Sensors and Materials, Vol. 38, No. 5 (2026) 2455–2475
MYU Tokyo

S & M 4447

Explainable Vision-sensing Signal Processing for Tongue-coating Assessment in Healthcare Monitoring: A Two-stage U-Net Segmentation and Feature-attention Classification Framework

Chenwei Zhang,[1*] Qi Qiao,[1] Zhiheng Pan,[1] and Huiying Hu[2,3]

[1]Faculty of Computer and Communication Engineering, Jiangsu Vocational College of Electronics and Information, No. 3 Meicheng East Road, Huaian 223003, China
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Jalan Ilmu 1/1, Shah Alam 40450, Malaysia
[3]Faculty of Information Engineering, Jiujiang Vocational University, No. 88 Lianxi Road, Jiujiang 332000, China

(Received December 18, 2025; accepted January 23, 2026)

Keywords: vision-based medical sensing, healthcare monitoring, tongue-coating assessment, U-Net segmentation, explainable deep learning

In intelligent healthcare monitoring, vision-based sensing offers a practical route for non-contact medical diagnostics, yet tongue-coating assessment in traditional Chinese medicine (TCM)-related clinical settings is often hindered by background interference, illumination variation, and limited interpretability. In this study, we present an explainable vision-sensing signal-processing framework built as a two-stage segmentation–classification pipeline. First, a U-Net model segments the tongue region from RGB tongue images to suppress non-tongue artifacts and standardize the sensing signal. Second, encoder features learned during segmentation are transferred to a feature-attention classifier to recognize four coating textures (thin, thick, peeled-like, and mirror-like). To support trustworthy clinical use, gradient-weighted class activation mapping (Grad-CAM) is employed to visualize discriminative regions behind predictions. Experiments demonstrate improved robustness and accuracy over a conventional VGG-based baseline, while providing interpretable evidence. The proposed method serves as a deployable component for vision-centric intelligent sensing and can be integrated into broader healthcare monitoring systems.

1. Introduction

Traditional Chinese medicine (TCM) tongue diagnosis is a clinically valuable method for assessing physiological and pathological conditions by observing the tongue body and coating. The appearance of the tongue coating—particularly its thickness and texture—reflects the functional state of the spleen, stomach, and internal organs. Although tongue inspection has been practiced for thousands of years, its clinical application still largely depends on subjective

*Corresponding author: e-mail: 31520181154413@stu.xmu.edu.cn
https://doi.org/10.18494/SAM6125

ISSN 0914-4935 © MYU K.K.

visual judgment by TCM practitioners. The lack of objectivity and inter-observer consistency limits the standardization of tongue-based diagnosis in modern medical settings.

In recent years, the rapid development of deep learning technologies has provided new opportunities for the quantitative analysis of medical images.(1) Convolutional neural networks (CNNs) have achieved remarkable success in object recognition, lesion detection, and medical image segmentation.(2) However, their development in TCM tongue diagnosis remains relatively limited compared with applications in Western medical imaging. Most existing deep-learning studies on tongue analysis do not explicitly separate the tongue region from the irrelevant facial and environmental background, leading to noisy feature extraction and suboptimal classification performance.

To address this limitation, we propose a two-stage deep learning framework for tongue coating texture classification. In the first stage, a U-Net based segmentation model is trained to automatically extract the tongue region from raw tongue images. This preprocessing step ensures that only clinically relevant areas are fed into subsequent classification modules. In the second stage, the encoder portion of the trained U-Net is used as a feature extractor, and newly inserted fully connected layers are fine-tuned for coating texture classification. This feature-transfer strategy effectively leverages the hierarchical representation power of the U-Net encoder and results in significant improvements compared with conventional CNN baseline models.

Another key issue in the clinical adoption of deep learning systems is interpretability. To enhance model transparency, we employ Grad-CAM to visualize the most discriminative regions contributing to each classification decision. These visual explanations reveal activation patterns that correspond to TCM-defined areas of clinical interest, such as the tongue tip, center, sides, and root. By bridging modern deep learning interpretation with classical TCM theory, the proposed framework has the potential to assist practitioners in objective and reproducible diagnostic evaluation.

The main contributions of this study are summarized as follows.

(1) Improved U-Net encoder for texture-aware tongue analysis
We construct an improved U-Net segmentation backbone and provide a stage-wise architectural specification (Fig. 1 and Table 1) together with a hybrid segmentation objective (Sect. 2.2). Why it works: the hierarchical multi-scale encoding and skip fusion preserve coating-related micro-textures while maintaining anatomically consistent tongue boundaries, which reduces background-induced feature bias and provides texture-aware representations for downstream recognition. Evidence: the network design is explicitly documented in Table 1 and its segmentation performance is reported in Sect. 3.2.

(2) Two-stage segmentation–classification pipeline with encoder sharing
We propose a two-stage framework in which tongue region-of-interest (ROI) extraction is first performed by the improved U-Net, and the encoder is then reused as the feature extractor for four-class coating recognition with a lightweight fully connected head (Fig. 1; Sect. 2.4). Why it works: explicit tongue localization suppresses irrelevant background/illumination artifacts, while encoder sharing transfers pixel-level supervised representations to the classification task, yielding a compact model without relying on external CNN backbones such as VGG. Evidence: classification comparisons and the effectiveness of encoder-based feature extraction are presented in Sects. 3.3 and 3.4.

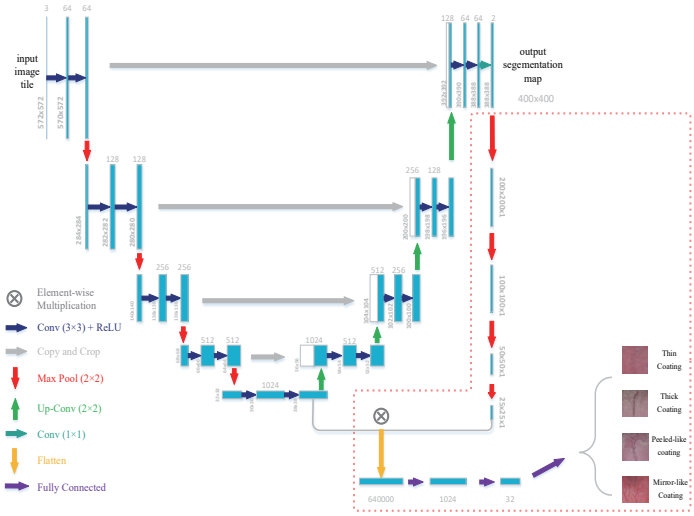

Fig. 1. (Color online) Overall architecture of the proposed two-stage framework. The pipeline first segments the tongue region using the improved U-Net and then performs coating classification using the shared encoder features and a lightweight fully connected head.

Table 1
Stage-wise specification of the proposed U-Net for tongue region segmentation (RGB input), consistent with Fig. 1.

