Swin-ResNet: Research and Implementation of a Tooth-Marked Tongue Classification Method Combining ResNet-50 and Swin Transformer

Conference Paper · April 2024

DOI: 10.1145/3644116.3644271

Xinshen Zhao

1 PUBLICATION **0** CITATIONS

Xiangyong Kong

University of Shanghai for Science and Technology

14 PUBLICATIONS **43** CITATIONS

Swin-ResNet: Research and Implementation of a Tooth-Marked Tongue Classification Method Combining ResNet-50 and Swin Transformer

Xinshen Zhao
School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
17521018375@163.com

Bo Zhao
School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
3313862514@qq.com

Qiyuan Zhang
School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
1153656619@qq.com

Rong Li
School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
lirong8882022@163.com

Xiangyong Kong*
School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
kxy@usst.edu.cn

Ping Wang*
Department of Nephrology and Rheumatology, Shanghai Children's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
wangping@shchildren.com.cn

ABSTRACT

Tongue diagnosis is an integral part of diagnosis by observation in Chinese medicine, and the tooth mark feature in the tongue image is an important objective indicator for the diagnosis of spleen deficiency. In order to solve the problem that the features of tooth marks are difficult to be recognized by naked eyes and reduce the error of subjective judgment, this paper designed a deep learning-based classification model for detecting and classifying tooth-marked tongues. By embedding the Swin-T module within the Resnet residual structure, we proposed and implemented a neural network integrating Resnet and Swin-Transformer (Swin-Resnet), which first down-sampled the input image by 7×7 convolution and max pooling, and then subsequently went through a stack of the improved Swin-Resnet modules for feature extraction, and finally a fully connected layer was utilized to complete high-precision classification of tooth-marked tongues. Meanwhile, we improved the proposed Swin-Resnet in terms of model running speed by transforming the convolutional layer of the original Resnet residual block into a convolutional block with smaller parameters and fewer network parameters while ensuring the network performance. Comparison experiments with other algorithms were conducted with sample datasets and the results of the experiments were evaluated, and Swin-Resnet achieved an average accuracy of 0.9959 for the three classifications, and the accuracies of 0.9832, 0.9792, and 0.9890 for each type of tooth-marked tongue classifications of Lightly Tooth, No Tooth, and Severe Tooth, respectively. The experiments showed that the classification method can identify dentition features with higher accuracy, which was important for improving the accuracy of objectivized analysis of tongue diagnosis and practical application in healthcare.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision problems; Object recognition.

KEYWORDS

Tongue Classification, Deep Learning, Tooth-Marked Tongue, Residual Block, Swin Transformer

ACM Reference Format:
Xinshen Zhao, Bo Zhao, Qiyuan Zhang, Rong Li, Xiangyong Kong*, and Ping Wang*. 2023. Swin-ResNet: Research and Implementation of a Tooth-Marked Tongue Classification Method Combining ResNet-50 and Swin Transformer. In *2023 4th International Symposium on Artificial Intelligence for Medicine Science (ISAIMS 2023), October 20–22, 2023, Chengdu, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3644116.3644271

1 INTRODUCTION

The practice of Traditional Chinese Medicine (TCM) has evolved over 2500 years, forming a comprehensive medical system that integrates diagnosis, treatment, and prognosis. Tongue diagnosis is a diagnostic method in traditional Chinese medicine (TCM). By observing features such as tongue texture, tongue coating, and tongue shape, TCM doctors can determine the functional status of each of the patient's internal organs, monitor the progression of the disease and the effectiveness of the treatment, and provide additional signs to help the doctor make a more accurate diagnosis, which can help in the prevention, diagnosis, and treatment of diseases [1].

As part of tongue diagnosis, tooth-marked tongue refers to an abnormal tongue pattern characterized by teeth marks [2]. As the name suggests, it refers to the traces of teeth that can be seen on the edge of the tongue, usually due to the tongue body is fat and suffered

ISAIMS 2023, October 20–22, 2023, Chengdu, China

ACM ISBN 979-8-4007-0813-8/23/10
https://doi.org/10.1145/3644116.3644271

from the edge of the teeth for a long time pressure and form [3, 4]. Tooth-marked tongue is commonly seen in patients with spleen deficiency and are an important objective diagnostic indicator for diagnosing spleen deficiency [5]. Recognition of tooth-marked tongue is very helpful for the diagnosis and treatment of traditional Chinese medicine. Traditional diagnosis of tooth-marked tongue is faced with the following challenges: the shape and distribution of tooth marks vary greatly depending on the individual, the variety of tooth marks is difficult to recognize, and the diagnosis is very much dependent on the doctor's subjective experience, therefore, there is an urgent need for an objective and accurate diagnostic method.

