ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

Spot Recognition in Tongue Images based on ResNet50

Dehong Zeng, Shuo Wang*, Xuanyi Liu, Chunlei Zhao, Xiyuan Zhang, Hao Zhang

North China University of Science and Technology, Tangshan 063000, China

Abstract

Tongue diagnosis plays a crucial role in traditional Chinese medicine (TCM) by offering valuable insights for clinical syndrome differentiation and health assessment through tongue image analysis. Among these features, tongue spots, as a common abnormal manifestation, may indicate underlying pathological changes. Nevertheless, manual tongue diagnosis in TCM heavily relies on subjective visual observation, posing challenges in standardization. To tackle this issue, our study introduces an automated method for recognizing tongue spots using deep residual networks and transfer learning. Leveraging a real clinical dataset of 5,371 tongue images, we preprocessed the images, applied data augmentation, and fine-tuned a ResNet50 model pretrained on ImageNet for binary classification of tongue spot features. We employed five-fold cross-validation to assess the proposed approach. The results demonstrated that ResNet50 achieved an average accuracy of 97.48% (±0.53%) and a macro-F1 score of 53.84% (±6.64%), surpassing the ResNet18 baseline model. These findings suggest that deep residual learning effectively captures distinctive features from tongue images, showing promise for automated tongue feature recognition. Our study offers practical insights for the advancement of intelligent TCM-assisted diagnostic systems and may pave the way for further investigations into detailed tongue image analysis.

Keywords

Tongue Image Recogition; Deep Learning Residual Network; Transfer Learning.

1. Introduction

1.1 Research Background

Tongue diagnosis plays a crucial role in clinical assessment within traditional Chinese medicine (TCM). Clinicians can gather significant insights regarding the functional status of internal organs and disease progression by examining tongue color, coating, texture, and localized abnormalities. Among the various tongue manifestations, tongue spots represent a common abnormal feature that may offer valuable information for syndrome differentiation and clinical evaluation. Consequently, the precise identification of tongue spot characteristics holds practical importance for the objective analysis of tongue images.

Conventional tongue diagnosis primarily depends on visual observation and clinical experience of physicians, leading to subjectivity and potential inconsistency among practitioners. Manual assessment, being time-consuming and challenging to standardize, hinders its widespread use in primary healthcare settings. The rapid advancement of artificial intelligence has facilitated the use of deep learning for medical image analysis, enhancing diagnostic objectivity and efficiency[2][3].

1.2 Research Objective

The main objective of this study is to develop an automatic recognition method for tongue spot features using ResNet50[1] and transfer learning. By training a deep convolutional neural network on real clinical tongue images, this study aims to improve the accuracy and reliability of tongue spot

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

classification and reduce dependence on manual observation.We also aims to verify the effectiveness of transfer learning in tongue image analysis. Since medical image datasets are often limited in size, using a pretrained network can provide better initialization and improve convergence during model training[3][6].

Therefore, an ImageNet-pretrained ResNet50 model was adopted and fine-tuned for the target task[1].

1.3 Main Contributions

This study makes several contributions to automatic tongue image analysis. First, it applies an ImageNet-pretrained ResNet50 model to tongue spot feature recognition using real clinical tongue images. Second, it introduces image preprocessing and data augmentation strategies to improve the robustness and generalization ability of the model. Third, it evaluates the method through five-fold cross-validation and reports the results using average values and standard deviations, which improves the reliability of the findings.

Overall, this study provides a practical exploration of deep learning-based tongue image recognition and offers a useful reference for the future development of intelligent TCM-assisted diagnosis systems.

2. Materials and Methods

2.1 Data Source

The dataset utilized in this study comprised 5,371 authentic clinical tongue images gathered from actual medical scenarios. In contrast to small laboratory datasets, real clinical tongue images typically exhibit greater variability in shooting angles, lighting conditions, image clarity, and tongue posture. These attributes not only complicate the recognition task but also enhance the practical significance of the study. As the images were sourced from real-world clinical environments, the dataset more accurately represents the complexities associated with tongue diagnosis in medical applications.

