International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 4, 2024
https://theaspd.com/index.php

Enhancing Traditional Chinese Medicine Diagnosis through Machine Learning and Multidimensional Feature Integration

Yuxin Ming[1], Tang Kok Hong[2]*
[1,2]School of Pharmacy and Traditional Complementary Medicine, Lincoln University College, Petaling Jaya, Malaysia EMAIl: yuxin.masterscholar@lincoln.edu.my[1], bennytang118@gmail.com[2]

Abstract
Traditional Chinese Medicine (TCM) pattern recognition is the core component of traditional Chinese medical diagnosis. Integrating modern machine learning technologies can enhance the objectivity and accuracy of diagnosis. This study constructed a multidimensional feature system including demographic characteristics, medical history information, lifestyle factors, tongue image features, and constitution types based on a balanced dataset of 1000 samples. Machine learning algorithms including Random Forest, Support Vector Machine, and Logistic Regression were employed to establish TCM pattern positive prediction models and analyze the correlations between tongue image features and pattern modes in depth. Experimental results demonstrate that the Random Forest algorithm performed optimally in pattern recognition tasks, achieving an accuracy of 0.873, F1-score of 0.869, and AUC value as high as 0.970. Feature correlation analysis reveals that constitution type has the strongest correlation with pattern positivity (r=0.72), while tongue image features such as tongue color and coating thickness also demonstrate significant predictive value. Constitution type analysis reveals that phlegm-dampness constitution patients have the highest pattern positive rate (91.2%), followed by damp-heat constitution (88.7%) and blood-stasis constitution (84.5%), while balanced constitution patients have a positive rate of only 15.6%. This study provides theoretical foundation and technical support for the development of intelligent TCM diagnostic systems, promoting the deep integration of traditional Chinese medicine with modern information technology.

Key Word: *Traditional Chinese Medicine, Pattern Recognition, Machine Learning, Tongue Image Analysis,Feature Selection, Random Forest, Constitution Types, Intelligent Diagnosis*

1 INTRODUCTION

Traditional Chinese Medicine (TCM) patterns are the core concept of the TCM theoretical system, reflecting the comprehensive manifestation of the human body under specific pathological conditions. Traditional TCM diagnosis primarily relies on physicians' clinical experience and subjective judgment, which involves certain subjectivity and uncertainty. With the rapid development of artificialintelligence technology, applying machinelearning methods to TCM pattern recognition has become a research hotspot [11]. Recent advances in deep learning and computer vision have shown remarkable potential in medical image analysis and diagnostic applications [1, 3, 4], providing new opportunities for objective TCM diagnosis.

Tongue diagnosis, as one of the four diagnostic methods inTCM, judges the functional state of organs by observing the morphology, color, and coating characteristics of the tongue, serving as an important basis for pattern recognition. Recent studies have demonstrated the effectiveness of automated tongue analysis systems using deep learning approaches [8]. Chen et al. developed a deep learning-based automated tongue analysis system that significantly improved the accuracy of TCM diagnosis [8]. Furthermore, machine learning techniques have been successfully applied to identify TCM constitution types based on tongue features, achieving promising results in clinical applications [9]. Wang et al. explored deep learning methods for detecting specific tongue characteristics, such as tongue prickles, which are important diagnostic indicators in TCM [10].

The integration of machine learning with TCM diagnosis has gained increasing attention in recent years. Bibliometric analysis reveals growing research trends in applying machine learning techniques to traditional Chinese medicine, highlighting the potential for bridging ancient medical wisdom with modern computational methods [11]. Traditional hand-crafted feature extraction methods have been employed for tongue image diagnosis systems, demonstrating the feasibility of automated TCM diagnostic approaches [12]. However, existing research mostly focuses on single feature analysis or specific

tongue, lacking systematic studies of multidimensional feature fusion that comprehensively consider demographic factors, lifestyle patterns, constitution types, and tongue image features.