Deep learning applications in tongue diagnosis primarily focus on tongue segmentation and extraction of tongue features for classification. TCM regards the tongue as a reflection of the heart, with various scholars exploring the relationship between tongue features and diseases as well as prescriptions [6, 7]. For reducing inter-doctor variability, Li et al. [8] proposed a CNN-based tongue image classification framework, incorporating an improved facial landmark detection method and U-Net for tongue segmentation, and utilizing ResNet-34 as the backbone network to extract features and perform classification. Sun et al. [9] improved the loss function based on the Inception-ResNet-V1 network, maximizing the feature distance between non-similar samples while minimizing the feature distance between samples of different classes, increasing the depth and width of the network for better feature extraction and reduced parametric complexity. Chun-Mei Huo et al. [10] introduced an efficient tongue image classification method by applying a pre-trained convolutional neural network, AlexNet, for the classification of tongue fissures and spots, enhancing the efficiency of tongue image classification. Song Chao et al. [11] employed a multi-branch network structure for tongue image classification, utilizing different transfer learning strategies in each branch network to improve the classification performance. Zhai Pengbo et al. [12] introduced an attention mechanism into tongue image classification to reduce sample noise, and established individual branch networks for each specific feature, constructing a simple network structure to improve the efficiency of tongue image classification.

This paper addresses the challenges of diverse and difficult-to-identify tooth mark shapes and sizes, the unsatisfactory performance of traditional tongue body classification, and the limited availability of tongue image data for gold standard. To tackle these issues, we propose a neural network-based tongue image feature recognition and classification method, called Swin-Resnet, which integrates Swin Transformer and Resnet. The research findings demonstrate that our classification method, by combining the ideas of W-MSA, SW-MSA, and residual connections, as well as improving the residual blocks of ResNet to reduce the number of convolutional neural network parameters, can effectively detect and identify tooth mark features and classify their severity compared to other algorithms. The remaining sections of this paper are organized as follows: Section 2 introduces the related work, Section 3 presents the theoretical framework and structure of the proposed model, Section 4 showcases the experimental results and corresponding analysis, and finally, Section 5 concludes the paper and provides future prospects for further research.

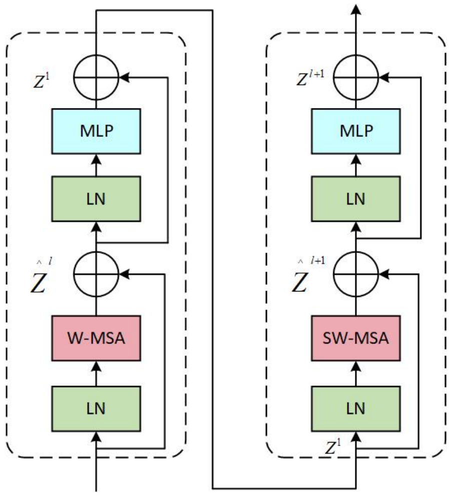


Figure 1: Swin Transformer block structure

2 RELATED WORK

2.1 Swin Transformer

In 2021, Liu et al. introduced the Swin Transformer model, which incorporates hierarchical feature maps and the shift window attention mechanism. This approach involves dividing the input image into smaller patches and processing them at multiple levels to extract and recombine image features through progressive encoding and decoding [13]. This hierarchical strategy allows the model to effectively capture features at different scales and resolutions, enabling better contextual understanding.

The Swin Transformer consists of four stages to construct feature maps of different sizes. Except for Stage 1, which involves a Linear Embedding layer, the remaining three stages use a Patch Merging layer for downsampling. These stages are followed by a stack of W-MSA (Window-based Multi-head Self-Attention) and SW-MSA (Shifted Window-based Multi-head Self-Attention) modules. In this paper, we utilize both the Patch Merging module and the Swin Transformer Block to form the Swin-T module. The structure of the Swin Transformer Block is illustrated in Figure 1.

2.1.1 Hierarchical Feature Maps. In the Swin-T module, the hierarchical feature maps are obtained through a downsampling process applied to the dental impression tongue image, analogous to the methods used in convolutional neural networks. This downsampling procedure effectively reduces the size of the feature maps, resulting in feature maps at 4 times, 8 times, and 16 times downsampling ratios. The hierarchical nature of these feature maps becomes evident as the depth of the feature layers increases, leading to progressively reduced height and width dimensions of the feature maps, as visually represented in Figure 2.

