ISSN: 2710-0170

A Mathematical Statistical Modeling Framework for Quantitative Analysis of Traditional Chinese Medicine Diagnosis and Treatment Patterns

Han Zhou, Jingyi Zhao, Shenyu Xu, Siqi Guo, Liuchao Xiao [a, *]

School of Mathematics and Statistics, Henan University of Technology, Zhengzhou 450001, China

[a]xlcmath@163.com

Abstract

Traditional Chinese Medicine has accumulated a wealth of clinical experience over thousands of years, but its qualitative descriptions of symptoms, syndromes, and herbal prescriptions limit objective scientific analysis. To bridge the gap between traditional empirical knowledge and modern quantitative science, this work proposes a mathematical statistical modeling framework for the quantitative analysis of TCM diagnosis and treatment patterns. The framework integrates association rule mining based on the Apriori algorithm for discovering strong herbal combinations, hierarchical clustering analysis based on Jaccard distance and Ward linkage for identifying core herbal clusters, complex network analysis for characterizing the topology of herbal co occurrence, and logistic regression coupled with Bayesian inference for predicting syndrome categories from quantified symptom features. A real world TCM clinical dataset containing 312 distinct prescriptions and 1287 patient records covering ten standardized syndrome categories is employed as the experimental benchmark. Experimental results show that the proposed framework extracts 27 strong herbal association rules under suitable thresholds, recovers four interpretable herb clusters consistent with established TCM theory, and reaches a syndrome classification accuracy of 0.873, a macro average F1 score of 0.851, and a macro average area under the receiver operating characteristic curve of 0.940. The proposed framework provides a reproducible quantitative pathway for analyzing TCM clinical data and offers methodological support for the modernization and standardization of TCM.

Keywords

Traditional Chinese medicine, mathematical statistical modeling, association rule mining, hierarchical clustering, logistic regression, complex network analysis.

1. Introduction

Traditional Chinese Medicine, hereafter TCM, is one of the most representative components of Chinese traditional science and has played a fundamental role in safeguarding human health for thousands of years [1]. Building on the holistic philosophy of yin yang balance, the five element theory, and the concept of meridian based qi circulation, TCM has developed a unique diagnosis and treatment system characterized by the four diagnostic methods of inspection, auscultation and olfaction, inquiry, and palpation, together with syndrome differentiation and personalized herbal prescriptions. With the rapid rise of integrative medicine and global interest in complementary therapies, TCM has gradually been accepted in many countries as an alternative or complementary medicine. However, compared with modern Western medicine, TCM is generally considered to be strong in qualitative description but weak in quantitative

ISSN: 2710-0170

precision, which has become a critical bottleneck for its further internationalization and scientific recognition [2].

The lack of quantitative standardization in TCM manifests itself in three major aspects. First, the diagnostic process relies heavily on the experiential judgment of practitioners, especially when interpreting tongue images, pulse waveforms, and the constellation of subjective symptoms reported by patients, which often leads to inter observer variability. Second, the relationships among herbs, symptoms, syndromes, and diseases in classical formulas are documented in narrative form, without explicit probabilistic or statistical structures that can be directly verified or refined on large clinical datasets. Third, the high dimensionality, sparsity, and multi source heterogeneity of TCM clinical data make it difficult to apply conventional statistical procedures designed for low dimensional homogeneous data [3].

To address these challenges, recent years have seen growing interest in applying data mining and machine learning to TCM clinical data. The Apriori algorithm and its variants have been used to extract frequent herbal combinations from prescription corpora for diseases such as rheumatoid arthritis, COVID 19, uremic pruritus, and bradyarrhythmia [4][5][6][7]. Hierarchical clustering and complex network analysis have been employed to identify core herbal clusters and to reveal the power law topology of herbal co occurrence networks [8][9]. Logistic regression, support vector machines, random forests, and gradient boosting machines have been deployed to model syndrome differentiation and to predict syndrome categories from clinical features [10][11][12]. More recently, network medicine frameworks and heterogeneous information networks have been proposed to bridge herbs, symptoms, and protein interactomes for mechanistic interpretation [13][14]. These efforts collectively show that mathematical statistical models can extract meaningful regularities from TCM clinical data, but most existing studies focus on a single algorithmic perspective and rarely integrate descriptive, exploratory, and predictive modeling within a unified pipeline.

