF value means the importance of the term for the document.

3.4 Model Development

To translate a text correctly, a Siamese LSTM network is merged with the DBSO algorithm to further improve accuracy in translation. The Siamese LSTM captures deep semantic relationships. Additionally, it dynamically fine-tunes the hyperparameter through DBSO to enhance the model's performance. A cross-linguistic consistency is ensured in terms of TCM term standardization.

3.4.1 Siamese Long Short-Term Memory (Siamese LSTM) for TCM Terminology Translation

The Siamese LSTM network was applied in this research to measure the semantic similarity between TCM terms and their synonymous counterparts. Given the inconsistencies of the nomenclature used across sources, this kind of Siamese LSTM model ensures good classification and normalization through deep contextual learning between synonymous terms. In translating the defined descriptions, the shared-weight LSTM structure ensures that there is consistency while conserving their semantic integrity.

- **Network Architecture**

The Siamese LSTM model is formed by two identical LSTM networks that analyze TCM term pairs, concerning to learn about the contextual similarities between them. Deep semantic meaning is extracted from each term as a sequential input and is passed through an LSTM unit. The final similarity score determines whether the terms are synonymous and should be standardized. In the case of an input pair $(v^{(z)}, v^{(l)})$, where $v^{(z)}$ and $v^{(l)}$ are two TCM terms, the updates for the LSTM are defined in the equations (4), (5), (6), (7), and (8).

$$j_q = \left(E_j r_{q-1} + D_j v_q + l_j\right) \quad (4)$$

$$u_q = \rho\left(E_u r_{q-1} + D_u v_q + l_u\right) \quad (5)$$

$$s_q = u_q s_{q-1} + j_q tanr\left(E_s r_{q-1} + D_s v_q + l_s\right) \quad (6)$$

$$p_q = \rho\left(E_p r_{q-1} + D_p v_q + l_p\right) \quad (7)$$

$$r_q = p_q \tan\left(s_q\right) \quad (8)$$

Where, $j_q, u_q, and\ p_q$ stand for input, forget, and output gates. s_q is the cell state, which includes information on the context from step to step. E and D are trainable weight matrices. l is used as the bias term.

The final similarity score for the Siamese LSTM is computed as the Euclidean distance between the hidden representations($r^{(z)}, r^{(l)}$), of the input term pair, using equation (9).

$$m = exp\left(-\left\|r_Y^{(z)} - r_Y^{(l)}\right\|_1\right) \ \epsilon\ [0,1] \tag{9}$$

where m is the similarity measure. The Siamese LSTM is incorporated into the CST model that classifies term pairs as synonymous or non-synonymous. Figure 2 illustrates the architecture of the Siamese LSTM.

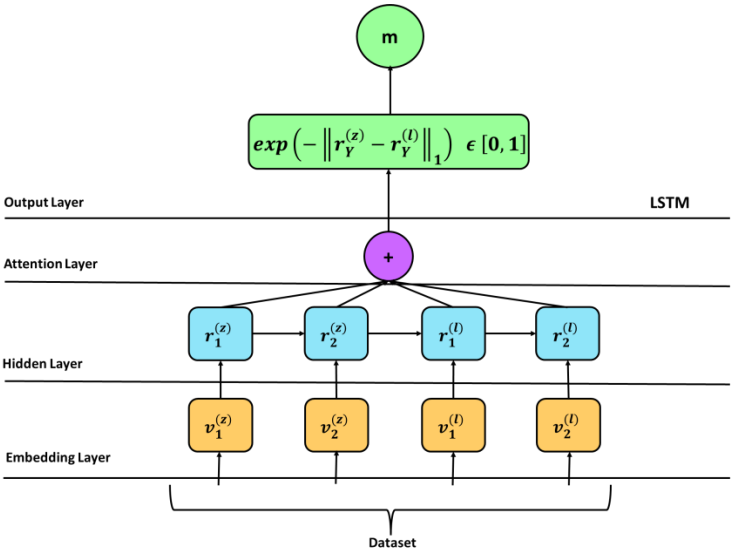

Figure 2 Architecture of Siamese Long Short-Term Memory (Siamese LSTM)

3.4.2 Dynamic Backtracking Search Optimization (DBSO) Algorithm

The DBSO algorithm is an upgraded version of the standard Backtracking Search Algorithm (BSA), proposed to further improve the validity and consistency of translation in TCM terminology normalization. Dynamic adaptation makes DBSO go beyond the traditional process by incorporating adaptive dynamic adjustments that help in the optimization of a term-classification translation task. It combines historical population data with an adaptive mutation strategy on the CST model for optimized text classification. As described below are the operations involved in the DBSO algorithm.