Path	Stage	Feature map size	Operations (in order)	No. of output channels	Notes/Remarks
Input	—	400 × 400	Input tongue image	3	Standardized RGB input and resizing 400 × 400
Encoder	E1	400 × 400	Conv(3 × 3)→ReLU → Conv(3 × 3)→ReLU → MaxPool(2 × 2)	64	Capture fine texture/edge cues (thin and peeled patterns)
Encoder	E2	200 × 200	Conv(3 × 3)→ReLU → Conv(3 × 3)→ReLU → MaxPool(2 × 2)	128	Multi-scale encoding for coating texture variations
Encoder	E3	100 × 100	Conv(3 × 3)→ReLU → Conv(3 × 3)→ReLU → MaxPool(2 × 2)	256	Encode mid-level micro-textures and regional patterns
Encoder	E4	50 × 50	Conv(3 × 3)→ReLU → Conv(3 × 3)→ReLU → MaxPool(2 × 2)	512	Robust semantic cues under illumination variation
Bottleneck	B	25 × 25	Conv(3 × 3)→ReLU → Conv(3 × 3)→ReLU	1024	Global context representation (25 × 25 × 1024)
Decoder	D4	50 × 50	Up-Conv(2 × 2) → Concat(skip E4)→ Conv(3 × 3)+ReLU × 2	512	Up-conv refinement + skip fusion for boundary recovery
Decoder	D3	100 × 100	Up-Conv(2 × 2) → Concat(skip E3)→ Conv(3 × 3)+ReLU × 2	256	Recover mid-scale structures and contours
Decoder	D2	200 × 200	Up-Conv(2 × 2) → Concat(skip E2)→ Conv(3 × 3)+ReLU × 2	128	Restore fine details with skip cues
Decoder	D1	400 × 400	Up-Conv(2 × 2) → Concat(skip E1)→ Conv(3 × 3)+ReLU × 2	64	Sharpen final boundaries at full resolution
Output	–	400 × 400	Conv(1 × 1)	2	2-class pixel-wise map (tongue/background)

(3) Explainability via Grad-CAM for clinically meaningful visualization
We integrate Grad-CAM as a visualization module and provide a complete mathematical formulation and implementation procedure (Sect. 2.5). Why it works: class-discriminative activation maps reveal which spatial regions contribute most to each coating prediction, enabling the qualitative verification of whether the model attends to anatomically plausible tongue and coating regions rather than background cues. Evidence: representative activation maps and corresponding qualitative analyses are reported in Sect. 3.4.

2. Methods

In this section, we present the proposed two-stage tongue-coating assessment framework, including tongue region segmentation, encoder-sharing classification, and interpretability via Grad-CAM. The dataset description and experimental environment are reported in Sect. 3.1, followed by quantitative and qualitative evaluations in Sect. 3.

2.1 Framework overview of the proposed two-stage pipeline

As illustrated in Fig. 1, the proposed framework follows a two-stage design. In Stage I, an improved U-Net is trained to segment the tongue region and generate a tongue ROI, suppressing background interference and preserving anatomically consistent boundaries. In Stage II, the encoder of the segmentation network is reused as a feature extractor for coating classification, and a lightweight fully connected head predicts four coating categories. To improve interpretability, Grad-CAM is applied to the encoder features to visualize discriminative regions that contribute to each predicted class. This two-stage design explicitly decouples tongue localization from coating recognition, allowing the classifier to focus on coating-related texture cues while reducing sensitivity to illumination changes and background artifacts.

2.2 U-Net-based tongue region segmentation

A U-Net architecture was employed to automatically segment the tongue region from each raw image.[3] U-Net is widely used in biomedical image segmentation owing to its encoder–decoder structure with skip connections, which enables the extraction of both high-level semantic features and fine-grained spatial details. This property makes it particularly suitable for tongue images, where coating boundaries and edge contours need to be preserved.

In this study, the input to the U-Net model was the preprocessed RGB tongue image, and the network output was a pixel-wise probability map $P \in [0,1]^{H \times W}$ indicating the likelihood of each pixel belonging to the tongue region. A binary mask M was then obtained by thresholding P (e.g., $\tau = 0.5$), which was used to identify the tongue region. The encoder path performs multi-scale feature extraction through repeated convolution and downsampling, while the decoder path progressively restores spatial resolution using upsampling and concatenation with corresponding encoder features. These skip connections help maintain coating-related texture information during reconstruction.

To improve segmentation accuracy under varying illumination and background interference, a hybrid loss function combining binary cross-entropy (BCE) and Dice loss was adopted.(4) BCE ensures pixel-wise classification consistency, whereas Dice loss enhances region-level overlap, particularly in areas with fuzzy boundaries. Formally, let $P_i \in [0,1]$ and $g_i \in \{0,1\}$ denote the predicted probability and ground-truth label of pixel i, respectively, and let N be the total number of pixels. The BCE loss is defined as

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[g_i \log(p_i) + (1-g_i)\log(1-p_i)\right]. \tag{1}$$

The Dice loss is given by

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon}, \tag{2}$$

where ϵ is a small constant for numerical stability. The overall segmentation objective is

$$\mathcal{L}_{seg} = \lambda \mathcal{L}_{BCE} + (1-\lambda)\mathcal{L}_{Dice}, \tag{3}$$

where λ controls the balance between pixel-wise supervision and region-level overlap. In our experiments, $\lambda = 0.5$. The BCE term provides stable pixel-wise supervision and encourages accurate probability estimation for each pixel, which is beneficial under illumination variation and background interference. However, BCE may be sensitive to foreground–background imbalance and does not directly optimize region overlap. In contrast, Dice loss explicitly maximizes the overlap between the predicted mask and the ground truth, improving boundary consistency and robustness when the tongue region occupies a variable proportion of the image. Therefore, combining BCE and Dice leverages both pixel-level discrimination and region-level agreement, yielding more reliable tongue-region segmentation for images with fuzzy contours and specular highlights. The Adam optimizer was used with an initial learning rate of 1×10^{-4}, and training was performed until validation performance converged.(5)

The resulting segmentation masks were used to extract clean tongue regions, which were then fed into the subsequent classification module. This segmentation step helps eliminate irrelevant background information and ensures that the classifier focuses solely on coating-related visual patterns.

2.3 Improved U-Net architecture for tongue image analysis

Although the original U-Net architecture provides strong performance in biomedical segmentation, its standard configuration is not fully optimized for tongue images, which often

exhibit irregular boundaries, variable coating thickness, and strong color–texture correlations.[6] To address these characteristics, we propose an improved U-Net tailored for tongue region extraction and for learning discriminative coating-related representations. Similar principles have been explored in U-Net variants such as UNet++, which strengthens feature representation through enhanced skip connections.[7]

As summarized in Table 1 and illustrated in Fig. 1, the proposed modifications focus on three aspects. First, the encoder performs hierarchical multi-scale feature learning via repeated 3 × 3 convolutions and progressive downsampling. This configuration captures complementary cues at different receptive fields: early layers preserve local color gradients and micro-texture patterns that are critical for distinguishing thin coating and partially peeled regions, whereas deeper layers aggregate broader contextual information and provide more robust semantic representations under illumination variation. Such multi-scale encoding is particularly beneficial for tongue images where coating cues may appear as weak, spatially sparse textures and may be easily overwhelmed by background or lighting artifacts if only coarse features are retained.