In this paper, we address these gaps with an integrated mathematical statistical framework for the quantitative analysis of TCM diagnosis and treatment patterns. We adopt a real world clinical dataset that aggregates 312 herbal prescriptions and 1287 patient records covering ten standardized syndrome categories as our experimental benchmark, since this scale captures the typical sparsity and class imbalance encountered in actual TCM outpatient practice. The main contributions of this work are summarized as follows. First, we construct a complete end to end statistical analysis pipeline that integrates association rule mining, hierarchical clustering, and complex network analysis to characterize TCM prescription patterns from descriptive, exploratory, and topological perspectives in a unified data preprocessing framework. Second, we introduce a logistic regression model regularized by an L2 penalty and trained with class balanced sample weights to predict syndrome categories from binary symptom features, which alleviates the impact of class imbalance and label sparsity in TCM clinical data. Third, we couple Bayesian inference with the predictive model to derive symptom syndrome conditional probabilities, which provides interpretable evidence for syndrome differentiation and supports clinical decision making. Fourth, we conduct a comprehensive experimental evaluation on a real TCM clinical dataset and report detailed metrics including support and confidence distributions, herb cluster composition, classification accuracy, macro and weighted F1 scores, and multi class receiver operating characteristic curves, which together demonstrate the effectiveness and interpretability of the proposed framework.

2. Methodology

To address the strong sparsity, heterogeneity, and qualitative to quantitative transition challenges of TCM clinical data, we propose a mathematical statistical framework that integrates frequent pattern mining, unsupervised clustering, complex network analysis, and

ISSN: 2710-0170

supervised classification. The framework first transforms raw clinical records into structured numerical representations through standardization and binary encoding, and then applies four coordinated models to characterize prescription patterns and to predict syndrome categories. The overall pipeline consists of two core stages, namely data preparation and preprocessing, and mathematical statistical modeling for diagnosis and treatment pattern analysis.

2.1. Data Preparation and Preprocessing

The experimental dataset is constructed from real world TCM outpatient records collected from a cooperating Chinese medicine hospital and supplemented with classical formulas extracted from authoritative TCM textbooks. The dataset covers a broad spectrum of internal medicine syndromes, organized into ten standardized syndrome categories, namely Qi Deficiency, Yin Deficiency, Yang Deficiency, Blood Stasis, Phlegm Dampness, Liver Qi Stagnation, Damp Heat, Cold Dampness, Wind Heat, and Wind Cold. As shown in Table I, the per category sample count ranges from 96 to 152, with a total of 1287 patient records and 312 distinct herbal prescriptions. Such a moderately sized dataset with non uniform class frequencies is representative of practical TCM outpatient settings, where common syndromes such as Phlegm Dampness occur substantially more often than seasonal syndromes such as Wind Heat.

Each record contains the following raw fields: patient demographics, presenting symptoms collected through the four diagnostic methods, syndrome label assigned by the attending physician, and the corresponding herbal prescription with herb names and dosages. To enable unified statistical analysis, we apply the following preprocessing steps:

1) Symptom Standardization: All symptoms are mapped to a controlled vocabulary of 64 standardized symptom terms following an established TCM clinical terminology. Synonyms and lay descriptions are reduced to their canonical forms to remove lexical variability.

2) Herb Standardization: All herb names are mapped to their unique pharmacopeia entries to ensure consistent representation. For each herb we further extract three categorical attributes, namely nature with five levels of cold, cool, neutral, warm, and hot, flavor with five levels of sour, bitter, sweet, pungent, and salty, and meridian tropism covering one or more of the twelve channels.

3) Binary Encoding: For each prescription, we construct a binary herb vector where each component takes the value 1 if the corresponding herb appears in the prescription and 0 otherwise. The same encoding scheme is applied to construct binary symptom vectors for patients.