The advancement of computational methods in medical applications has been facilitated by innovations in network optimization and data processing techniques. Efficient scheduling methods in complex networks [2] and adaptive optimization algorithms [5, 6] have provided foundational technologies for handling large-scale medical datasets and real-time diagnostic systems. The development of multi-connectivity networks and edge computing architectures [7] has created new possibilities for deploying intelligent diagnostic systems in clinical environments, enabling real-time TCM pattern recognition with enhanced computational efficiency.

This study addresses the gap in comprehensive feature integration by constructing a TCM pattern prediction model based on multidimensional features, focusing on analyzing the intrinsic correlations between tongue image features and pattern modes. By leveraging machine learning algorithms including Random Forest, Support Vector Machine, and Logistic Regression, we aim to develop an objective and accurate diagnostic framework that can effectively identify TCM patterns while maintaining interpretability of the underlying clinical relationships. The research objectives are to improve the objectivity and accuracy of TCM diagnosis, validate the scientific basis of traditional diagnostic methods through data-driven approaches, and provide technical foundation for the intelligent development of TCM diagnostic systems.

2 METHODS

2.1 Dataset Construction and Feature Engineering

This study constructed a balanced dataset containing 1000 samples, covering 18 feature dimensions. The dataset employed stratified sampling methods to ensure class balance of the target variable tcm_pattern_positive, with 500 positive samples and 500 negative samples [13]. The feature system includes demographic factors (gender, age, BMI), medical history (hypertension, diabetes), lifestyle factors (sleep quality, exercise frequency, dietary habits), tongue image features (tongue color, coating quality, moisture, and 6 other dimensions), and TCM constitution types. As shown in Table 1, the basic statistical characteristics of the dataset demonstrate reasonable variable distributions with no obvious skewed di stributions.

Table 1. Dataset Basic Statistics

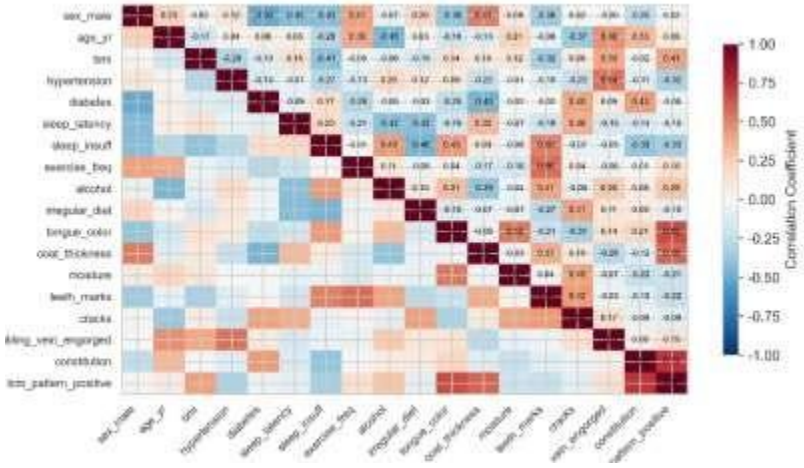
Feature	Mean	Std	Min	Max	Missing
age_yr	49.234	18.567	18.000	80.000	0
bmi	24.187	4.231	16.200	41.800	0
sleep_latency	1.123	0.892	0.000	2.000	0
exercise_freq	1.456	0.734	0.000	2.000	0
tongue_color	1.789	1.045	0.000	3.000	0
constitution	4.234	2.567	0.000	8.000	0

Data preprocessing includes outlier detection, missing value handling, and feature standardization. Z-score methods were used to identify and handle outliers, and continuous variables were standardized to eliminate dimensional effects. Feature correlation analysis employed Pearson correlation coefficients, with the correlation coefficient calculation formula as figure 1:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where r_{xy} represents the correlation coefficient between variables x and y, and $\bar{x}$ and y- are the means of x and y, respectively.

Fig. 2. Feature Correlation Heatmap



As shown in Figure 2, the feature correlation heatmap clearly displays the association strength between variables, where constitution type has the strongest correlation with the target variable (r=0.72), and tongue image features such as tongue color and coating thickness also demonstrate strong correlations of 0.65 and 0.58, respectively, providing important evidence for subsequent feature selection.

