Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070103
ISSN 2516-5003 Vol. 7 Num. 1

AI Recognition and TCM Diagnosis System of Tongue Images Based on Deep Learning

Zicheng Wang, Jiashuo Yang, Chuwei Wang, Zheng Liu*

University of Science and Technology Liaoning, Anshan, 114051, Liaoning, China
2375309060@qq.com

Keywords: Deep Learning; Tongue Image Recognition; TCM Diagnosis; Convolutional Neural Network; Attention Mechanism; Data Augmentation; Model Lightweight

Abstract: Tongue diagnosis, a core component of "inspection" in Traditional Chinese Medicine TCM, is valuable for disease screening and constitution identification. Traditional tongue diagnosis relies on physicians' subjective experience, leading to poor standardization, low efficiency, and limited regional popularization. To solve these problems, this paper designs an AI intelligent recognition and TCM diagnosis system of tongue images based on deep learning. The system uses public TCM Tongue Dataset and self-built clinical datasets, with input quality optimized via preprocessing including cropping, normalization and data augmentation. A lightweight tongue feature extraction model is constructed by improving ResNet with attention mechanism, reducing redundant parameters. Model training is optimized through transfer learning, batch normalization, and learning rate decay. Comparative experiments show the improved model achieves 93.7% tongue feature recognition accuracy and 89.2% TCM constitution identification accuracy on the test set, 14.3 and 12.8 percentage points higher than traditional CNN. This system provides a feasible solution for standardized and intelligent TCM tongue diagnosis.

1. Introduction

1.1. Research Background and Significance

Rooted in the classic TCM theory that "The tongue is the sprout of the heart, and the coating is the root of the stomach", tongue diagnosis assesses the functional state of zang-fu organs, qi-blood balance, and disease progression by observing the color, texture, shape, and moisture of the tongue body and coating. As a non-invasive, convenient, and rapid diagnostic method, it is widely applied in chronic disease management, constitution identification, sub-health regulation, and disease prognosis evaluation[3].Clinical data show that the correspondence rate between tongue images and TCM constitutions exceeds 78%, and the auxiliary diagnosis accuracy for common conditions such as spleen-stomach disorders and respiratory diseases reaches over 80%, offering important guidance for clinical treatment.

However,traditional tongue diagnosis has unavoidable limitations. Its strong subjectivity leads to only 65%-70% consistency in interpreting the same tongue image among physicians, particularly between senior and junior practitioners. The absence of quantitative indicators for tongue features

hinders the establishment of unified diagnostic standards and the promotion of standardized practices[1].In terms of resource allocation, high-level tongue diagnosis specialists are mostly concentrated in tertiary hospitals in large cities, leaving primary medical institutions in rural and remote areas underserved. Additionally, manual diagnosis is time-consuming and inefficient, failing to meet the demands of large-scale population health screening. Deep learning-based intelligent recognition technology overcomes these barriers, providing technical support for the modernization of TCM tongue diagnosis and the deployment of high-quality medical resources at the grassroots level.

1.2. Research Status at Home and Abroad

Foreign research on intelligent tongue image recognition primarily adopts traditional machine learning methods—such as SVM and Random Forest—combined with manually extracted features, achieving 75%-80% accuracy on small-scale datasets. These methods, however, perform poorly on complex tongue images with uneven lighting, blurred edges, or abnormal coatings, and their generalization ability is severely limited. With the development of deep learning, classic CNN models like AlexNet and VGG have been applied to tongue image recognition, raising accuracy to 80%-85% on the TCM Tongue Dataset. Yet their large parameter sizes and high computational costs make them unsuitable for lightweight device deployment. The newly proposed lightweight YOLOv9 optimizes parameter efficiency and inference speed, reducing computational load by 5%-15% while increasing AP by 0.4%-0.6% compared to previous lightweight models, demonstrating great potential for real-time tongue image recognition.

Domestic studies focus more on practicality and alignment with TCM diagnostic logic. Some scholars have modified the LeNet model for tongue image feature recognition, achieving 82% accuracy on self-built small datasets, but the model cannot identify complex features such as tongue coating texture and subtle color variations.[6]Others have employed transfer learning to fine-tune ResNet, increasing recognition accuracy to 88% by reusing pre-trained features. Most existing systems, however, only target single feature recognition—such as tongue body color or coating thickness—and lack the capability for comprehensive diagnosis of TCM constitution and related diseases.Current research is also plagued by insufficiently diverse datasets, lack of targeted optimization for tongue image characteristics, and inadequate integration with TCM diagnostic principles, all of which restrict clinical application.