Professional physicians annotated all images, encompassing both tongue surface features and syndrome-related labels, yielding a dataset conducive to model development. The dataset comprised 13 binary labels, incorporating characteristics of tongue images and organ-related indicators. This framework facilitates not only the present single-label classification task but also lays the groundwork for forthcoming multi-label investigations.

Furthermore, the dataset was meticulously structured. Image paths and corresponding labels were documented in CSV files, while the images were systematically stored in directories categorized by patient ID. This systematic organization facilitates efficient data administration, reproducibility of experiments, and training of models. Moreover, it establishes a robust technical framework for future research endeavors encompassing patient-centric investigations or intricate diagnostic modeling.

2.2 Target Label Selection

The Spot label, among the 13 annotated labels, was chosen as the focal point of this investigation. The task of recognition was framed as a binary classification issue, with the objective of the model being to determine if a tongue image exhibited spot characteristics. This particular label was selected due to its clinical relevance and a more evenly distributed class representation in contrast to various other labels within the datasets.

Based on the training statistics, the Spot label comprised 1,602 positive samples and 1,769 negative samples, yielding a positive ratio of 0.48. This distribution is advantageous for both model training and performance assessment as it mitigates the potential for significant class imbalance. Moreover, tongue spots exhibit a wide range of visual characteristics such as shape, size, color, and location, rendering them a suitable testbed for assessing a deep learning model's proficiency in fine-grained tongue image recognition.

Tongue spots, as a significant tongue manifestation in Traditional Chinese Medicine (TCM) diagnosis, present as localized abnormal areas on the tongue surface, potentially linked to specific pathological

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

conditions. Due to variations in size, number, color, and distribution, these spots pose a complex challenge for image classification models. Thus, focusing on the Spot label in this study was not only clinically relevant but also methodologically sound.

2.3 Image Preprocessing

Tongue spots, as a significant tongue manifestation in Traditional Chinese Medicine (TCM) diagnosis, present as localized abnormal areas on the tongue surface, potentially linked to specific pathological conditions. Due to variations in size, number, color, and distribution, these spots pose a complex challenge for image classification models. Thus, focusing on the Spot label in this study was not only clinically relevant but also methodologically sound.

All images were subsequently normalized using the mean and standard deviation values derived from the ImageNet dataset. Given that the backbone network was initialized with pretrained weights from ImageNet, this normalization strategy ensured consistency between the training images and the pretrained feature space. Additionally, standardized preprocessing mitigated the impact of irrelevant visual variations, allowing the network to concentrate more effectively on clinically significant features.

Image preprocessing in this study served as both a technical necessity and a crucial measure to enhance model stability. Clinical tongue images frequently exhibit variations in brightness, saturation, and contrast attributable to diverse camera settings and environmental conditions. By standardizing the preprocessing, the impact of these extraneous variables can be mitigated, enabling the network to prioritize medically significant visual characteristics like local color variances, textural designs, and spot-like formations.

2.4 Data Augmentation

To enhance the model's generalization capability, various data augmentation techniques were employed during training. These techniques encompassed random horizontal flipping, random rotation, and color jitter. Such operations augmented sample diversity and replicated typical image variations encountered in clinical practice, including variations in imaging angles and illumination conditions.

Random horizontal flipping improved the model's resilience to left-right spatial variability. Random rotation facilitated the model's adjustment to subtle changes in posture within tongue images. Color augmentation proved particularly beneficial due to the susceptibility of tongue image appearance to variations in lighting and camera configurations. These augmentation techniques collectively bolstered the model's ability to handle real-world image diversity. Nevertheless, during validation and testing, only resizing and normalization were implemented to maintain the objectivity of the evaluation outcomes.