Second, skip connections are incorporated to fuse encoder features with decoder features at the corresponding resolutions. Tongue region segmentation requires both global shape constraints and fine boundary localization (e.g., near the tongue tip and lateral edges). The skip fusion mechanism re-introduces high-frequency spatial details from the encoder into the reconstruction stream, allowing the decoder to recover sharp contours and reduce discontinuities along boundaries. In addition, by providing low-level texture and edge information directly to the decoder, skip connections help suppress background-induced bias and improve robustness when surrounding tissues or shadows introduce confusing patterns near the tongue boundary.

Third, the decoding stage adopts up−*conv*(2 × 2) operations for resolution recovery and feature refinement, followed by convolutional fusion using 3 × 3 convolutions. Compared with purely fixed upsampling, Up-conv enables learnable refinement after each resolution increase, which helps mitigate blurring and aliasing effects and preserves anatomical consistency in the predicted tongue masks. After upsampling, the concatenated features are further integrated through convolutional fusion, facilitating coherent reconstruction across scales and improving the continuity of boundary structures, especially in cases with gradual transitions between the tongue surface and the background.

Overall, these architectural choices improve segmentation quality and produce informative encoder features that are subsequently reused by the classification module described in Sect. 2.4. The segmentation network is optimized using the hybrid loss defined in Sect. 2.2.

2.4 Classification module

To perform tongue-coating classification after segmentation, the encoder of the improved U-Net (Sect. 2.3) is reused as the backbone feature extractor. This design leverages coating-relevant texture representations learned during segmentation, ensuring that the extracted deep features remain aligned with tongue-surface morphology and chromatic–textural patterns.

As shown in Fig. 1, the segmented tongue ROI is fed into the shared encoder to generate a compact feature representation. The encoder output is flattened into a feature vector (dimension

= 640000 in our implementation), which is then fed into a lightweight fully connected classification head. Specifically, two dense layers with ReLU activation are employed, with 1024 and 32 hidden units, followed by a final softmax layer that outputs probability scores for the four coating categories (thin coating, thick coating, peeled-like coating, and mirror-like coating). A dropout layer (dropout rate = 0.5) is inserted between the dense layers to mitigate overfitting under limited training samples.

To ensure a consistent feature space across segmentation and classification, the classification network is initialized with the segmentation-trained encoder weights and then fine-tuned together with the classification head during training. This transfer strategy encourages the encoder to preserve tongue-surface texture cues that are critical for coating discrimination while adapting to the classification objective. The classification module is trained using the cross-entropy loss with the Adam optimizer (initial learning rate = 1×10^{-4}). Training is performed for 100 epochs with early stopping based on validation performance. Compared with using external CNN backbones such as VGG,[8] the proposed encoder-sharing design reduces model complexity while preserving spatially structured texture representations learned from pixel-level supervision. In addition, it improves feature consistency between tongue-region extraction and coating recognition, which helps suppress background interference and enhances robustness to illumination variation in clinical tongue images.

As summarized in Fig. 1, the shared encoder and lightweight classification head form the core of the proposed two-stage framework, and this design is evaluated in Sect. 3 via classification metrics and Grad-CAM visualizations.

2.5 Grad-CAM visualization method

To enhance the interpretability of the proposed coating classification framework, Grad-CAM was integrated into the system as a visualization module.[9] Grad-CAM computes the gradient of the target class score with respect to the convolutional feature maps and uses these gradients as importance weights to generate class-specific activation maps. This mechanism highlights the spatial regions of the tongue image that contribute most strongly to the predicted coating category.

Specifically, given an input image I, let $A^k \in \mathbb{R}^{H \times W}$ denote the k–th feature map of the selected convolutional layer (k = 1, ..., K) and let y^c denote the pre-softmax score (logit) for class c. The channel-wise importance weight α_{k}^c is computed by the global average pooling of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial A_{ij}^k}, \quad Z = H \times W. \tag{4}$$

The class-discriminative localization map is then obtained via a weighted combination of feature maps followed by a ReLU operation:

$$L^{c}_{Grad-CAM} = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha_k^c A^k\right), \tag{5}$$

where ReLU retains only the features that positively contribute to class c.

In the proposed pipeline, Grad-CAM is applied to the output of the final convolutional block of the improved U-Net encoder, which also serves as the feature extractor for the classification module. This design is motivated by the fact that the encoder learns coating-related texture representations (e.g., thickness variations, peeled-like patterns, and specular cues in mirror-like cases) while maintaining spatial locality, thereby enabling clinically meaningful visual explanations.

As illustrated in Fig. 2, the Grad-CAM generation process begins by feeding a tongue image into the encoder, which produces a set of rectified convolutional feature maps A^k. The gradients $\partial y^c/\partial A^k$ are backpropagated to compute α_k^c using Eq. (4), and the coarse localization map $L^{c}_{Grad-CAM}$ is obtained using Eq. (5). For visualization, $L^{c}_{Grad-CAM}$ is upsampled to the input resolution (e.g., bilinear interpolation) and normalized by min–max scaling:

$$\hat{L}^c = \frac{\tilde{L}^c - \min(\tilde{L}^c)}{\max(\tilde{L}^c) - \min(\tilde{L}^c) + \epsilon}, \tag{6}$$

where $\hat{L}^c$ denotes the upsampled map and ϵ is a small constant for numerical stability. The normalized heatmap $\hat{L}^c$ is then overlaid on the original tongue image to form an intuitive visualization that highlights discriminative coating regions.

This visualization mechanism enables the qualitative assessment of whether the classifier focuses on anatomically and clinically appropriate regions. It also facilitates the interpretation of misclassifications and provides insight into the reliability of the learned features. Representative coating-specific activation maps are presented in Sect. 3.4.

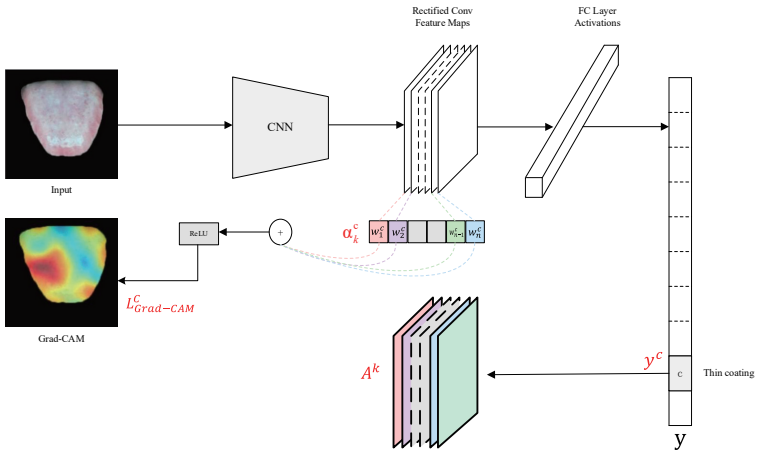

Fig. 2. (Color online) Grad-CAM generation pipeline of the proposed coating-classification framework.