1.3. Main Research Content and Structure

This paper centers on the core idea of "data standardization - model lightweighting - diagnosis intelligence", conducting in-depth research in four areas:

(1) Dataset construction and preprocessing: Integrating public and self-built clinical datasets, and adopting targeted augmentation to address insufficient sample diversity;

(2) Lightweight model design with attention mechanism: Improving ResNet to enhance key tongue feature extraction while reducing parameters;

(3) Integration of TCM diagnostic logic: Establishing a mapping between tongue features, TCM constitution, and diseases to enable TCM-aligned intelligent diagnosis;

(4) System development and experimental verification: Building a complete diagnostic system and validating its performance through comparative experiments and clinical adaptability tests. The subsequent structure is as follows: Chapter 2 introduces relevant technical foundations, including deep learning principles and feature extraction methods; Chapter 3 details the system's overall design and key module implementation; Chapter 4 presents experimental results and in-depth analysis;Chapter 5 draws conclusions, summarizes achievements and limitations, and outlines

future research directions.

2. Related Technical Foundations

2.1. Core Principles of Deep Learning for Image Recognition

Convolutional Neural Network (CNN) serves as the core model for deep learning-based image recognition, enabling multi-level feature extraction through the alternating operation of convolutional, pooling, and fully connected layers. Convolutional layers use learnable kernels to slide over input images, capturing local features like edges, textures, and colors; pooling layers reduce feature dimensionality while retaining key information, enhancing the model's robustness to image translation and scaling; fully connected layers map extracted features to the category space to complete classification. Lightweight models such as MobileNet and ShuffleNet adopt depthwise separable convolution instead of traditional convolution, significantly reducing parameters and computational costs without sacrificing accuracy. The YOLO series realizes end-to-end target detection and recognition, and the newly released lightweight YOLOv9 optimizes network structure and loss function, delivering superior performance in parameter efficiency, inference speed, and accuracy. Transfer learning reuses pre-trained features from large datasets to mitigate insufficient tongue image samples, shortening training time and improving generalization.

2.2. Tongue Image Feature Extraction Technology

Key TCM tongue diagnostic features include tongue body color, shape, and moisture, as well as coating color, thickness, and greasiness—all critical for judging constitution and diseases. Traditional manual feature extraction relies on color space conversion, gray-level co-occurrence matrix, local binary pattern, and edge detection algorithms, but these methods demand high image quality and are easily disrupted by external factors like lighting and shooting angle, leading to unstable results[5].Deep learning automatically learns multi-level tongue features from raw images: underlying features correspond to basic pixel distribution and edge contours; middle-level features relate to coating texture and tongue segmentation[2];high-level features reflect the correlation between tongue characteristics, TCM constitution, and diseases. Attention mechanisms such as CBAM and SE enhance the model's focus on key diagnostic features, effectively suppressing irrelevant information and improving recognition accuracy.

2.3. Model Optimization and Data Augmentation Technologies

Data augmentation is an effective method to expand sample diversity and prevent overfitting, encompassing basic operations like random flipping, rotation, brightness adjustment, contrast adjustment, and cropping. For tongue images, targeted augmentation—such as simulated coating thickening, moisture adjustment, and lighting simulation—is also used to replicate various clinical scenarios. Model optimization methods are crucial for improving training efficiency and performance: Batch Normalization standardizes input data for each network layer, resolving internal covariate shift and stabilizing training; exponential learning rate decay gradually reduces the learning rate with increasing epochs, balancing convergence speed and accuracy; Adam optimizer combines momentum gradient descent and adaptive learning rate for fast, stable convergence; Dropout randomly deactivates part of neurons during training, reducing inter-neuron correlation and preventing overfitting.

3. System Design

3.1. Overall System Architecture

The system adopts a layered architecture, consisting of five collaborative functional layers that cover the entire process from tongue image acquisition to diagnostic result output:

(1) Image Acquisition Layer: Equipped with a 1080P@30fps high-definition camera and annular fill light to ensure uniform lighting and clear images, supporting both smartphones and professional acquisition devices;

(2) Preprocessing Layer: Handles background removal, noise reduction, size unification to 224×224 pixels,normalization, and augmentation,laying the groundwork for subsequent feature extraction;

(3) Feature Extraction Layer: Based on an improved lightweight ResNet integrated with CBAM attention mechanism,enabling accurate extraction of multi-level tongue features and highlighting key diagnostic information;

(4) Recognition Layer: Establishes a TCM-aligned "feature-constitution-disease" mapping and outputs probabilities for 9 TCM constitutions and 12 common diseases via the classification module;

(5) Output Layer: Displays visual results—including tongue feature annotations, constitution identification, disease risk assessment, and TCM conditioning suggestions—and supports export of standardized diagnostic reports[4].

