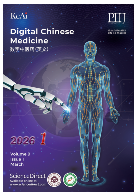
Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed



Knowledge graph-enhanced long-tail learning approach for traditional Chinese medicine syndrome differentiation

Weikang Kong, Chuanbiao Wen*, Yue Luo*

School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 611137, China

ARTICLE INFO

Article history
Received 09 October 2025
Accepted 10 January 2026
Available online 25 March 2026

Keywords
Syndrome differentiation
Medical knowledge graph
Graph neural networks
Long-tail learning
Data-efficient learning

ABSTRACT

Objective To address the dual challenges of long-tail distribution and feature sparsity in traditional Chinese medicine (TCM) syndrome differentiation within real clinical settings, we propose a data-efficient learning framework enhanced by knowledge graphs.

Methods We developed Agent-GNN, a three-stage decoupled learning framework, and validated it on the Traditional Chinese Medicine Syndrome Diagnosis (TCM-SD) dataset containing 54 152 clinical records across 148 syndrome categories. First, we constructed a comprehensive medical knowledge graph encoding the complete TCM reasoning system. Second, we proposed a Functional Patient Profiling (FPP) method that utilizes large language models (LLMs) combined with Graph Retrieval-Augmented Generation (RAG) to extract structured symptom-etiology-pathogenesis subgraphs from medical records. Third, we employed heterogeneous graph neural networks to learn structured combination patterns explicitly. We compared our method against multiple baselines including BERT, ZY-BERT, ZY-BERT + Know, GAT, and GPT-4 Few-shot, using macro-F1 score as the primary evaluation metric. Additionally, ablation experiments were conducted to validate the contribution of each key component to model performance.

Results Agent-GNN achieved an overall macro-F1 score of 72.4%, representing an 8.7 percentage points improvement over ZY-BERT + Know (63.7%), the strongest baseline among traditional methods. For long-tail syndromes with fewer than 10 samples, Agent-GNN reached a macro-F1 score of 58.6%, compared with 39.3% for ZY-BERT + Know and 41.2% for GPT-4 Few-shot, representing relative improvements of 49.2% and 42.2%, respectively. Ablation experiments confirmed that the explicit modeling of etiology-pathogenesis nodes contributed 12.4 percentage points to this enhanced long-tail syndrome performance.

Conclusion This study proposes Agent-GNN, a knowledge graph-enhanced framework that effectively addresses the long-tail distribution challenge in TCM syndrome differentiation. By explicitly modeling manifestation-mechanism-essence patterns through structured knowledge graphs, our approach achieves superior performance in data-scarce scenarios while providing interpretable reasoning paths for TCM intelligent diagnosis.

*Corresponding author: Yue Luo, E-mail: luoyue@cdutcm.edu.cn. Chuanbiao Wen, E-mail: wcb@cdutcm.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2026.02.005

Citation: KONG WK, WEN CB, LUO Y. Knowledge graph-enhanced long-tail learning approach for traditional Chinese medicine syndrome differentiation. Digital Chinese Medicine, 2026, 9(1): 57-67.

1 Introduction

Syndrome differentiation and treatment are the essence of traditional Chinese medicine (TCM). Syndrome identification, as a critical component, aims to comprehensively assess the body's current functional imbalance using the four diagnostic methods (inspection, auscultation and olfaction, inquiry, and palpation) [1]. Accurate syndrome identification is the prerequisite for subsequent principle establishment, prescription formulation, and medication administration. In recent years, with the rapid development of artificial intelligence, deep learning models have achieved remarkable progress in TCM syndrome identification tasks [2-4]. In particular, methods based on pre-trained language models such as BERT [5] and domain-specific medical language models [6] have set new performance records across multiple datasets. Recent advances in knowledge graph-enhanced large language models (LLMs) have also shown promising results in TCM syndrome differentiation [7, 8].

However, applying these idealized laboratory models to real, complex clinical environments still faces an insurmountable deployment gap. The first is the severe long-tail distribution problem. Medical long-tailed learning has emerged as a critical research area in recent years [9, 10]. In the Traditional Chinese Medicine Syndrome Diagnosis (TCM-SD) dataset [11] with 54 152 training samples, the top 10 syndromes account for 69.3% of samples, while 11 syndromes have fewer than 10 samples, with the minimum being only 7. This extreme imbalance severely limits the recognition capability of traditional end-to-end models for long-tail syndromes. Second, there is difficulty with cross-disease generalization. TCM theory emphasizes same syndrome in different diseases [1], meaning the same syndrome can appear in different diseases. Still, existing models struggle to capture this universal diagnostic logic that transcends disease manifestations. Third, the limitations of large language models. Although LLMs like GPT-4 [12] demonstrate remarkable few-shot reasoning capabilities, they suffer from hallucination outputs [13], unstable reasoning [14], and high costs in medical diagnostic scenarios.

The core of TCM syndrome differentiation lies in identifying structured combinations of symptoms and pathogenesis, which is a typical structured reasoning process from manifestation to essence [1]. For example, the diagnosis of “Qi deficiency and blood stasis syndrome” is based on the causal relationships between symptoms “fatigue, shortness of breath” and pathogenesis “Qi deficiency”, and between symptoms “chest tightness and stabbing pain” and pathogenesis “blood stasis”. Such structured reasoning paths can be explicitly modeled using knowledge graphs [15, 16].

To address the challenges of long-tail distribution and semantic gap in TCM diagnosis, this study aims to construct an interpretable yet data-efficient model. We propose a decoupled framework that combines structured knowledge extraction with graph neural reasoning. Specifically, the proposed Agent-GNN framework constructs a panoramic medical knowledge graph and employs a Functional Patient Profiling (FPP) method to extract structured patient subgraphs. By introducing etiology and pathogenesis as cross-syndrome shared semantic nodes, the framework enables effective feature transfer for rare syndromes. The model transforms black-box diagnosis into transparent multi-layer reasoning paths of “symptom → etiology → pathogenesis → syndrome”, providing auditable evidence consistent with TCM clinical logic.

2 Data and methods

This section proposes an intelligent TCM syndrome diagnosis framework (Agent-GNN) based on panoramic knowledge graphs and heterogeneous graph neural networks. The framework aims to address the low recognition rate of traditional models under a long-tail distribution by constructing a multi-layer heterogeneous knowledge graph covering "symptom-etiology-pathogenesis-syndrome" and by combining LLMs to extract FPP, thereby enabling practical reasoning and the identification of complex syndromes.

Figure 1 illustrates the three-phase architecture of the Agent-GNN framework. Stage 1 constructs the panoramic medical knowledge graph from the syndrome knowledge base and patient medical records. Stage 2 extracts FPP using Graph Retrieval-Augmented Generation (RAG) technology. Stage 3 performs syndrome reasoning through heterogeneous graph neural networks. The three stages form a complete closed loop from data to knowledge and from knowledge to reasoning.

2.1 Data source and preprocessing

This study employs the TCM-SD dataset, released on the Tianchi platform (https://tianchi.aliyun.com/dataset/137683), as the experimental benchmark. The data originate from real clinical electronic medical records and include complete diagnostic information, such as chief complaints, present illness history, physical examination, and diagnostic results after desensitization.