Furthermore, an essential concept within the Swin-T module is Windows Multi-Head Self-Attention (W-MSA). W-MSA introduces a computationally efficient self-attention mechanism by partitioning the image into windows of size 7x7. Traditional Multi-Head Self-Attention (MSA) often incurs a significant computational overhead. However, the incorporation of W-MSA effectively addresses the

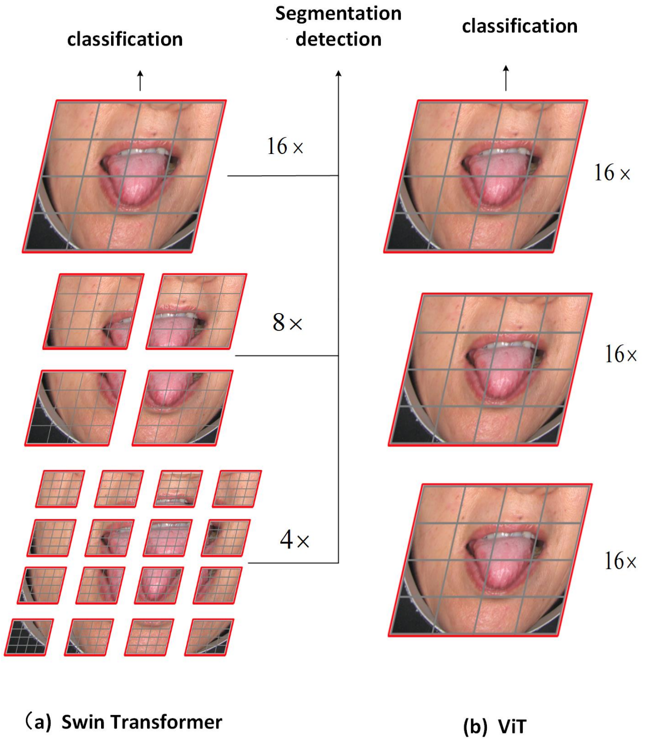


Figure 2: Hierarchical Feature Maps of Swin Transformer(a) and ViT(b)

computational complexities associated with MSA. The self-attention mechanism can be defined using the following formula:

<?TeX Attention $(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{dm}}}\right)\mathrm{V}$?>

However, due to the occlusion and non-interaction of information between the segmented windows, the context of the image cannot be connected, which leads to poor modeling, for which the Shifted Windows Multi-Head Self-Attention (SW-MSA) module is introduced.

2.1.2 Shifted Windows Multi-Head Self-Attention. SW-MSA serves as a pivotal constituent within the Swin-T module, contributing to its architectural design. SW-MSA leverages the concept of shifted windows to facilitate inter-window information propagation. A visual analysis elucidates the disparity between the W-MSA and SW-MSA modules. Specifically, the left panel demonstrates the W-MSA module, while the right panel showcases the SW-MSA module, as depicted in Figure 3. Notably, the window arrangement in the left panel exhibits a distinct shift, wherein each window undergoes a displacement by M/2 units towards the bottom-right corner, with M denoting the window size. This deliberate displacement engenders crucial interconnections between adjacent windows. It entails the model's sequential execution of self-attention computation for both the original and shifted windows. The introduction of cross-links among windows yields a noteworthy enhancement in the overall performance of the network architecture.

2.2 Resnet

Deep Residual Network (ResNet) was proposed by He et al. to solve the problem of vanishing gradients during network training caused by the deeper neural networks. The key innovation of ResNet lies in the introduction of residual blocks, which incorporate skip connections to directly propagate the original information to subsequent layers. This facilitates the exchange of features across different depths and enables effective gradient flow to earlier layers, thereby mitigating the performance degradation associated with increasing network depth. Consequently, ResNet alleviates the issue of gradient vanishing, accelerates the training speed of neural networks, and significantly enhances their image classification capabilities [14, 15].

3 PROPOSED DEEP NEURAL NETWORK

3.1 Improved Swin-Resnet module

Since the Resnet residue block has two structures: the basic structure of 1×1, 3×3 convolution and the bottleneck structure of 1×1, 3×3, 3×3, the Swin-T Block is embedded into two different network structures, as shown in Figure 4.

With the progressive deepening of models, the parameter count in convolutional neural networks (CNNs) increases exponentially, rendering them susceptible to overfitting, wherein the loss function reaches a minuscule value while the generalization performance deteriorates substantially. Put simply, when an excessive number of features are present, the derived hypothesis may fit the training set effectively yet fail to extend its applicability to novel examples. Moreover, the augmented parameter count exacerbates computational complexity, escalates memory consumption, and diminishes training efficiency. Achieving widespread adoption and dissemination of deep learning algorithms hinges on minimizing the number of network parameters while harnessing the computational prowess of GPUs. Hence, apart from data augmentation techniques applied to the training dataset, an alternative remedy entails parameter reduction in CNNs. In this study, we propose substituting the conventional ResNet residual blocks with compact modules encompassing 1×3, 3×1, 1×3, and 3×1 residual structures [16] (Figure 5).