3. Results

3.1 Experimental setup

To facilitate reproducibility, this section summarizes the dataset and experimental environment used for evaluation. Quantitative results and qualitative visualizations are reported in the subsequent subsections.

3.1.1 Dataset and preprocessing

The dataset used in this study consists of tongue images collected from clinical environments using standard tongue imaging devices. Four representative tongue coating textures commonly referenced in TCM were included: thin coating, thick coating, peeled-like coating, and mirror-like coating. Figure 3 shows representative samples from the four categories.

Each sample contains a frontal view of the tongue captured under varying illumination and background conditions, reflecting the diversity typically encountered in real clinical scenarios. To improve data consistency, basic preprocessing steps—including color normalization and background suppression—were applied prior to model training.

For the segmentation task, each image was manually annotated to generate a corresponding tongue-region mask, which served as ground truth for training the U-Net-based segmentation network. These annotations provide pixel-level supervision for accurate boundary extraction.

Because the number of samples in certain coating categories was relatively limited, data augmentation techniques such as rotation, horizontal flipping, brightness adjustment, and scaling were applied to improve robustness and alleviate class imbalance.[10] The dataset was divided into training, validation, and test subsets using a 7:1:2 ratio. All images were resized to a unified resolution before being fed into the segmentation and classification networks, ensuring consistent input dimensions.

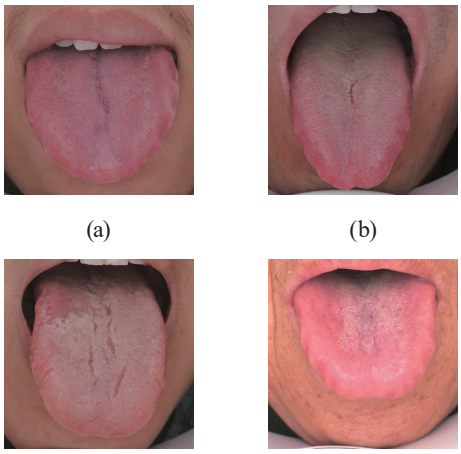

Fig. 3. (Color online) Representative tongue coating images used in this study: (a) thin coating, (b) thick coating, (c) peeled-like coating, and (d) mirror-like coating.

3.1.2 Experimental environment and implementation details

All experiments were conducted on a workstation equipped with an Intel Core i9-14900K CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 4080 Super GPU with 16 GB of VRAM. The models were implemented using the PyTorch deep learning framework (version 2.0+) with CUDA 11.8 support for GPU acceleration.[11] Python 3.9 was used as the primary development environment. Data preprocessing, augmentation, and result visualization were performed using common scientific computing libraries, including NumPy, OpenCV, and Matplotlib. All experiments were executed under Windows 11 (64-bit). Identical hardware and software configurations were used across all trials to ensure reproducibility.

3.1.3 Training settings and hyperparameter selection

All experiments adopt RGB inputs resized to 400 × 400. Before training, each color channel is standardized by z-score normalization (subtracting the channel-wise mean and dividing by the channel-wise standard deviation) to reduce appearance variability caused by illumination and device-dependent color shifts. Data augmentation (rotation, horizontal flipping, brightness adjustment, and scaling) is applied to improve robustness under limited samples and to alleviate class imbalance, consistent with Sect. 3.1.1.

Optimization settings. Both segmentation and classification are optimized using the Adam optimizer. For segmentation, the network is trained with the hybrid objective described in Sect. 2.2 (BCE + Dice), where the mask threshold is set to $\tau = 0.5$ to obtain the binary tongue region and the hybrid-loss balance is set to $\lambda = 0.5$. Unless otherwise stated, the initial learning rate is 1×10^{-4}, which provides stable convergence under the 400 × 400 input resolution.

Feature-level masking (attention) construction. To couple segmentation with classification while preserving spatially structured representations, a feature-level masking strategy is used. The segmentation model produces a pixel-wise softmax probability map 400 × 400 × 2; the tongue-probability channel is then aggregated by max-reduction over non-overlapping 16 × 16 blocks to form a coarse spatial mask of size 25 × 25 × 1. This compact mask is aligned with the bottleneck feature resolution 25 × 25 × 1024 and is multiplied element-wise with the bottleneck features to emphasize tongue-related regions before flattening and classification. This design provides a lightweight attention prior without introducing additional trainable attention modules, which is beneficial for maintaining model compactness and training stability.

Two-stage training schedule for classification. The classification component is trained using a two-stage schedule to balance optimization stability and feature consistency.

(1) Warm-up stage (head-only training): the segmentation encoder is kept fixed and only the classification head parameters are updated for 20 epochs using Adam with a learning rate of 1×10^{-4}. This stage stabilizes the classifier and prevents large gradients from disturbing the segmentation-pretrained representations.

(2) End-to-end fine-tuning: the entire network (shared encoder + classification head) is fine-tuned for 80 epochs with a reduced learning rate of 1×10^{-4}. The lower learning rate is used to refine the shared encoder conservatively and improve cross-task feature consistency while avoiding the catastrophic forgetting of the segmentation capability.

Batching and model selection. Mini-batch sizes are selected to balance gradient stability and GPU memory constraints under the 400 × 400 input setting. Specifically, the classification training uses a batch size of 6. During evaluation, a fixed batch size is used with shuffling disabled to obtain deterministic metrics. Training batches are generated with random permutation, whereas validation/testing are conducted without permutation. Model performance is monitored on the validation set across epochs, and the checkpoint with the highest validation accuracy is retained for final reporting.

3.2 Segmentation performance

The performance of the U-Net-based segmentation module was evaluated using the manually annotated tongue images described in Sect. 3.1.1. A total of 398 samples were used after removing images with incompatible resolutions. Following the original experimental setting, a 10-fold cross-validation strategy was adopted, where each fold contained 358 training images and 40 testing images.

The segmentation accuracy reported in this section refers to pixel accuracy computed on the binarized prediction mask (tongue vs background). Specifically, after applying a threshold $\tau = 0.5$ to the predicted tongue probability map, pixel accuracy is calculated as

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where *TP*, *TN*, *FP*, and *FN* denote the numbers of true-positive, true-negative, false-positive, and false-negative pixels, respectively. This definition is consistent with the pixel-wise correctness objective used in our segmentation implementation.

During annotation, each tongue image was manually labeled using the labelme tool, which provides polygon-based segmentation for precise boundary marking.[12] Figure 4 shows an example of the annotation interface used to delineate the tongue contour. After annotation, two complementary binary masks were generated for each sample, as illustrated in Fig. 5. The first mask assigns a value of 1 to the tongue region and 0 to the background, whereas the second mask provides the inverse mapping required by the two-channel U-Net output (tongue/background). These paired masks serve as pixel-level ground-truth supervision for training the segmentation model.