The dataset contains 54 152 samples, randomly split into training (43 180), validation (5 486), and test sets (5 486) at a ratio of 8 : 1 : 1, covering 148 TCM syndrome categories. To comprehensively evaluate model performance under an imbalanced distribution, we divide syndromes into three tiers: Head ($n \geqslant 100$, 43 categories), Mid ($10 \leqslant n < 100$, 94 categories), and Tail ($n < 10$, 11 categories). The data exhibit typical long-tail distribution characteristics (Figure 2). The data exhibit typical long-tail characteristics, with 43 high-frequency syndromes accounting for 91.8% of samples, while 11 low-frequency syndromes account for only 0.2%.

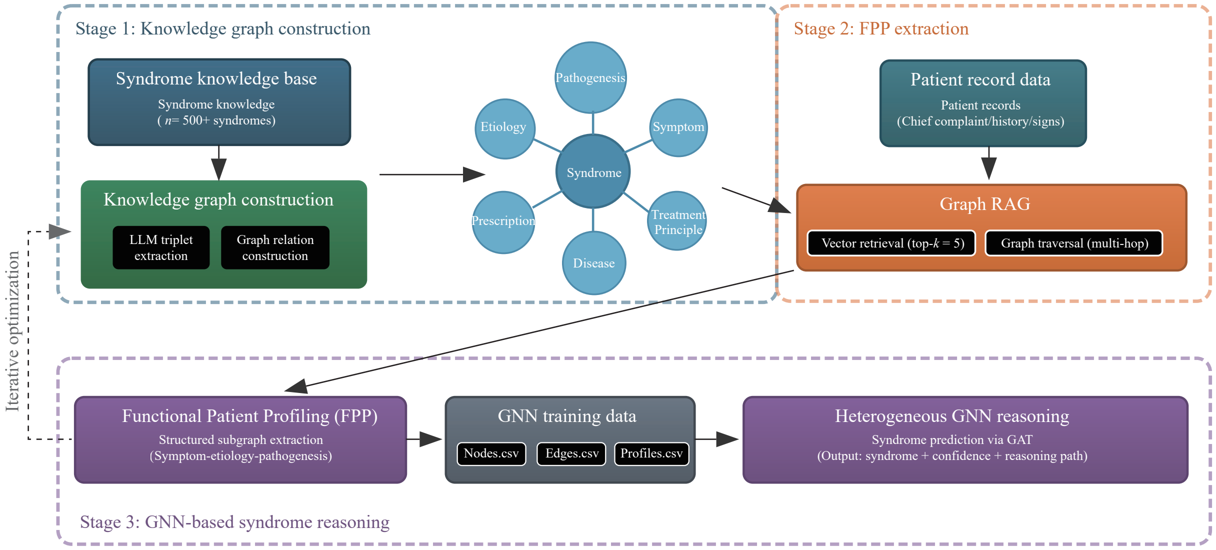

Stage 1: Knowledge graph construction
Syndrome knowledge base
Syndrome knowledge
(n= 500+ syndromes)
Knowledge graph construction
LLM triplet extraction
Graph relation construction
Pathogenesis
Etiology
Symptom
Syndrome
Prescription
Treatment Principle
Disease
Stage 2: FPP extraction
Patient record data
Patient records
(Chief complaint/history/signs)
Graph RAG
Vector retrieval (top-k = 5)
Graph traversal (multi-hop)
Iterative optimization
Functional Patient Profiling (FPP)
Structured subgraph extraction
(Symptom-etiology-pathogenesis)
GNN training data
Nodes.csv
Edges.csv
Profiles.csv
Heterogeneous GNN reasoning
Syndrome prediction via GAT
(Output: syndrome + confidence + reasoning path)
Stage 3: GNN-based syndrome reasoning

Figure 1 Overall system architecture of the Agent-GNN framework

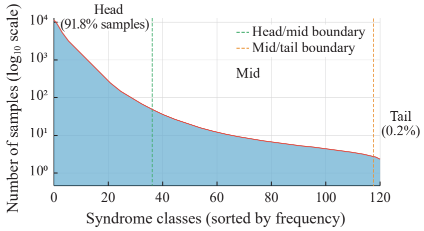


Figure 2 Long-tail distribution of syndrome samples in the TCM-SD dataset

To improve data quality, the following preprocessing was performed before experiments. First, invalid samples with missing chief complaints or diagnostic labels were removed. Second, abnormal samples with text length less than 50 or greater than 4 000 characters were filtered to reduce noise interference. Finally, semantically overlapping syndrome labels were normalized, for example, merging "damp-heat accumulation syndrome" and "damp-heat internal accumulation syndrome" into the standard term "damp-heat internal accumulation syndrome".

2.2 Panoramic prior medical knowledge graph construction

This study constructs a multi-level heterogeneous knowledge graph that systematically integrates core elements of TCM syndrome differentiation. The construction integrates classical TCM literature, the GB/T 15657-1995 standard for syndrome classification [17], and clinical prior knowledge from the TCM-SD dataset. Based on TCM syndrome differentiation theory, the graph defines seven entity types (symptoms, etiology, pathogenesis, syndromes, diseases, prescriptions, and treatments) and six types of directed semantic relationships. To ensure feature space stability, the study establishes a standardized mapping mechanism: mapping etiology to 52 standard categories, including "six external pathogens and seven emotions", and expanding pathogenesis to 138 standard nodes based on the "nineteen pathogenesis" theory.

Knowledge graph completion and quality evaluation have been identified as key factors affecting the performance of TCM intelligent systems [18]. Therefore, knowledge extraction adopts a semi-automated process. First, the GPT-4 model [12] is used to extract structured triplets from multi-source texts according to a predefined schema. To address the potential "hallucination" problem in LLMs [13], this study introduces a multi-agent expert simulation review mechanism based on the LlamaIndex framework. The system configures three agents with "chief physician" roles to perform independent logical verification and to extract fact-check results. Experiments show that the Fleiss' kappa coefficient among agents reaches 0.86 [19], indicating high consistency in knowledge review. Based on this high-confidence result, the system further performs entity alignment and deduplication to complete graph fusion.

The final panoramic medical knowledge graph contains 1 226 nodes and 3 512 directed edges. Topological statistics show that "symptom-to-syndrome" connections are the densest, accounting for 35.5% of total edges, followed by "pathogenesis-to-syndrome" at 25.4%. Notably, although pathogenesis nodes are relatively few, each pathogenesis node connects to an average of 6.5 syndrome nodes, forming a dense semantic bridge connecting clinical manifestations and pathological essence. The schema layer structure and a multi-hop reasoning example are shown in Figure 3. The graph enables GNNs to learn deep associations through multi-hop paths (e.g., symptom → etiology → pathogenesis → syndrome → treatment), thereby facilitating causal reasoning and knowledge transfer across syndromes.