To calculate the parameters of the convolutional layer, the calculation formula is as follows:

$$T\,(conv) = C \times H \times W \times K \tag{1}$$

where, C×H×W is the size of the convolution kernel, and K is the number of the convolution kernel.

$$T\,(basic) = (64 \times 3 \times 3 \times 64) \times 2 = 18 \times 64^2 \tag{2}$$

$$T\,(bottleneck) = 256{\times}1{\times}1{\times}64{+}64{\times}3{\times}3{\times}64{+}64{\times}1{\times}1{\times}256 = 17{\times}64^2 \tag{3}$$

$$T\,(small) = (64 \times 1 \times 3 \times 64) \times 2 + (64 \times 3 \times 1 \times 64) \times 2 = 12 \times 64^2 \tag{4}$$

Compared with the original Basic and Bottleneck residual structures, the number of parameters of this structure is reduced by 33.3% and 29.4%, respectively. It reduces the number of network parameters and improves the training efficiency while keeping the feature extraction capability unchanged.

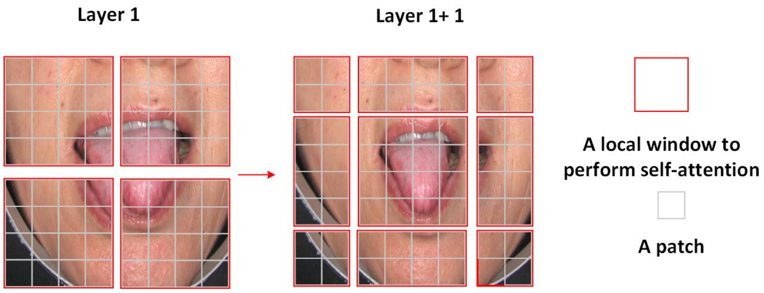


Figure 3: W-MSA modules and SW-MSA modules

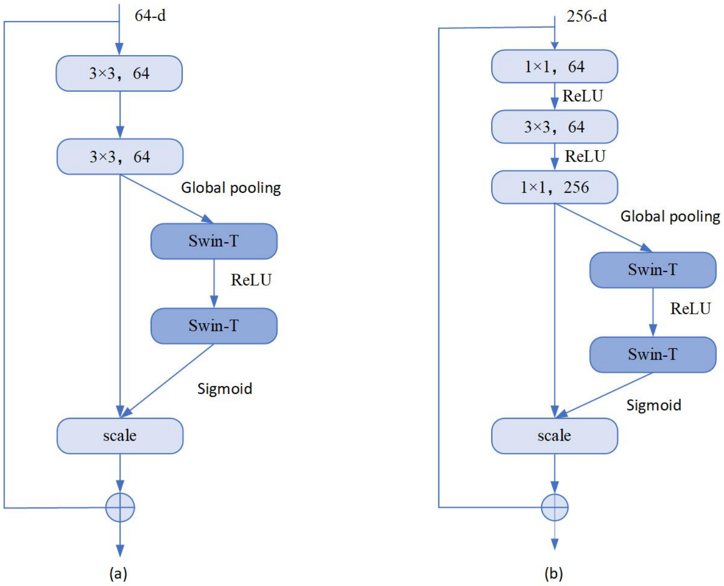


Figure 4: (a)Swin-T embedded in Basic structure (b)Swin-T embedded in Bottleneck structure

3.2 Improved Swin-Resnet network architecture

In this study, the Swin-T module is embedded into the improved Resnet residual block to form the improved Swin-Resnet module, and the Swin-Resnet module is stacked to form the Swin- Resnet50 classification network, forming a chimeric neural network (Swin-Resnet) that integrates Resnet and Swin Transformer as shown in Figure 6.

In the present study, a downsampling technique was employed to resize the original input images to a size of 224×224, which served as the network input. Using a combination of 7×7 convolutions and max pooling operations, the feature map size was reduced to 56×56 with a dimension of 64. Subsequently, the network architecture was divided into four stages: Stage 1, Stage 2, Stage 3, and Stage 4. Our network structure comprised an improved stack of Swin-ResNet modules, consisting of 3, 4, 6, and 3 modules, respectively. Following Stage 1, the feature map size was 56×56 with a dimension of 256. After each subsequent stage, the feature map size was halved while the dimension doubled. Finally, the feature map underwent average pooling and flattening processes, projecting the multidimensional features to a one-dimensional space. Three types of classification results were obtained using fully connected layers.

4 EXPERIMENTAL METHODS

4.1 Data preprocessing and enhancement

Our research group collaborated with ShuGuang Hospital, affiliated with Shanghai University of Traditional Chinese Medicine, to collect a dataset of 1,500 clinical tongue images. The images were captured using specialized acquisition equipment, following established collection standards. To enhance the accuracy of subsequent research on tongue image features, the annotation process involved the assistance of multiple professional practitioners of traditional Chinese medicine, ensuring consistency in the annotations. For this study, we selected a total of 400 cases of tongues without tooth marks and 756 cases of tongues with tooth marks, with a resolution of 2,816×2,112. To account for the varying severity of tooth marks, we categorized the tongues with tooth marks into three types for classification: tongues without tooth marks, tongues with lightly tooth marks, and tongues with severe tooth marks, referred to as