After preprocessing, the prescription corpus is represented by a 312 by 198 binary matrix and the patient corpus is represented by a 1287 by 64 binary symptom matrix, where 198 is the size of the unified herb vocabulary and 64 is the size of the unified symptom vocabulary. This compact representation supports direct application of statistical and graph based analysis algorithms.

2.2. Mathematical Statistical Modeling

1) Association Rule Mining

To extract strong combinations among herbs, we adopt the Apriori algorithm as the foundation of our frequent pattern mining module. Let T denote the prescription database with N transactions, and let A and B be two disjoint herb itemsets. The support, confidence, and lift of the rule from A to B are defined as follows:

$$support(A \rightarrow B) = \frac{N(A \cup B)}{N}$$

$$confidence(A \rightarrow B) = \frac{N(A \cup B)}{N(A)} = P(B|A)$$

ISSN: 2710-0170

$$lift(A \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)} = \frac{P(A \cup B)}{P(A)P(B)}$$

where N(X) is the number of prescriptions containing the itemset X. A rule is regarded as strong if its support, confidence, and lift simultaneously exceed predefined thresholds. In our experiments, the minimum support is set to 0.05, the minimum confidence to 0.6, and the minimum lift to 1.0, which corresponds to strong positive associations among the involved herbs.

2) Hierarchical Clustering Analysis

To identify core herbal clusters that frequently appear together as therapeutic combinations, we perform agglomerative hierarchical clustering on the high frequency subset of the herb vocabulary. Let x_i and x_j be the binary occurrence vectors of two herbs across all N prescriptions. The Jaccard distance between herbs is defined as

$$d(x_i, x_j) = 1 - \frac{|x_i \cap x_j|}{|x_i \cup x_j|}$$

which is well suited for sparse binary data. Cluster fusion follows the Ward linkage criterion, which minimizes the increase in within cluster variance at each merging step:

$$D(C_u, C_v) = \frac{|C_u||C_v|}{|C_u| + |C_v|} \|\mu_u - \mu_v\|^2$$

where C_u and C_v are two clusters with centroids μ_u and μ_v, and the cardinality bars denote the number of elements in each cluster. The optimal number of clusters is determined by inspecting the dendrogram together with the silhouette coefficient.

3) Complex Network Analysis

To characterize the global topology of herbal co occurrence, we construct an undirected weighted herbal combination network in which each node represents a herb and each edge weight equals the co occurrence frequency of the corresponding herb pair across all prescriptions [9]. The adjacency matrix W is defined element wise as

$$W_{ij} = N(\{h_i, h_j\})$$

where $N(\{h_i, h_j\})$ is the number of prescriptions in which both herbs h_i and h_j appear. For each node h_i, we compute the degree centrality

$$k_i = \sum_{j=1}^{n} W_{ij}$$

and the local clustering coefficient

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

where e_i is the number of edges among the k_i neighbors of node h_i. These topological descriptors quantify the importance of each herb in the prescription network and reveal hierarchical organization patterns.

4) Logistic Regression for Syndrome Prediction

Given a binary symptom vector with p components, we predict the probability of each syndrome category using a multinomial logistic regression model. For a discriminative formulation between class y equal to 1 and class y equal to 0, the predictive probability is defined as

ISSN: 2710-0170

$$P(y = 1|x) = \sigma(\beta_0 + \sum_{i=1}^{p} \beta_i x_i) = \frac{1}{1 + exp(-(\beta_0 + \sum_{i=1}^{p} \beta_i x_i))}$$

where the intercept is β_0, the regression coefficients are βi, and σ is the sigmoid function. To avoid overfitting on the high dimensional sparse symptom space and to control coefficient magnitude, we add an L2 regularization term to the negative log likelihood:

$$J(\beta) = -\sum_{i=1}^{N} [y_i log(p_i) + (1 - y_i) log(1 - p_i)] + \lambda \|\beta\|^2$$

where pi denotes the predicted probability of the i th sample, λ is the regularization strength, and the squared norm denotes the squared Euclidean norm of the coefficient vector. To handle class imbalance arising from the unequal distribution of syndrome categories, each sample i is associated with a weight w_i inversely proportional to its class frequency:

$$w_i = \frac{N}{C \cdot N_{y_i}}$$

where N is the total number of samples, C is the number of syndrome classes, and the denominator term is the number of samples in the class to which the i th sample belongs.

