MONGOLIAN MEDICINE PRESCRIPTION RECOMMENDATION USING GRAPH ATTENTION NETWORKS
Leveraging Semantic Associations for Precise Predictions

by

Shuqin HAN[a], Sandan BAO[b*], and Haibin LI[a]

[a]Inner Mongolia University of Technology, Hohhot, China
[b]International Mongolian Hospital of inner Mongolia, Hohhot, China

Original scientific paper
https://doi.org/10.2298/TSCI2602107H

The objective of this study is to address the challenges faced by traditional Mongolian medicine in the modern era. The complex knowledge system and challenges related to inheritance in Mongolian medicine represent significant obstacles to the modern development of this discipline. The present study introduces a graph attention network (GAT) model to address these issues. The GAT model establishes graphs of symptoms, Mongolian medicine, and symptoms-Mongolian medicine. The GAT model employs graph convolution operations to effectively capture the intricate relationships among symptoms and Mongolian medicines. This facilitates the model capacity to discern representations that are both discriminative of symptoms and Mongolian medicines. Consequently, the model is capable of matching appropriate Mongolian medicinal prescriptions according to the input symptoms. A series of experimental evaluations were conducted on a dataset derived from the Encyclopedia of Mongolian Medicine. These evaluations demonstrated that the proposed GAT model outperforms existing models in terms of prescription recommendation accuracy. Specifically, the model achieves an accuracy of 37.59%, representing significant improvements compared to other models. These findings suggest that the GAT model can effectively leverage the relationships among symptoms and Mongolian medicines to provide reliable prescription recommendations, offering a promising solution for the modernization of Mongolian medicine.

Key words: *traditional Mongolian medicine, prescription recommendations, deep learning, graph neural network*

Introduction

Mongolian medicine, a comprehensive body of knowledge that embodies the collective wisdom of the Mongolian people, has a long-standing history that dates back centuries. It has evolved within the context of Mongolian society, emerging from the long-term endeavors of the people in their daily lives, production activities, and battles against various diseases. This traditional medical system is distinguished by its comprehensive theoretical underpinnings and extensive practical expertise. It has been a fundamental component of healthcare for a considerable number of individuals, providing distinctive treatment modalities that have effectively mitigated patient suffering. For instance, its holistic view of the human body, which

* Corresponding author, e-mail: lhbnm2003@126.com

emphasizes the interconnection of different physiological components, is distinct from Western medical approaches [1].

Nevertheless, in the modern era, traditional Mongolian medicine faces a series of formidable challenges. The intricacy of its knowledge system, which encompasses a vast array of theories related to the body's constitution, disease etiology, and treatment principles, poses a significant hurdle to its widespread understanding and dissemination. Furthermore, the process of inheritance is often accompanied by a series of challenges. The conventional approach to the dissemination of medical knowledge, which is predominantly characterized by oral instruction and apprenticeships, has become progressively ineffective in the contemporary context, which is characterized by a proliferation of distractions and a accelerating pace of life. This has resulted in a potential loss of valuable medical wisdom that has been accumulated over generations [1].

The fundamental principle of Mongolian medicine is predicated on the concept of the three roots (Heyi, Xila, Badagan) and seven constituents (essence, blood, flesh, fat, bones, marrow, vital energy, and essential spirit) [2]. It is hypothesized that these elements contribute to maintaining bodily balance and health. The interplay among the three roots is instrumental in elucidating physiological and pathological phenomena. In the event of disruption by various pathogenic factors, an imbalance in the levels of the three roots has been demonstrated to trigger diseases. In clinical practice, Mongolian doctors adhere to the principle of *etiological differentiation*. Information is gathered through comprehensive methods, including patient interviews, meticulous observations, and precise palpations. Through meticulous analysis of this information, healthcare professionals can accurately ascertain the etiology, nature, and manifestation of the disease, and subsequently diagnose it in terms of the relationships among the three roots. Treatment strategies are subsequently devised to restore the equilibrium of the three roots, encompassing the selection of suitable medications. Consequently, the accurate diagnosis of the relationships among the three roots based on symptoms and the recommendation of suitable medications are two essential tasks for an intelligent Mongolian medical auxiliary diagnosis and treatment system [3, 4].