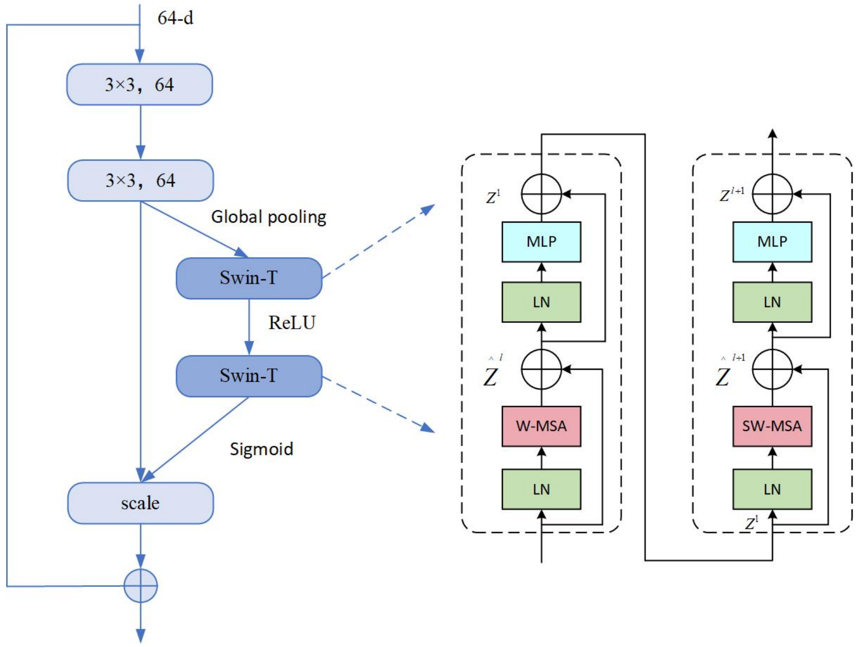


Figure 5: Improved Swin-Resnet module structure diagram

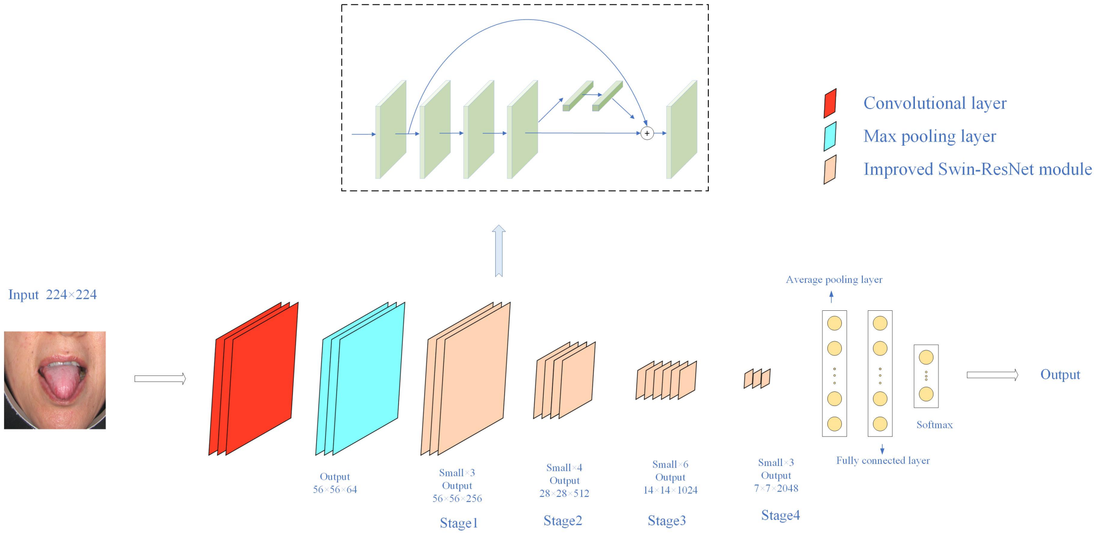
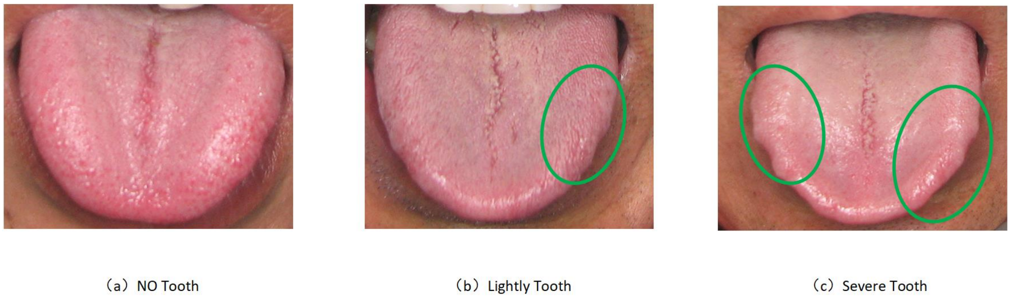


Figure 6: Improved Swin-Resnet network architecture

No Tooth, Lightly Tooth and Severe Tooth, as illustrated in Figure 7.