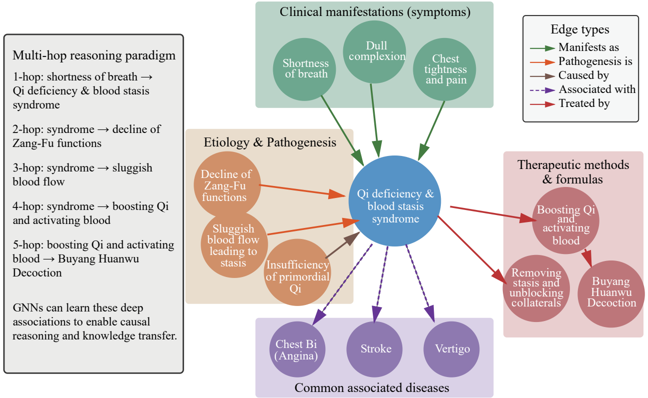

Multi-hop reasoning paradigm
1-hop: shortness of breath → Qi deficiency & blood stasis syndrome
2-hop: syndrome → decline of Zang-Fu functions
3-hop: syndrome → sluggish blood flow
4-hop: syndrome → boosting Qi and activating blood
5-hop: boosting Qi and activating blood → Buyang Huanwu Decoction
GNNs can learn these deep associations to enable causal reasoning and knowledge transfer.
Clinical manifestations (symptoms)
Shortness of breath
Dull complexion
Chest tightness and pain
Edge types
Manifests as
Pathogenesis is
Caused by
Associated with
Treated by
Etiology & Pathogenesis
Decline of Zang-Fu functions
Sluggish blood flow leading to stasis
Insufficiency of primordial Qi
Qi deficiency & blood stasis syndrome
Therapeutic methods & formulas
Boosting Qi and activating blood
Removing stasis and unblocking collaterals
Buyang Huanwu Decoction
Chest Bi (Angina)
Stroke
Vertigo
Common associated diseases

Figure 3 Structure of TCM syndrome knowledge graph and multi-hop reasoning paradigm

The figure illustrates the knowledge graph structure using "Qi deficiency and blood stasis syndrome" as an example. The graph contains seven entity types: 148 syndrome nodes, 624 symptom nodes, 52 etiology nodes, 138 pathogenesis nodes, 85 disease nodes, and 179 treatment/formula nodes (total: 1 226 nodes, 3 512 directed edges). Six types of semantic relationships connect these entities, distinguished by edge color and line style: "manifests as" (symptom → syndrome, green), "pathogenesis is" (etiology/pathogenesis → syndrome, orange), "caused by" (etiology → pathogenesis, brown), "associated with" (syndrome → disease, purple dashed line), "treated by method" (syndrome → treatment principle, red), and "treated by formula" (treatment principle → prescription, red). The left panel shows a multi-hop reasoning paradigm demonstrating how GNNs can learn deep associations through 5-hop paths to enable causal reasoning and knowledge transfer.

2.3 FPP knowledge graph extraction

2.3.1 Motivation and concept definition TCM syndrome classification tasks typically represent patient medical records as discrete symptom vectors (one-hot) or high-dimensional semantic vectors encoded through pre-trained models (such as BERT). However, this representation, based on surface semantics, ignores the deep pathological mechanisms underlying symptoms. For example, for a mixed hemorrhoid patient simultaneously presenting "anal swelling and pain" and "red tongue with yellow coating", traditional methods only treat them as independent input features, unable to capture the pathological reasoning path formed through "damp-heat internal accumulation" (etiology) and "damp-heat downward pour" (pathogenesis). This semantic gap leads to difficulty in effective generalization when the model encounters long-tail phenomena with scarce training samples, due to insufficient support for causal logic.

To address this, we propose FPP method. This profile maps unstructured medical record descriptions into structured subgraphs containing "symptom-etiology-pathogenesis-syndrome". This subgraph implies potential pathological states through knowledge graph constraints, providing transferable intermediate feature representations for subsequent models.

2.3.2 Graph RAG-enhanced construction process To accurately extract functional profiles, we design a Graph RAG strategy. The process includes three formalized steps: (i) symptom entity recognition and standardization, (ii) panoramic etiology-pathogenesis inference, and (iii) functional subgraph construction. A comparison between our Graph RAG approach and traditional methods is presented in Figure 4.

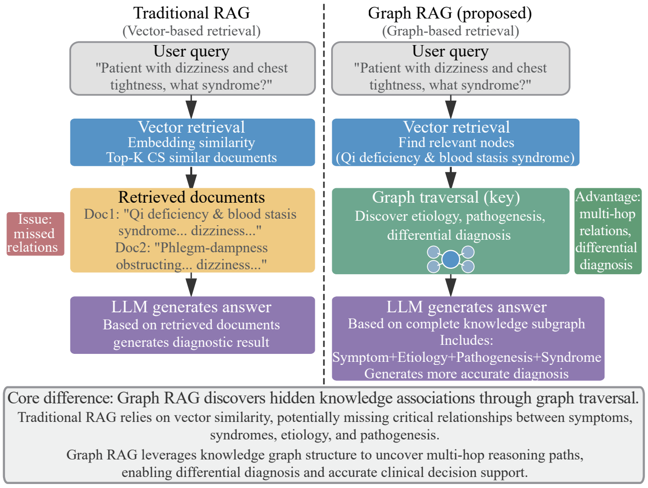

Traditional RAG
(Vector-based retrieval)
User query
"Patient with dizziness and chest tightness, what syndrome?"
Vector retrieval
Embedding similarity
Top-K CS similar documents
Issue: missed relations
Retrieved documents
Doc1: "Qi deficiency & blood stasis syndrome... dizziness..."
Doc2: "Phlegm-dampness obstructing... dizziness..."
LLM generates answer
Based on retrieved documents generates diagnostic result
Graph RAG (proposed)
(Graph-based retrieval)
User query
"Patient with dizziness and chest tightness, what syndrome?"
Vector retrieval
Find relevant nodes
(Qi deficiency & blood stasis syndrome)
Graph traversal (key)
Discover etiology, pathogenesis, differential diagnosis
Advantage: multi-hop relations, differential diagnosis
LLM generates answer
Based on complete knowledge subgraph
Includes:
Symptom+Etiology+Pathogenesis+Syndrome
Generates more accurate diagnosis
Core difference: Graph RAG discovers hidden knowledge associations through graph traversal.
Traditional RAG relies on vector similarity, potentially missing critical relationships between symptoms, syndromes, etiology, and pathogenesis.
Graph RAG leverages knowledge graph structure to uncover multi-hop reasoning paths, enabling differential diagnosis and accurate clinical decision support.

Figure 4 Comparison of Graph RAG working principle with traditional RAG

Left, traditional RAG directly retrieves text fragments. Right, Graph RAG retrieves structured knowledge through graph topology, ensuring semantic alignment.

(i) Symptom entity recognition and standardization. Given the patient's medical record text D (including chief complaint, and present illness history), the LLM first extracts the raw symptom set S_{raw}. To eliminate differences between colloquial descriptions and standard terminology, the system introduces a pre-trained encoder [20] to map symptoms into high-dimensional vectors. It performs vector similarity retrieval in the symptom node subspace V_{symptom} of the knowledge graph to obtain the best-matched standardized node S_{std}, as formulated in Equation (1):

$$S_{\text{std}} = \arg\max_{S\in V_{\text{symptom}}} \text{sim}\left[\text{Encoder}\left(S_{\text{raw}}\right),\text{Encoder}\left(S\right)\right] \tag{1}$$

This process ensures semantic alignment between input features and graph knowledge.

(ii) Panoramic etiology-pathogenesis inference. This is the core innovation of this module. Unlike traditional methods that focus only on symptoms themselves, this module leverages the structured prior of the knowledge graph to activate potential etiology and pathogenesis nodes connected to symptoms automatically. Since medical records often do not explicitly record pathogenesis, the system defines a retrieval function based on graph topology to infer the potential pathological state set $\{E, M\}$ that causes the current symptom combination, as shown in Equation (2):

$$E, M = \text{GraphRetrieve}\left(S_{\text{std}}, \text{KG}\right) \tag{2}$$

where E and M represent the retrieved etiology and pathogenesis sets, respectively, and KG denotes the knowledge graph.