- ***Initialization***

The initial population Y and historical population $OLDY$ are generated using equations (10) and (11).

$$Y_{cr} = om_r + rand \times (pm_r - om_r) \tag{10}$$

$$OLDY_{cr} = om_r + rand \times (pm_r - om_r) \tag{11}$$

where $c = 1,2,3,\ldots,N$, stands for population size, and $r = 1,2,3,\ldots,D$ is variable dimension. pm_r and om_r are the minimum and maximum for each of the dimensions. $rand$ has a uniform distribution between 0 and 1.

- ***Selection I (Historical Population Update)***

At the start of every generation, the historical population is computed with dynamics with the former population using equation (12)

$$OLDY_{cr} = \begin{cases} Y_{cr}, & if\ e < m, \quad e, m \sim W(0,1) \\ PREVIOUS\ OLDY_{cr}, & otherwise \end{cases} \tag{12}$$

Where $W(0,1)$ is a uniformly random variable, the historical population is further modified using a permutation function in equation (13).

$$OLDY \leftarrow PERMUTING(OLDY) \tag{13}$$

This action adds some unpredictability to hold the memory in the preceding generation, and so the TCM terminology search direction improved.

- ***Mutation (Dynamic Adaptive Strategy)***

The mutation progression presents dynamic scaling using an adaptive factor K_g, which improves exploration and exploitation capabilities, using equation (14).

$$MUTATION_{cr} = Y_{cr} + K_g \times (OLDY_{cr} - Y_{cr}) \tag{14}$$

where K_g is defined as in the equation (15).

$$K_g = K_{BASE} + \beta \times randz \tag{15}$$

Where K_{BASE} is an initial scaling factor, β is an adaptively tuned coefficient, and $randz$ represents a normally distributed random number. The dynamic adaptation ensures robustness in normalizing synonymous TCM terms across different linguistic structures.

- ***Crossover (Hybrid Search Enhancement)***

A crossover procedure is useful to enhance the search path by producing a new population S, using equation (16).

$$S_{cr} = \begin{cases} MUTATION_{cr}, & if\ MAP_{cr} = 1 \\ Y_{cr}, & if\ MAP_{cr} = 0 \end{cases} \tag{16}$$

where MAP is a binary matrix defining the crossover operation, and only meaningful changes are introduced while the high accuracy of translations is preserved.

- ***Selection II (Greedy Selection for Optimal Term Matching***)

To ensure the highest quality translations on TCM normalization, the aptness of individual solutions is verified, and their best solution is conserved from the equation (17).

$$Y_{cr} = \begin{cases} S_{cr}, & if\ \ t(S_c) < t(Y_c) \\ Y_{cr}, & otherwise \end{cases} \tag{17}$$

where $t(S_c)$ and $t(Y_c)$ are the trials and the current population's fitness. The fitness function is optimized to improve translation accuracy and consistency. The flow for DBSO is illustrated in Figure 3.

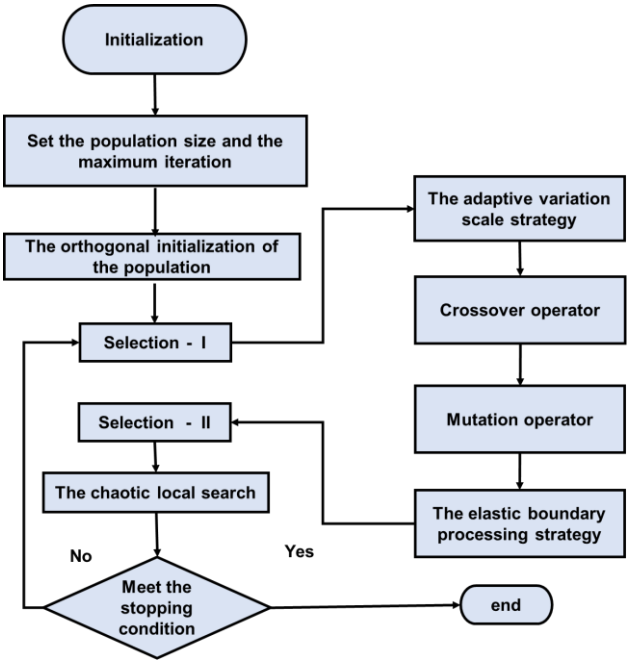

Figure 3 Flow of Dynamic Backtracking Search Optimized (DBSO)