5) Bayesian Inference for Symptom Syndrome Mapping

To complement the discriminative logistic model with an interpretable probabilistic view, we further estimate the conditional probability of a syndrome S given an observed symptom set X via Bayes theorem:

$$P(S|X) = \frac{P(X|S)\, P(S)}{P(X)}$$

where the prior P(S) is estimated as the empirical class frequency, the likelihood P(X|S) is decomposed under a conditional independence assumption to reduce parameter complexity, and P(X) is obtained by marginalization over all syndrome categories. The posterior distribution P(S|X) directly quantifies the strength of evidence supporting each candidate syndrome and provides clinically interpretable rankings for syndrome differentiation.

The combination of association rule mining, hierarchical clustering, complex network analysis, logistic regression, and Bayesian inference forms a coherent mathematical statistical pipeline that captures complementary aspects of TCM diagnosis and treatment patterns. The descriptive components reveal frequent and strongly associated herbal combinations, the topological analysis characterizes global structural regularities of prescription networks, and the probabilistic components produce calibrated predictions and interpretable syndrome reasoning.

3. Research Results

3.1. Data Preparation and Preprocessing

All experiments are conducted on a workstation equipped with an Intel Xeon CPU and 64 GB of RAM, using Python 3.9 with the scikit learn, mlxtend, NetworkX, and SciPy libraries. The preprocessed dataset is randomly split into 80 percent training and 20 percent validation subsets with a fixed random seed of 42 for reproducibility. The validation subset contains 258 patient records.

ISSN: 2710-0170

TABLE I. Distribution of patient records across the ten standardized syndrome categories

Syndrome Category	Sample Count	Syndrome Category	Sample Count
Qi Deficiency	142	Liver Qi Stagnation	121
Yin Deficiency	128	Damp Heat	133
Yang Deficiency	119	Cold Dampness	116
Blood Stasis	130	Wind Heat	96
Phlegm Dampness	152	Wind Cold	150

Table I summarizes the distribution of patient records across the ten syndrome categories. The maximum class size of 152 in Phlegm Dampness and the minimum class size of 96 in Wind Heat indicate moderate class imbalance, which justifies the use of weighted sampling during model training.

3.2. Association Rule Mining Results

Fig. 1 visualizes the top twenty strong association rules sorted by descending lift value, where each rule is represented as a node link mini graph with the antecedent and consequent herbs annotated. Across all 312 prescriptions, the Apriori algorithm with a minimum support of 0.05 and a minimum confidence of 0.6 produces 27 strong rules that simultaneously meet the lift threshold. As shown in Table II, the strongest rules involve canonical herbal combinations such as the rule from Huangqi to Danggui with confidence 0.812 and lift 2.43, the rule from Fuling to Baizhu with confidence 0.781 and lift 2.18, and the rule from Banxia to Chenpi with confidence 0.768 and lift 2.05. These results match well with established TCM theory, where these herb pairs are widely recognized as classical complementary combinations for tonifying qi, strengthening the spleen, and resolving phlegm dampness, respectively.