For example, through graph path constraints, the system can establish explicit associations between “anal swelling and pain” and “damp-heat” pathogenesis, filling the reasoning gap from manifestation to essence.

(iii) Functional subgraph construction and generation. The system integrates activated symptom, etiology, and pathogenesis nodes and their associations into a structured graph fragment G_{sub}, which is re-injected into the LLM prompt as enhanced context. Guided by this high-quality graph knowledge, the LLM generates the final FPP by maximizing sequence probability, as formulated in Equation (3):

$$\text{FPP} = \text{LLM}_{\text{generate}}\left(D|G_{\text{sub}}\right) \tag{3}$$

where D represents the original medical record and G_{sub} denotes the constructed functional subgraph. This closed-loop feedback mechanism effectively reduces terminology ambiguity and suppresses LLM “hallucination” outputs, ensuring the generated profile has clinical logic consistency.

2.4 Heterogeneous graph neural network reasoning

The FPP is essentially a heterogeneous graph containing multiple node types (symptoms, etiology, pathogenesis, syndromes) and edge types. Traditional homogeneous graph neural networks [21] struggle to capture heterogeneous semantics across different node types effectively. Therefore, this study designs a Hierarchical Message Passing Architecture [22, 23], as shown in Figure 5A. This architecture is explicitly optimized for TCM logic "syndrome differentiation to seek cause, reviewing cause to determine treatment". It fundamentally alleviates the data sparsity problem in long-tail syndromes by leveraging a feature reuse mechanism for etiology-pathogenesis nodes.

(i) Node representation initialization: to introduce domain prior knowledge, we employ ZY-BERT [20] pre-trained on a large-scale TCM literature corpus to generate initial semantic vectors for all nodes. As shown on the left side of Figure 5A, for any node v, its initial representation is formed by concatenating BERT-encoded text features (768-dimensional) with learnable type embedding (128-dimensional):

$$h_v^{(0)} = \left[\text{BERT}_{768}\left(\text{text}_v\right) \,||\, \text{TypeEmbed}_{128}\left(\text{type}_v\right)\right] \tag{4}$$

where || denotes the vector concatenation operation. The BERT encoder outputs a 768-dimensional semantic vector, and the type embedding adds a 128-dimensional learnable vector specific to each node type (symptom, etiology, pathogenesis, or syndrome), ensuring that nodes of different types have differentiated initial representations in the embedding space.

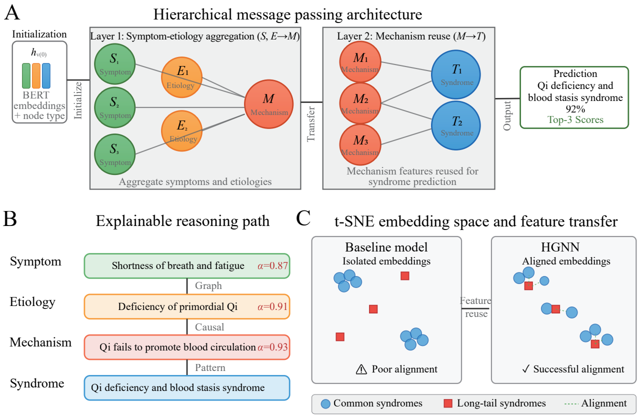

A
Hierarchical message passing architecture
Initialization
$h_{v(0)}$
BERT embeddings + node type
Initialize
Layer 1: Symptom-etiology aggregation (S, E→M)
S1 Symptom
S2 Symptom
S3 Symptom
E1 Etiology
E2 Etiology
M Mechanism
Aggregate symptoms and etiologies
Transfer
Layer 2: Mechanism reuse (M→T)
M1 Mechanism
M2 Mechanism
M3 Mechanism
T1 Syndrome
T2 Syndrome
Mechanism features reused for syndrome prediction
Output
Prediction
Qi deficiency and blood stasis syndrome
92%
Top-3 Scores
B
Explainable reasoning path
Symptom
Shortness of breath and fatigue α=0.87
Graph
Etiology
Deficiency of primordial Qi α=0.91
Causal
Mechanism
Qi fails to promote blood circulation α=0.93
Pattern
Syndrome
Qi deficiency and blood stasis syndrome
C
t-SNE embedding space and feature transfer
Baseline model
Isolated embeddings
⚠ Poor alignment
Feature reuse
HGNN
Aligned embeddings
✓ Successful alignment
● Common syndromes
■ Long-tail syndromes
····· Alignment

Figure 5 Hierarchical heterogeneous graph neural network (HGNN) reasoning architecture

A, two-stage message passing mechanism. Layer 1 aggregates symptoms (S) and etiologies (E) into mechanisms (M), and Layer 2 transfers mechanism features to syndromes (T). B, visualization of attention weights (α) demonstrating a complete 4-step interpretable reasoning path: $S \rightarrow E \rightarrow M \rightarrow T$. C, t-SNE visualization showing that the etiology-mechanism reuse strategy successfully aligns the syndromes (red squares) with semantically related common syndromes (blue circles).

(ii) First layer: syndrome differentiation to seek mechanism (S, $E \rightarrow M$). This layer aims to infer mechanisms (M) from symptoms (S) and etiologies (E), in line with TCM clinical thought: “syndrome differentiation to seek mechanism” (Figure 5B). For each pathogenesis node m in the functional profile, we aggregate associated symptom and etiology nodes:

$$h_m^{(1)} = \sigma\left(\sum_{v \in N_m^{S,\mathcal{E}}} \alpha_{mv} \cdot W_1 \cdot h_v^{(0)}\right) \tag{5}$$

where $N_m^{S,\mathcal{E}}$ represents the set of symptom and etiology nodes connected to the pathogenesis node m, α_{mv} is the attention weight automatically learned by the model reflecting the importance of node v to pathogenesis m, W_1 is the learnable transformation matrix of the first layer, and σ is the rectified linear unit (ReLU) activation function. The attention weight α_{mv} is calculated through the self-attention mechanism:

$$\alpha_{mv} = \text{softmax}\left(\text{LeakyReLU}\left(a^{\mathrm{T}} \cdot \left[W_1 \cdot h_v^{(0)} \,||\, W_1 \cdot h_m^{(0)}\right]\right)\right) \tag{6}$$

where a is the learnable attention parameter vector. This design allows the model to autonomously learn which symptom-etiology combinations are most indicative of a particular pathogenesis.

(iii) Second layer: determining syndrome based on mechanism and feature transfer ($M \rightarrow T$). This is the core module for solving long-tail problems. For the syndrome node t to be predicted, the second-layer update rule is:

$$h_t^{(2)} = \sigma\left(\sum\nolimits_{m \in N_t^M} \beta_{tm} \cdot W_2 \cdot h_m^{(1)}\right) \tag{7}$$

where N_t^M is the set of pathogenesis nodes connected to syndrome t, β_{tm} is the attention weight from pathogenesis to syndrome, and W_2 is the second layer transformation matrix. Similarly, β_{tm} is also calculated through the attention mechanism. The final syndrome prediction probability is obtained by projecting $h_t^{(2)}$ through a fully connected layer followed by softmax normalization.

(iv) Core mechanism: cross-syndrome feature transfer. Only pathogenesis nodes M activated in the patient profile participate in message passing. As shown in Figure 5C, this design constructs a "semantic bridge": even if a rare syndrome (such as "fluid retention in pericardium syndrome") has extremely few training samples ($n = 3$), as long as its associated pathogenesis nodes (such as "fluid retention") appear frequently in other common syndromes (such as phlegm-dampness obstruction syndrome, $n = 120$), the model can learn the embedding representations of these shared pathogenesis nodes and derive discriminative features for rare syndromes.