2.5 Model Architecture

This study utilized ResNet50 as the foundational network for identifying tongue spots. ResNet50, a profound convolutional neural network incorporating residual connections, mitigates issues of gradient vanishing and network degradation in deep structures. Due to its robust feature extraction capabilities, ResNet50 finds extensive application in image classification and medical image analysis endeavors[1].

In the present binary classification task, the original 1000-class fully connected layer of ResNet50 was substituted with a new linear layer featuring two output nodes. This modification enabled the adaptation of the network from the ImageNet classification task to the recognition of tongue spots. The remainder of the network architecture was retained, thereby allowing the model to preserve its deep residual feature extraction capabilities while simultaneously learning task-specific patterns from tongue images.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

2.6 Transfer Learning Strategy

Transfer learning was a key component of this study. Instead of training the network from scratch, pretrained weights from the ImageNet dataset were used as model initialization[3][6]. This strategy enabled the model to inherit general visual representations, including low-level and mid-level features such as edges, textures, and color patterns, which are also useful in tongue image analysis.

After loading the pre-trained weights, the entire network underwent fine-tuning using the tongue image dataset. This process enabled the model to transition smoothly from general natural image recognition to the precise classification of tongue spots, enhancing training efficiency, hastening convergence, and boosting the model's overall effectiveness in the intended task.

2.7 Training Settings

The Adam optimizer was employed during model training due to its efficiency and stable convergence behavior in deep learning tasks. Adam integrates first-order and second-order gradient moment estimation, rendering it often more appropriate for medical image classification tasks with limited data compared to conventional optimization methods. The initial learning rate in this study was established at 0.001.

A cosine annealing learning rate schedule was employed during training. This approach progressively decreases the learning rate according to a cosine curve, enabling the model to explore the parameter space more freely in the initial training phases and subsequently converge more smoothly in later epochs. This scheduling strategy can help mitigate unstable oscillations and enhance the final convergence quality of the model.

The loss function employed in this study was cross-entropy loss, a standard choice for classification tasks. The batch size was established at 32, and the model underwent training for 20 epochs. These parameters were chosen to optimize the trade-off between computational expense and classification accuracy. Overall, the training configuration was straightforward yet effective, providing a stable foundation for comparing the two residual network models.

2.8 Evaluation Metrics

To comprehensively evaluate the model, we employed five-fold cross-validation. The dataset was partitioned into five subsets. During each iteration, four subsets were allocated for training, and one subset was reserved for validation. This procedure was reiterated five times, and the mean outcomes were documented. The explicit formulas are delineated below:

2.8.1 Accuracy

Accuracy refers to the proportion of correctly predicted samples among all samples, and it is used to measure the overall classification performance of a model. A higher accuracy indicates better overall predictive performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2.8.2 Precision

Recall refers to the proportion of actual positive samples that are correctly identified by the model, and it is used to measure the model’s ability to detect positive cases. A higher recall indicates fewer missed detection.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

2.8.3 Recall

Precision refers to the proportion of correctly predicted positive samples among all samples predicted as positive, and it is used to measure the reliability of positive predictions. A higher precision indicates fewer false positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

2.8.4 Macro F1

Macro-F1 score is the arithmetic mean of the F1 scores calculated separately for each class, and it is used to evaluate the overall balance between precision and recall across classes. A higher Macro-F1 score indicates better and more balanced classification performance.

$$\text{Macro F1} = \frac{1}{2} \sum_{c \in \{0,1\}} \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4}$$

In these metrics, TP indicates that the model predicts a sample as positive when it is indeed positive. TN signifies that the model predicts a sample as negative when it is actually negative. FP denotes that the model predicts a sample as positive despite it being negative, while FN indicates that the model predicts a sample as negative even though it is positive.

The primary evaluation metrics employed were accuracy and macro-F1 score. Accuracy assessed the overall proportion of correctly classified samples. In contrast, the macro-F1 score evaluated the balance between precision and recall across classes, serving as a valuable complement to accuracy, particularly when class distributions are not perfectly equal.By combining cross-validation with multiple evaluation metrics, the study was able to provide a more comprehensive and statistically robust analysis of model performance.