The advent of the new technological and industrial revolutions has witnessed the remarkable rise of AI [5, 6]. The potential for AI to transform various fields, including Mongolian medicine, is significant. Graph neural networks [7, 8] represent a state-of-the-art deep learning technology based on graph-structured data. These networks have been shown to possess powerful representational learning capabilities [9]. These tools have been demonstrated to effectively capture complex relationships within data structures. The modeling of the Mongolian medical prescription process as a graph, with symptoms and medications as interconnected nodes, suggests the potential for graph neural networks to facilitate technological and knowledge-based innovation in Mongolian medicine research [10].

Given the complexity of Mongolian medicine diagnostic and treatment methods, this paper proposes the utilization of GAT [11]. The GAT architecture is designed to construct graphs of symptoms and Mongolian medicine, as well as graphs of Mongolian medicine and symptoms, in the representation learning layer. By employing graph convolution operations across these multiple graphs, GAT can learn the latent representations of symptoms and Mongolian medicines. In the prediction layer, a multi-layer perceptron (MLP) [12] is employed to fuse the learned symptom representations. The objective of this fusion process is to obtain an implicit representation of the disease cause, which is then used to recommend appropriate Mongolian medicine prescriptions.

To illustrate the practical process of Mongolian medicine diagnosis and treatment, Figure 1 presents the case of *hemorrhagic headache*. During the collection of symptoms, medical professionals may observe indications such as *throbbing headache*, *nosebleed*, *brownish urine*, and *feverish feeling*. Following a thorough diagnostic evaluation, the underlying cause of the patient's symptoms was identified as a headache induced by elevated blood temperature. Consequently, the treatment primary focus is on the clearance of blood heat and the alleviation of pain. In the initial phase of treatment, which is rooted in traditional medicine practices, a range of medications are prescribed. These include bezoar, Gentiana, and a dispersion of safflower, each with its own unique properties and application. Each of these medications has been associated with specific symptoms. For example, bezoar targets throbbing pain, Gentiana addresses nosebleed and clears blood heat, and the thirteen-flavor safflower dispersion has detoxifying and heat-clearing effects, corresponding to brownish urine and feverish symptoms. This example vividly demonstrates that identifying the etiology of the disease serves as a crucial link between symptoms and prescriptions. However, it is not uncommon for different Mongolian doctors to arrive at different diagnoses and prescriptions for the same disease due to variations in their personal experiences and knowledge levels [13]. This variability underscores the necessity for an objective and standardized approach, which has led to the development of a model capable of automatically generating prescriptions through the analysis of classical medical records. It is imperative to underscore that the objective of this study is not to supplant human practitioners but rather to furnish candidate prescriptions that will facilitate the prescription-making process. These recommendations are informed by the practical constraints inherent to medical practice and the indispensable role of human expertise therein.