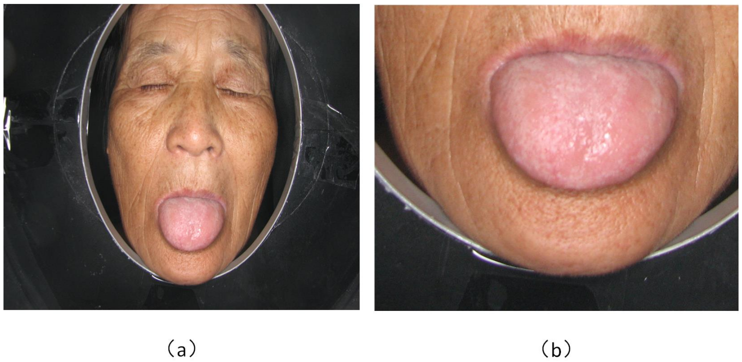
The selected images were initially standardized in their naming conventions. Following this, a pre-trained tongue segmentation network model was utilized to crop the acquired tongue images, as depicted in Figure 8. This step served two key purposes: firstly, it aimed to minimize the influence of facial regions present in the images, thus focusing solely on the tongue area. Secondly, given the high original resolution of the images (2,816×2,112), downsizing them was necessary to alleviate computational complexities associated with training and reduce the training duration. The resultant images were resized to 1016×837, allowing them to be directly inputted into the neural network.

To effectively exploit the limited dataset and achieve comparable significance levels to those of larger datasets, as well as to enhance the robustness and generalizability of the trained models, data augmentation techniques were employed. These techniques involved manipulating the position of tooth marks on the tongues, introducing noise, and applying operations such as rotation, translation, and flipping to the images. During the data augmentation process,



（a）NO Tooth （b）Lightly Tooth （c）Severe Tooth

Figure 7: Recognition of different degrees of tooth marks



（a） （b）

Figure 8: (a)Original image (b)Processed image

a careful balance was maintained amongst the three categories to ensure an equivalent distribution of tongue image features across the training, validation, and testing sets.

4.2 Evaluation methodology and performance indicators

The evaluation of model performance holds paramount importance in this study. To assess the effectiveness of the models, six widely used classification model evaluation metrics were employed: Accuracy, Precision, Recall, Specificity, AUC (Area Under the ROC Curve), and AUPR (Area Under the Precision-Recall Curve). These metrics provided comprehensive insights into the model's performance. In the context of discriminating between tongues without tooth marks and tongues with tooth marks, four distinct scenarios were considered based on the classifier's predictions on the test dataset: True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP). For example, when the true label indicated tongues with tooth marks and the predicted label also identified it as such, it was classified as a true positive. However, if the predicted label incorrectly classified it as tongues without tooth marks, it was considered a false negative. Similarly, when the true label indicated tongues without tooth marks and the predicted label accurately categorized it as such, it was classified as a true negative. Conversely, if the predicted label falsely labeled it as tongues with tooth marks, it was designated as a false positive. The specific formulas utilized for calculating these metrics are presented below for reference:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

4.3 Experimental results and analysis

In order to assess the effectiveness of the Swin-Resnet model for classifying the tooth-marked tongue dataset, this study conducted a comparative analysis of its performance against other classification models, including Swin Transformer, Efficientnet-v2 [17], and Regnet [18, 19]. The evaluation of these models was based on their accuracy on the test dataset, which served as a measure of their classification performance. The precision comparison and performance curves of Swin-Resnet and the alternative models can be observed in Figure 9.

The experimental findings clearly indicate that Swin-Resnet outperforms the other models in terms of tooth-marked tongue classification, exhibiting superior performance and greater stability in the training process. In terms of model performance, Swin-Resnet demonstrated the highest level of accuracy, achieving an average precision of 0.9959 for the three categories of tooth-marked tongue.