The proposed DBSO algorithm improves the Siamese LSTM performance over TCM terminology classification by optimizing search parameters, dealing with multimodal data in an efficient way, and ensuring accurate translation of synonymous terms. DBSO enhances the accuracy of classification, reduces overfitting, and accelerates convergence by fine-tuning the key hyperparameters of learning rate, dropout rate, batch size, and number of hidden units. Adaptive mutation and crossover strategies through robust optimization improve cross-linguistic TCM term mapping and enhance the support for global integration in healthcare. Algorithm for DBSO-Siamese LSTM is presented below as Algorithm 1.

Algorithm 1: Dynamic Backtracking Search Optimized Siamese Long Short-Term Memory (DBSO-Siamese LSTM) algorithm

Step 1: Initialization

Initialize population Y and historical population $OLDY$ using random values

Set hyperparameter search space

Set max iterations T_{max} and termination criteria

Step 2: Siamese LSTM Model Definition

Define Siamese LSTM with shared-weight architecture

For each input term pair $(v^{(z)}, v^{(l)})$:

Process through LSTM layers:

$$j_q = \rho\left(E_j r_{q-1} + D_j v_q + l_j\right)$$

$$u_q = \rho\left(E_u r_{q-1} + D_u v_q + l_u\right)$$

$$s_q = u_q s_{q-1} + j_q tanr\left(E_s r_{q-1} + D_s v_q + l_s\right)$$

$$p_q = \rho\left(E_p r_{q-1} + D_p v_q + l_p\right)$$

$$r_q = p_q \tan\left(s_q\right)$$

Compute similarity score:

$$m = exp\left(-\left\|r_Y^{(z)} - r_Y^{(l)}\right\|_1\right) \in [0,1]$$

Step 3: DBSO Hyperparameter Optimization

While the termination condition is not met:

Generate population Y and OLDY:

$$Y_{cr} = om_r + rand \times (pm_r - om_r)$$

$OLDY_{cr} = om_r + rand \times (pm_r - om_r)$
Selection I - Update historical population
If e < m (e, m ~ W (0,1)):
$OLDY_{cr}$ = Y_{cr}
Else:
$OLDY_{cr}$ = previous $OLDY_{cr}$
Shuffle OLDY using permutation function: OLDY ← PERMUTING(OLDY)
$MUTATION_{cr}$ = Y_{cr} + K_g × ($OLDY_{cr}$ - Y_{cr})
$K_g = K_{BASE} + \beta \times randz$
If MAP_{cr} = 1:
S_{cr} = $MUTATION_{cr}$
Else:
S_{cr} = Y_{cr}
If $t(S_c) < t(Y_c)$:
Y_{cr} = S_{cr}
Else:
Y_{cr} remains unchanged
Step 4: Evaluation of the Model
Evaluate Siamese LSTM performance using selected hyperparameters
Store the best hyperparameter set
Return the optimized Siamese LSTM model

4. Result and Analysis

The evaluation of the DBSO-Siamese LSTM model for normalizing of TCM terminology. The analysis covers training and validation performance, classification metrics, and comparative research against an existing model. In addition to classification capability through the ROC curve, accuracy, recall, and F1-score were used to represent the quantitative insight into the model's performance. A comparative evaluation further highlighted improvements achieved by the proposed approach.

4.1 Experimental Setup

The experiment setup runs on a 16GB RAM, 500GB SSD-based multi-core Intel Core i7 processor paired with an optional Nvidia GTX 1060 DL-enhanced GPU. The software environment employed Python 3.11.2 for model development built within TensorFlow/Keras and Scikit-learn features for extracting the features along with classification. Then settings are managed in Jupyter Notebook/PyCharm on MySQL/PostgreSQL storage management.