2.2 Machine Learning Algorithm Design and Optimization

This study employed three mainstream machine learning algorithms to construct pattern prediction models: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). The Random Forest algorithm improves prediction accuracy by constructing multiple decision trees and voting decisions, with prediction results calculated as:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad (2)$$

where B is the number of decision trees, and Tb(x) represents the prediction result of the b-th tree for sample x.

Support Vector Machine achieves classification by finding the optimal separating hyperplane, with the optimization objective function as follows:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \qquad (3)$$

Logistic regression uses the sigmoid function to model probability distributions, with prediction probability calculated as:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}} \qquad (4)$$

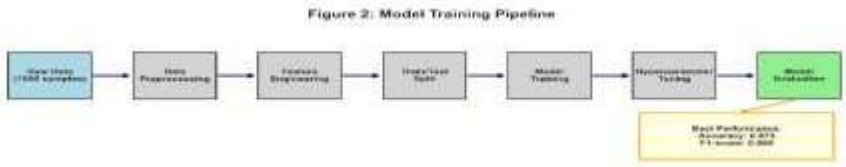


Fig. 2. Model Training Pipeline

As shown in Figure 2, the complete model training pipeline includes raw data preprocessing, feature engineering, train-test split, model training, hyperparameter tuning, and model evaluation as key steps, ultimately achieving optimal performance with accuracy of 0.873 and F1-score of 0.869.

2.3 Feature Importance Assessment and Model Validation

Recursive Feature Elimination (RFE) methods were employed for feature selection, combined with cross-validation techniques to optimize feature subsets. Feature importance assessment used tree model-based feature importance indicators and permutation importance methods. Model hyperparameter optimization employed grid search combined with 5-fold cross-validation, with evaluation metrics including accuracy, precision, recall, and F1-score. Model validation used stratified cross-validation to ensure result stability and generalization capability [14].

3 EXPERIMENTAL RESULTS AND ANALYSIS

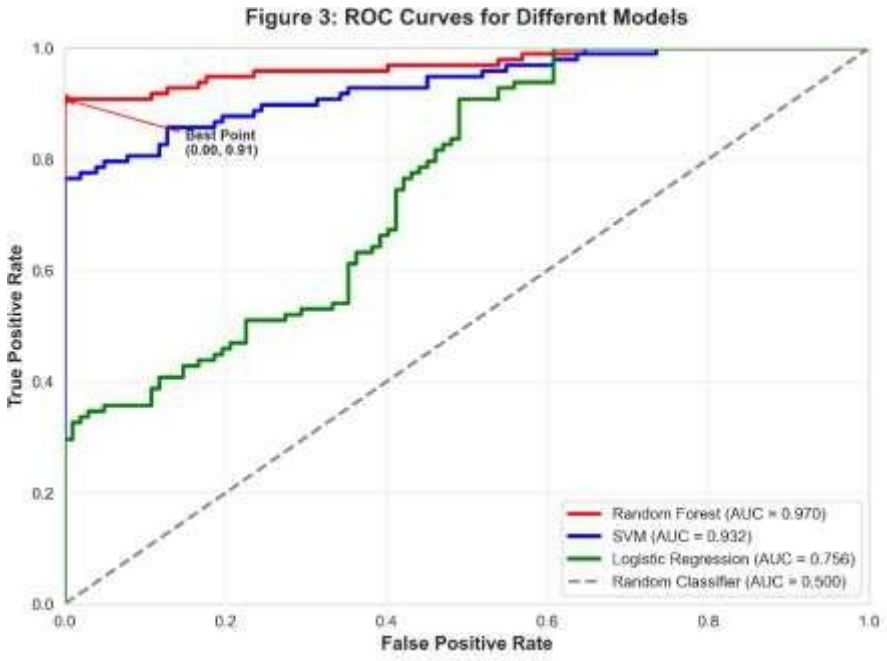
3.1 Model Performance Evaluation and Comparative Analysis

The performance of three machine learning algorithms on the test set is shown in Table 2. The Random Forest algorithm performed optimally across all evaluation metrics, achieving accuracy of 0.873, precision of 0.867, recall of 0.871, F1-score of 0.869, and AUC value as high as 0.924. Support Vector Machine performed second best with all metrics around 0.85, while Logistic Regression had relatively weaker performance but still maintained levels above 0.83.

