Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.26043

Identification, Prediction, and Classification of Traditional Chinese Medicine Syndromes Based on Machine Learning Algorithms

Jiajing Zheng

Second College of Basic Medical Sciences for Clinical Medicine, Guizhou University of Transitional Chinese Medicine, Guizhou, China
19816750346@163.com

Abstract. The CNN-SVM AdaBoost fusion model constructed in this article combines the deep feature extraction ability of convolutional neural networks for multi-source four diagnostic information with the high-dimensional discriminative boundary advantage of support vector machines, and further corrects difficult cases using AdaBoost's iterative weighting mechanism to form an end-to-end syndrome identification system. Compared with classical logistic regression, gradient boosting, K-NN, and the best performing random forest, this model improved the accuracy from 87.4% to 95.8%, F1 score from 87% to 95.5%, AUC from 0.924 to 0.983 in the four classification task, with leading margins of 8.4%, 8.5%, and 0.059, respectively; Compared to the remaining baseline, there is a gap of more than 10%. What is particularly valuable is that the model has a high recall rate of 96% and an F1 score of 95.5%, which enables it to maintain sensitive capture and robust discrimination of minority categories in traditional Chinese medicine scenarios with fuzzy syndrome boundaries and imbalanced samples, significantly reducing the risk of misdiagnosis and missed diagnosis. Practice has proven that the collaboration between deep learning and traditional machine learning in the field of traditional Chinese medicine can break through the bottleneck of a single algorithm, provide reliable tools for accurate diagnosis and treatment, and have direct and profound significance for improving the quality of clinical diagnosis and treatment.

Keywords: Convolutional Neural Networks, Support Vector Machines, AdaBoost, Traditional Chinese Medicine Syndrome Identification System.

1. Introduction

Traditional Chinese Medicine (TCM) syndrome identification is the core process of TCM diagnosis and treatment, which essentially involves comprehensively judging the patient's internal pathological state through multi-source information such as observation, hearing, questioning, and cutting [1]. However, traditional methods heavily rely on the personal experience of physicians, are subjective and difficult to standardize, especially in the rehabilitation stage, where patient symptoms often exhibit dynamic evolution and a mixture of deficiency and excess, making identification more

DOI: 10.54254/2755-2721/2025.26043

complex [2]. At the same time, patients in the rehabilitation period often have multiple chronic diseases coexisting, and their syndrome manifestations are intertwined. Traditional four diagnostic information acquisition may be limited or incomplete [3]. The massive amount of multimodal data accumulated in the modern medical environment provides unprecedented opportunities to break through this bottleneck. These data dimensions complement each other, with tongue imaging and facial diagnosis reflecting surface morphological changes, pulse diagnosis containing circulatory dynamics information, inquiry recording subjective symptoms, and Western medicine indicators providing objective physiological and biochemical parameters [4]. How to effectively integrate these heterogeneous and high-dimensional information, construct an objective, accurate, and scalable intelligent identification model for traditional Chinese medicine syndromes in rehabilitation patients, has become a key scientific issue in promoting the modernization and personalization of traditional Chinese medicine rehabilitation, and achieving precise intervention of "disease syndrome combination". It has important clinical value and application prospects.

Machine learning algorithms play the role of the core engine in this topic, and their value is mainly reflected in the deep analysis and intelligent fusion of complex multimodal data [6]. Traditional statistical methods are unable to handle heterogeneous high-dimensional data such as tongue images, pulse signals, text speech, and structured indicators. Machine learning, especially deep learning techniques, can automatically extract key discriminative features: Convolutional neural networks (CNN) can efficiently analyze visual information such as tongue coating, tongue texture, color, and facial morphology; Recurrent Neural Networks (RNNs) or Transformers are adept at handling semantic associations and temporal evolution of symptoms in inquiry texts/speech; Time series models (such as LSTM) can capture subtle rhythmic features of pulse waveforms; Meanwhile, structured indicators of Western medicine can be efficiently incorporated into the feature space [7]. More importantly, machine learning integrates information from different senses and dimensions through multimodal fusion technology, excavates the nonlinear and deep level interaction between them, transcends the limitations of single mode or artificial rules, and builds a comprehensive mapping model of the overall functional state of "syndrome". Its powerful pattern recognition ability can learn complex mapping relationships between syndromes and multi-source data, optimize identification boundaries, achieve automated and high-precision identification of dynamic and multidimensional syndromes in rehabilitation patients, greatly improve diagnostic efficiency and consistency, and provide intelligent decision support for personalized rehabilitation plans. This article improves and optimizes support vector machines based on convolutional neural networks, and integrates them with AdaBoost for the identification, prediction, and classification of traditional Chinese medicine syndromes.