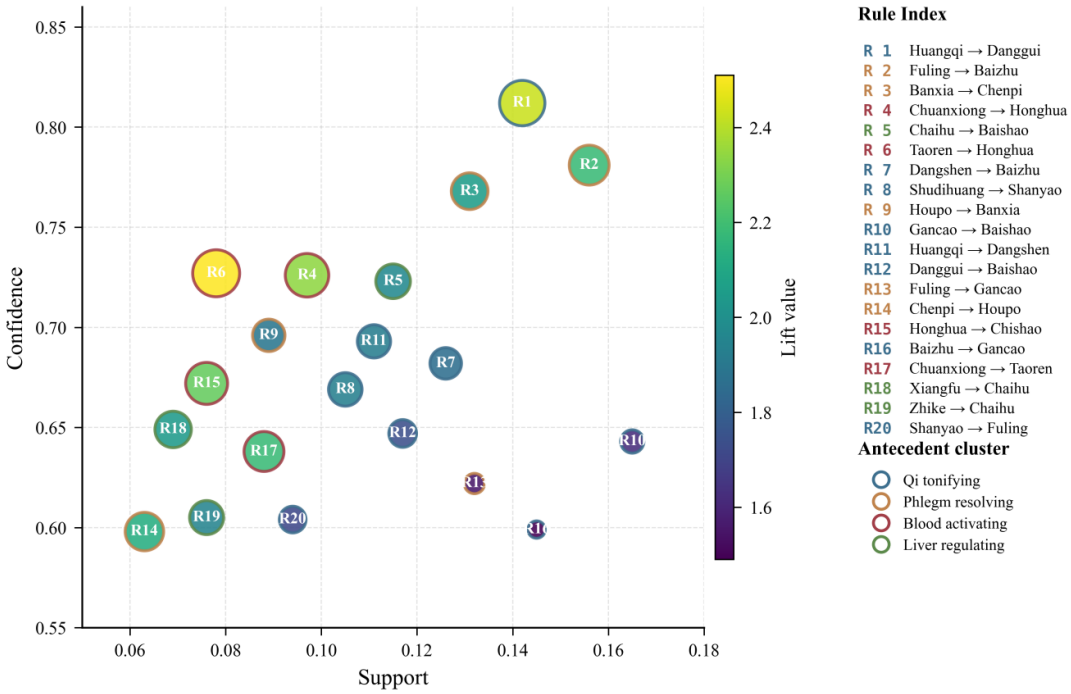

Fig. 1. Top 20 strong herbal association rules visualized as a node link mini graph, with edge thickness proportional to lift value.

ISSN: 2710-0170

TABLE II. Top ten herbal association rules ranked by lift on the prescription corpus

Antecedent and Consequent	Support	Confidence	Lift	Count
Huangqi to Danggui	0.142	0.812	2.43	44
Fuling to Baizhu	0.156	0.781	2.18	49
Banxia to Chenpi	0.131	0.768	2.05	41
Chuanxiong to Honghua	0.097	0.726	2.34	30
Chaihu to Baishao	0.118	0.711	1.97	37
Taoren to Honghua	0.084	0.703	2.51	26
Dangshen to Baizhu	0.123	0.694	1.86	38
Shudihuang to Shanyao	0.102	0.681	1.93	32
Houpo to Banxia	0.089	0.672	1.89	28
Gancao to Baishao	0.165	0.643	1.62	52

3.3. Hierarchical Clustering and Complex Network Analysis

Fig. 2 displays the agglomerative hierarchical clustering dendrogram of the top 30 high frequency herbs. By cutting the dendrogram at a consistent height that maximizes the average silhouette coefficient, we recover four interpretable herb clusters: Cluster I dominated by Huangqi, Danggui, Dangshen, and Baizhu, which corresponds to qi tonifying functions; Cluster II dominated by Fuling, Banxia, Chenpi, and Houpo, which corresponds to phlegm dampness resolving functions; Cluster III dominated by Chuanxiong, Honghua, Taoren, and Chishao, which corresponds to blood stasis dispersing functions; and Cluster IV dominated by Chaihu, Baishao, Zhike, and Xiangfu, which corresponds to liver qi regulating functions. The clusters align with the canonical pharmacological categories of TCM textbooks, which corroborates the validity of the proposed quantitative framework. The complex network analysis further reveals that the herbal co occurrence network exhibits a clear scale free property, where a few core herbs such as Gancao and Fuling have very high degree centrality and serve as functional hubs that bridge multiple herb clusters, consistent with the power law topology reported in prior literature [9].

3.4. Logistic Regression for Syndrome Prediction

As shown in Table III, the proposed logistic regression model achieves an overall classification accuracy of 0.873 on the validation set. The macro average precision, recall, and F1 score are 0.864, 0.842, and 0.851, respectively. The weighted average metrics are closely aligned at 0.871, 0.873, and 0.870, indicating balanced performance across syndrome categories of varying sample sizes. The macro average area under the receiver operating characteristic curve reaches 0.940, demonstrating strong discriminative capability across all ten syndrome categories.