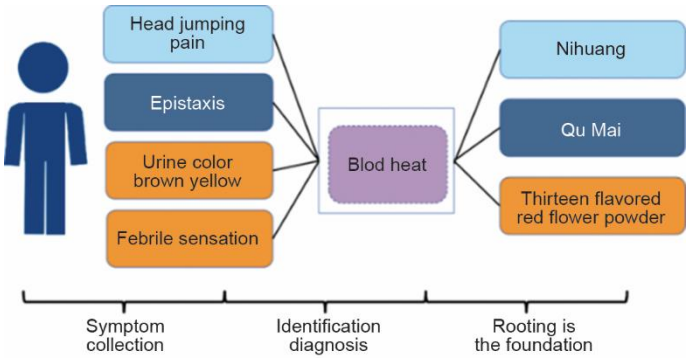

Figure 1. Example of Mongolian medicine treatment process for *bloody headache*

Methodology

Problem definition

The task of recommending Mongolian prescription formulations is defined as recommending a set of Mongolian medicines suitable for treating a given set of symptoms. Let $S = \{s_1, s_2,\ldots, s_m\}$, $H = \{h_1, h_2,\ldots, h_N\}$, represent the sets of all symptoms and all Mongolian medicines, respectively, where M and N represent the scale of symptoms and Mongolian medicines, respectively. The $p = \{s_1, s_2,\ldots, s_k; h_1, h_2,\ldots, h_l\}$ represents a prescription consisting of a group of symptoms $S = \{s_1, s_2,\ldots, s_k\}$ and corresponding Mongolian medicines $h = \{h_1,$

h_2,…, h_l}, with the set of all Mongolian medicine prescriptions denoted as P. The goal of the model is to predict the probability vector $\hat{y}$ for all Mongolian medicines given a set of symptoms s.

Modeling framework

This paper proposes the use of GAT models to implement the task of recommending Mongolian medicinal prescriptions. The basic structure, as illustrated in fig. 2, includes three parts: the input layer, the graph attention layer, and the prediction layer.

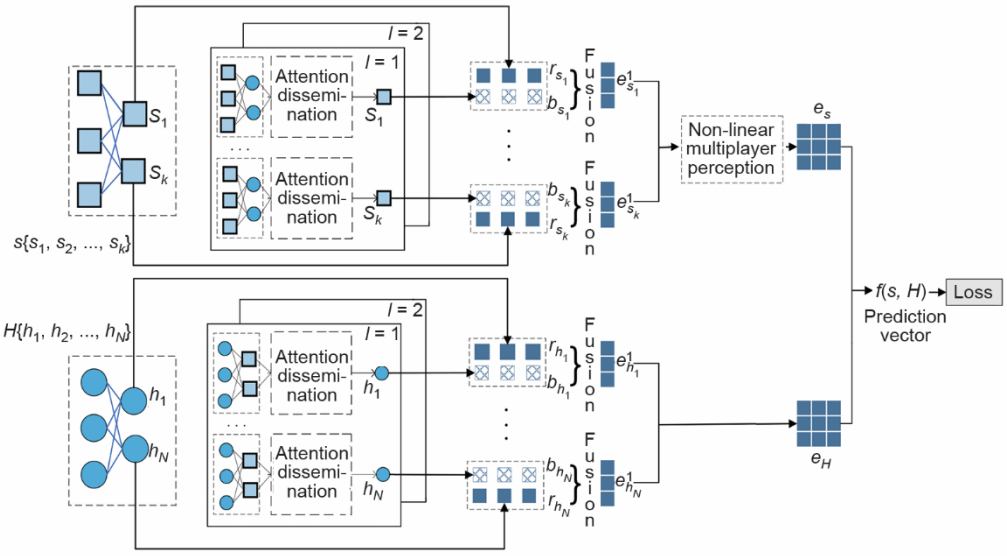

Figure 2. The overall architecture of the GAT model

Input layer

The objective of the input layer in the GAT model is to encode the entity features of symptoms and Mongolian medicines, generating low-dimensional embedding vectors as input for the graph attention layer. Separate symptom-symptom graphs and Mongolian medicine-Mongolian medicine graphs are constructed as input to the model. Figure 3, respectively, show examples of symptom-symptom graphs and Mongolian medicine-Mongolian medicine graphs. Let the initial node representation be $H_s^0 = x_{s_i}, H_h^0 = x_{h_i}, x_{s_i}, x_{h_i} \in R^{1\times d_0}$, where x_{s_i}, x_{h_i} represents the initial vectors for symptom and Mongolian medicine nodes, respectively, and d_0 is the dimensionality of the node initial representation.