2. Data sources

This dataset contains multimodal TCM syndrome differentiation data of 527 rehabilitation patients, including 14 feature variables and 1 target variable. Tongue features (tongue color, coating thickness, coating color) are obtained through image recognition, pulse features (pulse rate, pulse rhythm, pulse strength) are obtained from signal processing, facial features (complexion, glossiness) are based on facial image analysis, inquiry features (fatigue level, appetite score, sleep quality) are obtained from natural language processing, and Western medicine examination indicators (systolic blood pressure, diastolic blood pressure, heart rate, blood glucose) are integrated. The target variables are four types of TCM syndrome classification (qi deficiency syndrome, blood deficiency syndrome, yin deficiency syndrome, yang deficiency syndrome), which can be directly used for

DOI: 10.54254/2755-2721/2025.26043

automatic identification of TCM syndromes. Machine learning model development. Partial datasets are shown in Table 1.

Table 1. Selected partial dataset

Tongue color	Coating thickness	Coating color	Pulse rate	Pulse rhythm	Pulse strength	Face color	Face gloss	Fatigue level	Blood pressure sys	Blood pressure dia	Heart rate	Blood sugar	Syndrome type
1	0.35	2	89.59	0	0.68	0	0.62	0.86	133.49	82.44	77.64	5.06	1
4	0.15	0	58.15	1	0.57	0	0.89	0.67	132.34	81.74	73.44	6.18	1
2	0.88	1	78.14	0	0.38	3	0.89	0.48	105.56	65.20	85.00	4.26	2
2	0.83	0	74.98	0	0.87	0	0.69	0.84	130.39	80.53	99.37	5.12	2
4	0.73	0	68.19	0	0.66	3	0.44	0.51	141.10	87.15	78.92	6.66	3
0	0.42	1	75.81	0	0.41	3	0.96	0.25	110.87	68.48	76.68	4.80	0

3. Method

3.1. Convolutional neural network

Convolutional neural networks are a type of deep model designed specifically for processing data with grid structures, which first shone in image recognition. Its core idea is to apply several learnable convolution kernels on the input in a sliding window manner, with each kernel resembling a small and focused searchlight, capturing specific patterns such as edges, textures, or higher-order shapes only within a local receptive field. When these local responses are stacked in space, a feature map is formed, which preserves key information while compressing the data dimension. The subsequent introduction of nonlinear activation allows the network to express complex functions, while pooling operations further reduce redundancy through downsampling, making features robust to transformations such as translation and scaling. Multiple sets of convolution activation pooling hierarchical combinations gradually abstract the original pixels into more meaningful representations, laying the foundation for subsequent decision-making. The network structure of convolutional neural network is shown in Figure 1.

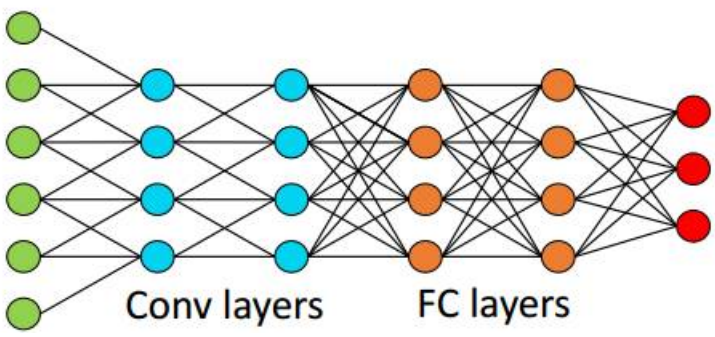

Figure 1. The network structure of convolutional neural network

DOI: 10.54254/2755-2721/2025.26043

In order to complete the mapping from pixels to categories, the network usually connects several fully connected layers at the end, flattens and integrates the two-dimensional feature map into a global description, and finally outputs the probability distribution through functions such as softmax. The entire system adjusts all convolutional kernels and fully connected weights end-to-end through backpropagation, so that the error signal is transmitted layer by layer from top to bottom, driving each layer filter to adaptively learn the most beneficial features for the task.