3. Experimental Results

3.1 Experimental Environment

All experiments were conducted using Python 3.11.14 and the PyTorch framework. GPU acceleration facilitated faster training and enhanced computational efficiency. Additionally, several supporting libraries were utilized, including torchvision for image transformation and pretrained model loading, scikit-learn for calculating evaluation metrics, and matplotlib for visualizing results. Collectively, these tools established a comprehensive technical environment for data processing, model training, evaluation, and analysis.

3.2 Overall Performance

The experimental results showed that the transfer learning-based ResNet50 model performed well in the tongue spot recognition task. Under five-fold cross-validation, the model achieved an average accuracy of 97.48% with a standard deviation of 0.53%. This result indicates that the model had strong overall classification ability and stable performance across different validation folds.In terms of macro-F1 score, ResNet50 achieved 53.84% with a standard deviation of 6.64%. Although this value was lower than the accuracy, it still suggests that the model had a reasonable ability to balance precision and recall across classes. Taken together, these results show that ResNet50 can effectively extract discriminative features from tongue images and perform well on the tongue spot binary classification task[2][4][5][7].

3.3 Cross-Validation Results

By our experiments,we get our results.Please see Table 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

Table 1. Cross-Validation Results

Fold	Accuracy	Macro F1
Fold 1	97.63%	49.40%
Fold 2	96.56%	49.12%
Fold	**Accuracy**	**Macro F1**
Fold 3	97.27%	66.45%
Fold 4	97.98%	54.75%
Fold 5	97.98%	49.49%
Mean±std	**97.48% ± 0.53%**	**53.84% ± 6.64%**

A detailed examination of the five-fold validation results reveals that model performance remained relatively stable. The accuracies reported for the five folds were 97.63%, 96.56%, 97.27%, 97.98%, and 97.98%, respectively. These values fell within a narrow range, indicating that the proposed method exhibits robust performance across different data splits. In contrast, the macro-F1 scores for the five folds were 49.40%, 49.12%, 66.45%, 54.75%, and 49.49%, respectively. Compared to the accuracy results, these values displayed greater variability. This observation suggests that, although the model achieved high overall accuracy, class-wise performance was influenced by variations in sample distribution across the folds. Nonetheless, the overall trend reinforces the effectiveness of the proposed method.

3.4 Result Summary

To further evaluate the training process and optimization performance of the proposed ResNet50 model in this study, it is necessary to analyze the variation of the loss function during training.

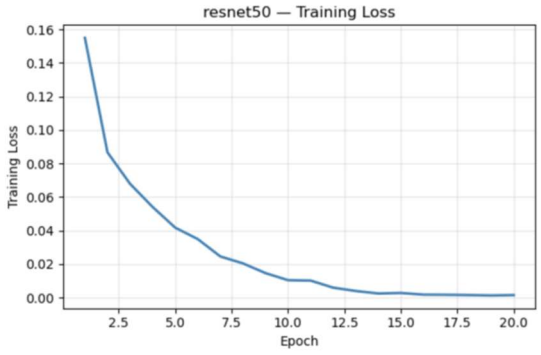

Figure 1. ResNet50 Training Loss Curve

The experimental results illustrate the efficacy of the proposed ResNet50-based approach for identifying tongue spots in real clinical images. This method exhibited high average accuracy, consistent cross-validation performance, and outperformed the baseline model, indicating that combining deep residual learning with transfer learning offers a practical solution for automated tongue feature analysis. Discrepancies between accuracy and macro-F1 scores suggest that the task remains partially unresolved. While the model excels in overall classification, enhancing the balanced

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

recognition of positive and negative samples requires further investigation. Nevertheless, these initial findings demonstrate promising practical utility and establish a robust experimental foundation for future optimization and expansion.