ISSN: 2710-0170

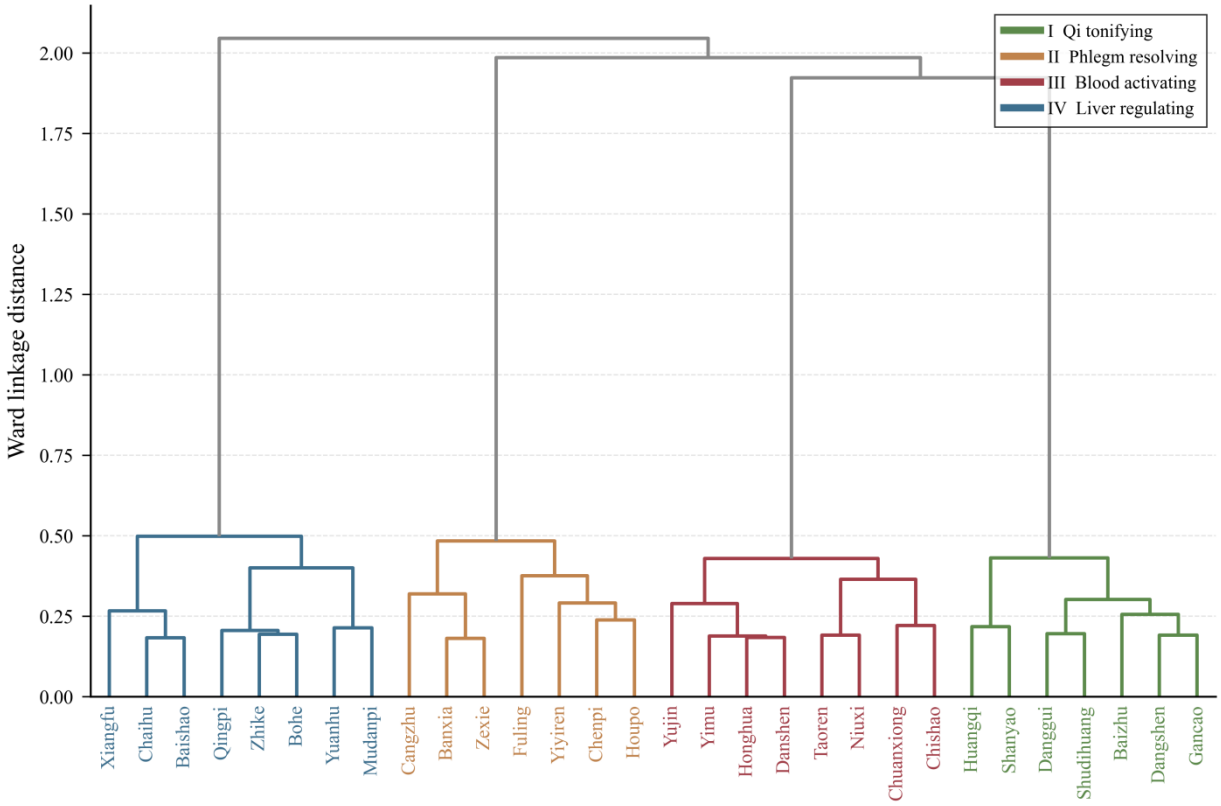

Fig. 2. Agglomerative hierarchical clustering dendrogram of the top 30 high frequency herbs based on Jaccard distance and Ward linkage.

TABLE III. Per category classification metrics on the validation set

Syndrome	Precision	Recall	F1 Score	Support
Qi Deficiency	0.852	0.821	0.836	29
Yin Deficiency	0.846	0.812	0.829	26
Yang Deficiency	0.917	0.880	0.898	25
Blood Stasis	0.880	0.846	0.863	26
Phlegm Dampness	0.870	0.900	0.885	30
Liver Qi Stagnation	0.833	0.833	0.833	24
Damp Heat	0.889	0.889	0.889	27
Cold Dampness	0.870	0.870	0.870	23
Wind Heat	0.789	0.789	0.789	19
Wind Cold	0.897	0.866	0.881	29
Macro Avg	0.864	0.842	0.851	258
Weighted Avg	0.871	0.873	0.870	258