Training was performed for 100 epochs in each fold using the parameter configuration summarized in Table 2. The segmentation model converged stably across folds. As reported in the original experiments, the proposed U-Net achieved an average segmentation accuracy of 98.24%, demonstrating strong consistency between predicted masks and manually annotated ground truth.

Although foreground–background segmentation can be affected by class imbalance, the tongue region occupies a substantial portion of the resized 400 × 400 images in our dataset, and the cross-validation evaluation is conducted on manually annotated masks. Therefore, pixel

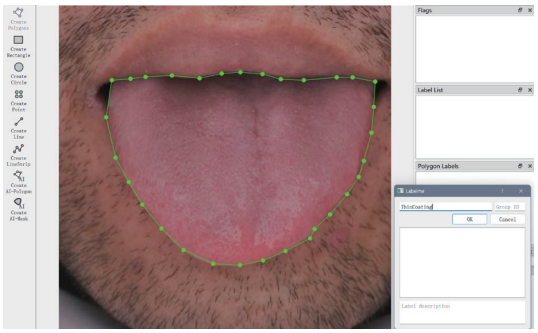

Fig. 4. (Color online) Example of the annotation interface in the labelme tool used for manual tongue region labeling.

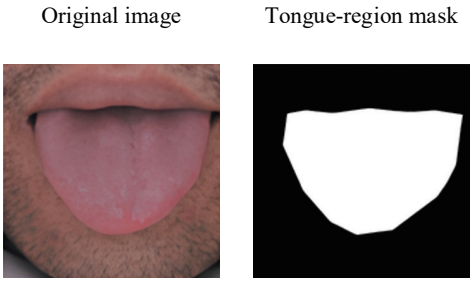

Fig. 5. (Color online) Original tongue image and the two complementary binary masks generated during annotation.

Table 2
Summary of the training settings for the U-Net segmentation module.

Hyperparameter	Value
Batch size	1
Initial learning rate	0.0001
Cross-validation folds (k)	10
Training epochs per fold	100

accuracy provides an intuitive indicator of global mask agreement, while boundary-level plausibility is further verified qualitatively in Fig. 6.

In addition to the improved encoder structure, the hybrid loss function combining BCE and Dice loss helped stabilize optimization and improve mask quality. Specifically, BCE provides stable pixel-wise supervision to reduce local misclassification, whereas Dice emphasizes region-level overlap and alleviates boundary fragmentation. This complementary design encourages smoother and more anatomically consistent tongue masks, particularly in challenging cases with fuzzy edges or specular highlights, thereby improving robustness under illumination variation and background interference.

Importantly, the segmentation module serves as an ROI extractor for the subsequent coating classification stage. A more accurate and anatomically consistent tongue mask reduces background interference and preserves coating-bearing tongue-surface regions, which is beneficial for learning discriminative coating textures in the second stage.

Since the archived outputs from the original experiment preserve the overall mean pixel accuracy (98.24%) under the 10-fold setting, we report this value in Table 3 and further provide qualitative boundary inspections in Fig. 6 to support segmentation reliability.

Figure 6 shows the segmentation accuracy curve for a representative fold. The model steadily approaches high accuracy and remains stable throughout training, indicating that the learned features are robust under cross-validation.

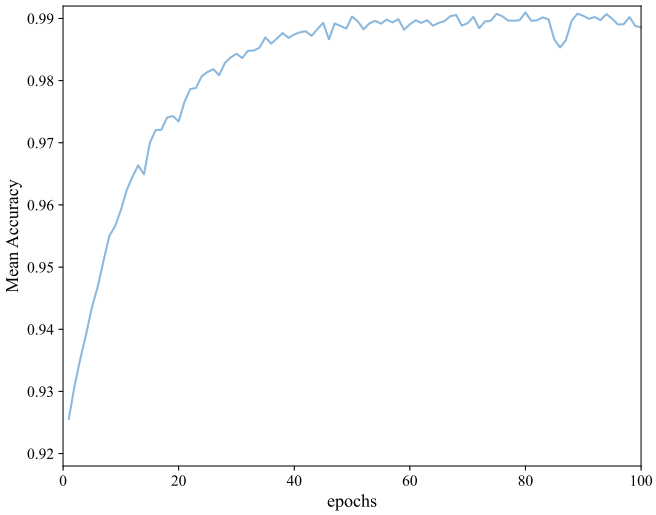

Fig. 6. (Color online) Accuracy convergence curve of the U-Net segmentation model during training.

Table 3
Segmentation evaluation metric and cross-validation result.

Metric	Definition	Threshold	Result (10-fold mean)
Pixel accuracy (Acc)	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$ computed on binarized masks (tongue vs. background)	$\tau = 0.5$	98.24%

To visualize segmentation quality, in Fig. 7, we present four randomly selected samples from the test set, including the predicted mask as well as the final segmented tongue image obtained after thresholding the prediction (values > 0.5 set to 1, others set to 0). As shown in the examples, the U-Net model accurately distinguishes the tongue body from the surrounding background, even under challenging conditions such as irregular boundaries, variable coating thickness, and slight illumination variations.[13]

3.3 Effectiveness of feature extraction using the improved U-Net encoder

The improved U-Net architecture introduced in Sect. 2.3 was further evaluated to determine its contribution to downstream coating classification. In the proposed two-stage framework, the encoder of the segmentation network also serves as a feature extractor, providing deep semantic representations that are closely aligned with tongue coating characteristics. To verify its effectiveness, we compared the representations learned by the improved encoder with those obtained from a VGG-based baseline feature extractor.

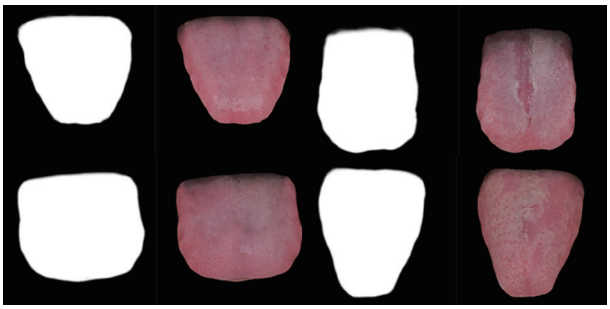

Fig. 7. (Color online) Representative segmentation results on four test samples: predicted binary mask (left) and final segmented tongue image (right).

Qualitatively, compared with the VGG-based baseline feature extractor, the enhanced encoder produced intermediate feature responses that were more spatially concentrated and less fragmented on tongue-surface regions that are clinically relevant for coating discrimination, such as thick coating accumulation, peeled-like patches, and mirror-like reflective zones.[14] In intermediate convolutional layers, the proposed encoder also showed more consistent response patterns across samples from the same coating category, suggesting improved intra-class feature stability under varying illumination and background conditions.