Through two-layer message passing, the model transfers learned pathogenesis feature embeddings from data-abundant to data-scarce syndromes. This is essentially a graph-structure-based transfer learning approach [24], learning general features of pathogenesis from data-abundant common syndromes and transferring them to data-scarce long-tail syndromes, thereby fundamentally alleviating the overfitting problem in long-tail syndromes and outperforming traditional end-to-end models that treat each syndrome independently.

2.5 Interpretable reasoning

For any prediction result, we can trace the attention weight distribution in the graph neural network [25] to construct a complete reasoning path. As shown in Figure 5B, for a patient predicted as "Qi deficiency and blood stasis syndrome", the system outputs a four-step reasoning chain consistent with TCM theory.

Initially, the system performs symptom recognition by identifying key symptoms from the medical record, specifically associating "fatigue and shortness of breath" with Qi deficiency and "chest tightness and stabbing pain" with blood stasis. Subsequently, based on the medical history, it infers the etiology node as "chronic disease". Through first-layer message passing, the model specifically activates two pathogenesis nodes: M_1 "Qi deficiency" (exhibiting high attention weight on fatigue symptoms) and M_2 "blood stasis" (showing high attention weight on chest pain). Finally, via second-layer aggregation, these pathogenesis nodes (M_1 and M_2) jointly point to the final diagnosis of "Qi deficiency and blood stasis syndrome". This multi-layer reasoning path not only provides explicit evidence chains with attention weights for clinical decision-making—offering auditable evidence consistent with the TCM "principle-method-prescription-medicine" system, but also fundamentally differs from the black-box nature of end-to-end deep learning models.

2.6 Clinical case visualization and analysis

To more intuitively demonstrate the effectiveness of the hierarchical message-passing architecture in handling complex clinical cases, we selected a typical sample (ID: 450031) from the TCM-SD dataset for visual analysis.

The patient was diagnosed with "diabetes (Xiaoke disease)" and "Qi deficiency and blood stasis syndrome". Although typical diabetes symptoms like "polydipsia" were present, the core complaint was a 15-year disease course aggravated by "numbness in the right lower limb". As shown in the visualization, the model first captured key node features through Medical-BERT initialization. The comprehensive reasoning process for this patient is illustrated in Figure 6.

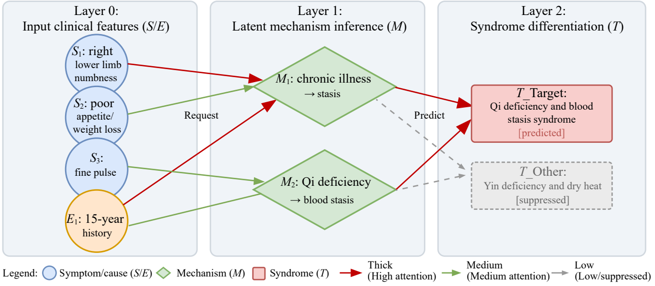

Layer 0:
Input clinical features (S/E)
Layer 1:
Latent mechanism inference (M)
Layer 2:
Syndrome differentiation (T)
S1: right lower limb numbness
S2: poor appetite/ weight loss
S3: fine pulse
E1: 15-year history
M1: chronic illness → stasis
M2: Qi deficiency → blood stasis
Request
Predict
T_Target: Qi deficiency and blood stasis syndrome [predicted]
T_Other: Yin deficiency and dry heat [suppressed]
Legend:
Symptom/cause (S/E)
Mechanism (M)
Syndrome (T)
Thick (High attention)
Medium (Medium attention)
Low (Low/suppressed)

Figure 6 Diabetes patient (ID 450031) reasoning process

Colors represent different node types. Arrow thickness indicates attention weight intensity. Layer 0 (input), the model identifies key features like "right lower limb numbness" and "15-year history". Layer 1 (mechanism inference), high attention weights (thick red arrows) activate latent mechanism nodes M_1 (chronic illness → stasis) and M_2 (Qi deficiency → blood stasis). Layer 2 (syndrome differentiation), these mechanisms jointly predict the target syndrome "Qi deficiency and blood stasis" with high confidence, while correctly suppressing the irrelevant "Yin deficiency & dry heat" syndrome (dotted line).

2.7 Baseline models

To comprehensively evaluate the effectiveness of the proposed Agent-GNN framework, we compared it against robust baselines across three categories: text-, graph-, and LLM- based methods. Each baseline represents a distinct technical paradigm for syndrome differentiation.

In the text-based category, BERT-Base serves as the fundamental semantic encoder, utilizing bert-base-Chinese to perform classification directly via [CLS] (classification) token embeddings. We also included ZY-BERT [11], a model pre-trained on massive TCM corpora using

Domain-Adaptive Pretraining (DAPT). Its knowledge-enhanced variant, ZY-BERT + Know, concatenates syndrome definitions from an external knowledge base with the medical record text at the input layer. Comparing these variants allows us to quantify the impact of introducing unstructured external knowledge versus our structured graph approach.

In the graph-based and LLM categories, we selected Graph Attention Network (GAT) [22] and GPT-4 as representative baselines. The GAT model constructs a heterogeneous topology using a star-shaped symptom graph, where symptom entities extracted via Jieba tokenization are connected to a virtual syndrome node. For the LLM baseline, we employed GPT-4 in a 3-shot learning setting, utilizing semantic retrieval to select the three most similar historical cases as context. This represents the upper bound of current general-purpose generative reasoning capabilities without specialized training.

2.8 Implementation details and training strategy

Experiments were implemented in PyTorch 2.11 on an NVIDIA GeForce RTX 5060 GPU (8 GB). For a fair comparison, all models utilized bert-base-chinese as the backbone text encoder with a maximum sequence length of 512 tokens. In the Agent-GNN framework, the top 3 standardized symptom and pathogenesis nodes were retrieved via vector similarity. The reasoning module featured 2 graph attention layers (256-dim, 8 heads) with a 0.4 dropout rate to prevent overfitting. The GAT baseline utilized a homogeneous symptom co-occurrence graph with matching depth and hyperparameters, while ZY-BERT + Know concatenated 128-token knowledge snippets with medical records.

A stratified 5-fold cross-validation strategy was adopted to ensure robust evaluation, maintaining consistent class distributions even for tail categories ($n < 10$). Optimization was performed using the AdamW optimizer with an initial learning rate of 2×10^{-5}, weight decay of 1×10^{-4}, and a batch size of 32. Training was limited to 50 epochs with early stopping (patience = 5) based on validation performance. Final reported metrics represent the macro F1-scores across all folds on the independent test set.

3 Experiments and results

Before discussing specific performance metrics, we first present the overall experimental pipeline. As shown in Figure 7, our systematic evaluation framework encompasses four integrated stages: from initial data collection and knowledge graph construction to Graph RAG prediction and final GNN model training. This pipeline ensures a rigorous assessment across multiple dimensions, including predictive accuracy, knowledge graph quality, and retrieval effectiveness.