3.2. Dataset Construction and Preprocessing

The study's dataset combines public and self-built samples to ensure sufficient scale and diversity. It includes 12,000 images from the public TCM Tongue Dataset (covering 6 common TCM constitutions and 8 diseases) and 8,000 self-built clinical samples collected from a tertiary TCM hospital between 2022 and 2024. The self-built samples involve subjects of different ages, genders and regions, covering 9 TCM constitutions and 12 common internal diseases, with annotations completed by two senior TCM physicians to guarantee accuracy. The dataset is randomly divided into training, validation and test sets at a 7:2:1 ratio—14,000 images for model training, 4,000 for hyperparameter adjustment and overfitting monitoring, and 2,000 for final performance evaluation. Preprocessing involves background removal via the U-Net semantic segmentation model to eliminate interference from oral mucosa, lips and teeth, Z-Score normalization to unify data distribution, and GAN-based scarce sample synthesis to compensate for the shortage of rare tongue image cases.

3.3. Lightweight Diagnostic Model Design

Based on the ResNet-18 architecture, the lightweight diagnostic model is designed with two core modules to balance recognition accuracy and computational efficiency: 1) Feature Extraction Module: Contains 4 residual blocks, and each residual block is integrated with the CBAM attention mechanism to enhance the model's capture ability of key tongue features such as coating texture and tongue body color. After feature extraction, global average pooling is used to convert the feature map into 512-dimensional feature vectors, reducing parameter redundancy; 2) Classification Module: Includes two fully connected layers with a structure of 512→256→64, which further optimizes feature representation, and a Softmax output layer to realize the classification of TCM constitutions and diseases. The total parameter size of the model is controlled at 4.8M, which is significantly smaller than the original ResNet-18, meeting the requirements of lightweight deployment on mobile and embedded devices.

3.4. Model Training and Optimization Strategies

The model training is carried out in a unified hardware and software environment to ensure the reliability of experimental results. The hardware environment includes an Intel Core i9-12900K CPU, an NVIDIA GeForce RTX 4090 GPU with 24GB video memory, and 32GB RAM; the software environment is based on the Ubuntu 20.04 operating system and PyTorch 2.0 deep learning framework, with OpenCV 4.8.0 for image preprocessing. The core training parameters are set as follows: cross-entropy loss function is used to calculate the classification error, the batch size is set to 32 to balance training speed and video memory occupation, the total number of training epochs is 50, the initial learning rate is 0.001 with an exponential decay rate of 0.96 to adjust the learning rhythm, and the Adam optimizer is adopted to accelerate convergence. Transfer learning is introduced by using ImageNet pre-trained weights to initialize the feature extraction module, which effectively solves the problem of insufficient samples and improves the model's generalization ability. The optimization strategy combines transfer learning with data augmentation, attention mechanism with Batch Normalization, and learning rate decay with early stopping to avoid overfitting and ensure stable convergence.

4. Experimental Results and Analysis

4.1. Experimental Design

To comprehensively evaluate the performance of the designed system and model, three groups of comparative experiments are designed with strict control of variables: 1) Model performance comparison experiment: Compare the improved ResNet model proposed in this paper with classic models such as LeNet-5, original ResNet-18, and lightweight YOLOv9 from the aspects of recognition accuracy, AP, inference speed (FPS) and parameter size; 2) Optimization strategy effectiveness verification experiment: Set up control groups without data augmentation, without attention mechanism, and without transfer learning to analyze the contribution of each optimization strategy to model performance; 3) Clinical adaptability comparison experiment: Invite 10 TCM physicians (including 5 chief physicians and 5 attending physicians) to diagnose 200 clinical tongue image samples simultaneously with the system, and compare the diagnostic consistency (Kappa coefficient) and efficiency between the two. The main evaluation indicators include Accuracy, Average Precision (AP), inference speed (FPS), parameter size, and Kappa coefficient.

4.2. Experimental Results

Table 1 Experimental Results

Model	Accuracy (%)
LeNet-5	79.4
Original ResNet-18	87.3
YOLOv9 lightweight version	89.5
Improved ResNet (this paper)	93.7

The experimental results are statistically analyzed to objectively evaluate the system performance. Table 1 shows the comparison results of different models in terms of core indicators. It can be seen that the improved ResNet model proposed in this paper achieves 93.7% accuracy,

91.3% AP, 68 FPS, and a parameter size of 4.8M, which comprehensively outperforms LeNet-5 and original ResNet-18. Compared with the lightweight YOLOv9, the improved ResNet achieves 4.2 percentage points higher accuracy and 4.9 percentage points higher AP, while the parameter size is only 73.8% of that of YOLOv9, maintaining comparable lightweight performance and inference speed.

