

A Novel Skin Disease Identification Algorithm Based on Attention Mechanism and Transformer Architecture

Qinlan Wu¹, Hailin Zou², and Xijian Chao³

¹ Ph.D. Candidate, Associate Professor and Vice Dean, School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, 338000, China, E-mail: qinlanwu@outlook.com (corresponding author).

² Associate Professor, School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, 338000, China, E-mail: hailinzou@heu.uu.me

³ Professor and Ph.D. candidate, School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, 338000, China, E-mail: chaoxijian@heu.uu.me

Project Management

Received February 8, 2026; revised July 16, 2026; September 15, 2026; accepted October 14, 2026

Available online October 26, 2026

Abstract: Deep learning-based skin disease recognition technology has been continuously developing, but existing methods still fall short in terms of detail representation and computational efficiency due to limitations such as limited image samples, small lesion areas, and significant fine-grained differences. Against this backdrop, this study proposes a skin disease recognition algorithm based on an attention mechanism and a Transformer architecture. While retaining the local texture-feature modeling capability of convolutional networks, a Transformer recognition framework that integrates convolutional features with an attention mechanism is constructed to jointly model both local fine-grained features and global semantic information. Experimental validation was conducted using the International Society for Skin Imaging Collaboration (ISIC) 2018 dataset. The proposed method achieved good classification performance on this dataset, with a Macro-F1 score of 0.90 and a Weighted-F1 score of 0.95, and demonstrated superior recognition performance in comparisons with various Transformer models. The results show that this method can effectively integrate local fine-grained features and global semantic information, which is significant in improving the accuracy and stability of automatic skin disease recognition and provides a feasible technical path for intelligent assisted diagnosis of skin diseases.

Keywords: Skin disease identification; multi-scale local feature modeling; lesion perception attention mechanism; transformer architecture; medical image analysis.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).
DOI 10.32738/JEPPM-2026-209

1. Introduction

1.1. Research Background

There are many types of skin diseases with different clinical manifestations. Different diseases are highly similar in visual features such as color, texture, and boundary (Gupta et al., 2025). The same disease can also vary in form across individuals or stages, exhibiting typical fine-grained discriminative characteristics (Wei et al., 2021). Clinical diagnosis relies on doctors' experience and subjective observation, and is easily affected by factors such as unclear early features, background interference, or image quality. With the development of dermoscopy imaging and digital medicine, the use of computer vision technology to achieve automatic and stable recognition of skin disease images has become an important research direction for intelligent assisted diagnosis and medical decision support. This type of intelligent recognition technology can not only improve the efficiency of skin disease screening, but also provide technical support for medical institutions in management decisions such as diagnosis triage, resource allocation and screening process optimization. Among them, deep learning methods have shown clear advantages in medical image analysis, and Convolutional Neural Network (CNN) based models have been widely used for tasks such as skin disease classification, segmentation and detection (Hassan et al., 2026). Dermatological images often contain small and irregular lesions with strong background interference. Although Transformer models can capture long-range dependencies (Sattarzadeh et al., 2024; Ma et al., 2022), directly applying them may lead to insufficient local feature representation. Therefore, this study proposes an Attention-based Transformer for Skin Disease Recognition (ATS-Former) to improve recognition performance and stability. The research focuses on the following questions: First, how to effectively integrate multi-scale convolutional feature modeling and lesion perception attention mechanism into Transformer architecture in order to enhance the expressive power of fine-grained features in skin disease

images. Second, whether the ATS-Former model can achieve better recognition results in skin disease recognition tasks compared to existing Transformer models. Third, whether this model provides effective technical support for intelligent assisted diagnosis in improving the stability and reliability of automatic skin disease recognition.

1.2. Related work

The advancement of automatic skin disease identification technology has been driven by advances in deep learning for medical image analysis. Automatic skin disease identification methods based on computer vision have received widespread attention. In view of the strong subjectivity, insufficient doctor resources and medical obstacles such as the epidemic in the diagnosis of skin diseases, Göçeri (2021) developed a user-friendly intelligent diagnosis application based on MATLAB R2019. The application uses digital photos of skin diseases in public databases to realize image classification and recognition. To improve early detection of high-risk melanoma, Saba et al. (2021) conducted a systematic comparison of handcrafted features (including the seven clinical rules, texture, and Gabor features) and deep learning-based features for skin cancer detection. In response to the issues of cluttered and non-standardized clinical skin photographs, Yanagisawa et al. (2023) proposed an automatic segmentation framework grounded on DeepLabv3+ CNN to construct a standardized dataset. To address the challenge of early melanoma detection, Jaber et al. (2025) developed a robust CNN-based binary classification method and introduced the Harris Hawks Optimization (HHO) algorithm to optimize the network.

With the continuous advancement of attention mechanisms and the Transformer architecture in computer vision, self-attention recognition has been gradually introduced to medical image analysis tasks. In response to the issue of diverse lesion types and low contrast with normal tissue, Du et al. (2022) embedded a dual module comprising of dense multiplicative connections and local pyramid attention into the Swin Transformer (Swin T) to improve the robustness of polyp segmentation. Pang et al. (2023) developed the Slim UNETR framework, which uses a Self-Attention Mechanism (SAM) decomposition and lightweight representation aggregation as its core modules, thereby compressing model size while maintaining accuracy. Pacal et al. (2024) improved the Swin T by adopting a hybrid shift-window SAM to better capture overlapping lesions and long-range dependencies, thereby balancing diagnostic speed and accuracy.

In summary, while previous studies have explored integrating attention mechanisms and Transformers into skin disease recognition, most have adopted general structures and failed to optimize for the limited number of skin disease image samples, small lesion regions, and significant fine-grained differences. This results in insufficient representation of local detail and reduced computational efficiency. Therefore, this study proposes the ATS-Former algorithm, which integrates attention mechanisms into the Transformer architecture to jointly model fine-grained features and improve the model's performance and stability for complex skin disease recognition.