4.2 Performance Analysis

The performance of the DBSO-Siamese LSTM model is assessed using the training and validation accuracy and loss curves to measure the learning efficiency and generalization. The curve for accuracy gives the classification performance of the model, and the figure for loss shows the decline in error. The smooth decline of loss and the increase in accuracy confirm stable convergence and reduced overfitting. Figure 4 shows (a) Training and Validation Accuracy and (b) Training and Validation Loss, which indicates the good performance of the model in TCM terminology normalization.

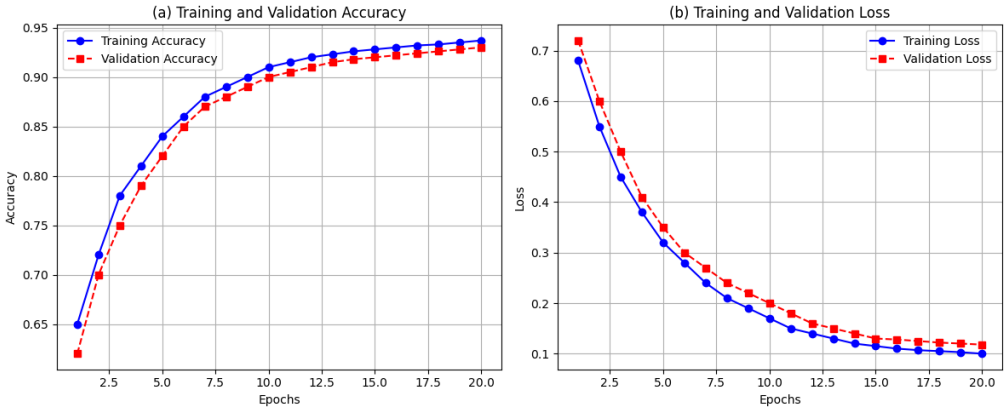

Figure 4 DBSO-Siamese LSTM's Training and Validation (a) Accuracy and (b) Loss

For further verification of model performance, the following set of metrics was employed. Accuracy is the ratio of correctly classified synonymous TCM terms and non-synonymous TCM terms. Recall measures the model's capability to capture appropriately all actual synonymous instances. F1-Score gives a balanced measure between correctness and recall for indicating effectiveness in global classification. The mathematical representations of the aforesaid metrics are provided in equations (18), (19), and (20).

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{18}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{19}$$

$$F1 - Score = 2 \times \frac{Precision}{Recall} \tag{20}$$

Where, F_n= False Negatives, T_p= True Positives, T_n= True Negatives. F_p = False Positives. The resultant values for these evaluation metrics are provided in Table 1.

Table 1 Performance Metrics of DBSO-Siamese LSTM

Metrics	Values (%)
Accuracy	92.23
Recall	91.78
F1-Score	92.15

The proposed DBSO-Siamese LSTM model achieved a 92.23% accuracy, which further proved its adequacy in synonymously classifying TCM terms. The high sensitivity of identifying true synonymous terms is indicated by a 91.78% recall, whereas the 92.15% F1-score ensures accurate and consistent terminology standardization.

Receiver Operating Characteristic (ROC) curve

A Receiver Operating Characteristic (ROC) curve was plotted, to further evaluate the DBSO-Siamese LSTM model's performance in classification. The ROC curve illustrates the capability of the model to differentiate between synonymous and nonsynonymous TCM terms for various classification thresholds. The Area Under the Curve (AUC) measures the complete performance of the model. Figure 5 demonstrates the ROC curve for the DBSO-Siamese LSTM model.

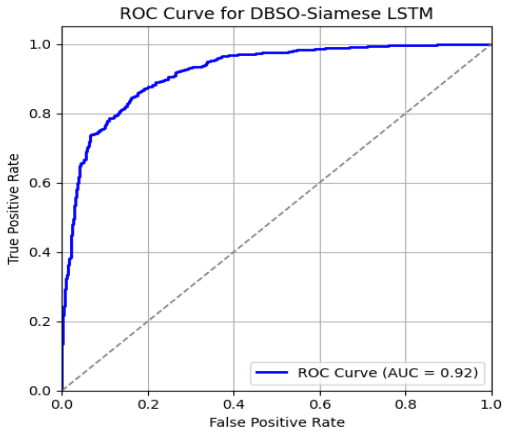

Figure 5 ROC Curve for DBSO-Siamese LSTM

4.3 Comparative Analysis

A comparative analysis of the proposed DBSO-Siamese LSTM model was carried out with VCPC–WE [16], considering the accuracy of translation and the performance in classification. The comparison was concentrated on accuracy, recall, and F1-score since they define the aptness of normalizing synonymous TCM terms. The performance comparison outcomes are given in Table 2. Figure 6 exhibits a graphical representation of comparison results.