Graph attention layer

This paper utilizes a single-layer GAT to process the symptom-symptom graph (S – S) and the medicine-medicine graph (H – H), resulting in the representations of symptoms r_s and r_h defined:

$$r_{s_i} = \tanh\left(\sum_{k\in N_{s_i}^{S-S}} \omega_{ij}^{S-S} x_{s_i} V_s\right) \tag{1}$$

$$r_{h_i} = \tanh\left(\sum_{q\in N_{h_i}^{H-H}} \omega_{ij}^{H-H} x_{h_i} V_h\right) \tag{2}$$

where $N_{s_i}^{S-S}$ represents the set of neighbouring nodes of symptom s_i in the symptom-symptom graph and $N_{h_i}^{H-H}$ – the set of neighbouring nodes of Mongolian medicine h_i in the Mongolian medicine-Mongolian medicine graph.

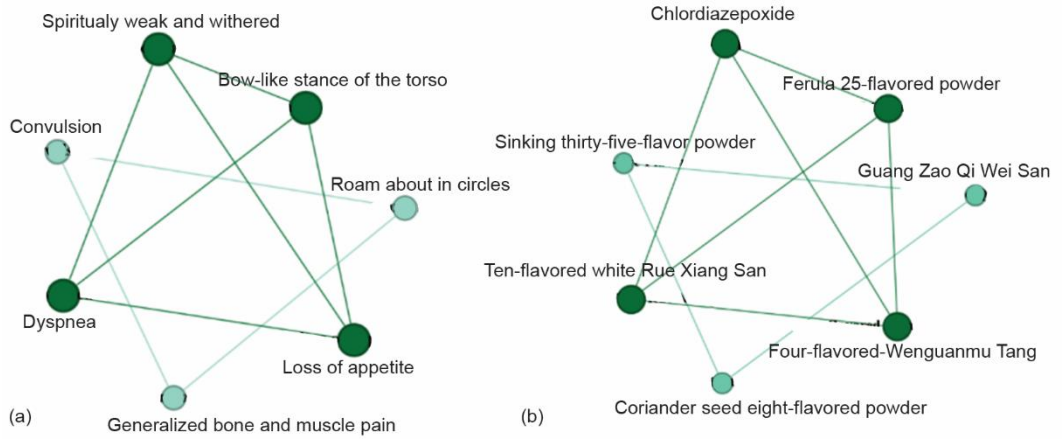

Figure 3. Schematic diagram of the symptom-symptom graph (a) and schematic diagram of the Mongolian medicine - Mongolian medicine graph (b)

The V_s and V_h, respectively, represent the attention weight parameters of the symptom-symptom graph and the Mongolian medicine-Mongolian medicine graph, while ω_{ij}^{S-S} and ω_{ij}^{H-H}, respectively, represent the weights of the edges between neighbouring nodes in the symptom-symptom graph and the Mongolian medicine-Mongolian medicine graph, defined:

$$\omega_{ij}^{S-S} = \frac{\exp(x_s x_k^T)}{\sum_{k' \in N_{s_i}^{S-S}} \exp(x_s x_{k'}^T)} \tag{3}$$

$$\omega_{ij}^{H-H} = \frac{\exp(x_h x_q^T)}{\sum_{q' \in N_{h_i}^{H-H}} \exp(x_h x_{q'}^T)} \tag{4}$$

To endow the model with the flexibility to model various types of spaces, besides distinguishing between symptom and Mongolian medicine nodes, it also shares the topological structure of the symptom-traditional Mongolian medicine bipartite graph, fig. 4. Parameters are trained separately, allowing the model to capture the distinct propagation patterns of the two types of nodes through different parameter values.

The core of the graph convolution operation is the process of message passing between nodes. Graph attention neural networks learn the importance of any two adjacent entities h_i and h_j in the graph by introducing an attention weight matrix. The presence of a relationship between nodes is determined based on the adjacency matrix:

$$A = (a_{ij}) + I_N \tag{5}$$

where $A = (a_{ij})$ denotes the adjacency matrix and a_{ij} is the $(i, j)^{\text{th}}$ entry of A.