4. Discussion

4.1 Interpretation of Results

The strong performance of ResNet50 may be explained by several factors. First, the dataset size was relatively sufficient for a binary classification task, and the Spot label had a comparatively balanced class distribution. Second, transfer learning provided strong initial feature representations, allowing the model to converge faster and learn meaningful visual patterns more effectively. Third, data augmentation improved robustness by increasing variation in the training data.

At the same time, the gap between accuracy and macro-F1 indicates that the classification performance was not fully balanced across classes. This may be related to the complexity of tongue spot appearance. In clinical images, spots may vary greatly in size, shape, and color, and some cases may be visually confused with other local tongue abnormalities. These factors increase the difficulty of positive sample recognition and influence the balance of precision and recall.

4.2 Advantages of the Proposed Method

The proposed method has several advantages. It uses a mature and powerful deep residual network with proven effectiveness in image recognition. It also benefits from transfer learning, which reduces the dependence on massive labeled medical datasets and improves training efficiency. In addition, the method is relatively simple to implement and can be adapted to other tongue feature recognition tasks in future studies.Another advantage is that the experiments were conducted on real clinical data rather than only on small laboratory datasets. This improves the practical relevance of the results and suggests that the method has potential for future application in intelligent TCM-assisted diagnosis systems.

4.3 Limitations of the Study

Despite the promising results, this study still has several limitations. First, it focused only on the binary recognition of the Spot label and did not consider the simultaneous recognition of multiple tongue features. Second, although the overall accuracy was high, the macro-F1 result suggests that classification balance still needs improvement. Third, the current study mainly adopted a standard deep residual network and did not explore more advanced mechanisms such as attention modules or class-balanced optimization strategies.Future studies may address these limitations by introducing attention mechanisms, using weighted loss functions, or extending the current framework to multi-label tongue image recognition. External validation with data from other clinical centers may also help further verify the generalization ability of the model.

5. Conclusion

This study presents an automatic recognition method for tongue spot features utilizing ResNet50 and transfer learning. A binary classification model was developed and assessed using 5,371 authentic clinical tongue images, employing five-fold cross-validation. The findings revealed that the proposed method attained an average accuracy of 97.48% (±0.53%) and a macro-F1 score of 53.84% (±6.64%), demonstrating robust recognition performance in tongue spot classification.

These findings illustrate the robust capability of deep residual learning in automating tongue image analysis, offering technical backing for intelligent Traditional Chinese Medicine (TCM)-assisted diagnosis. Subsequent research could expand the suggested framework to encompass more intricate tasks in recognizing tongue features and wider applications in clinical settings.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0036

Acknowledgments

The authors express their gratitude for the financial support received from North China University of Science and Technology. This research is also funded by the University through the Undergraduate Innovation and Entrepreneurship Training Program, with the project number is X2025138.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[2] Q. Liu, Y. Li, P. Yang, Q. Liu, C. Wang, K. Chen, and Z. Wu, "A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation," Digit. Health, vol. 9, p. 20552076231191044, 2023.

[3] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," BMC Med. Imag., vol. 22, no. 1, p. 69, Apr. 2022.

[4] T. Jiang, X. Guo, L. Tu, Z. Lu, J. Cui, B. Ma, L. Wang, J. Xu, and J. Xu, "Tongue image quality assessment based on a deep convolutional neural network," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–14, May 2021.

[5] Artificial intelligence in tongue diagnosis: classification of tongue lesions and normal tongue images using deep convolutional neural network," BMC Med. Imag., vol. 24, no. 1, p. 66, Mar. 2024.

[6] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine-tuning?" IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1299–1312, May 2016.

[7] W. Ding, Y. Huang, Y. Luo, Y. Wang, M. Geng, Q. Zhao, and W. Zhong, "Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark," Comput. Struct. Biotechnol. J., vol. 18, pp. 973–980, 2020.