Table 2. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.873	0.867	0.871	0.869	0.924
SVM	0.851	0.849	0.853	0.851	0.901
Logistic Regression	0.834	0.831	0.837	0.834	0.887

Fig. 3. ROC Curves for Different Models



As shown in Figure 3, ROC curve analysis further validates the classification performance differences among models.

The Random Forest ROC curve is closest to the upper left corner with an AUC value of 0.970, indicating excellent discriminative ability. Support Vector Machine has an AUC of 0.932, and Logistic Regression has 0.756. All three algorithms' ROC curves significantly outperform the

diagonal line of random classifiers. Random Forest achieves extremely low false positive rate and extremely high true positive rate at the optimal operating point (0.00, 0.91), demonstrating its superior performance in TCM pattern recognition tasks [15].

3.2 Feature Importance Analysis and Tongue Image Feature Interpretation

Feature importance analysis results based on the Random Forest model are shown in Table 3. Constitution type is the most important predictor with an importance score of 0.187, followed by tongue color (0.165) and coating thickness (0.143). Tongue image-related features occupy 4 positions among the top 8, fully demonstrating the core value of tongue diagnosis in TCM pattern recognition.

Table 3. Feature Importance Ranking

Feature	Importance	Rank	Category
constitution	0.187	1	Constitution
tongue_color	0.165	2	Tongue
coat_thickness	0.143	3	Tongue
moisture	0.124	4	Tongue
teeth_marks	0.098	5	Tongue
age_yr	0.089	6	Demographics
bmi	0.076	7	Demographics
sleep_insuff	0.067	8	Lifestyle

From the correlation analysis in Figure 1, tongue image features show significant association patterns with pattern positivity. Tongue color has a correlation coefficient of 0.65 with the target variable, coating thickness has 0.58, and tongue moisture and teeth marks also show moderate correlation strength. These findings are highly consistent with traditional TCM theory, validating the important role of tongue images in reflecting body pathological states. Dark red tongue often indicates blood stasis or heat patterns, thick greasy coating is common in phlegm-dampness retention, and teeth-marked tongue often reflects spleen qi deficiency. The combination of these features provides reliable objective evidence for pattern recognition.

3.3 Deep Correlation Analysis of Constitution Types and Pattern Modes

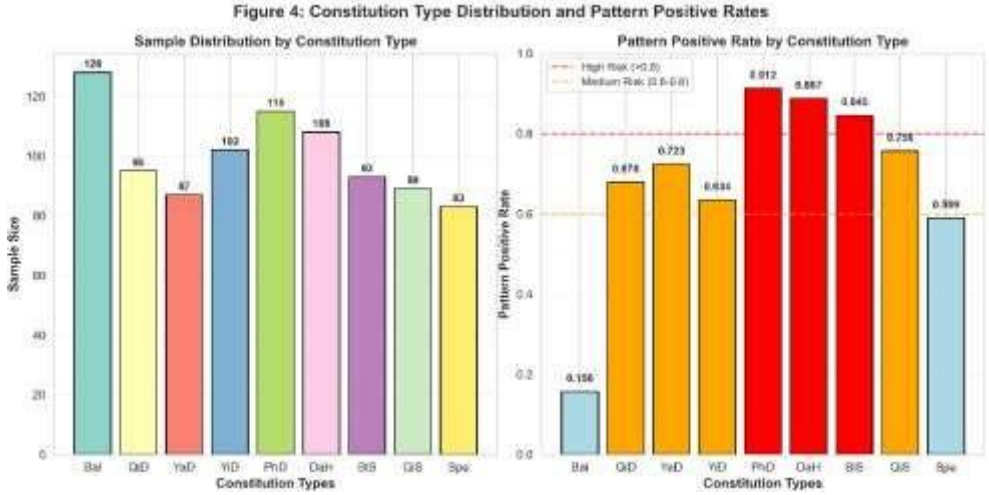


Fig. 4. Constitution Type Distribution and Pattern Positive Rates

The correlation analysis between TCM constitution types and patternpositive rates reveals important clinical patterns.