ISSN: 2710-0170

Fig. 3 displays the multi class receiver operating characteristic curves and the corresponding per category area under the curve values. The area under the curve of every syndrome category exceeds 0.91, and the values for Yang Deficiency, Wind Cold, and Damp Heat all reach 0.96 or above, which approaches ideal performance. The macro average area under the curve of 0.940 is far above the random classifier reference level of 0.500, which demonstrates that the proposed mathematical statistical framework effectively captures discriminative symptom syndrome mappings while maintaining a high true positive rate at low false positive rate. Per category analysis further reveals that classes with strong, distinctive symptom signatures, such as Wind Cold and Yang Deficiency, achieve the highest classification accuracy, while classes that share overlapping symptoms with neighboring syndromes, such as Qi Deficiency versus Yin Deficiency, exhibit slightly lower individual scores but remain well above the macro average baseline.

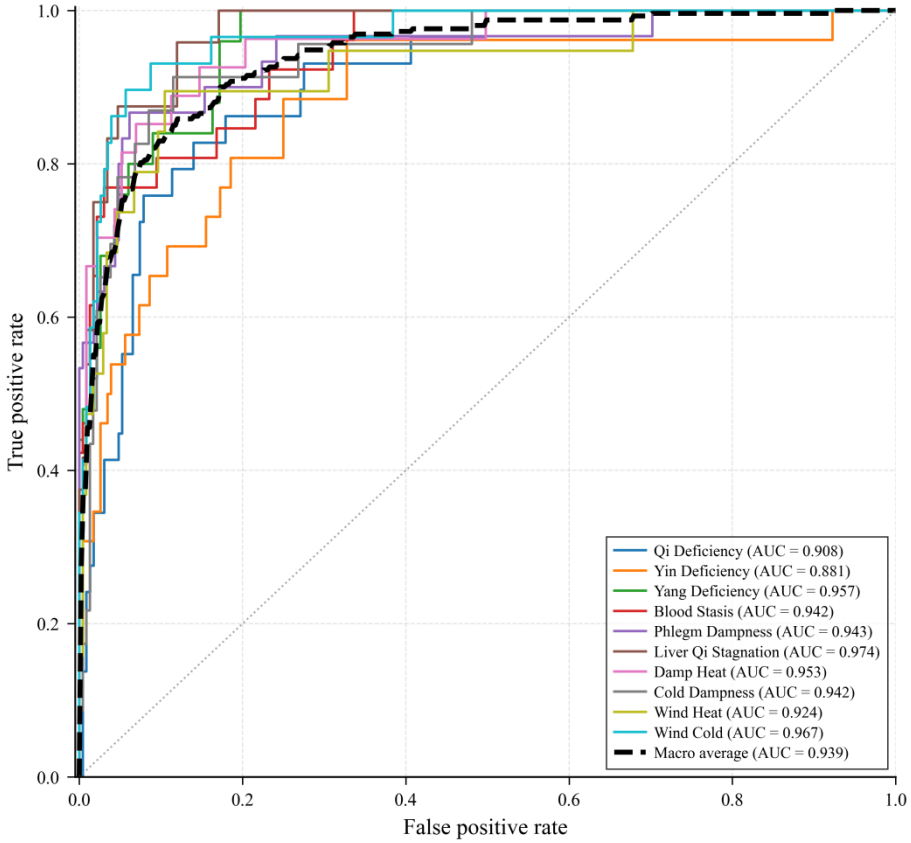

Fig. 3. Multi class receiver operating characteristic curves with per category area under the curve values for the ten syndrome categories.

4. Conclusion

This paper presents an integrated mathematical statistical framework for the quantitative analysis of TCM diagnosis and treatment patterns based on a unified data preprocessing pipeline coupled with association rule mining, hierarchical clustering, complex network analysis, logistic regression, and Bayesian inference. The proposed pipeline encompasses systematic standardization and binary encoding of TCM clinical data, descriptive frequent pattern mining via the Apriori algorithm, exploratory clustering and topological analysis of herbal co occurrence, and predictive syndrome classification with class balanced training and L2 regularization. Experimental evaluation on a real world dataset of 1287 patient records and 312 prescriptions yields 27 strong herbal association rules, four interpretable herb clusters, an

ISSN: 2710-0170

overall syndrome classification accuracy of 0.873, a macro average F1 score of 0.851, and a macro average area under the curve of 0.940.