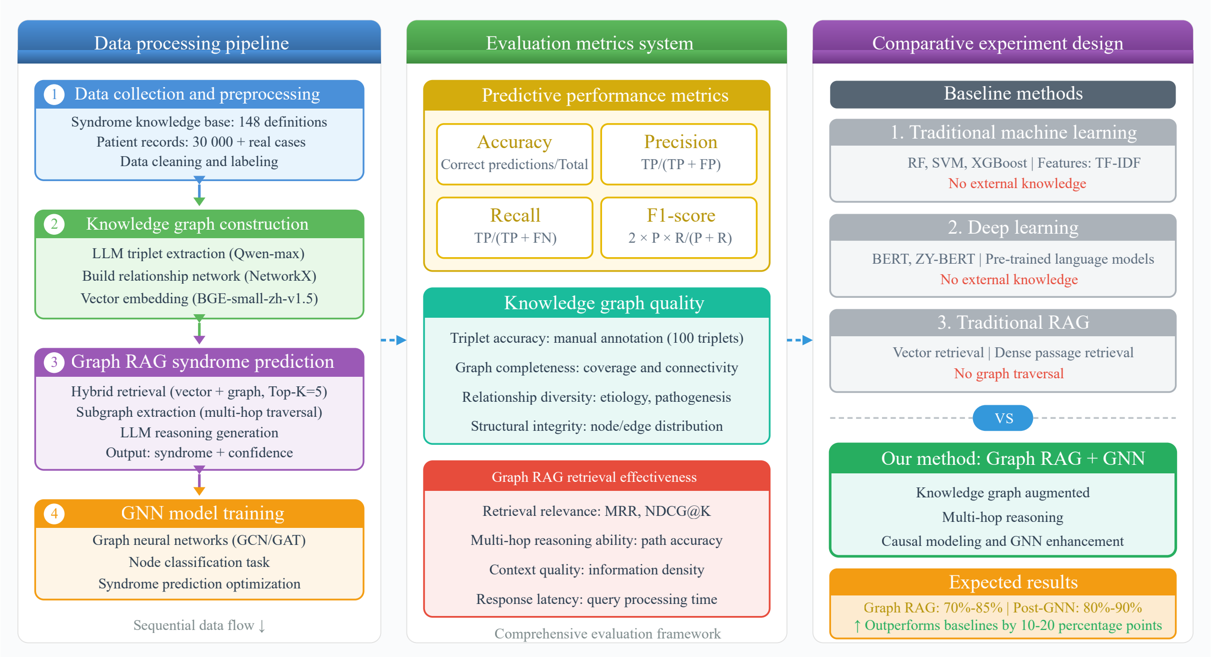

Data processing pipeline
1 Data collection and preprocessing
Syndrome knowledge base: 148 definitions
Patient records: 30 000 + real cases
Data cleaning and labeling
2 Knowledge graph construction
LLM triplet extraction (Qwen-max)
Build relationship network (NetworkX)
Vector embedding (BGE-small-zh-v1.5)
3 Graph RAG syndrome prediction
Hybrid retrieval (vector + graph, Top-K=5)
Subgraph extraction (multi-hop traversal)
LLM reasoning generation
Output: syndrome + confidence
4 GNN model training
Graph neural networks (GCN/GAT)
Node classification task
Syndrome prediction optimization
Sequential data flow ↓
Evaluation metrics system
Predictive performance metrics
Accuracy
Correct predictions/Total
Precision
TP/(TP + FP)
Recall
TP/(TP + FN)
F1-score
2 × P × R/(P + R)
Knowledge graph quality
Triplet accuracy: manual annotation (100 triplets)
Graph completeness: coverage and connectivity
Relationship diversity: etiology, pathogenesis
Structural integrity: node/edge distribution
Graph RAG retrieval effectiveness
Retrieval relevance: MRR, NDCG@K
Multi-hop reasoning ability: path accuracy
Context quality: information density
Response latency: query processing time
Comprehensive evaluation framework
Comparative experiment design
Baseline methods
1. Traditional machine learning
RF, SVM, XGBoost | Features: TF-IDF
No external knowledge
2. Deep learning
BERT, ZY-BERT | Pre-trained language models
No external knowledge
3. Traditional RAG
Vector retrieval | Dense passage retrieval
No graph traversal
VS
Our method: Graph RAG + GNN
Knowledge graph augmented
Multi-hop reasoning
Causal modeling and GNN enhancement
Expected results
Graph RAG: 70%-85% | Post-GNN: 80%-90%
↑ Outperforms baselines by 10-20 percentage points

Figure 7 Complete experimental pipeline and evaluation framework

TP, true positive. FP, false positive. FN, false negative. P, precision. R, recall. RF, Random Forest. SVM, Support Vector Machine. XGBoost, eXtreme Gradient Boosting. TF-IDF, Term Frequency-Inverse Document Frequency. BERT, Bidirectional Encoder Representations from Transformers. GCN, Graph Convolutional Network. GAT, Graph Attention Network. GNN, Graph Neural Network. RAG, Retrieval-Augmented Generation. MRR, Mean Reciprocal Rank. NDCG@K, Normalized Discounted Cumulative Gain at K.

3.1 Main performance comparison

Table 1 presents the overall performance of Agent-GNN compared with the baseline models on the independent test set.The results highlight three key observations. First, traditional deep learning methods exhibit a pronounced discrepancy between accuracy and macro-F1 (15% – 20% gap), indicating severe deficiency in identifying long-tail syndromes. Second, GPT-4 (52.8%) underperforms compared to domain-specific BERT (54.2%) in zero-shot settings, verifying that general LLMs lack domain-specific medical knowledge. Third, Agent-GNN achieves an overall macro-F1 score of 72.4%, surpassing the strongest baseline (ZY-BERT + Know) by 8.7 percentage points and GPT-4 Few-shot by 14.0 points, demonstrating the superiority of structured knowledge guidance.

Table 1 Overall performance comparison on the TCM-SD dataset

Model	Accuracy (%)	Macro-F1 (%)
BERT [5]	76.8	54.2
ZY-BERT [11]	79.4	58.9
ZY-BERT + Know	81.5	63.7
GAT [22] (w/o etiology-pathogenesis)	82.3	65.6
GPT-4 Few-shot [12]	76.5	58.4
Agent-GNN	85.2	72.4

w/o, without.

3.2 Long-tail performance analysis

To further analyze robustness, we evaluated performance across three frequency tiers: Head ($n \geqslant 100$), Mid ($10 \leqslant n < 100$), and Tail ($n < 10$). The Tail tier performance represents the core breakthrough. While traditional methods struggle with scarce samples (31.2% – 42.8%), Agent-GNN achieves a macro-F1 score of 58.6%. This represents a relative improvement of 49.2% over ZY-BERT + Know and 42.2% over GPT-4 Few-shot. For the 11 most extreme long-tail syndromes (7 – 10 samples), Agent-GNN maintains 47.2%, whereas GPT-4 drops to 22.5%, proving exceptional data efficiency (Table 2).

Table 2 Performance comparison across different frequency tiers

Model	Head ($n > 100$)	Mid ($10 \leqslant n < 100$)	Tail ($n < 10$)
BERT [5]	78.4	52.3	31.2
ZY-BERT [11]	81.7	56.8	34.6
ZY-BERT + Know	84.2	62.1	39.3
GAT [22]	85.6	64.7	42.8
GPT-4 Few-shot [12]	79.8	55.6	41.2
Agent-GNN	88.3	75.2	58.6

3.3 Ablation study

We conducted an ablation study to quantify the contribution of each module. As shown in Table 3, the inclusion of etiology-pathogenesis nodes yields the most significant impact, contributing 6.8% to overall performance and a remarkable 12.4% to Tail tier performance. This confirms that the intermediate reasoning layer is the critical component for solving the long-tail problem. These quantitative results provide a solid foundation for the mechanistic analysis in the next chapter.