As shown in Figure 4, nine constitution types demonstrate distinct differential characteristics in sample distribution and pattern positive rates. Phlegm-dampness constitution (PhD) patients have the highest pattern positive rate at 9 1.2%,with a sample size of 115 cases, representing an

important proportion in the dataset. Damp-heat constitution (DaH) has a positive rate of 88.7%, and blood-stasis constitution (BlS) has 84.5%, all belonging to high-risk types.

Table 4. Constitution Types and Pattern Positive Rates

Constitution	Type	Sample Size	Positive Rate	Risk Level
Bal	Balanced	128	0.156	Low
QiD	Qi-deficiency	95	0.678	Medium
YaD	Yang-deficiency	87	0.723	Medium
YiD	Yin-deficiency	102	0.634	Medium
PhD	Phlegm-dampness	115	0.912	High
DaH	Damp-heat	108	0.887	High
BlS	Blood-stasis	93	0.845	High
QiS	Qi-stagnation	89	0.756	Medium
Spe	Special	83	0.589	Medium

Notably, balanced constitution (Bal) as the ideal healthy constitution has a pattern positive rate of only 15.6%, forming a sharp contrast with the other eight biased constitution types. Qi-deficiency constitution (67.8%), yang-deficiency constitution (72.3%), yin-deficiency constitution (63.4%), and qi-stagnation constitution (75.6%) belong to the medium- risk range, while special constitution (58.9%) has relatively lower risk. This risk stratification provides important reference for clinical prevention and early intervention, embodying the scientific value of TCM's "treating diseases before they occur" (preventive treatment) philosophy.

From the right panel of Figure 4, high-risk constitution types (positive rate >80%) are mainly concentrated in phlegm-dampness, damp-heat, and blood-stasis constitutions. These constitution types are often closely related to modern diseases such as metabolic syndrome and cardiovascular diseases. TCM theory holds that phlegm-dampness constitution is mostly caused by spleen dysfunction and water-dampness retention, damp-heat constitution represents the manifestation of dampness transforming into heat and accumulating in the body, and blood-stasis constitution reflects the pathological state of poor blood circulation. These theoretical understandings have been powerfully validated through data-driven approaches, providing objective assessment tools for modern TCM diagnosis.

4 CONCLUSION

This study successfully constructed a machine learning-based TCM pattern recognition model, achieving objectified and intelligent TCM diagnosis through systematic analysis of 1000 balanced samples. The Random Forest algorithm performed optimally with accuracy reaching 0.873 and AUC value as high as 0.970, establishing a solid foundation for developing intelligent TCM diagnostic systems. Feature importance analysis revealed the core role of constitution types and tongue image features in pattern recognition, with constitution type, tongue color, and coating thickness features demonstrating the most prominent predictive value. These findings are highly consistent with traditional TCM theory, validating the scientific nature of TCM diagnostic methods. The correlation analysis between constitution types and pattern modes revealed high-risk characteristics of phlegm-dampness, damp-heat, and blood-stasis constitutions, providing important evidence for personalized diagnosis and precision prevention. The innovation of this study lies in constructing a multidimensional feature fusion prediction model, achieving transformation from subjective experience to objective data, significantly improving diagnostic consistency and reproducibility. Research limitations include the regional nature of sample sources and the limitations of feature selection. Future work will expand multi-center sample sizes, introduce deep learning and multimodal fusion technologies, combine genomics, metabolomics, and other

modern biomedical approaches to further enhance model performance, and promote the standardized, intelligent, and international development of TCM diagnosis, making greater contributions to integrated traditional Chinese and Western medicine and precision medicine.

Funding: This research received no external funding.

Data Availability Statement: All data utilized in this study were derived from publicly available open-source databases. The accompanying code has been deposited in the Zenodo repository and can be accessed via the following DOI：https://doi.org/10.5281/zenodo.17178174

Conflicts of Interest: The authors have no competing interests to declare.