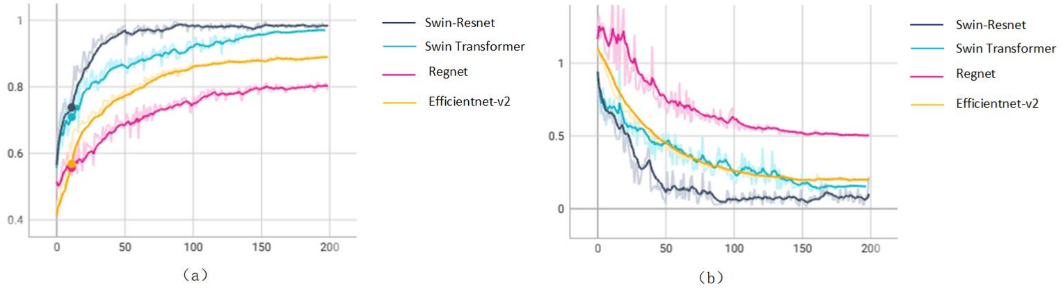


Figure 9: (a)Variation of model's accuracy (b)Changes in model's loss value during training

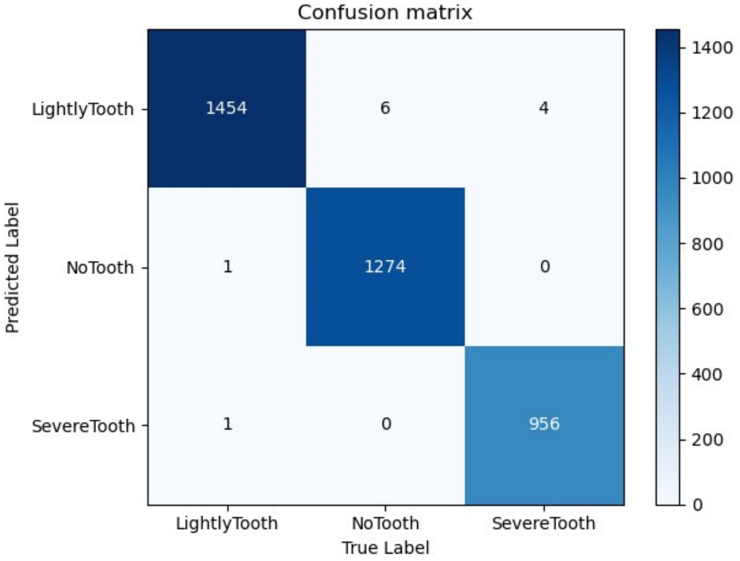


Figure 10: Confusion matrix for the prediction results of the Swin-Resnet algorithm

Table 1: Results of Swin-Resnet's evaluation of different severity classifications of tooth-marked tongue

Tooth-Marked Tongue	Precision	Recall	Specificity	AUC	AUPR
LightlyTooth	0.983	0.989	0.996	0.989	0.918
NoTooth	0.979	0.985	0.989	0.974	0.881
SevereTooth	0.989	0.986	0.986	0.990	0.965

Additionally, Swin-Resnet stands out for its lightweight architecture, with fewer parameters compared to the other models. In comparison, Efficientnet-v2 exhibited relatively weaker capability in recognizing tooth-marked tongue images, resulting in an accuracy of only 0.8038 and highlighting sub-optimal model performance. Similarly, Regnet showed moderate performance in the classification of tooth-marked tongue images, with an accuracy of 0.8796. Compared to Swin-Resnet, Swin Transformer has a larger number of parameters, but it performs slower and performs poorly in the tooth-marked tongue classification task.

The evaluation of our tooth-marked tongue classification model was conducted using a dedicated test dataset, and the experimental results were used to generate a confusion matrix (Figure 10). By employing evaluation metrics, we thoroughly assessed the model and obtained evaluation results for different severity levels of tooth-marked tongue, as presented in Table 1.

5 CONCLUSIONS AND PROSPECT

This study applied the latest deep learning techniques to the task of tooth-marked tongue feature extraction and classification. Addressing challenges such as the diverse and difficult-to-recognize shapes of tooth marks and limited tongue image data, we designed an image classification method named Swin-Resnet, which combined Resnet and Swin Transformer. By incorporating ideas like residual connections and Shifted Windows Multi-Head Self-Attention, we overcame the limitations of traditional convolutional neural networks in terms of receptive fields, enabling effective local and global feature interactions. Furthermore, we optimized Swin-Resnet in terms of parameter storage, memory operations, and computational complexity. We replaced the original Resnet residual blocks with a residual structure composed of four smaller modules, reducing the model's parameters and improving its runtime speed. Comparative experiments demonstrated that our Swin-Resnet outperformed other models in the task of classifying the severity of tooth-marked tongues. Lastly, we evaluated the classification performance of the model on different severity levels of tooth-marked tongues using a confusion matrix and various evaluation metrics, obtaining favorable results.

In the future, this research will continue in the following directions:

1. Due to limitations in the types and quantity of Traditional Chinese Medicine tongue diagnosis images obtained, the current sample size of the tongue image dataset established in this paper is still insufficient, with limited coverage of tongue types, and it only focuses on the classification of tooth-marked tongues. We plan to collaborate with more Traditional Chinese Medicine institutions to expand the variety of tongue image datasets, extending the research to more tongue abnormality tasks, thus better assisting doctors in diagnosis and treatment.