Research contributions and innovations are as follows: (1) A skin disease recognition framework, ATS-Former, based on an attention mechanism and a Transformer architecture, was proposed. In doing so, convolutional feature modeling and SAM were integrated. This enables collaborative modeling of local fine-grained features and global semantic relationships in skin disease images. (2) A multi-scale local feature modeling strategy was designed. Using multi-scale convolutional mapping, fine-grained texture changes and structural features in skin disease were extracted from images, thereby enhancing the model's ability to express complex lesion morphology. (3) A Lesion-aware Attention Enhancement (LAE) mechanism was constructed. Using dual-channel and Spatial Attention (SA), key lesion areas were adaptively enhanced, reducing background interference and improving the stability and discriminative ability of skin disease recognition.

2. Methods and Materials

2.1. Local Representation Modeling Methods for Fine-Grained Features of Skin Diseases

Discrimination cues in dermatological images often arise from subtle changes in color gradients, texture distributions, and irregular lesion boundaries in local regions. These features exhibit inconsistencies in spatial scale and morphology (Gupta et al., 2024; Adnanto et al., 2025). For details, please refer to Fig. 1.

As shown in Fig. 1, single-scale feature extraction struggles to simultaneously capture both fine lesion details and larger morphological structures, while directly using fixed image patches for sequence modeling can easily disrupt the spatial continuity and local semantic consistency of lesions. Therefore, this study introduces a multi-scale, structure-preserving feature serialization strategy during the local representation modeling stage to strengthen the model's ability to capture fine-grained features of dermatological diseases. $x \in \mathbb{R}^{H \times W \times C}$ means the input skin disease image, where H and W denote the height and width of the image, and C denotes the number of input channels. To comprehensively capture the lesion characteristics at diverse scales, the study employs a parallel multi-scale convolutional mapping approach to model the input image's local features, and the process is shown in Eq. (1).

$$F_s = f_c^{(s)}(x), \quad s \in \{1, 2, \dots, S\} \quad (1)$$

In Eq. (1), $F_s \in \mathbb{R}^{H_s \times W_s \times D_s}$ represents the feature map at the corresponding scale. $f_c^{(s)}$ represents the convolution mapping function at the s -th scale. S represents the number of scales. H_s , W_s and D_s refer to the spatial resolution and channel dimension at the scale, respectively. Feature mapping at different scales can focus on the texture of small lesions, local structural changes and relatively macroscopic lesion morphology information, respectively (Wen et al., 2025). To achieve collaborative modeling of Multi-Scale Features (MSFs), the features at each scale are aligned and fused to obtain a unified local feature representation $F = \phi(\{F_1, F_2, \dots, F_S\})$, $F \in \mathbb{R}^{H' \times W' \times D}$ represents the fused local feature map. ϕ represents the MSF fusion operator. After obtaining the fused feature F . The study further introduces a structure-preserving serialization modeling method to map the two-dimensional feature map into a feature sequence representation, as shown in Eq. (2) (Zhou et al., 2022; Zhu et al., 2023).

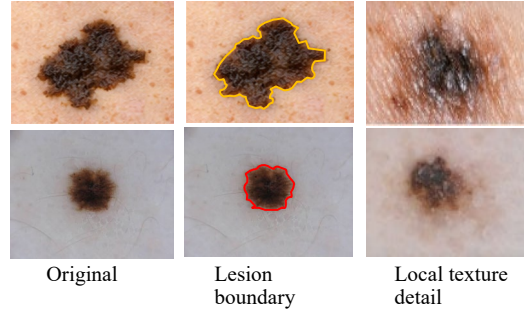


Fig. 1. Discrimination cues in dermatological images (image source: <https://gallery.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%5D> accessed October 14, 2026)

$$\begin{cases} Z = \text{Flatten}(F) \\ E_0 = ZW_e + b_e \end{cases} \quad (2)$$

In Eq. (2), $Z \in \mathbb{R}^{H' \times W' \times D}$ represents the feature sequence length. $W_e \in \mathbb{R}^{d \times D}$ represents the learnable embedding matrix, d represents the embedding feature dimension, $b_e \in \mathbb{R}^d$ represents the bias term, and $E_0 \in \mathbb{R}^{H' \times W' \times D}$ represents the initial local features. The MSF fusion and structure-preserving serialization process is shown in Fig. 2.

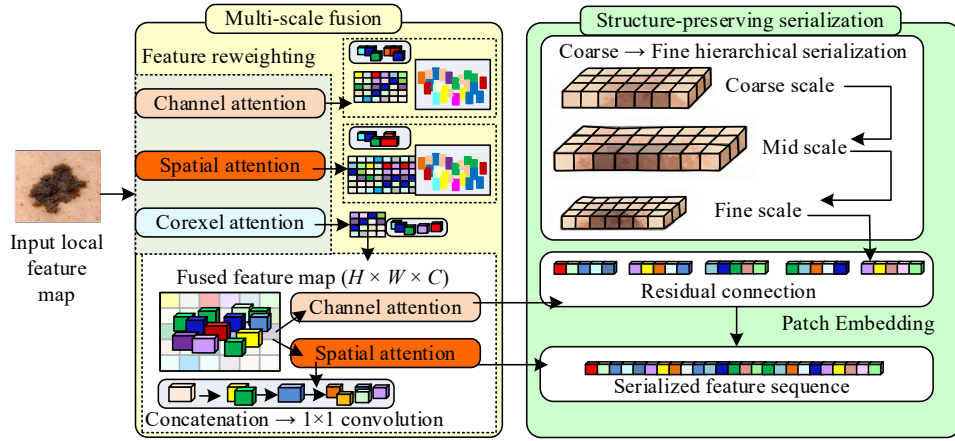


Fig. 2. Schematic diagram of multi-scale local feature fusion and structure-preserving serialization

In Fig. 2, after MSF fusion, the two-dimensional spatial features are smoothly transformed into a serialized representation through structure-preserving feature serialization and embedding mapping. The fused feature map is unfolded according to spatial location, so that each feature vector corresponds to a specific local region in the image, thereby preserving the spatial continuity and structural semantics of lesions at the sequence level.