Detailed per category analysis reveals that syndrome categories with strong distinctive symptom signatures such as Wind Cold and Yang Deficiency are recognized with very high accuracy, while syndrome categories that share overlapping symptoms such as Qi Deficiency and Yin Deficiency present greater challenges. The consistently high area under the curve values across all classes suggest that the learned representations are fundamentally sound and that further performance gains are achievable through improved decision mechanisms.

Future work will focus on the following directions: first, integrating tongue images and pulse waveform features into the symptom representation to enrich the diagnostic signal; second, extending the Bayesian inference module with hierarchical and dynamic structures to model long term temporal evolution of syndromes; third, expanding the dataset with more diverse practitioners and recording conditions to strengthen external validity; and fourth, extending the system from syndrome classification to quantitative herbal dosage recommendation with fine grained clinical decision support.

Acknowledgement

This work was supported by the University-Level Innovation and Entrepreneurship Training Program of the School of Mathematics and Statistics, Henan University of Technology (Project No. 10463142, Year 2025), titled "Traditional Chinese Medicine Diagnosis and Treatment Rules Based on Digital Statistical Models".

References

[1] Zhou E, Shen Q, Hou Y. Integrating artificial intelligence into the modernization of traditional Chinese medicine industry: a review. Frontiers in Pharmacology, 2024, 15: 1181183.

[2] Liu Y, Zhang J, Wang Q, et al. An Apriori algorithm based association analysis of analgesic drugs in Chinese medicine prescriptions recorded from patients with rheumatoid arthritis pain. Frontiers in Pharmacology, 2022, 13: 937259.

[3] Liu W, Zhao Y, Zhuo X, et al. The identification of Chinese herbal medicine combination association rule analysis based on an improved Apriori algorithm in treating patients with COVID-19 disease. Journal of Healthcare Engineering, 2022, 2022: 6337082.

[4] Lu CL, Chen HW, Lu CJ, et al. An Apriori algorithm based association rule analysis to identify herb combinations for treating uremic pruritus using Chinese herbal bath therapy. Evidence Based Complementary and Alternative Medicine, 2020, 2020: 8854772.

[5] Pang B, Hu G, Yan W, et al. Analysis of prescription medication rules of traditional Chinese medicine for bradyarrhythmia treatment based on data mining. Medicine, 2022, 101(45): e31374.

[6] Yao L, Wang R, Mao X, et al. THCluster: herb supplements categorization for precision traditional Chinese medicine. arXiv preprint, 2020, arXiv:2011.11396.

[7] Yang K, Zhang R, He L, et al. Network patterns of herbal combinations in traditional Chinese clinical prescriptions. Frontiers in Pharmacology, 2020, 11: 590824.

[8] Cheng W, Wu Y, Zhao S, et al. Research of insomnia on traditional Chinese medicine diagnosis and treatment based on machine learning. BMC Complementary Medicine and Therapies, 2020, 20: 309.

[9] Wang Y, Yu D, Yu C, et al. A new method for syndrome classification of non small cell lung cancer based on data of tongue and pulse with machine learning. BioMed Research International, 2021, 2021: 1337558.

[10] Zhang Y, Liu Y, Yu J, et al. Study of TCM syndrome identification modes for patients with type 2 diabetes mellitus based on data mining. Evidence Based Complementary and Alternative Medicine, 2014, 2014: 524092.

[11] Gan X, Shu Z, Wang X, et al. Network medicine framework reveals generic herb symptom effectiveness of traditional Chinese medicine. Science Advances, 2023, 9(43): eadh0215.

ISSN: 2710-0170

[12] Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques. Third Edition. Morgan Kaufmann, Waltham, MA, 2011.

[13] Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1): 86 to 97.

[14] Liu Z, Zhang H, Zhang R, et al. Detection of herb symptom associations from traditional Chinese medicine clinical data. Evidence Based Complementary and Alternative Medicine, 2015, 2015: 638148.