Table 2 Performance Comparison of Models

Model	Recall (%)	F1-Score (%)	Accuracy (%)
VCPC–WE [16]	85.04	85.52	86.01
DBSO-Siamese LSTM [Proposed]	91.78	92.15	92.23

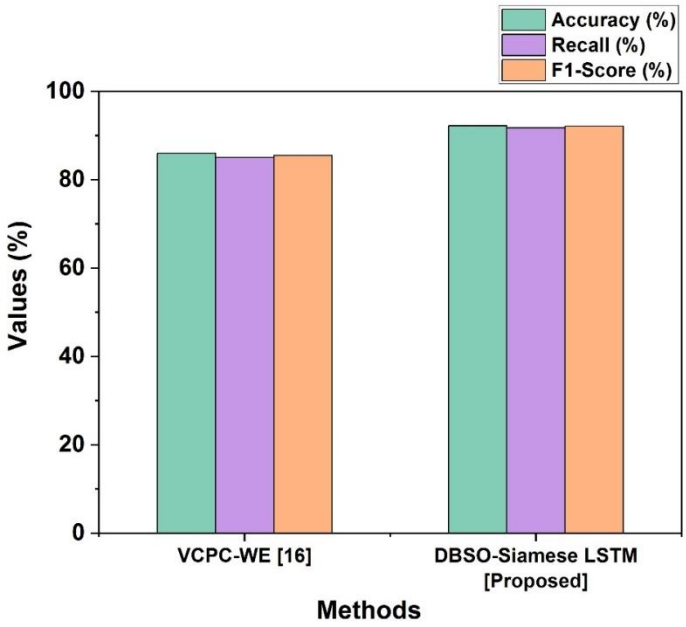

Figure 6 Comparison of Model Performance Metrics

The proposed DBSO-Siamese LSTM model outperforms VCPC–WE with 92.23% accuracy. Furthermore, the recall and F1-score were achieved at 91.78% and 92.15% for proposed models that significantly improved in retrieving and normalizing synonyms in TCM terms. The experiments support the strengths of integrating

DBSO optimization with Siamese LSTM in leveraging higher translation accuracy as well as improved classification performance in TCM terminology standardization.

5. Discussion

Existing research in the translation of TCM terminology has some inherent drawbacks, including non-standardization, poor linguistic adaptability, and suboptimal classification. Earlier approaches, like VCPC–WE [17], were not able to normalize synonymous and non-synonymous terms well, which caused many misclassifications. In addition lacked an effective optimization mechanism increased error rates and decreased recall. The unstructured feature extraction process also decreased the reliability of translations. The proposed model integrates DBSO with a Siamese LSTM network that enhances the accuracy of classification and translation precision. The DBSO-Siamese LSTM model effectively learns deep semantic relationships between terms, reduces misclassification errors, and thereby outperforms previous methods by a large margin in terms of accuracy, recall, and F1-score. This framework results in higher consistency in translation, better synonym recognition, and standardization of TCM terminology; hence, this is more reliable solution for medical text classification.

6. Conclusion

The integration of TCM into modern health systems faces inconsistency in translation due to terminology. A robust framework for the standardization of TCM terminology translation was developed using a DBSO-Siamese LSTM model. The research collected dataset includes medical records, literature, and sources verified by experts and preprocessing using tokenization and stop-word removal was followed by TF-IDF-based feature extraction. The model improved classification accuracy by capturing semantic relations between synonymous terms. Hyperparameters were optimized by incorporation of the DBSO algorithm for better performance in the model. Experimenting results proved an accuracy of 92.23% recall of 91.78%, and an F1-score of 92.15%, of the DBSO-Siamese LSTM model surpassed the existing model in performance. The model though effective requires further refinement in terms of performance using a much more diverse dataset that covers other linguistic variations. Hence, in future work, the idea would include incorporating domain-adaptive learning techniques to improve adaptability across various medical texts while enlarging applicability in global healthcare systems.