2.2. Lesion Perception: Attention-Driven Feature Selection and Enhancement Mechanism

After obtaining the structure-preserving local feature sequence representation, it is necessary to further address the high degree of mixing between the lesion area and the healthy skin and background areas in the skin disease image shown in Fig. 3.

Skin disease lesions usually account for a small proportion and are similar in color and texture to the surrounding skin. If all feature sequences are modeled with equal weights, a large amount of irrelevant information will be introduced, thereby weakening the model's discriminative ability and stability (Ahmed et al., 2023). To this end, LAE, based on local representation modeling, is introduced to adaptively weight the key channels and spatial positions within the feature sequence. The LAE structure reweights features across channel and spatial dimensions. Channel descriptors are first obtained via global average pooling, as shown in Eq. (3).

$$z_c = \frac{1}{N} \sum_{i=1}^N E_0(i) \quad (3)$$

In Eq. (3), $z_c \in \mathbb{R}^d$ represents the channel-level statistical features. N represents the feature sequence length, corresponding to the number of spatial locations in the image. The $E_0(i)$ means the i -th feature vector in the sequence. Next, the channel description vectors are modeled based on a fully connected network with a nonlinear mapping to generate Channel Attention (CA) weights, as shown in Eq. (4).

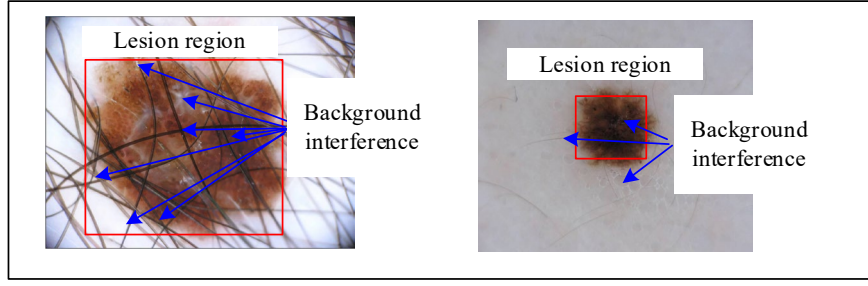


Fig. 3. Illustration of mixed lesion and background regions in dermatological images

$$M_c = \sigma(f_c(z_c)) \quad (4)$$

In Eq. (4), $M_c \in \mathbb{R}^d$ represents the CA weight vector. $f_c(\cdot)$ represents the CA mapping function. σ means the sigmoid activation function. After completing feature reweighting along the channel dimension, the distribution characteristics of the lesion region within the feature sequence are further characterized in the spatial dimension. The specific mathematical expression is shown in Eq. (5).

$$z_s = [\text{AvgPool}(E_0); \text{MaxPool}(E_0)] \quad (5)$$

In Eq. (5), $[\cdot; \cdot]$ represents the feature concatenation operation. $z_s \in \mathbb{R}^{H' \times W' \times 2}$ represents the fused spatial description features. Based on this, SA weights are generated through one-dimensional convolutional mapping, and CA and SA are applied synergistically to the original feature sequence to obtain the feature representation after lesion perception enhancement, as shown in Eq. (6).

$$\begin{cases} M_s = \sigma(f_s(z_s)) \\ E_1 = M_s \odot (E_0 \odot M_c) \end{cases} \quad (6)$$

In Eq. (6), $M_s \in \mathbb{R}^{H' \times W' \times 1}$ represents the attention weight of each spatial location. $f_s(\cdot)$ represents the SA mapping function. The $\odot$ stands for the element-wise multiplication operation. The $E_1 \in \mathbb{R}^{H' \times W' \times d}$ represents the feature sequence after LAE. The dual attention mechanism highlights disease-relevant regions and features while reducing redundant information (Zhu et al., 2021). The details of the enhancement effect are denoted in Fig. 4.

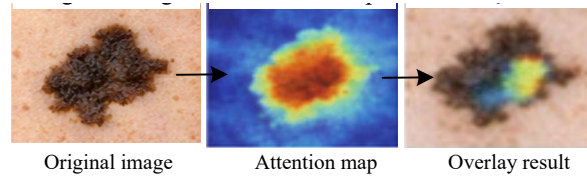


Fig. 4. Schematic diagram of the synergistic enhancement effect of CA and SA

In Fig. 4, the synergistic effect of CA and SA leads the model to exhibit a stronger activation response at the feature level in regions highly correlated with skin disease discrimination, while background regions and features with low correlation and disease discrimination are effectively suppressed. It can be observed in Fig. 4 that the attention-enhanced features are more spatially concentrated in potential lesion areas, and their response boundaries are more consistent with lesion morphology.

2.3. A Transformer-Based Global Semantic Association Modeling Method for Lesion Regions

After completing Local Fine-grained Modeling (LFM) and LAE, the study introduces a Transformer-based encoding structure to model global semantic associations over the lesion perception feature sequence. The Transformer encoder first projects the input features to the query Q , key K and value V subspace through linear mapping, and its calculation process is shown in Eq. (7).

$$\begin{cases} Q = E_1 W_Q \\ K = E_1 W_K \\ V = E_1 W_V \end{cases} \quad (7)$$

In Eq. (7), W_Q , W_K , and W_V represent the learnable projection matrices. Based on this, the correlation between different spatial locations within the feature sequence is modeled using a scaled dot-product attention mechanism, as shown in Eq. (8).

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

In Eq. (8), QK^T represents the similarity matrix between any two spatial locations in the feature sequence. $\sqrt{d_k}$ refers to the scaling factor. The model's capacity to express varied semantic relationships is enhanced by extending the aforementioned process with a multi-head SAM. The specific mathematical expression is shown in Eq. (9).

$$E_2 = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_o \quad (9)$$

In Eq. (9), $E_2 \in \mathbb{R}^{H' \times W' \times d}$ represents the feature sequence representation after global semantic association modeling. head_i denotes the output of the i -th attention head. The h means the number of attention heads. The W_o represents the output mapping. Combining the above, the flowchart of the ATS-Former algorithm is shown in Fig. 5.