3.2. Support vector machine

Support Vector Machine is a discriminative model that aims to maximize the "interval". It transforms the learning problem into finding an optimal hyperplane in the feature space, so that samples of different categories are clearly separated and the distance from the nearest point to the plane is as far as possible. Intuitively, these "nearest points" form the "support vectors" that support the entire decision boundary, firmly locking in the position and direction of the hyperplane like wedges. When the data is linearly separable, the algorithm only needs to solve a convex quadratic programming to obtain a unique and globally optimal partition plane; When there is noise or overlap in the data, a soft interval strategy allows a small number of samples to be misclassified, balancing the interval width and training error through penalty parameters; When the original space cannot be linearly partitioned, kernel techniques implicitly map the data to a higher dimensional or even infinite dimensional space, where an almost perfect linear boundary can often be found, while the computation is still completed in the original dimension, cleverly balancing expressive power and efficiency [8]. The network structure of support vector machine is shown in Figure 2.

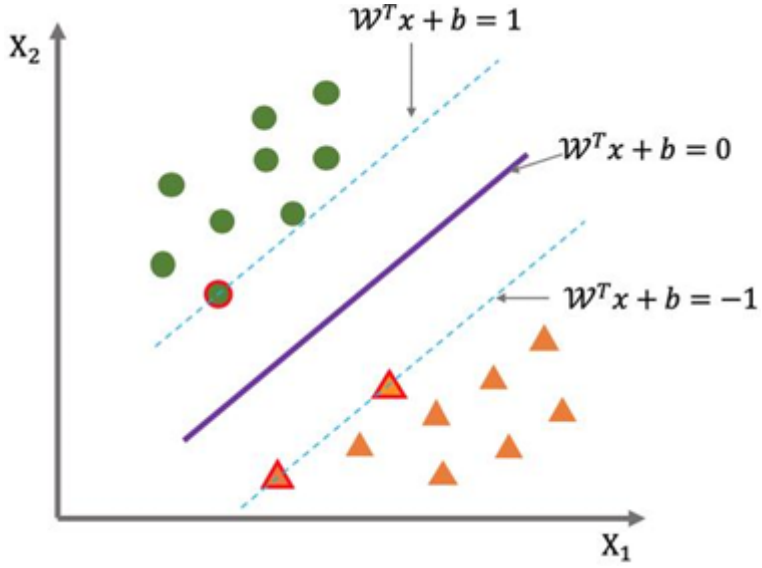

Figure 2. The network structure of support vector machine

3.3. Adaboost

Adaboost first assigns the same initial weight to each sample in the training set, and then trains a weak classifier in each round - a simple model that is slightly better than random guessing is sufficient. After training, the algorithm checks which samples have been misclassified by the current weak classifier and significantly increases their weights, while reducing the weights of correctly classified samples. Therefore, the next round of weak classifiers is forced to focus on the previously

DOI: 10.54254/2755-2721/2025.26043

most difficult to classify "stubborn" samples. At the same time, each round of the weak classifier will also obtain a "discourse power" coefficient based on its own error rate, where the fewer errors, the higher the weight [9]. After multiple rounds of focusing and weighting, Adaboost weights the outputs of all weak classifiers according to their respective discourse weights to obtain a powerful ensemble decision boundary. The entire process does not require explicit design of complex features. Simply repeatedly calling weak learning algorithms can allow simple models to accumulate layer by layer and ultimately approximate any complex true distribution. The network structure of Adaboost is shown in Figure 3.

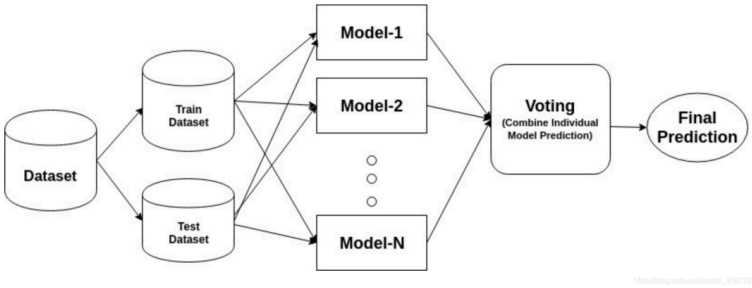

Figure 3. The network structure of Adaboost

3.4. CNN-SVM-AdaBoost

The CNN-SVM AdaBoost model proposed in this article first uses a convolutional neural network as a feature extraction engine. Through multi-layer convolution, activation, and pooling, the original high-dimensional input is compressed layer by layer into a highly abstract discriminative feature map. Then, the top output of the CNN is directly fed into a support vector machine, which constructs a more robust and sparse decision hyperplane in the high-dimensional feature space using its maximum interval criterion. This integrates the representation ability of CNN with the global optimum and small sample advantage of SVM; To further improve performance, the overall framework uses AdaBoost as the outer skeleton and considers "CNN-SVM" as a weak learner. In each iteration, the sample weights are readjusted based on the weighting error of the previous round, so that the subsequent CNN-SVM pays more attention to difficult to distinguish samples. The weak classifiers in each round are weighted and voted according to their accuracy. Finally, the weak classifiers are stacked into strong classifiers through serial boosting, achieving the unity of deep features, maximum spacing, and adaptive focusing advantages. In complex pattern recognition tasks, accuracy, sparsity, and generalization ability are balanced [10].