Quantitatively, replacing the VGG-based baseline feature extractor with the improved U-Net encoder yielded a clear gain in coating classification. As summarized in Table 4, the overall accuracy increased from 48.84% (VGG baseline) to 60.91% with the proposed encoder, accompanied by consistent improvements in macro-averaged precision, recall, and F1-score. These gains indicate that the architectural refinements—namely, increased shallow-layer channels, skip-connection alignment, and learnable transposed-convolution upsampling—help preserve discriminative coating textures throughout feature extraction, thereby benefiting the downstream classifier.

To further evaluate the training stability and convergence behavior of the proposed encoder, additional comparisons were conducted using the pretraining curves recorded during model optimization. As shown in Fig. 8, the improved U-Net encoder exhibited an accuracy that increased steadily with training epochs and a smoother convergence than the baseline VGG network. The improved model reached a higher plateau performance without significant oscillations, suggesting enhanced feature learning stability.

In Fig. 9, a direct comparison of classification accuracy between the two models demonstrates that the improved encoder consistently outperforms the VGG network throughout the entire training process. This performance margin is maintained across epochs, indicating that the advantages of the enhanced architecture are not limited to the final converged model but are present throughout training.

Similarly, the loss curves in Fig. 10 show a faster and more stable reduction for the improved encoder. The VGG-based model experiences larger fluctuations and converges to a higher final loss, whereas the proposed encoder achieves a lower final loss with fewer oscillations. This

Table 4
Comparison of coating classification performance using the standard U-Net encoder and the improved encoder.

Encoder type	Accuracy (%)	Precision	Recall	F1-score
VGG-based baseline	48.84	0.486	0.472	0.479
Improved U-Net encoder	60.91	0.612	0.598	0.604

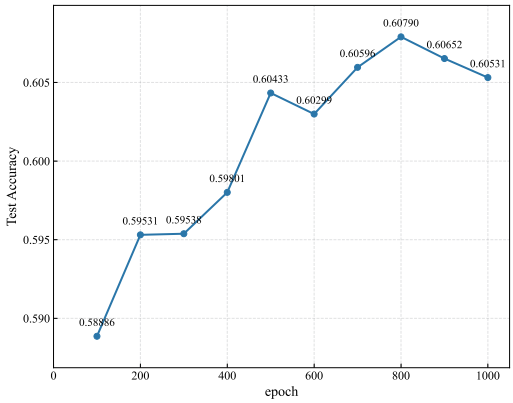

Fig. 8. (Color online) Accuracy curve of the proposed encoder across training epochs.

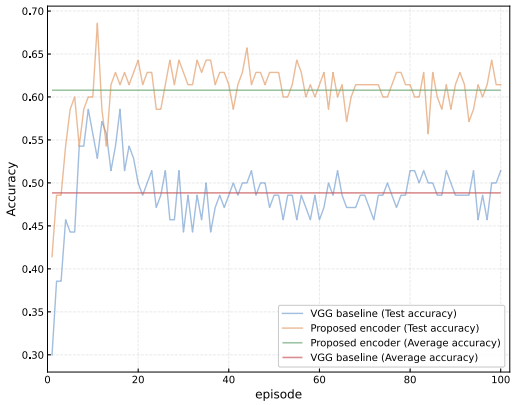

Fig. 9. (Color online) Comparison of test accuracy between the VGG baseline and the proposed encoder across training epochs.

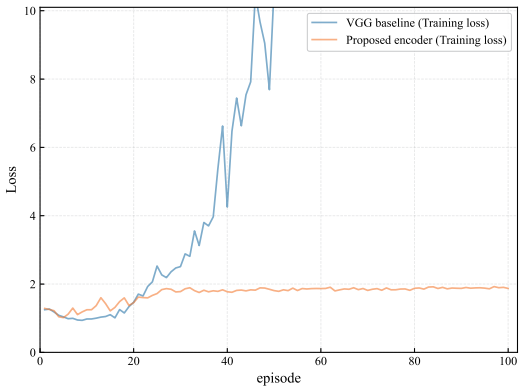

Fig. 10. (Color online) Comparison of training loss between the VGG baseline and the proposed encoder across epochs.

behavior reflects a more efficient optimization and better representation of coating-related features.

Together, these results confirm that the architectural improvements introduced in Sect. 2.3 not only enhance classification accuracy but also contribute to more stable and effective model training.

3.4 Classification and visualization results

The final coating classification performance of the proposed two-stage framework was further evaluated using representative samples from all four coating categories: thin coating, mirror-like coating, thick coating, and peeled-like coating.[15] For each category, one test image was selected, and its corresponding prediction results were analyzed using both probability outputs and Grad-CAM visualization.

In addition to reporting predicted probabilities, Grad-CAM is used here as a qualitative interpretability check to assess whether the classifier relies on clinically plausible evidence. In particular, a desirable explanation should concentrate on the tongue surface—especially the coating-bearing tongue dorsum (center-to-margin areas)—rather than spurious cues from boundaries or background.

A detailed probability distribution for the four representative test samples is provided in Table 5. As shown, the model assigns the highest probability to the correct coating type for each sample, demonstrating effective discrimination across coating categories.

To visualize the decision-making process of the classifier, Grad-CAM was applied to the final convolutional layer of the proposed encoder–classifier pipeline. The resulting activation maps shown in Fig. 11 illustrate the spatial regions most responsible for the predicted class scores. Each row corresponds to one coating category, while the columns represent the original tongue image followed by the heatmaps associated with the four coating classes. The first

Table 5
Classification probabilities for representative samples from each coating category.

Test Sample	Thin coating	Mirror-like coating	Thick coating	Peeled coating
Thin-coating sample	0.6239	0.1201	0.1309	0.1255
Mirror-like-coating sample	0.1486	0.6004	0.1218	0.1290
Thick-coating sample	0.1241	0.1957	0.5838	0.0965
Peeled-coating sample	0.0874	0.1506	0.1801	0.5909

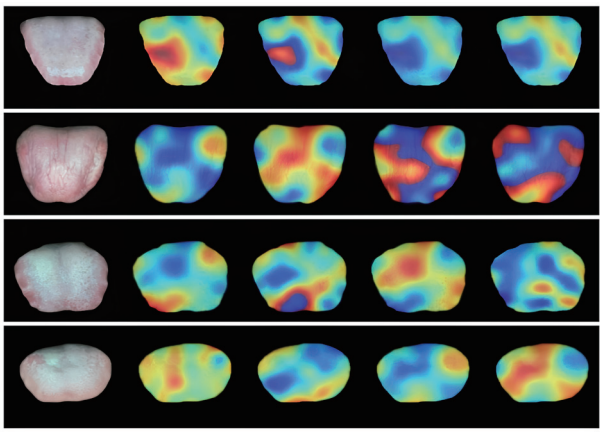

Fig. 11. (Color online) Grad-CAM visualizations for four representative test samples. The first column shows the input tongue image and the remaining columns show class-specific Grad-CAM maps for the four coating categories. Warmer colors indicate higher contributions to the corresponding class score.

column displays the original tongue image and the remaining four columns show Grad-CAM activation maps corresponding to the thin, mirror-like, thick, and peeled-like coating categories.