Table 3 Ablation study of key model components

Model variant	Overall	Head	Mid	Tail
Agent-GNN	72.4	88.3	75.2	58.6
w/o Etiology-pathogenesis nodes	65.6	86.1	68.4	46.2
w/o Class-balanced loss	66.8	87.2	69.7	49.3
w/o Attention mechanism	68.9	86.7	72.1	52.4

w/o, without. All values are macro-F1 scores (%). Overall (n = 148) represents performance across all 148 syndromes. Head ($n \geqslant 100$), Mid ($10 \leqslant n < 100$), and Tail ($n < 10$) represent different frequency tiers.

4 Discussion

4.1 Mechanism verification: feature reuse and interpretability

The primary challenge in long-tail diagnosis is the model's inability to generalize from categories with "single-digit samples". In our Tail tier analysis ($n < 10$), purely data-driven baselines such as BERT struggle substantially, achieving a macro-F1 score of only 42.8%. This underperformance stems from overfitting to high-frequency surface text patterns rather than learning robust pathological features. In contrast, Agent-GNN maintains a macro-F1 score of 58.6% in the same tier, demonstrating a relative improvement of 36.9%. This robustness indicates that by decoupling syndromes into shared mechanisms, the model effectively breaks the reliance on large-scale labeled data for rare diseases.

To intuitively demonstrate this breakthrough, we analyze a typical case of "Qi-blood dysregulation syndrome" (Tail tier, 7 training samples). The patient presented with "white patches all over the body (2 years)" and diagnostic signs of "macula, red tongue, thin white coating, thin pulse". Traditional BERT baselines were misled by the high-frequency keyword "white patches", incorrectly associating it with "blood stasis syndrome" based on surface-level textual similarity. This failure highlights the limitation of statistical correlations in capturing deep semantic connections in sparse data scenarios.

Conversely, Agent-GNN successfully identified the correct syndrome by activating the latent "Qi-blood disharmony" pathogenesis node (Figure 8). This success is attributed to the "etiology-pathogenesis feature reuse" mechanism. Although the target syndrome is rare, its underlying pathogenesis (Qi-blood disharmony) is frequently observed in common Head tier syndromes, such as "Qi-blood deficiency". By constructing a semantic bridge ($S \rightarrow E \rightarrow M \rightarrow T$), Agent-GNN transfers robust feature representations from data-rich common diseases to data-scarce rare ones. This explicitly validates the effectiveness of structured prior knowledge in solving the few-shot problem, aligning with the theoretical foundations of transfer learning [24].

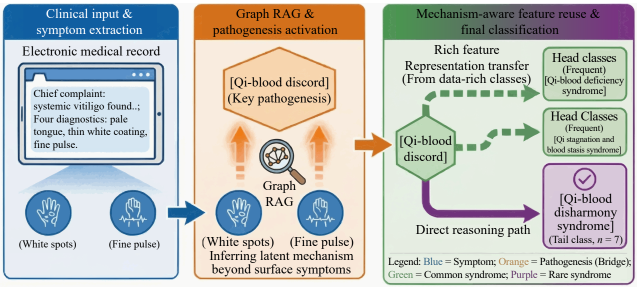

Clinical input & symptom extraction
Electronic medical record
Chief complaint: systemic vitiligo found..; Four diagnostics: pale tongue, thin white coating, fine pulse.
(White spots)
(Fine pulse)
Graph RAG & pathogenesis activation
[Qi-blood discord] (Key pathogenesis)
Graph RAG
(White spots)
(Fine pulse)
Inferring latent mechanism beyond surface symptoms
Mechanism-aware feature reuse & final classification
Rich feature Representation transfer (From data-rich classes)
Head classes (Frequent) [Qi-blood deficiency syndrome]
Head Classes (Frequent) [Qi stagnation and blood stasis syndrome]
[Qi-blood discord]
Direct reasoning path
[Qi-blood disharmony syndrome] (Tail class, n = 7)
Legend: Blue = Symptom; Orange = Pathogenesis (Bridge); Green = Common syndrome; Purple = Rare syndrome

Figure 8 Interpretable reasoning and feature reuse path for the syndrome "Qi-blood disharmony"

Furthermore, our approach aligns with recent trends in dual-channel knowledge attention mechanisms for TCM [7], while extending the capability to handle long-tail distributions.

4.2 Generalization capability and cross-domain transfer

Real-world clinical deployment frequently encounters data migration challenges across different hospitals and academic schools, a phenomenon termed "domain shift" [26]. The decoupled architecture of Agent-GNN offers a cost-effective solution for addressing this critical issue.

Unlike end-to-end black-box models that require retraining from scratch when confronted with new data distributions [27], our framework supports modular adaptation through its knowledge-guided design. When migrating to a new medical center or specialty, the system primarily requires a knowledge-driven update—adding specialty-specific pathogenesis nodes or adjusting edge weights in the Panoramic Knowledge Graph. Subsequently, the GNN layers only require lightweight fine-tuning rather than complete retraining.

This modular paradigm aligns with recent advances in parameter-efficient fine-tuning (PEFT) for large models [28], substantially promoting resource efficiency by reducing both computational costs and data requirements. Furthermore, the explicit separation of domain knowledge from learned representations facilitates interpretable model adaptation, a crucial consideration for clinical acceptance [29].

4.3 Limitations and future directions

Despite promising results, this study has several limitations that warrant discussion. First, the accuracy of FPP is contingent upon the reasoning capability of the underlying LLM; smaller models might miss subtle clinical cues embedded in unstructured medical records [30]. Recent study has demonstrated that LLM performance varies significantly across different medical domains and complexity levels [31], suggesting the need for careful model selection in clinical applications. Second, Graph RAG and multi-hop reasoning inevitably increase inference latency compared to simple BERT classifiers [32]. While this trade-off between accuracy and efficiency is acceptable for non-urgent diagnostic scenarios, real-time clinical applications may require optimization strategies.

To address these challenges, future work will focus on two primary directions. First, we will explore knowledge distillation techniques to compress the reasoning capabilities of large models into lightweight student networks [33], enabling deployment on resource-constrained devices while preserving diagnostic accuracy. Recent advances in medical model compression have demonstrated promising results in balancing model size and performance [34]. Second, we plan to validate the generalization capability of Agent-GNN on multi-center datasets to assess its robustness across diverse clinical settings and patient populations. Additionally, incorporating federated learning paradigms could further enhance privacy-preserving model adaptation across institutions.

5 Conclusion

This study addresses the critical challenge of long-tail distribution in TCM syndrome differentiation by proposing Agent-GNN. This novel framework bridges the gap in reasoning between clinical manifestations and the pathological essence. By constructing a "Panoramic Prior Knowledge Graph" and implementing a FPP mechanism, we not only transformed the "black-box" diagnosis process into a transparent reasoning path consistent with TCM theory but also successfully decoupled complex syndromes into shared pathogeneses, enabling effective feature reuse from common to rare diseases. Experimental results on the TCM-SD dataset demonstrate that our method surpasses strong baselines, including GPT-4, in overall accuracy and achieves a breakthrough improvement in identifying long-tail syndromes with extremely few samples. Specifically, the pathogenesis-feature reuse mechanism enabled a 49.2% relative improvement over data-driven baselines on rare syndromes. This study provides a new, interpretable, and highly generalizable paradigm for applying AI to TCM. It demonstrates that structured domain knowledge can fundamentally alleviate data scarcity in complex medical diagnosis, establishing a solid foundation for more robust, logic-driven clinical decision support systems.