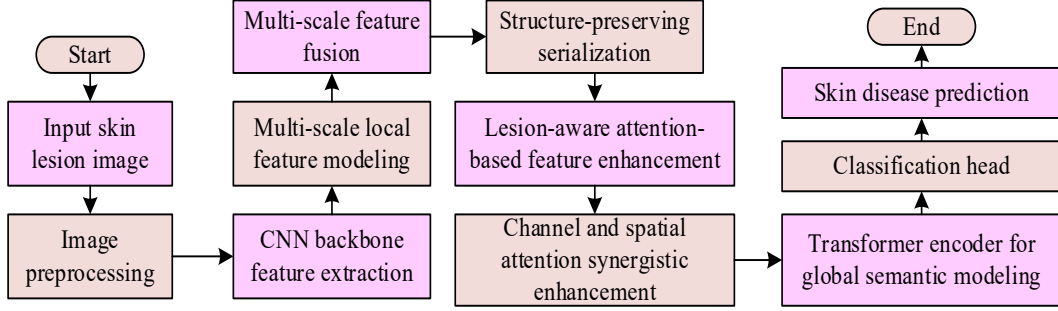


Fig. 5. Flowchart of the ATS-Former algorithm

Fig. 5 shows the ATS-Former pipeline. After preprocessing and CNN-based feature extraction, LFM captures multi-scale local features, which are fused and serialized. LAE enhances relevant features via channel and SA. The Transformer encoder then models global relationships, and the classification head produces the final result.

3. Results

3.1. Experimental Setup

To test the effectiveness of the developed skin disease identification algorithm, experiments were conducted on the publicly available ISIC skin disease image dataset. The experimental hardware and software environment is shown in Table 1.

Table 1. Hardware and software environment

Category	Configuration
Operating system	Ubuntu 20.04 LTS
CPU	Intel Core i9 series processor
GPU	NVIDIA RTX series GPU
GPU memory	24 GB
System memory	64 GB RAM
Deep learning framework	PyTorch
CUDA version	CUDA 11.7
cuDNN version	cuDNN 8.4
Programming language	Python 3.8

In the preprocessing stage, the input images were uniformly resized to 224×224 , and the local feature embedding dimension was set to 256. The Transformer encoder consisted of 4 layers, with 8 attention heads in each layer. Training used the Adam optimizer with an initial learning rate of 1×10^{-4} , followed by a learning-rate decay schedule. The batch size was 32, the maximum number of training epochs was 100, and the loss function was cross-entropy.

The experiment was conducted using the publicly available ISIC 2018 dermatology image dataset. This dataset contains 10,015 dermoscopic images covering seven common skin diseases: Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevi (NV), and Vascular Lesions (VASC). The specific details are shown in Fig. 6.

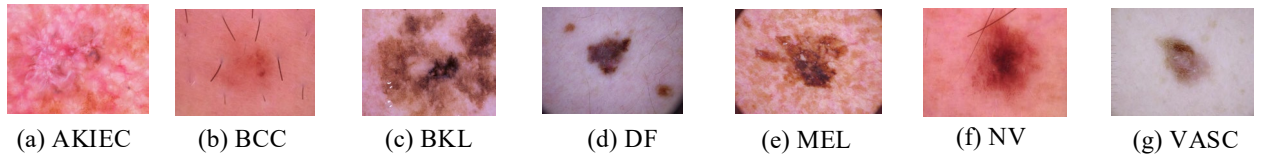


Fig. 6. Example dermoscopic images of seven skin disease categories in the ISIC 2018 dataset

The dataset is characterized by class imbalance and subtle feature differences. To mitigate the impact of class imbalance on model evaluation, the experiment used macro-average metrics, such as Macro-average F1-score (Macro-F1) and Macro-average Area Under the ROC Curve (Macro-AUC), to assess model performance.

In the experiment, it was divided into a training set of 7,011 images, a validation set of 1,002 images, and a test set of 2,002 images. During the model training phase, basic data augmentation was applied to the input images, including random horizontal flipping, random rotation, and brightness adjustment. Simultaneously, the model was trained multiple times with different random initializations, and the average performance metric was reported as the final result to assess stability.

3.2. Quantitative Results and Comparative Analysis

To comprehensively assess the skin disease recognition performance of the proposed method, Vision Transformer (ViT), Swin T, Convolutional Network Next (ConvNeXt), and Transformer-based U-Net (TransUNet) were selected as comparison models, and all models were reproduced using the official default configurations. On the ISIC 2018 test set, performance was compared using four metrics: Macro-F1, Weighted-F1, Macro-AUC, and Matthews Correlation Coefficient (MCC). The specific results are shown in Fig. 7.

In Fig. 7(a), ATS-Former's Macro-F1 was 0.90, which is 7.14%, 3.45%, 4.65%, and 0.78% higher than ViT, Swin T, ConvNeXt, and TransUNet, respectively. This indicates that ATS-Former enables balanced recognition of both common and rare skin diseases. Fig. 7(b) shows that ATS-Former's Macro-AUC was 0.97, which is 4.30% and 2.11% higher than ViT and Swin T, respectively. The MCC results indicate stable discrimination across thresholds, supporting reliable screening and triage decisions. Furthermore, the study also compared performance at the dataset category level, and the results are shown in Fig. 8.