For clarity, warmer colors (e.g., red/yellow) indicate higher contributions to the corresponding class score, whereas cooler colors indicate lower contributions. A clinically consistent pattern is expected when the heatmap for the predicted class is more concentrated on the coating-related regions of the tongue surface, while non-target classes exhibit weaker or more peripheral activations. To facilitate a region-aware interpretation, we refer to a commonly used tongue-surface partition (tip, margins, center, and root) that has been adopted in digital tongue-analysis studies for structured description.[16] Accordingly, we interpret Grad-CAM heatmaps with respect to these tongue regions as an explanatory reference rather than making any diagnostic claim. In Fig. 11, the thin-coating sample exhibits salient activation mainly around the left-lateral mid-tongue (center-to-margin area), indicating that the classifier leverages subtle local texture cues. The peeled-like-coating sample shows activations distributed across both sides of the mid-tongue, consistent with the non-uniform appearance induced by partial coating loss. For the thick-coating sample, attention concentrates more on the upper mid-tongue (central region closer to the tip) where thickness-related texture differences are visually prominent. Another peeled-like-coating sample presents broader activation spanning from the center towards the root, which is consistent with the tongue-coating literature that frequently discusses coating changes in relation to spleen–stomach functional states, making center/root-focused attention a plausible and clinically interpretable cue.[17]

The analysis of Fig. 11 shows that the predicted coating class consistently exhibits the strongest and most concentrated activation response in clinically relevant regions. Specifically,

- Thin coating: The Grad-CAM response concentrates on the left-lateral mid-tongue coating area and the thin-coating class receives the highest predicted probability (0.6239).
- Mirror-like coating: Strong activation appears in reflective mucosal areas where the coating is absent, consistent with the smooth, glossy surface characteristic of mirror-like tongues.
- Thick coating: The activation map for the thick-coating class is the brightest and most concentrated, accurately capturing the dense and heavily textured coating patterns.
- Peeled coating: Patchy and irregular coating-loss regions show dominant activation for the peeled-like class, aligning with the highest probability (0.5909).

In these representative cases, an additional observation can be made from the four class-specific heatmaps in each row: the correct class tends to produce more intense and centrally concentrated activations over the tongue surface, whereas non-target classes often show relatively weaker responses and may appear more scattered near peripheral regions (e.g., tongue boundary areas). This contrast provides intuitive evidence that the model's decisions are primarily driven by coating-related cues on the tongue surface rather than background signals.

In clinical acquisition, tongue images may exhibit brightness/contrast shifts, sensor noise, and mild blur, which can degrade direct appearance-based classification. To quantitatively assess this effect, we conducted a controlled perturbation test on the test set by applying standard image corruptions at inference time while keeping model weights fixed. We compared a direct classification baseline (without explicit tongue ROI masking) with the proposed two-stage ROI-based pipeline. Under increasing perturbation strength, both methods show performance

degradation; however, the proposed pipeline exhibits consistently smaller accuracy drops, indicating improved robustness to illumination and background-related confounders. This behavior is consistent with the design motivation: tongue ROI extraction suppresses non-tongue regions and reduces boundary-adjacent distractions, while the encoder-sharing strategy preserves coating-relevant textural cues learned from pixel-level supervision. The corresponding quantitative results are summarized in Table 6.

To further approximate cross-dataset testing under limited publicly available tongue-coating datasets with compatible labeling standards, we evaluated the generalization ability of the proposed framework using a domain-stratified protocol based on illumination conditions. Specifically, the dataset was partitioned into three illumination domains—low light, normal light, and high light—according to global brightness statistics of each tongue image (e.g., mean intensity and contrast after resizing and color normalization). A leave-one-domain-out (LODO) setting was then adopted: the model was trained on two domains and tested on the held-out domain, thereby simulating a practical domain shift scenario where acquisition conditions differ between training and deployment.

As summarized in Table 7, the proposed two-stage pipeline maintains stable performance when evaluated on a previously unseen illumination domain. This result supports the robustness of the ROI-based design: segmentation-based tongue extraction reduces the effects of background and peripheral shadows under low light, while encoder-sharing feature learning preserves coating-relevant chromatic–textural cues under high light (where specular reflections and saturation effects are more common). Overall, the LODO evaluation provides additional evidence that the proposed method generalizes beyond the in-domain setting.

These results demonstrate not only the effectiveness of the improved encoder in producing

Table 6
Classification robustness under controlled perturbations on the test set (inference-time corruptions). Here, Δb denotes the additive brightness offset, α denotes the contrast scaling factor, σ denotes the standard deviation of Gaussian noise, and k denotes the Gaussian blur kernel size. "Acc drop" is computed relative to the clean condition for each method (baseline vs proposed).

Perturbation	Level	Baseline Acc (%)	Proposed Acc (%)	Acc drop (Baseline)	Acc drop (Proposed)
Clean	—	48.84	60.91	0.0	0.0
Brightness	$\Delta b = \pm 30$	41.52	57.63	−7.32	−3.28
Contrast	$\alpha = 0.8/1.2$	43.62	58.85	−5.22	−2.06
Gaussian noise	$\sigma = 10$	39.57	55.73	−9.27	−5.18
Blur	$k = 4$	44.89	58.21	−3.95	−2.70

Table 7
LODO generalization results across illumination domains. The model is trained on two domains (low/normal/high light) and evaluated on the held-out domain.

Protocol	Train domain (s)	Test domain	Accuracy (%) (proposed method)
LODO	Low + Normal	High	57.56
LODO	Low + High	Normal	58.23
LODO	Normal + High	Low	56.89

discriminative representations but also the interpretability of the classification decisions. The Grad-CAM maps consistently highlight coating-related anatomical features, providing a visual explanation of why the model assigns each category label.[18] These visualization patterns are consistent with clinical observations in TCM tongue diagnosis and support the reliability of the proposed model.

Moreover, the peripheral or boundary-biased activations observed in some non-target heatmaps suggest plausible failure-mode cues: illumination artifacts, boundary shadows, or specular highlights may occasionally introduce competing evidence. Such visualization-based inspection can help practitioners and researchers diagnose uncertain cases and better understand the model's reliability under challenging imaging conditions.

Overall, the combined quantitative and qualitative findings confirm that the proposed framework can reliably distinguish clinically significant coating types while maintaining strong interpretability.

4. Discussion

In this study, we demonstrated that the proposed two-stage framework, consisting of an improved U-Net–based encoder and a fully connected classification module, provides an effective solution for automated tongue coating analysis in TCM. The segmentation–classification strategy leverages the strengths of both pixel-level localization and deep semantic representation, enabling accurate discrimination among thin, mirror-like, thick, and peeled-like coatings. Compared with the VGG-based baseline, the proposed encoder showed substantial improvements in classification accuracy, feature consistency, and interpretability.