Fundings

Sichuan TCM Culture Coordinated Development Research Center Project (2023XT131), National Key Science and Technology Project of China (2023ZD0509405), and National Natural Science Foundation of China (82174236).

Author contributions

Weikang Kong: conceptualization, methodology, software, investigation, formal analysis, and writing – original draft. Chuanbiao Wen and Yue Luo: supervision, project administration, funding acquisition, resources, and writing – review & editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

Chuanbiao Wen is an editorial board member for *Digital Chinese Medicine* and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

[1] ZHU WF. Diagnostics of Traditional Chinese Medicine. Beijing: China Press of Traditional Chinese Medicine, 2012.

[2] YANG L, CHEN Q, LIU J, et al. TCMCokg: a knowledge graph for traditional Chinese medicine constitution. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2021: 1429-1434.

[3] WANG Y, ZHOU K, GAO D, et al. A multi-task learning framework for traditional Chinese medicine syndrome differentiation. Artificial Intelligence in Medicine, 2021, 120: 102173.

[4] ZHANG Y, WANG X, LI Y, et al. A dual-channel knowledge-guided attention method for TCM syndrome differentiation. IEEE Access, 2021, 9: 34447–34459.

[5] DEVLIN J, CHANG MW, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[6] YANG X, CHEN AK, POURNEJATIAN N, et al. A large language model for electronic health records. NPJ Digital Medicine, 2022, 5: 194.

[7] LIU BT, HUANG HQ, LIU XM, et al. Dual-channel knowledge attention for traditional Chinese medicine syndrome differentiation. Scientific Reports, 2025, 15: 13487.

[8] YANG Y, MA T, LI R, et al. Jingfang: an LLM-based multi-agent system for precise medical consultation and syndrome differentiation in traditional Chinese medicine. arXiv, 2025. doi: 10.48550/arXiv.2502.04345.

[9] CHEN WX, YANG KX, YU ZW, et al. A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review, 2024, 57(6): 137.

[10] WU Z, GUO KH, LUO ET, et al. Medical long-tailed learning for imbalanced data: bibliometric analysis. Computer Methods and Programs in Biomedicine, 2024, 247: 108106.

[11] REN MC, HUANG HY, ZHOU YX, et al. TCM-SD: a benchmark for probing syndrome differentiation via natural language processing. Chinese Computational Linguistics. Cham: Springer International Publishing, 2022: 247–263.

[12] LEE P, BUBECK S, PETRO J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 2023, 388(13): 1233–1239.

[13] JI ZW, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12): 1–38.

[14] THIRUNAVUKARASU AJ, TING DSJ, ELANGOVAN K, et al. Large language models in medicine. Nature Medicine, 2023, 29(8): 1930–1940.

[15] XU H, LI T, CHEN J, et al. TCMBank: the largest traditional Chinese medicine database. Database, 2019, 2019: baz026.

[16] WU Y, ZHANG FL, YANG K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Research, 2019, 47(D1): D1110–D1117.

[17] National Technical Committee for Standardization of Traditional Chinese Medicine. GB/T 15657-1995 Classification and Codes of Diseases and Zheng of Traditional Chinese Medicine. Beijing: Standards Press of China, 1995.

[18] LIU C, LI Z, LI JM, et al. Research on traditional Chinese medicine: domain knowledge graph completion and quality evaluation. JMIR Medical Informatics, 2024, 12: e55090.

[19] FLEISS JL. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76(5): 378–382.

[20] GU Y, TINN R, CHENG H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 2022, 3(1): 1–23.

[21] KIPF TN, WELLING M. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations, 2017: 35–44.

[22] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks. International Conference on Learning Representations, 2018. doi: 10.17863/CAM.48429.

[23] WANG X, JI HY, SHI C, et al. Heterogeneous graph attention network. The World Wide Web Conference, 2019: 2022-2032.

[24] PAN SJ, YANG Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[26] ZHANG Y, LIU X, WANG J, et al. A survey of generalization and adaptation in medical imaging foundation models. Preprints, 2024. doi: 10.20944/preprints202507.0942.v1.

[27] MA QH, ZHANG J, QI L, et al. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 11642-11651.

[28] HAN Z, FU C, CHEN B, et al. Parameter-efficient fine-tuning for large models: a comprehensive survey. arXiv, 2024. doi: 10.48550/arXiv.2403.14608.

[29] AHMEDT-ARISTIZABAL D, ALI S, FOOKES C, et al. Graph neural networks in medical imaging: methods, applications, and future directions. IEEE Reviews in Biomedical Engineering, 2024, 17: 53–69.

[30] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge. Nature, 2023, 620(7972): 172–180.

[31] ERIKSEN AV, MÖLLER S, RYG J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI, 2024, 1(1): AIp2300031.

[32] WU J, ZHU Y, YANG Z, et al. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. arXiv, 2024. doi: 10.48550/arXiv.2408.04187.

[33] XU X, ZHOU Y, LIU Y, et al. A survey on knowledge distillation of large language models. arXiv, 2024. doi: 10.48550/arXiv.2402.13116.

[34] ZHOU YH, DU SY, LI HL, et al. Reprogramming distillation for medical foundation models. Medical Image Computing and Computer Assisted Intervention-MICCAI 2024. Cham: Springer, 2024: 533–543.

(Editor-in-Charge Jie Deng)

基于知识图谱增强的中医证候诊断长尾学习方法

孔伟康, 温川飙*, 罗悦*

成都中医药大学智能医学学院, 四川 成都 611137, 中国

【摘要】目的 针对真实临床环境中中医证候诊断面临的长尾分布与特征稀疏性双重挑战，本研究提出一种知识图谱增强的数据高效学习框架。**方法** 研究开发了Agent-GNN三阶段解耦学习框架，并在包含54 152条临床记录、涵盖148个证候类别的TCM-SD数据集上进行验证。首先，构建编码完整中医推理体系的全景医学知识图谱。其次，提出功能性患者画像（FPP）方法，利用大语言模型结合图检索增强生成技术从病历中提取结构化的症状–病因–病机子图。最后，采用异构图神经网络显式学习结构化组合模式。研究将Agent-GNN与多种基线模型进行对比，包括BERT、ZY-BERT、ZY-BERT + Know、GAT和GPT-4少样本学习，采用宏平均F1值作为主要评价指标。此外，通过消融实验验证各关键模块对模型性能的贡献。**结果** Agent-GNN实现了72.4%的整体宏平均F1值，较表现最优的传统方法ZY-BERT + Know（63.7%）提升8.7个百分点。对于样本量少于10的长尾证候，Agent-GNN的宏平均F1值达到58.6%，而ZY-BERT + Know和GPT-4少样本学习分别为39.3%和41.2%，相对提升幅度分别达49.2%和42.2%。消融实验证实，病因病机节点的显式建模为长尾证候性能提升贡献了12.4个百分点。**结论** 本研究提出的Agent-GNN知识图谱增强框架有效解决了中医证候诊断中的长尾分布难题。通过结构化知识图谱显式建模表象–机理–本质模式，该方法在数据稀缺场景下表现出更出色的性能，为中医智能诊断提供了可解释的推理路径。

【关键词】证候诊断；医学知识图谱；图神经网络；长尾学习；数据高效学习