2. Tongue diagnosis, due to its convenience and non-invasiveness, is considered an ideal choice for remote diagnosis. Based on the algorithm's portability, we will attempt to package the model and deploy it on mobile devices to provide real-time diagnosis and generate diagnostic reports and recommendations for users' uploaded tongue images, thereby improving the health status of sub-healthy populations.

3. Creating a tongue image dataset requires a significant amount of manual labeling work. Therefore, we plan to research how to utilize semi-supervised or unsupervised learning to achieve AI-based automated diagnosis, which will save considerable time and human resources while enhancing the model's generalization ability.

4. The Swin-Resnet classification model has high accuracy for the task of tooth-marked tongue classification, and tongue images often have correlation with other medical images, such as MRI, CT, etc., so we can try to extend it to multi-modal medical image analysis, combining different types of medical image data for comprehensive analysis and diagnosis, to improve the comprehensive effect of disease detection and analysis.

REFERENCES

[1] Wu Xin, Xu Hong, Lin Zhuosheng, *et al.* Research Summary of Deep Learning in Tongue Image Classification. Journal of Computer Science and Exploration, 2023, 17(02): 303-323.

[2] Rui Yingying, Kong Xiangyong, Liu Yanan, *et al.* Teeth Mark Tongue Image Recognition Based on Mask Scoring R-CNN. Chinese Journal of Medical Physics, 2021, 38(04): 523-528.

[3] Pan Ciming, Zhu Peichao, SISHIR Sharma, *et al.* Discussion on the Origin and Clinical Significance of Teeth Mark Tongue. Shaanxi Journal of Traditional Chinese Medicine, 2021, 42(09): 1267-1269.

[4] Yang Jiaxin, Han Dong, Dong Xinming, *et al.* Objective Study of Tooth Marked Tongue in Traditional Chinese Medicine Based on Morphological Feature Extraction. Journal of Laser & Optoelectronics Progress, 2022, 59(11): 365-373.

[5] Liang Rong, Wang Zhaoping, Jin Fenfang. Correlation Study between Teeth Marked Tongue and Spleen Deficiency Syndrome Symptoms in Health Examination Subjects. Chinese Journal of Medicine, 2003, 18(7): 400-403. DOI: 10.3969/j.issn.1673-1727.2003.07.006.

[6] ZHOU H, HU G Q, ZHANG X F. Study on TCM constitution classification method based on tongue image depth feature fusion[J].Beijing Biomedical Engineering, 2020, 39(3): 5-10.

[7] LI X L, YANG D, WANG Y, *et al.* Automatic tongue image segmentation for real-time remote diagnosis[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine, 2019: 409-414.

[8] L ,ZHANG ZZHU X, et a. Automatic classification framework of tonoue feature based on convolutonal neura networksiJ. Micromachines. 2022. 13(4):501.

[9] Sun Meng, Zhang Xinfeng. Research on Tongue Image Classification Method Based on Triplet Loss. Beijing Biomedical Engineering, 2020, 39(02): 131-137.

[10] Huo C M, Zheng H, Su H Y, *et al.* Tongue shape classification integrating image preprocessing and Convolution Neural Network[C],2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS). 2017:42-46.

[11] Song Chao. Research on Tongue Image Feature Classification Method Based on Deep Transfer Learning. (Master's thesis, Nanjing University of Finance and Economics, 2021). DOI: 10.27705/d.cnki.gnjcj.2020.000323.

[12] Zhai Pengbo, Yang Hao, Song Tingting, *et al.* Multi-stage Tongue Image Classification Algorithm with Attention Mechanism Fusion. Journal of Computer Engineering and Design, 2021, 42(06): 1606-1613. DOI: 10.16208/j.issn1000-7024.2021.06.014.

[13] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[14] He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. European conference on computer vision. 9908:630-645.https://doi.org/10.1007/978-3-319-46493-0 38.

[15] He K , Zhang X , Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). https://doi.org/10.1109/CVPR.2016.90.

[16] Jiang Y, Chen L, Zhang H, Xiao X (2019) Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. Plos One. 14(3).https://doi.org/10.1371/journal.pone.0214587.

[17] Tan M. X., Le Q. V. (2021). "EfficientNetV2: Smaller models and faster training," in International conference on machine learning (Vienna, Austria: Proceedings of the 38th International Conference on Machine Learning, PMLR), vol. 139. , 7102–7110.

[18] Radosavovic I *et al.* Designing Network Design Spaces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 10425-10433. https://doi.org/10.48550/arXiv.2003.13678.

[19] Mahbub M, Biswas M, Miah AM, Shahabaz A, Kaiser MS (2021) Covid-19 detection using chest x-ray images with a regNet structured deep learning model. In: Proceedings of the international conference on applied intelligence and informatics (Cham: Springer), Nottingham, pp 358–370. https:// doi. org/ 10. 1007/ 978 3-030-82269-9_28