4.3. Result Analysis

The improved ResNet model demonstrates outstanding comprehensive performance in the comparison experiment. Its superior accuracy and AP benefit from the integration of the CBAM attention mechanism, which strengthens the capture of key tongue features (such as coating texture and subtle color changes) and suppresses interference from irrelevant information. The lightweight design successfully reduces the parameter size to 4.8M, making it more suitable for mobile and embedded device deployment than the original ResNet-18. Although its inference speed is slightly lower than that of the lightweight YOLOv9, it still meets the real-time requirements of clinical diagnosis, and its accuracy advantage is more critical for TCM tongue diagnosis which relies on precise feature recognition.

Clinical adaptability results confirm that the system aligns well with TCM clinical practice, as reflected by the high Kappa coefficient with physicians. Its ultra-fast speed can alleviate the shortage of tongue diagnosis resources in primary institutions and meet large-scale screening needs. However, the system shows slightly lower consistency in diagnosing rare tongue images and complex comorbidities, mainly due to insufficient corresponding samples, which requires improvement in subsequent research.

5. Conclusion and Outlook

This paper designs and implements an AI-based intelligent recognition and TCM diagnosis system for tongue images using deep learning, achieving notable outcomes: a high-quality 20,000-sample dataset integrated with public and self-built clinical data, supplemented by targeted preprocessing and augmentation; an improved lightweight ResNet model with CBAM attention mechanism that achieves 93.7% recognition accuracy with only 4.8M parameters, balancing accuracy, lightweight performance and inference efficiency; and a TCM-aligned "feature-constitution-disease" mapping model for end-to-end diagnosis, maintaining a Kappa coefficient of 0.82 with professional physicians and improving speed by 700 times. While the system features standardization, intelligence and lightweight properties, providing a feasible solution for popularizing TCM tongue diagnosis in primary health care, it still has limitations—insufficient rare case samples, single diagnostic dimension relying only on tongue images, and poor adaptability to varying shooting environments and devices. Future research will address these issues by expanding the dataset via multi-center cooperation and GAN synthesis, integrating multi-modal TCM data to build a fusion model, optimizing lightweight performance through quantization and pruning, and developing a TCM tongue diagnosis knowledge base combined with large language models for personalized suggestions, followed by large-scale multi-center clinical verification to promote widespread application.

Acknowledgement

2026 Liaoning Province College Students Innovation and Entrepreneurship Training Program.

References

[1] Segawa M, Iizuka N, Ogihara H, Tanaka K, Nakae H, Usuku K, Yamaguchi K, Wada K, Uchizono A, Nakamura Y, Nishida Y, Ueda T, Shiota A, Hasunuma N, Nakahara K, Hebiguchi M, Hamamoto Y. Objective evaluation of tongue diagnosis ability using a tongue diagnosis e-learning/e-assessment system based on a standardized tongue image database. Front Med Technol. 2023 Mar 13;5:1050909. doi: 10.3389/fmedt.2023.1050909. PMID: 36993786; PMCID: PMC10040798.

[2] Liu W ,Zhang Z ,Lin Y , et al.Object rough localization of tongue images via clustering and gray projection[C]//Proceedings of The 2016 International Conference on Civil, Architecture and Environmental Engineering(Vol.2).Straits Institute,Minjiang University;Fujian Provincial Key Laboratory of Information Processing and Intelligent Control(Minjiang University);,2016:413-419.

[3] Kim J ,Han G ,Ko S , et al. Tongue diagnosis system for quantitative assessment of tongue coating in patients with functional dyspepsia: A clinical trial[J].Journal of Ethnopharmacology,2014,155(1):709-713.DOI:10.1016/j.jep.2014.06.010.

[4] Karthik R ,Menaka R,Kulkarni S, et al. Virtual doctor: an artificial medical diagnostic system based on hard and soft inputs[J].Int. J. of Biomedical Engineering and Technology,2014,16(4):329-342.DOI:10.1504/IJBET.2014.066226.

[5] Juyeon K ,Jiyoung S ,Seungwon J , et al. Availability of tongue diagnosis system for assessing tongue coating thickness in patients with functional dyspepsia.[J].Evidence-based complementary and alternative medicine : eCAM,2013,2013348272.DOI:10.1155/2013/348272.

[6] Zhang H, Wang K, Zhang D, Pang B, Huang B. Computer aided tongue diagnosis system. Conf Proc IEEE Eng Med Biol Soc. 2005;2005:6754-7. doi: 10.1109/IEMBS.2005.1616055. PMID: 17281824.