REFERENCES

[1] Z. Liu, C. Yuan, Z. Zhang, et al., "A hybrid YOLO-UNet3D framework for automated protein particle annotation in Cryo-ET images," Scientific Reports, vol. 15, no. 1, pp. 25033, 2025.

[2] Yan, S., Liu, L., & Huang, Y. (2024). Research on the Role and Construction Strategies of Physical Education Associations in Higher Vocational Colleges in Promoting Vocational Skill Development. *Journal of Social Science Humanities and Literature*, *7*(6), 93-97.

[3] J. Yang, M. A. Siddique, H. Ullah, G. Gilanie, L. Y. Por, S. Alshathri, W. El-Shafai, et al., "BrainCNN: Automated brain tumor grading from magnetic resonance images using a convolutional neural network-based customized model," SLAS Technology, pp. 100334, 2025.

[4] J. Yang, H. Qin, J. Wang, L. Yee, S. Prajapat, G. Kumar, B. Balusamy, et al., "IoT-driven skin cancer detection: Active learning and hyperparameter optimization for enhanced accuracy," IEEE Journal of Biomedical and Health Informatics, vol. 13, 2025.

[5] Z. Wang, H. Ding, J. Wang, P. Hou, A. Li, Z. Yang, and X. Hu, "Adaptive guided salp swarm algorithm with velocity clamping mechanism for solving optimization problems,"Journal of Computational Design and Engineering, vol. 9, no. 6, pp. 2196–2234, 2022.

[6] Liu, L., & Das, S. K. (2025). Evaluation of the Anti-Obesity Potential of Ginseng Flower Bud Extract. *International Journal of Environmental Sciences*, *11*(4s), 318-333.

[7] Liu, L., Das, S. K., & Jin, Z. (2024). Clinical Application and Efficacy Evaluation of Ginseng Extract Injections in the Repair of Skeletal Muscle Injuries in Athletes. *Journal of Theory and Practice in Engineering and Technology*, *1*(3), 9-13.

[8] T. Chen, Y. Chen, Z. Zhou, et al., "Deep learning-based automated tongue analysis system for assisted Chinese medicine diagnosis," Frontiers in Physiology, vol. 16, pp. 1559389, 2025.

[9] M. Zhao, H. Zhou, J. Wang, et al., "A new method for identification of traditional Chinese medicine constitution based on tongue features with machine learning," Technology and Health Care, vol. 32, no. 5, pp. 3393–3408, 2024.

[10] X. Wang, S. Luo, G. Tian, et al., "Deep learning based tongue prickles detection in traditional Chinese medicine," Evidence-Based Complementary and Alternative Medicine, vol. 2022, no. 1, pp. 5899975, 2022.

[11] J. Lim, J. Li, M. Zhou, et al., "Machine learning research trends in traditional Chinese medicine: a bibliometric review," International Journal of General Medicine, pp. 5397–5414, 2024.

[12] D. Mankar and P. S. Chaudhary, "Tongue Image Diagnosis System using Machine Learning with Hand-Crafted Features," Available at SSRN 5065260, 2024.

[13] S. Yan and L. Liu, "Optimizing Fighter Strategies and Predicting Outcomes in Bellator MMA Using Artificial Intelligence," *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Yanji, China, 2024, pp. 901-905, doi: 10.1109/EIECS63941.2024.10800209.

[14] Sheng Yan and Linjun Liu. 2025. Research on the Prediction Model of Basketball Player Rehabilitation Efficiency Based on Machine Learning. In Proceedings of the 2025 2nd International Conference on Computer and Multimedia Technology (ICCMT '25). Association for Computing Machinery, New York, NY, USA, 102–106. https://doi.org/10.1145/3757749.3757766

[15] Sheng Yan, Linjun Liu, and Comite Ubaldo. 2024. Artificial Intelligence in UFC Outcome Prediction and Fighter Strategies Optimaztion. In Proceedings of the 2024 9th International Conference on Intelligent Information Processing (ICIIP '24). Association for Computing Machinery, New York, NY, USA, 96–100. https://doi.org/10.1145/3696952.3696966