4. Result

In terms of hardware and software configuration, the hardware includes Intel Core i9-13900K 3.0 GHz, 128 GB DDR5-5600 memory, NVIDIA RTX 4090 24 GB video memory, and 2 TB PCIe 4.0 NVMe SSD; The software is Windows 11 Pro 23H2 and MATLAB R2024a. In terms of model parameter settings, the weak learner uses 100 decision tree stumps (maxsplit=20, learnRate=0.1) and performs 10 fold cross validation through resampling; Train optimizer Adam with an initial learning rate of 0.001, $\beta1=0.9$,$\beta2=0.999$,mini-batch 64, Maximum epoch 150, early stop Patience 10; GPU acceleration adopts CUDA 12.2+cuDNN 8.9. In terms of dataset partitioning, the dataset is divided into training and testing sets in a 7:3 ratio.

DOI: 10.54254/2755-2721/2025.26043

In terms of comparative models, this article uses logistic regression Gradient Boosting, Random forest and K-NN, in addition, this article evaluates the predictive performance of the model through Accuracy, Precision, Recall, F1, and AUC. The results of the comparative experiment are shown in Table 2. The confusion matrix of the CNN-SVM AdaBoost test set is shown in Figure 4.

Table 2. The results of the comparative experiment

Medel	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
Logistic Regression	81.2	80.5	80.7	80.6	0.872
Gradient Boosting	84.6	83.8	84.1	83.9	0.901
Random Forest	87.4	86.9	87.2	87	0.924
K-NN	78.9	78	78.2	78.1	0.851
CNN-SVM-AdaBoost	95.8	95.1	96	95.5	0.983

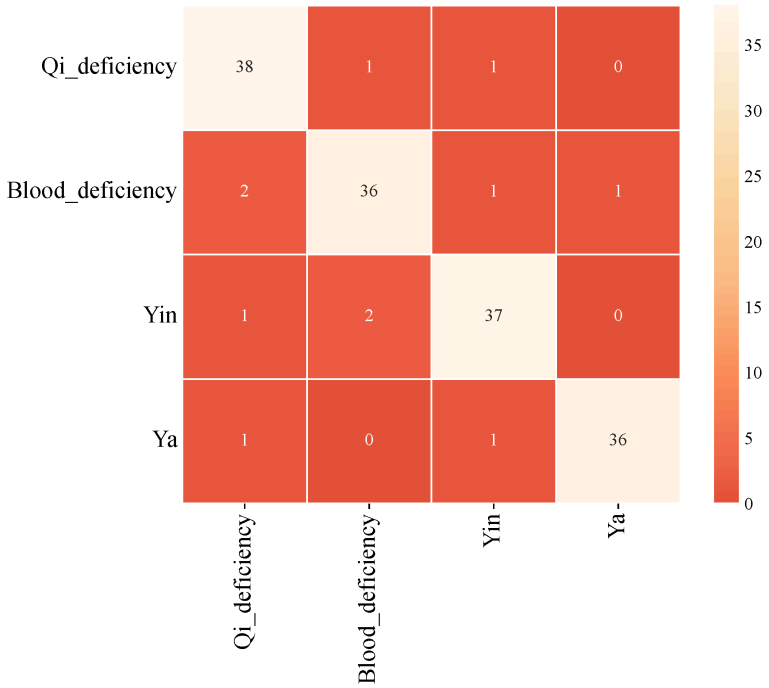

Figure 4. The confusion matrix of the CNN-SVM AdaBoost test set

The comparison of bar charts for each model is shown in Figure 5. From the four classification results, the traditional method Random Forest leads with an accuracy of 87.4%, an F1 score of 87%, and an AUC of 0.924. However, the CNN-SVM AdaBoost proposed in this paper directly raises various indicators to 95.8%, 95.5%, and 0.983, which are 8.4%, 8.5%, and 0.059 AUC higher than the suboptimal model Random Forest, respectively; Compared with logistic regression, gradient boosting, and K-NN, its advantage is over 10%. Especially when the recall rate (96%) and F1 (95.5%) remain high at the same time, it indicates that the model is both sensitive and robust in capturing a small number of syndrome categories, significantly reducing the misdiagnosis and missed diagnosis gap of traditional methods in complex TCM syndrome boundary fuzzy scenarios.