The update mechanism of GAT is:

$$b_{s_i} = \sigma\left(\sum_{j \in N_i} \omega_{ij} x W\right) \tag{6}$$

$$b_{h_i} = \sigma\left(\sum_{j \in N_i} a_{ij} x W\right) \tag{7}$$

where b_s and b_h, respectively, are the first-layer vector representations of the symptom s_i and Mongolian medicine h_i. The N_i denotes the collection of neighbour nodes for node i. The ω_{ij} and a_{ij} represent the attention weight coefficient matrices between the symptom and Mongolian medicine nodes i and j, respectively, with W indicating the parameter matrix for the first layer, and σ denoting the non-linear activation function. The importance of the neighbour nodes j for the symptom and Mongolian medicine target node i is determined by the edge weights ω_{ij} and a_{ij}, as illustrated in fig. 5.

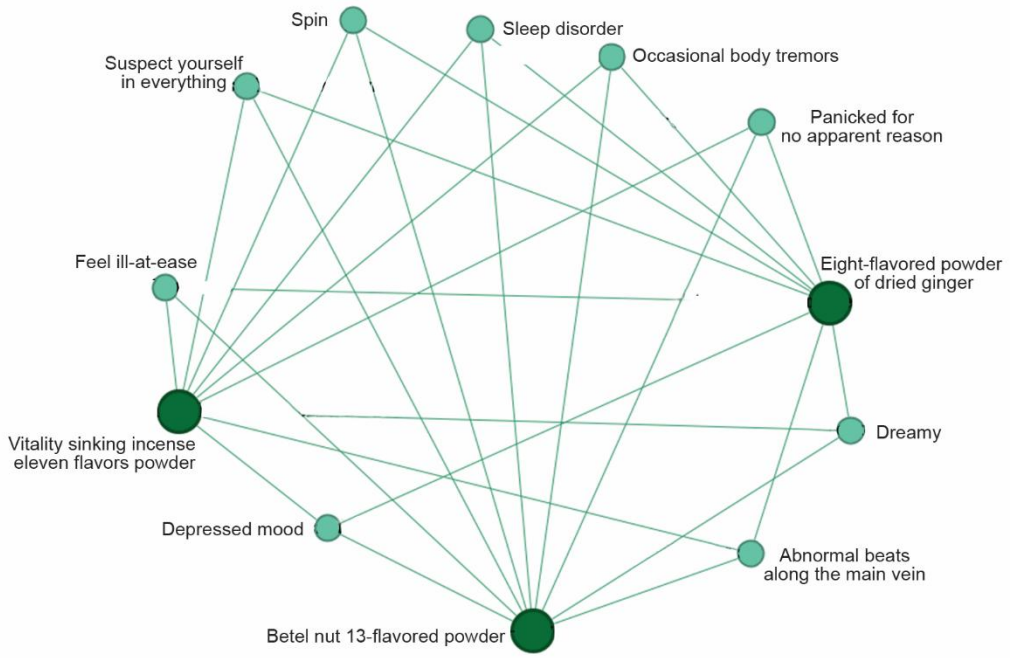

Figure 4. Symptoms - Mongolian medicine bipartite graph

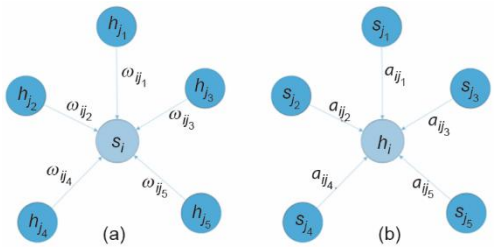

Figure 5. A bipartite GAT for symptoms (a) and Mongolian medicine (b)

The definitions of the weights ω_{ij} and a_{ij} are:

$$\omega_{ij} = \frac{\exp(x_s x_h^T)}{\sum_{h' \in N_i} \exp(x_s x_{h'}^T)} \tag{8}$$

$$a_{ij} = \frac{\exp(x_h x_s^T)}{\sum_{s' \in N_i} \exp(x_h x_{s'}^T)} \tag{9}$$