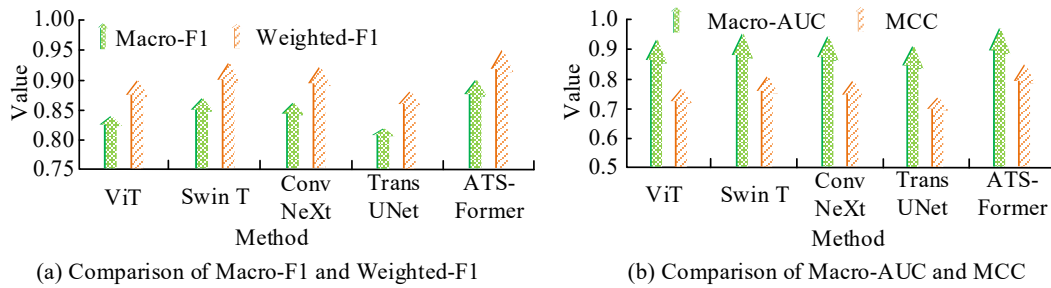


Fig. 7. Performance comparison of different methods

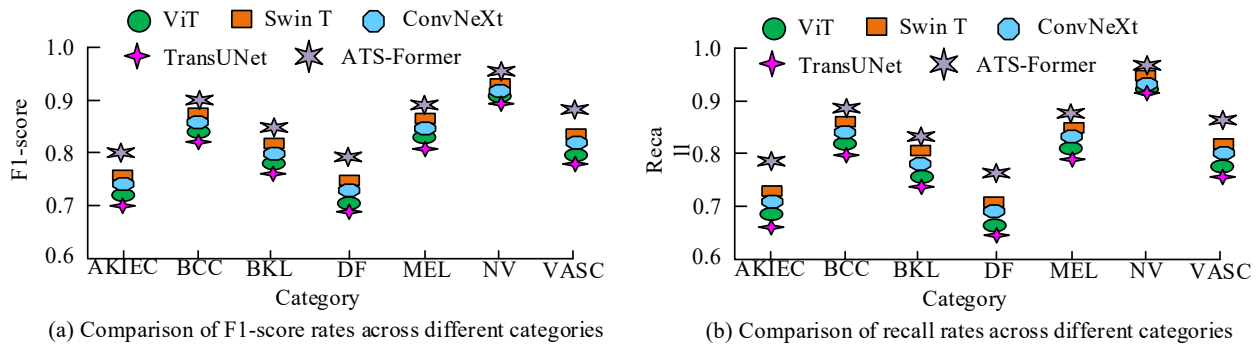


Fig. 8. Comparison of recognition performance of different models for each category

Fig. 8(a) shows that ATS-Former achieves higher F1 scores across key classes, while Fig. 8(b) demonstrates consistently higher recall, indicating more balanced performance. ROC curves and confidence distributions further validate its discriminative ability and decision stability. Therefore, this study compared the ROC curves and confidence scores of different methods, as shown in Fig. 9.

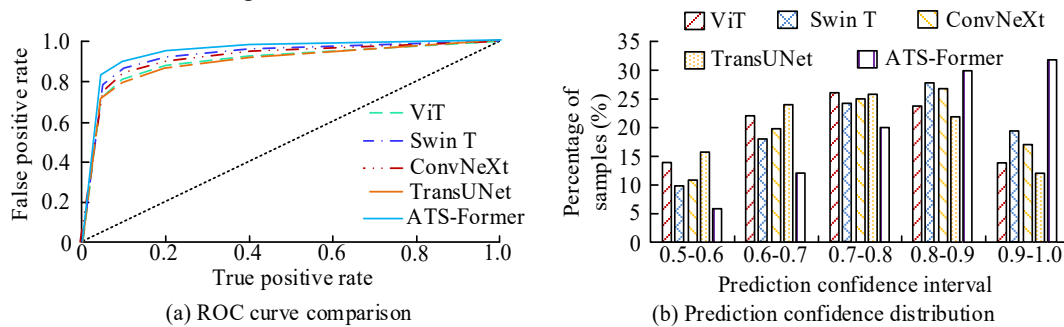


Fig. 9. Comparison of discriminative ability and prediction stability

In the ROC curves of Fig. 9(a), when the False Positive Rate (FPR) was ≤ 0.10 , the true positive rate of ATS-Former reached 0.83 and 0.90, significantly higher than ViT, Swin T, ConvNeXt, and TransUNet. This shows that ATS-Former maintains strong lesion detection under strict thresholds, enabling efficient identification of high-risk cases. Fig. 9(b) shows the distribution of prediction confidence of different methods. ATS-Former had a 61.93% proportion of predicted samples in the high confidence interval (≥ 0.80), which is significantly better than ViT (37.85%), Swin T (47.59%), ConvNeXt (44.06%), and TransUNet (34.04%), with improvements of 63.61%, 30.16%, 40.57%, and 81.93%, respectively. At the same time, its low confidence (< 0.70) sample proportion was significantly reduced, indicating that the prediction results are more concentrated and the uncertainty is lower. From a clinical decision-making perspective, higher predictive confidence enables more reliable risk stratification. High-confidence cases can be prioritized for specialist review, while low-risk cases may undergo routine follow-up. Thus, ATS-Former provides more stable decision support across different diagnostic thresholds and improves clinical screening efficiency. Furthermore, the study compared the specificity, precision, Negative Predictive Value (NPV), and FPR of each method to further evaluate misdiagnosis control and predictive reliability, as denoted in Table 2.

As shown in Table 2, ATS-Former achieved a specificity of 95.21%, which is 3.33%, 1.85%, 2.41%, and 3.92% higher than ViT, Swin T, ConvNeXt, and TransUNet, respectively. Higher specificity means the model can more accurately exclude non-target categories in clinical screening scenarios, thereby reducing unnecessary further examinations or misdiagnosis. This is significant for reducing medical costs and optimizing screening processes. Combining precision and NPV metrics further illustrates the model's advantages in predictive reliability. Higher precision indicates greater confidence in making positive predictions, while higher NPV indicates greater safety in classifying samples as non-target categories, thereby improving decision reliability in clinical screening. Furthermore, ATS-Former's FPR was reduced by 4.79%, 39.06%, and 26.53% compared with ViT and Swin T, respectively. Lower FPR means reduced false alarms and unnecessary medical interventions in practical applications, thereby reducing medical resource consumption and improving diagnostic efficiency. Overall, ATS-Former not only achieves excellent predictive accuracy but also offers advantages in misdiagnosis control, screening cost control, and clinical decision reliability.