DOI: 10.54254/2755-2721/2025.26043

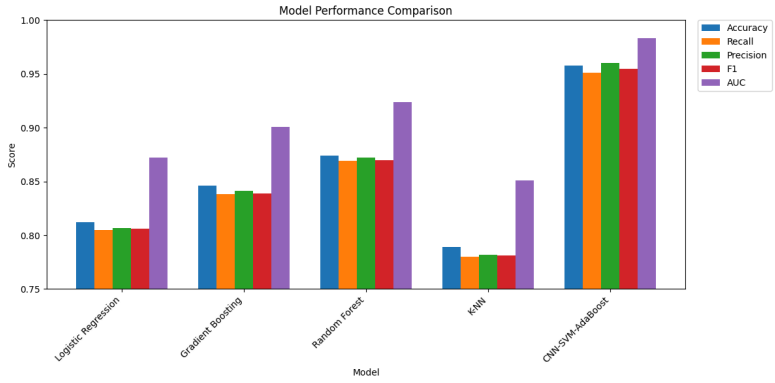

Figure 5. The comparison of bar charts for each model

5. Conclusion

This study deeply integrates convolutional neural networks with support vector machines, supplemented by AdaBoost's serial ensemble, to construct a CNN-SVM AdaBoost model for face oriented TCM syndrome identification. The experiment takes the four diagnostic data of observation, hearing, questioning, and cutting as input, and compares it with logistic regression in the four classification task Gradient Boosting, Random forest and K-NN as controls: The best performing random forest in the traditional approach achieved an accuracy of 87.4%, an F1 score of 87%, and an AUC of 0.924, while our model improved to 95.8%, 95.5%, and 0.983, leading by 8.4 percentage points, 8.5 percentage points, and an AUC of 0.059, respectively; Compared to the remaining three baselines, the advantage has expanded to over 10%. It is particularly crucial that the model maintains a high level of recall rate of 96% and F1 95.5%, and also has robust recognition ability for rare syndromes with small sample sizes, significantly compressing the misdiagnosis and missed diagnosis space caused by fuzzy syndrome boundaries.

This result not only verifies the feasibility of synergistic effect between deep learning and classical classifiers, but also provides an interpretable and practical intelligent tool for traditional Chinese medicine "syndrome differentiation", laying the algorithmic foundation for precise diagnosis and treatment.

References

[1] Li, FuFeng, et al. "Computer-assisted lip diagnosis on traditional Chinese medicine using multi-class support vector machines." BMC complementary and alternative medicine 12.1 (2012): 127.
[2] Zhao, Changbo, et al. "Advances in patient classification for traditional Chinese medicine: a machine learning perspective." Evidence-Based Complementary and Alternative Medicine 2015.1 (2015): 376716.
[3] Tang, Yuqi, et al. "Research of insomnia on traditional Chinese medicine diagnosis and treatment based on machine learning." Chinese Medicine 16.1 (2021): 2.
[4] Ming, Yuxin, and Tang Kok Hong. "Machine Learning Drives Intelligent Diagnosis of Traditional Chinese Medicine." Journal of Theory and Practice in Clinical Sciences 2 (2025): 60-67.
[5] Wang, Xuewei, et al. "A self-learning expert system for diagnosis in traditional Chinese medicine." Expert systems with applications 26.4 (2004): 557-566.
[6] Zhang, Sheng, et al. "Advances in the application of traditional Chinese medicine using artificial intelligence: a review." The American journal of Chinese medicine 51.05 (2023): 1067-1083..
[7] Duan, Yu-yu, et al. "Application and development of intelligent medicine in traditional Chinese medicine." Current medical science 41.6 (2021): 1116-1122.
[8] Qiusi, Mao. "Research on the improvement method of music education level under the background of AI technology." Mobile information systems 2022.1 (2022): 7616619.

DOI: 10.54254/2755-2721/2025.26043

[9] Wang, Shaohui, et al. "Practical implementation of artificial intelligence-based deep learning and cloud computing on the application of traditional medicine and western medicine in the diagnosis and treatment of rheumatoid arthritis." Frontiers in pharmacology 12 (2021): 765435.
[10] Li, Wenyu, et al. "Opportunities and challenges of traditional Chinese medicine doctors in the era of artificial intelligence." Frontiers in Medicine 10 (2024): 1336175.