So far, two types of node representations have been introduced and the two types of node representations are fused to get the final node representation in the following way:

$$e_{s_i}^1 = \tanh(b_{s_i} + r_{s_i}) \tag{10}$$

$$e_{h_i}^1 = \tanh(b_{h_i} + r_{h_i}) \tag{11}$$

where the 1 in the upper right corner of $e_{s_i}^1$ and $e_{h_i}^1$ indicates the training result of the first layer, which can be trained in multiple layers, keeping the same idea with the training of the first layer. The model learns the final representation $e_{h_i}^*$ of each Mongolian medicine through multiple layers to obtain the matrix representation e_H composed of all Mongolian medicines. Similarly, it can also learn the representation $e_{s_i}^*$ of each symptom and synthesize the representations of multiple symptoms to complete the process of three-root diagnosis.

Given the set of symptoms, *s*, and the current candidate montage, *h*, the attention network is to generate an implicit evidence representation. Specifically:

– The vector representation of the set of symptoms is combined into a matrix $E_s \in R^{|s| \times d}$, and the candidate montage, *h*, is denoted as $e_h \in R^{1 \times d}$.
– Attention weight $a_h = soft\max[E_s(e_h)^T]$ is calculated.
– Finally, a non-linear multilayer perceptual machine is used to complete the whole etiological identification process:

$$e_s = ReLU\left[W^{mlp} \cdot Sum(E_s \odot a_h') + b^{mlp}\right] \tag{12}$$

where $a_h' \in R^{|s| \times d}$ is the representation matrix obtained by copying *d* copies of a_h, $\odot$ – the element-level product, e_s – the representation of the six-root relationship generalized from the set of synthesized and analysed symptoms, *s*, and *ReLU* – the non-linear activation function.

Prediction layer

The node embeddings, e_s and e_H, which are finally learned by the graph attention convolutional layer, are connected to make predictions about the relationship between symptom *s* and montage *h*:

$$f(s, H) = sigmoid[rowsum(e_s \odot e_H)] \tag{13}$$

where *f*(*s*, *H*) is a vector of predicted probabilities for all montages and *rowsum* – the operation of summing the row elements of the matrix.

Model training and inference

In the Mongolian medicine prescription recommendation task, given a set of symptom sets, the model generates a set of Mongolian medicines for treating the set of symptoms. For each prescription, the training objective is to minimize the gap between the real prescription and the recommended Mongolian medicine. This objective is similar to the multi-label classification task, so a multi-label classification loss function is considered. The goal of mul-

tilabel loss function is to minimize the error between the predicted value and the true label. In multilabel classification, the predicted value of each label can be a probability value and each label is independent of each other. Therefore, the multilabel loss function can be obtained by adding up the binary cross entropy loss for each label, *i.e.*:

$$Loss = \arg\min \sum_{(s,h')\in P} WMSE[h', f(s,H)] + \lambda_\Theta \left\|\Theta\right\|_2^2 \tag{14}$$

$$WMSE[h', f(s,H)] = \sum_{i=1}^{|H|} w_i [h_i' - f(s,H)_i]^2 \tag{15}$$

$$w_i = \frac{\max_k freq(k)}{freq(i)} \tag{16}$$

where h' denotes the true vector corresponding to the symptom s, weighted mean square error (*WMSE*) – the weighted root mean square error, and λ_Θ – the regularization coefficient.

Inference: after the model is trained, given the set of symptoms, s, the probability $f(s, H)$ of all montages is predicted and the top k are selected as the final recommended montage prescriptions.

Experiment and analysis

Experimental data

In this paper, the model is validated using the symptoms and Mongolian medicines provided in the Encyclopedia of Mongolian Medicine [14] as a dataset, which contains historically comprehensive data on Mongolian medicines. In this paper, the data is processed and a total of 725 prescriptions are extracted from it and divided into training set, validation set and test set.