Table 2. Comparison of additional evaluation metrics on the ISIC 2018 dataset

Method	Specificity (%)	Precision (%)	NPV (%)	FPR (%)	p-value
ViT	92.14 $\pm$ 0.42	86.37 $\pm$ 0.55	88.92 $\pm$ 0.48	7.86 $\pm$ 0.36	<0.01
Swin T	93.48 $\pm$ 0.39	88.65 $\pm$ 0.47	90.11 $\pm$ 0.44	6.52 $\pm$ 0.31	<0.05
ConvNeXt	92.97 $\pm$ 0.41	87.84 $\pm$ 0.50	89.63 $\pm$ 0.46	7.03 $\pm$ 0.33	<0.05
TransUNet	91.62 $\pm$ 0.44	85.71 $\pm$ 0.59	87.46 $\pm$ 0.52	8.38 $\pm$ 0.41	<0.05
ATS-Former	95.21 $\pm$ 0.31	90.38 $\pm$ 0.36	91.84 $\pm$ 0.35	4.79 $\pm$ 0.24	-

Note: The results are reported as mean $\pm$ standard deviation over three independent runs. Statistical significance was evaluated using a paired t-test between ATS-Former and the baseline models.

3.3. Ablation Experiments and Robustness Analysis

To further explore the contributions of each key module in ATS-Former to improved model performance, experiments were conducted on the ISIC 2018 dataset. The ablation validation findings are shown in Table 3.

In Table 3, removing the LAE module reduced Macro-F1 by 4.44% and MCC by 5.88%, indicating that lesion-perception attention effectively enhances the model's focus on key regions and its predictive stability. Removing the LFM module reduced Macro-F1 by 3.33%, demonstrating that local fine-grained texture modeling is a fundamental support for skin disease discrimination. Removing the GTE module reduced Macro-AUC by 1.03%, and MCC also decreased, indicating that the Transformer's global semantic modeling plays a complementary role in improving class discriminability and discriminative consistency. The three modules work together to constitute the performance advantage of ATS-Former.

Table 3. Ablation results of various modules on ISIC 2018

LFM	LAE	GTE	Macro-F1	Weighted-F1	Macro-AUC	MCC
×	√	√	0.87	0.93	0.95	0.81
√	×	√	0.86	0.92	0.95	0.8
√	√	×	0.88	0.94	0.96	0.83
√	√	√	0.90	0.95	0.97	0.85

To compare the theoretical computational complexity of different models, this paper analyzed the dominant complexity of each method. The main complexity of ViT and TransUNet comes from global self-attention, with a computational cost typically of $O(LN^2d)$. Swin T uses a window attention mechanism, and its complexity can be approximated as $O(LNwd)$, lower than standard global attention. The complexity of ConvNeXt is mainly determined by convolution operations, and can be represented as $O(\sum_{l=1}^L H_l W_l k_l^2 C_l^{in} C_l^{out})$. Here, L represents the number of network layers, d represents the length of the input sequence, N represents the feature dimension, w represents the number of tokens within the window, H_l and W_l represent the spatial dimensions of the feature map of the l -th layer, k_l represents the convolution kernel size. C_l^{in} and C_l^{out} represent the number of input and output channels, respectively. ATS-Former consists of multi-scale local modeling, a LAE module, and a Transformer encoder. The dominant complexity of the global modeling part is also, while the local modeling and attention enhancement modules mainly introduce controllable convolution and linear-mapping overhead. Overall, ATS-Former is comparable to ViT and TransUNet in theoretical complexity but achieves a better balance between performance and complexity through more effective local feature extraction and lesion enhancement.

Based on this, the study applied Gaussian noise and height perturbation to the input image to evaluate the robustness of ATS-Former in real clinical acquisition scenarios, as shown in Fig. 10.

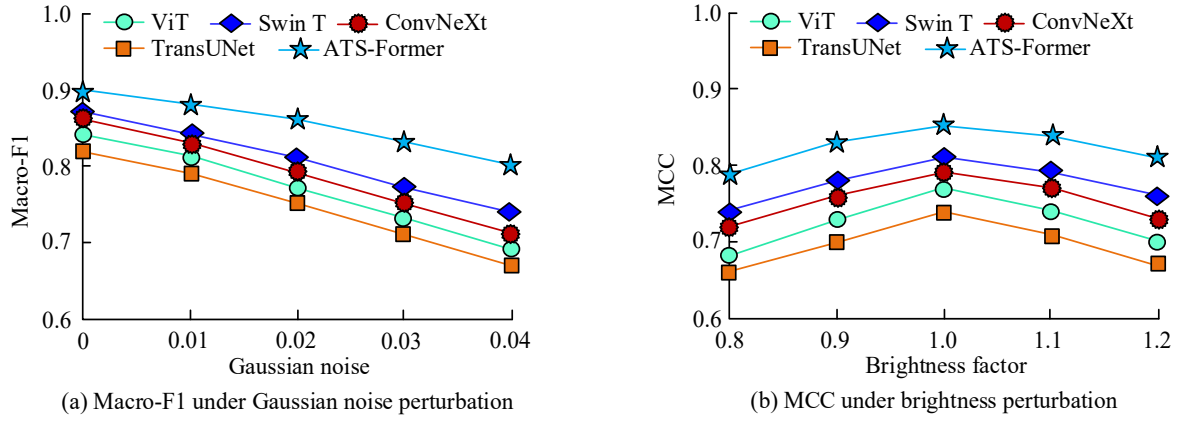


Fig. 10. Robustness comparison of different methods under common image perturbations

As shown in Fig. 10(a) and (b), ATS-Former exhibited the smallest performance degradation under both Gaussian noise and brightness perturbations, indicating greater robustness than baseline models. To further demonstrate the recognition effects of different models, some representative skin disease images were selected for visual comparison, and the results are shown in Fig. 11.

As shown in Fig. 11, when lesion boundaries were blurred or texture differences were small, models such as ViT and TransUNet were prone to misclassification, while ATS-Former correctly identified lesion categories significantly more often. This indicates that the proposed method is more effective in skin disease identification tasks.