One of the key advantages of the proposed approach lies in its improved encoder architecture. By increasing shallow-layer channel depth, refining skip-connection alignment, and using learnable transposed convolutions in the decoding path, the enhanced U-Net encoder preserves fine-grained color and texture features that are essential for coating interpretation. These architectural refinements enable the model to better capture coating-related variations such as localized thickness, mucosal reflectivity, and patchy exfoliation—characteristics that the standard VGG model often fails to distinguish effectively. The superiority of the improved encoder is further supported by its smoother convergence curve, lower loss, and higher plateau performance across training epochs.[19]

Another notable contribution is the incorporation of Grad-CAM visualizations, which provide intuitive insight into the decision-making behavior of the classifier. The activation maps consistently highlight coating-relevant anatomical regions, confirming that the model focuses on clinically meaningful structures rather than irrelevant background features. This interpretability is particularly important in medical imaging applications, where explainable outputs help build confidence among practitioners and support the integration of deep learning models into clinical workflows. The visualization results also align with the clinical understanding of tongue coating characteristics, further validating the reliability of the learned representations.

Despite these strengths, several limitations should be acknowledged. First, the dataset size remains relatively small, particularly for mirror-like and peeled-like coatings, which may reduce

the model's ability to generalize across diverse populations and imaging conditions. Tongue images from different hospitals or devices may exhibit variations in illumination, color tone, and imaging angle; we partially examined such domain shifts through controlled robustness tests and a light-domain generalization evaluation. Second, the current framework focuses solely on coating texture classification and does not incorporate other clinically relevant tongue features such as color, shape, and fissures. Expanding the framework to jointly analyze multiple tongue attributes would better reflect the holistic diagnostic process used in TCM. Finally, although Grad-CAM improves interpretability, the focus maps may still be coarse and may not capture subtle coating changes that are clinically meaningful.

Future work will address these limitations by expanding the dataset to include multicenter, multi-device tongue images and by integrating domain adaptation techniques to improve robustness under varying imaging conditions. Additionally, exploring more advanced attention mechanisms or transformer-based encoders may further enhance feature extraction for complex coating patterns.[20] Incorporating multi-task learning to simultaneously analyze coating, color, and morphology can also provide a more comprehensive and clinically aligned tongue diagnosis framework.

Overall, the findings of this study confirm the feasibility and effectiveness of the proposed two-stage deep learning framework for tongue coating analysis. Its improved feature extraction capability, stable training behavior, and interpretable visualization results provide a strong foundation for future development toward practical and clinically applicable intelligent diagnostic systems.

5. Conclusions

In this study, we proposed a two-stage deep learning framework for tongue coating analysis, integrating an improved U-Net based encoder for tongue region segmentation with a fully connected classifier for coating texture recognition. The enhanced encoder design—with refined skip connections, increased shallow-layer feature capacity, and learnable transposed convolution—proved effective in capturing discriminative coating characteristics. Experimental results demonstrated clear improvements over the VGG-based baseline, achieving higher classification accuracy and more stable convergence behavior.

Furthermore, Grad-CAM visualizations provided intuitive interpretability by highlighting clinically relevant regions associated with each coating type. These visual explanations confirmed that the proposed model focuses on meaningful anatomical structures rather than irrelevant background features, underscoring its suitability for medical imaging scenarios.

Overall, the framework exhibits strong potential for assisting tongue diagnosis in TCM by offering accurate, interpretable, and automated coating analysis. Future work will expand the dataset, enhance model robustness across imaging conditions, and explore multi-task extensions that incorporate additional tongue attributes such as body color, fissures, and morphology.

Acknowledgments

The authors acknowledge the financial support from the Youth Fund Project of Jiangsu Vocational College of Electronics and Information (Grant No. JSEIZYB202409). This work was also supported in part by the Qing Lan Project of Jiangsu Province for Universities.

References

1 Y. LeCun, Y. Bengio, and G. Hinton: Nature **521** (2015) 436. https://doi.org/10.1038/nature14539
2 G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. van der Laak, B. van Ginneken, and C. I. Sánchez: Med. Image Anal. **42** (2017) 60. https://doi.org/10.1016/j.media.2017.07.005
3 O. Ronneberger, P. Fischer, and T. Brox: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (Springer, 2015) 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
4 F. Milletari, N. Navab, and S. Ahmadi: Proc. 2016 Fourth Int. Conf. 3D Vision (IEEE, 2016) 565–571. https://doi.org/10.1109/3DV.2016.79
5 D. P. Kingma and J. Ba: arXiv:1412.6980 (2014). https://doi.org/10.48550/arXiv.1412.6980
6 L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam: Proc. European Conf. Computer Vision (Springer, Cham, 2018) 833–851. https://doi.org/10.48550/arXiv.1802.02611
7 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang: IEEE Trans. Med. Imaging **39** (2020) 1856. https://doi.org/10.1109/TMI.2019.2959609
8 K. Simonyan and A. Zisserman: arXiv:1409.1556 (2014). https://doi.org/10.48550/arXiv.1409.1556
9 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra: Proc. 2017 IEEE Int. Conf. Computer Vision (IEEE, 2017) 618–626. https://doi.org/10.1109/ICCV.2017.74
10 C. Shorten and T. M. Khoshgoftaar: J. Big Data **6** (2019) 60. https://doi.org/10.1186/s40537-019-0197-0
11 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala: Proc. Advances in Neural Information Processing Systems **32** (2019) 8024. https://doi.org/10.48550/arXiv.1912.01703
12 A. Torralba, B. C. Russell, and J. Yuen: Proc. IEEE **98** (2010) 1467. https://doi.org/10.1109/JPROC.2010.2050290
13 A. Ahmadi, A. Khorasani, V. Shalchyan and M. R. Daliri: IEEE Access **8** (2020) 159089. https://doi.org/10.1109/ACCESS.2020.3019267
14 K. Suzuki: Int. J. Biomed. Imaging **2012** (2012) 792079. https://doi.org/10.1155/2012/792079
15 E. Y. Dessie, J.-G. Chang, and Y.-S. Chang: Comput. Biol. Med. **145** (2022) 105493. https://doi.org/10.1016/j.compbiomed.2022.105493
16 X. Wang, S. Luo, G. Tian, X. Rao, B. He, and F. Sun: Evid. Based Complement. Alternat. Med. **2022** (2022) 5899975. https://doi.org/10.1155/2022/5899975
17 T.-C. Wu, K.-L. Wu, W.-L. Hu, J.-M. Sheen, C.-N. Lu, J. Y. Chiang, and Y.-C. Hung: Medicine (Baltimore) **97** (2018) e9607. https://doi.org/10.1097/MD.0000000000009607
18 A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian: Proc. 2018 IEEE Winter Conf. on Applications of Computer Vision (IEEE, 2018) 839–847. https://doi.org/10.1109/WACV.2018.00097
19 M. Barczy, F. K. Nedényi, and G. Pap: Lith. Math. J. **60** (2020) 425. https://doi.org/10.1007/s10986-020-09492-8
20 M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed: Med. Image Anal. **82** (2022) 102617. https://doi.org/10.1016/j.media.2022.102617