Evaluating indicator

In this paper, two commonly used evaluation metrics for recommender systems, accuracy and recall, are used to assess the effectiveness of the model. The specific definitions are:

$$Precision = \frac{\left|top(s,k) \cap h\right|}{k} \tag{17}$$

$$Recall = \frac{\left|top(s,k) \cap h\right|}{\left|h\right|} \tag{18}$$

where $top(s, k)$ refers to the top k Mongolian medicines with the highest prediction scores given the symptom s. *Precision* refers to the probability that the top k montanas belong to the true remedy. *Recall* refers to the proportion of recommended remedies among the real remedies. The maximum length of the recommendation list corresponding to the previous evaluation metrics is limited to 20, as the vast majority of Monk's Remedies prescriptions fall within this range.

Experimental result

In this paper, two commonly used evaluation metrics for recommender systems, accuracy and recall, are used to assess the effectiveness of the model. Table 1 shows the comparison of the proposed model in this paper with the existing models.

Table 1. The experimental result

Model	*P*	*R*
Multi-label	10.83	29.72
seq2seq	26.03	13.52
HCKGETM	27.83	21.97
GAT	37.59	22.21

First, the GAT model proposed in this paper performs the best in terms of the accuracy of prescription recommendation. Specifically, GAT improves 15.2%, 11.56%, and 9.76% over Multi-label, seq2seq, and HCKGETM [14] in terms of accuracy, and –7.51%, 8.69%, and 0.24% over Multi-label, seq2seq, and HCKGETM in terms of recall.

Secondly, Multi-label has high recall but poor accuracy, the possible reason is that the length generated by the model is fixed, unlike the seq2seq model which can generate random length remedies. seq2seq model has improved accuracy but poor recall, considering the fact that seq2seq learns the node features of symptoms and mongooses only based on the randomly generated feature vectors continuously and did not take into account the relationship between them, thus reducing the effectiveness of the model.

Then, HCKGETM outperforms the Multi-label, seq2seq model, and this phenomenon indicates that the fusion of the knowledge graph into the topic model makes the semantics of the symptoms and remedies in the model obtain richer associations. Finally, the experimental results demonstrate the superiority of graph neural networks in modelling the higher-order semantic associations among Montana entities.

Conclusions

In this study, we introduced a novel approach for Mongolian medicine prescription recommendation, enabling automatic prediction of prescriptions from textual symptom descriptions. To support this research, a dataset of Mongolian medicine formulas was meticulously constructed using the Encyclopedia of Mongolian Medicine as a foundation. To this end, a GAT was employed to comprehensively learn the features of symptoms and Mongolian medicines. Given the intricate nature of Mongolian medicine theories and prescriptions, graphs were constructed to facilitate analysis. These graphs were created using two distinct approaches: an initial approach, which built graphs symptom-by-symptom, and a secondary approach, which built graphs symptom-by-Mongolian medicine and symptom-by-Mongolian medicine. These graphs were instrumental in capturing the complex semantic relationships among them, facilitating a more accurate representation of the characteristics of Mongolian medicine symptoms.

In the prediction layer, a MLP with a non-linear transformation was employed to simulate the process by which Mongolian medicine practitioners identify the etiology of diseases during treatment. This mechanism effectively filtered out suitable Mongolian medicine formulas. The experimental results on the constructed dataset demonstrated the superiority of the proposed method in comparison to existing models. The GAT based model demonstrated a noteworthy level of accuracy, with a percentage of 37.59% in terms of prescription recommendation, thereby surpassing the performance of other comparative models. This finding not only substantiates the efficacy of our model but also underscores the promise of graph neural networks in effectively processing intricate relationships within the domain of Mongolian medicine.