References

[1] Wang, W.Y., Zhou, H., Wang, Y.F., Sang, B.S. and Liu, L., 2021. Current policies and measures on the development of traditional Chinese medicine in China. *Pharmacological research*, *163*, p.105187.https://doi.org/10.1016/j.phrs.2020.105187

[2] Lee, D.Y., Li, Q.Y., Liu, J. and Efferth, T., 2021. Traditional Chinese herbal medicine at the forefront battle against COVID-19: Clinical experience and scientific basis. *Phytomedicine*, *80*, p.153337.https://doi.org/10.1016/j.phymed.2020.153337

[3] Huang, N., Huang, W., Wu, J., Long, S., Luo, Y. and Huang, J., 2024. Possible opportunities and challenges for traditional Chinese medicine research in 2035. *Frontiers in Pharmacology*, *15*, p.1426300.https://doi.org/10.3389/fphar.2024.1426300

[4] Cheung, H., Doughty, H., Hinsley, A., Hsu, E., Lee, T.M., Milner-Gulland, E.J., Possingham, H.P. and Biggs, D., 2020. Understanding Traditional Chinese Medicine to strengthen conservation outcomes. https://doi.org/10.1002/pan3.10166

[5] Ying, C., Shuyu, Y., Jing, L., Lin, D. and Qi, Q., 2021. Errors of machine translation of terminology in the patent text from English into Chinese. *ASP Transactions on Computers*, *1*(1), pp.12-17.https://doi.org/10.52810/TC.2021.100022

[6] Zheng, Y., Yifu, S. and Xiumin, T., Exploration of Computer-Assisted Translation Technology in Translating Technical Terms in Traditional Chinese Medicine under the Perspective of AI Vision. https://dx.doi.org/10.23977/jaip.2023.060706

[7] Wang, X., Chen, G.H., Song, D., Zhang, Z., Chen, Z., Xiao, Q., Jiang, F., Li, J., Wan, X., Wang, B. and Li, H., 2023. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.https://doi.org/10.48550/arXiv.2308.08833

[8] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M. and Dufour, R., 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*. https://doi.org/10.48550/arXiv.2402.10373

[9] Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y. and Xie, W., 2024. Towards building multilingual language model for medicine. *Nature Communications*, *15*(1), p.8384. https://doi.org/10.1038/s41467-024-52417-z

[10] Hartl, D., de Luca, V., Kostikova, A., Laramie, J., Kennedy, S., Ferrero, E., Siegel, R., Fink, M., Ahmed, S., Millholland, J. and Schuhmacher, A., 2021. Translational precision medicine: an industry perspective. *Journal of translational medicine*, *19*(1), p.245. https://doi.org/10.1186/s12967-021-02910-6

[11] De Maria Marchiano, R., Di Sante, G., Piro, G., Carbone, C., Tortora, G., Boldrini, L., Pietragalla, A., Daniele, G., Tredicine, M., Cesario, A. and Valentini, V., 2021. Translational research in the era of precision medicine: where we are and where we will go. *Journal of Personalized Medicine*, *11*(3), p.216. https://doi.org/10.3390/jpm11030216

[12] Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.N., Laleh, N.G., Löffler, C.M.L., Schwarzkopf, S.C., Unger, M., Veldhuizen, G.P. and Wagner, S.J., 2023. The future landscape of large language models in medicine. *Communications medicine*, *3*(1), p.141.https://doi.org/10.1038/s43856-023-00370-1

[13] Tan, Y., Zhang, Z., Li, M., Pan, F., Duan, H., Huang, Z., Deng, H., Yu, Z., Yang, C., Shen, G. and Qi, P., 2024. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Computers in Biology and Medicine*, *172*, p.108290. https://doi.org/10.1016/j.compbiomed.2024.108290

[14] Kim, K., Cho, K., Jang, R., Kyung, S., Lee, S., Ham, S., Choi, E., Hong, G.S. and Kim, N., 2024. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean Journal of Radiology*, *25*(3), p.224. https://doi.org/10.3348/kjr.2023.0818

[15] https://www.kaggle.com/datasets/guoxiangzu/traditional-chinese-medicine-data?utm_source=chatgpt.com

[16] Ma, Y., Sun, Z., Zhang, D. and Feng, Y., 2022. Traditional chinese medicine word representation model augmented with semantic and grammatical information. *Information*, *13*(6), p.296. https://doi.org/10.3390/info13060296