3.4. Extended Validation

To further evaluate the model's generalization capability across different data distributions, a cross-dataset validation experiment was conducted. After training on the ISIC 2018 dataset, the model was directly evaluated on the HAM10000 dataset without retraining or fine-tuning. The HAM10000 dataset contains dermoscopic images collected from various medical institutions and using different acquisition devices, resulting in differences in image resolution, lesion distribution, and sample composition compared with ISIC 2018. All models were trained three times using different random seeds, and the mean and standard deviation were reported. The statistical significance of differences in performance among methods was evaluated using a paired t-test. Results are shown in Table 4.

As shown in Table 4, compared with ViT, Swin T, ConvNeXt, and TransUNet, ATS-Former's Macro-F1 score was improved by 6.10%, 3.57%, 4.82%, and 7.41%, respectively. Statistical tests show that this performance improvement is statistically significant ($p < 0.05$). This indicates that the proposed multi-scale feature modeling and LAE mechanism can, to some extent, mitigate performance degradation due to changes in data distribution, thereby improving the model's cross-dataset generalization.

4. Discussion and Conclusion

To address the challenges of large-scale variation in lesion regions, subtle fine-grained feature differences, and insufficient global semantic modeling in dermatological images, this study proposes the ATS-Former model, which integrates multi-scale local feature modeling, LAE, and Transformer-based global semantic modeling. Experimental results demonstrate that the proposed method achieves superior recognition performance compared with baseline models such as ViT, Swin T, ConvNeXt, and TransUNet on the ISIC 2018 dataset, and maintains good stability in cross-dataset validation. By introducing multi-scale local feature enhancement and lesion-aware mechanisms into the Transformer architecture, the proposed model

effectively improves the representation of fine-grained features in skin disease images, thereby enhancing recognition performance under complex discrimination conditions.

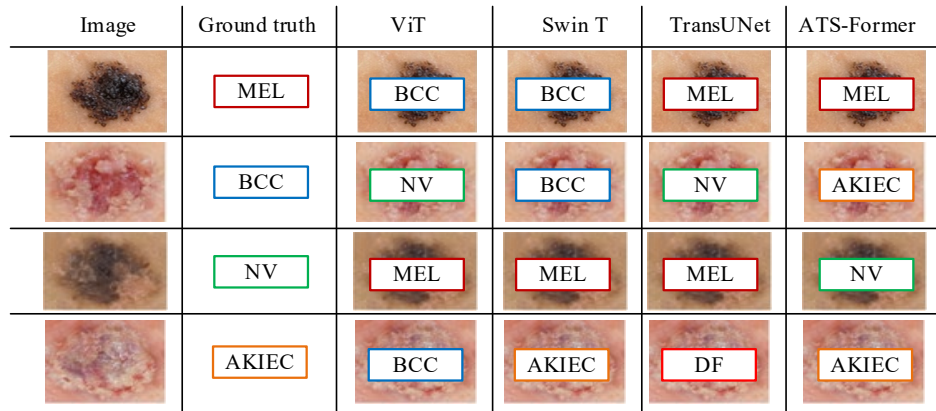


Fig. 11. Qualitative comparison of different methods on representative dermoscopic images

Table 4. Comparison of generalization performance across datasets (train: ISIC2018, test: HAM10000)

Method	Macro-F1	Macro-AUC	MCC
ViT	0.82 ± 0.012	0.90 ± 0.010	0.76 ± 0.013
Swin T	0.84 ± 0.011	0.92 ± 0.009	0.78 ± 0.012
ConvNeXt	0.83 ± 0.013	0.91 ± 0.011	0.77 ± 0.014
TransUNet	0.81 ± 0.014	0.90 ± 0.012	0.75 ± 0.015
ATS-Former	0.87 ± 0.010	0.94 ± 0.008	0.81 ± 0.011

From a theoretical perspective, this study proposes a hybrid modeling framework that integrates local structural information with global semantic relationships, providing a new feature representation approach for fine-grained medical image recognition tasks. Compared with single convolutional architectures or standard Transformer models, the proposed method improves recognition of small lesions and weakly textured regions through the coordinated design of multi-scale feature modeling and attention mechanisms.

From a practical perspective, ATS-Former can be deployed as an auxiliary decision-support tool for preliminary screening of skin diseases in telemedicine or mobile healthcare systems. In real-world screening scenarios, the model's prediction confidence can be used to define diagnostic thresholds and support risk-based case triage. For instance, high-confidence suspected high-risk lesions can be prioritized for specialist review, while low-risk cases can be scheduled for routine follow-up. In this way, hospital managers can use the model to support more efficient diagnostic triage and medical resource allocation, thereby improving screening efficiency and reducing clinicians' workload.

Despite the promising results, several limitations remain. First, the experiments were primarily conducted on the ISIC 2018 dataset, with additional cross-dataset validation on the HAM10000 dataset. Further validation on larger multi-center datasets is required to fully assess the model's generalization capability. Second, the current study focuses on image-based classification and does not incorporate clinical metadata or multimodal information. Future work may integrate multi-source data and lightweight model designs to further enhance the feasibility of deploying the proposed method in real-world clinical environments.

Author Contributions

Qinlan Wu contributed to conceptualization, methodology, supervision, project administration, and funding acquisition. Hailin Zou contributed to software, validation, data collection, and formal analysis. Xijian Chao contributed to investigation, visualization, draft preparation, and manuscript editing.

Funding

This research is supported by Jiangxi University of Engineering Research Project, Innovation and Practice of 'Multi-Linked' Big Data Development Technology Course, 2025-JGJG-23. Jiangxi University of Engineering Research Project, Research on the Reform and Innovation of the Practice Teaching Model for Industry-Academia Collaborative Education in Vocational Colleges, 2024-JGJG-61, Xinyu Municipal Bureau of Science and Technology: Optimization and Improvement of the AI-based Dynamic Evaluation Platform for College Classroom Teaching Quality, 20253090988.

Institutional Review Board Statement

The data used in this study came from the publicly available dataset ISIC 2018 (International Skin Imaging Collaboration). No new human trials or patient data collection were involved in the study. Therefore, no ethical approval was required.

Declaration of Artificial Intelligence (AI) Tools

No artificial intelligence (AI) tools were used in the development, analysis, writing, or revision of this manuscript.