In the future, the incorporation of artificial intelligence, particularly graph neural network-based techniques, into the process of Mongolian medicine prescription recommendations will mark a substantial milestone in the field of medicine. This initiative presents novel prospects for the advancement of traditional Mongolian medicine. The automation of prescription-recommendation processes has the potential to assist Mongolian medicine practitioners in making more informed decisions, particularly in cases where human expertise is limited. Furthermore, it establishes a foundation for the development of innovative data-mining strategies in Mongolian medicine, facilitating the identification of latent patterns and relationships within extensive medical data sets. This initiative is expected to catalyze the modernization and development of Mongolian medicine while ensuring its long-term preservation. It is anticipated that this research will serve as a catalyst for further exploration in this domain, culminating in the development of more sophisticated and pragmatic applications that can benefit both the Mongolian medicine community and global healthcare as a whole.

References

[1] Wurchaih, H., *et al.*, Medicinal Wild Plants Used by the Mongol Herdsmen in Bairin Area of Inner Mongolia and its Comparative Study between TMM and TCM, *Ethnobiology Ethnomedicine*, *15* (2019), 32
[2] Zhou, B., *et al.*, Assessment of Pulmonary Infectious Disease Treatment with Mongolian Medicine Formulae Based on Data Mining, Network Pharmacology and Molecular Docking, *Chinese Herbal Medicines*, *14* (2022), 3, pp. 432-448
[3] Bo, A., *et al.*, Mechanism of Mongolian Medical Warm Acupuncture in Treating Insomnia by Regulating miR-101a in Rats with Insomnia, *Experimental and Therapeutic Medicine*, *14* (2017), 1, pp. 289-297
[4] Gula, A., History, Current Situation, and Future Development of Mongolian Medicine, *Journal of Traditional Chinese Medical Sciences*, *8* (2021), Suppl. 1, pp. S17-S21
[5] Xu, Y., *et al.*, Artificial Intelligence: A Powerful Paradigm for Scientific Research, *The Innovation*, *2* (2021), 100179
[6] He, J.-H., Transforming Frontiers: The Next Decade of Differential Equations and Control Processes, *Advances in Differential Equations and Control Processes*, *32* (2025), 1, 2589
[7] Li, H. B., *et al.*, Correlation Analysis Based on Neural Network Copula Function, *Thermal Science*, *27* (2023), 3A, pp. 2081-2089
[8] Zamfirache, I. A., *et al.*, Q - Learning, Policy Iteration and Actor - Critic Reinforcement Learning Combined with Metaheuristic Algorithms in Servo System Control, *Facta Universitatis, Series: Mechanical Engineering*, *21* (2023), 4, pp. 615-630
[9] Zhou, J., *et al.*, Graph Neural Networks: A Review of Methods and Applications, *AI Open*, *1* (2020), pp. 57-81
[10] Yu, B., *et al.*, Graph Neural Network Based Model for Multi-Behavior Session-Based Recommendation, *GeoInformatica*, *26* (2022), 2, pp. 429-447
[11] Zhang, M., *et al.*, Personalized Graph Neural Networks with Attention Mechanism for Session-Aware Recommendation, *IEEE Transactions on Knowledge and Data Engineering*, *34* (2020), 8, pp. 3946-3957
[12] ***, Editorial Board Encyclopedia of Mongolian Studies, *Encyclopedia of Mongolian Studies: Medicine Volume*, Inner Mongolia People's Publishing House, Hohhot, China, 2012
[13] Cai, M. Y., *et al.*, Bibliometric Analysis of Mongolian Medicine and Medicinal Materials in China since 2000, *Heliyon*, *10* (2024), e35499
[14] Ma, T., *et al.*, Kr-gcn: Knowledge-aware Reasoning with Graph Convolution Network for Explainable Recommendation, *ACM Transactions on Information Systems*, *41* (2023), 1, pp. 1-27

Paper submitted: October 10, 2024
Paper revised: March 21, 2025
Paper accepted: March 25, 2025
2026 Published by the Vinča Institute of Nuclear Sciences, Belgrade, Serbia.
This is an open access article distributed under the CC BY-NC-ND 4.0 terms and conditions.