References

- Adnanto, H. F., Sari, A. P., and Junaidi, A. (2025). Autoimmune skin disease image classification using EfficientViT-M1 with AdamW optimizer. *bit-Tech*, 8(2), 2599-2609. <https://doi.org/10.32877/bt.v8i2.3300>.
- Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiaq, T., Rafa, N., and Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11), 13521-13617. <https://doi.org/10.1007/s10462-023-10466-8>.
- Du, H., Wang, J., Liu, M., Wang, Y., and Meijering, E. (2022). SwinPA-Net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4), 5355-5366. <https://doi.org/10.1109/TNNLS.2022.3204090>.
- Göçeri, E. (2021). An application for automated diagnosis of facial dermatological diseases. *İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi*, 6(3), 91-99. <https://orcid.org/0000-0002-2329-4107>.
- Gupta, P., Nirmal, J., and Mehendale, N. (2024). A survey of recent advances in analysis of skin images. *Evolutionary Intelligence*, 17(5), 4155-4178. <https://doi.org/10.1007/s12065-024-00977-w>.
- Gupta, P., Nirmal, J., and Mehendale, N. (2025). A survey on computer vision approaches for automated classification of skin diseases. *Multimedia Tools and Applications*, 84(11), 8673-8705. <https://doi.org/10.1007/s11042-024-19301-w>.
- Hassan, F., Deif, M. A., Zaghloul, A., and Elgohary, R. (2026). Engineering predictive AI/ML pipelines for EGFR-TKI resistance in NSCLC: A systematic review. *Journal of Engineering, Project, and Production Management*, 16(5), 2025-315 (in press). <https://doi.org/10.32738/JEPPM-2025-315>.
- Jaber, N. J. F. and Akbas, A. (2025). Melanoma skin cancer detection based on deep learning methods and binary Harris Hawk optimization. *Multimedia Tools and Applications*, 84(22), 25709-25722. <https://doi.org/10.1007/s11042-024-19864-8>.
- Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., and Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7), 1200-1217. <https://doi.org/10.1109/JAS.2022.105686>.
- Pacal, I., Alaftekin, M., Zengul, F. D. (2024). Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and SwiGLU-based MLP. *Journal of Imaging Informatics in Medicine*, 37(6), 3174-3192. <https://doi.org/10.1007/s10278-024-01140-8>.
- Pang, Y., Liang, J., Huang, T., Chen, H., Li, Y., Li, D., and Wang, Q. (2023). Slim UNETR: Scale hybrid transformers to efficient 3D medical image segmentation under limited computational resources. *IEEE Transactions on Medical Imaging*, 2023, 43(3), 994-1005. <https://doi.org/10.1109/TMI.2023.3326188>.
- Saba T. (2021). Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features. *Microscopy Research and Technique*, 84(6), 1272-1283. <https://doi.org/10.1002/jemt.23686>.
- Sattarzadeh, A. R., Pathirana, P. N., Kutadinata, R., and Huynh, V. T. (2024). Extracting long-term spatiotemporal characteristics of traffic flow using attention-based convolutional transformer. *IET Intelligent Transport Systems*, 18(10), 1797-1814. <https://doi.org/10.1049/itr2.12468>.
- Wei, X. S., Song, Y. Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., and Belongie, S. (2021). Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 8927-8948. <https://doi.org/10.1109/TPAMI.2021.3126648>.
- Wen, Y., Zeng, Y., Liu, S., Xiong, L., Li, H., Wang, Y., and Liang, Z. (2025). DSF-Net: A dual-stream fusion network integrating structural and detailed features for fundus-based diabetic retinopathy classification. *The Visual Computer*, 41(15), 12839-12851. <https://doi.org/10.1007/s00371-025-04187-y>.
- Yanagisawa, Y., Shido, K., Kojima, K., Yamasaki, K. (2023). Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images. *Journal of Dermatological Science*, 109(1), 30-36. <https://doi.org/10.1016/j.jdermsci.2023.01.005>.
- Zhou, J., Liu, L., Wei, W., and Fan, J. (2022). Network representation learning: from preprocessing, feature extraction to node embedding. *ACM Computing Surveys (CSUR)*, 55(2), 1-35. <https://doi.org/10.1145/3491206>.
- Zhu, H., Zhang, C., Wei, Y., Huang, S., and Zhao, Y. (2023). ESA: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10), 6131-6143. <https://doi.org/10.1109/TCSVT.2023.3253548>.
- Zhu, W., Sun, L., Huang, J., Han, L., and Zhang, D. (2021). Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. *IEEE Transactions on Medical Imaging*, 40(9), 2354-2366. <https://doi.org/10.1109/TMI.2021.3077079>.



Qinlan Wu has earned her master's and Ph.D. in Software Engineering, and is currently an Associate Professor and Vice Dean at the School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, China. She serves as an expert on the Information Engineering Testing Professional Committee of the China Electronics Chamber of Commerce, as well as a Member of the Xinyu Big Data Development Strategy Expert Pool and the Xinyu Digital Economy Talent Expert Pool. Her research interests include data analysis and mining, computer technology, big data applications, information security, software engineering and software development.



Hailin Zou has earned a master's degree majoring in mathematics and applied mathematics. He is currently serving as an Associate Professor at the School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, China. He has long been engaged in the field of computer technology application, focusing on teaching reform and scientific research innovation, and serves as the host of university-level teaching reform projects and a core researcher of scientific research projects. His research interests include artificial intelligence and data analysis.



Xijian Chao is currently a Professor at the School of Artificial Intelligence and Data Science, Jiangxi Polytechnic Institute, Xinyu, China, and a Ph.D. candidate. He serves as a Member of the Jiangxi Provincial Teaching Steering Committee for Computer Majors, is a Member of the Xinyu Big Data Development Strategy Expert Pool, and an Expert of the Xinyu Digital Economy Talent Expert Pool. He has long been engaged in the fields of artificial intelligence, big data applications and virtual reality, presided over many national social science fund projects and provincial-level projects and led the promotion of teaching reform and school-enterprise cooperation. His research interests include artificial intelligence, technology application, big data analysis, and virtual reality.