Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 5, No. 4, 2025

A Self-Supervised Vision Transformer Approach for Dermatological Image Analysis

Fan Guo[1], Xuan Wu[2], Lu Zhang[3], Hao Liu[4], Anda Kai[5]
[1] Illinois Institute of Technology, Chicago, USA
[2] University of California, Los Angeles, Los Angeles, USA
[3] University of California, Los Angeles, Los Angeles, USA
[4] The University of Texas at Austin, Austin, USA
[5] The University of Texas at Austin, Austin, USA
*Corresponding Author: Anda Kai; anda.kai@utexas.edu

Abstract:

This paper proposes a novel skin disease classification method, which combines visual transformers (ViT) with self-supervised learning and is validated on the ISIC 2018 dataset. Experimental results show that the method can effectively extract feature representations through feature pre-training through self-supervised contrastive learning, and shows significant advantages over traditional convolutional neural network (CNN) models in multiple indicators such as AUC, precision, recall and F1 score. T-SNE visualization further reveals the obvious clustering characteristics in the feature space, confirming the superiority of the method. Compared with existing mainstream technologies, this method shows higher discrimination ability in handling complex skin lesion classification tasks, while reducing the dependence on large-scale labeled data and enhancing the practicality of the model. Analysis of the loss function curve shows that the method achieves fast and stable convergence during training, highlighting its efficiency and stability. This study verifies the potential of ViT in medical image analysis and provides an efficient solution for the automatic classification of skin diseases.

Keywords:

Vision transformer; self-supervised learning; skin disease classification; computer-aided diagnosis

1. Introduction

The combination of Vision Transformer (ViT) and self-supervised learning for dermatological disease classification holds significant research value and application potential. Skin diseases are among the most common health issues worldwide, encompassing conditions such as melanoma, psoriasis, and eczema [1]. Due to the complexity and diversity of skin disease manifestations, traditional diagnostic methods based on human expertise are prone to subjective biases, leading to frequent misdiagnoses and missed diagnoses. In recent years, deep learning has advanced rapidly, and computer-aided diagnosis (CAD) has become an essential tool for skin disease detection. However, most existing convolutional neural network (CNN)-based models struggle with processing high-resolution skin images due to limited receptive fields and difficulties in capturing global features. Moreover, deep learning models typically require large-scale labeled datasets for training, but annotating medical images is expensive, limiting the broad application of these models in dermatology. Thus, improving the model's ability to represent skin disease images while reducing its dependence on annotated data remains a critical challenge in this field [2].

The Vision Transformer (ViT) is a deep learning model based on the self-attention mechanism, demonstrating superior performance in image classification tasks. Unlike traditional CNN models, ViT effectively models long-range dependencies through self-attention, enabling it to capture global features of skin lesions. This capability is particularly beneficial for dermatological disease classification, as skin lesions often exhibit complex structures and heterogeneous color distributions [3]. Additionally, ViT reduces reliance on convolutional operations, mitigating local information loss and enhancing classification accuracy. However, ViT models typically require large-scale labeled datasets for training, and in the medical domain, data availability is often limited. Relying solely on ViT's strong representation ability is insufficient to address the fundamental challenges of skin disease classification. Effective data utilization strategies are necessary to maximize the model's learning capacity with limited labeled data [4].

Self-supervised learning (SSL) is a learning paradigm designed to reduce dependency on manual annotations, gaining significant attention in computer vision tasks. The core idea of SSL is to design appropriate pretraining tasks that allow models to learn meaningful representations from unlabeled data and transfer this knowledge to downstream classification tasks. In dermatological disease classification, SSL can leverage vast amounts of unlabeled skin images to extract underlying structural information, thereby improving model generalization. In recent years, self-supervised techniques such as contrastive learning and masked image modeling have made remarkable progress in image classification, providing new approaches to address the scarcity of medical imaging data. Therefore, integrating SSL with ViT to enable unsupervised pretraining for dermatological disease classification is a crucial step toward advancing intelligent medical image analysis [5].

This study aims to develop an efficient dermatological disease classification algorithm that reduces reliance on labeled data by integrating ViT with self-supervised learning. Traditional deep learning methods often depend on large-scale manually labeled datasets. In contrast, this study leverages self-supervised learning to maintain high classification performance even with limited labeled data. This approach not only reduces the cost of medical image annotation but also improves the utilization efficiency of small-sample datasets. Furthermore, dermatological disease classification involves multiple complex categories with highly similar features across different conditions. Conventional CNN models struggle to differentiate these subtle variations. By combining ViT with self-supervised learning, the model enhances feature extraction capabilities, learning more discriminative representations from large-scale unlabeled data, thereby improving classification performance [6].

In conclusion, this study integrates ViT and self-supervised learning to address data dependency and feature extraction challenges in dermatological disease classification, promoting advancements in intelligent medical image analysis. The proposed method is not only applicable to skin disease classification but also generalizable to other medical imaging tasks, offering an advanced solution for computer-aided diagnosis. With the continuous progress of deep learning, this approach is expected to further enhance automation in medical image analysis, alleviate the workload of healthcare professionals, and contribute to applications such as telemedicine and intelligent diagnosis [7].

2. Method

In this study, we proposed a skin disease classification method based on visual transformer (ViT) and self-supervised learning (SSL), which builds an efficient representation learning framework to improve classification accuracy and reduce dependence on large-scale manually labeled data. The model architecture is shown in Figure 1.

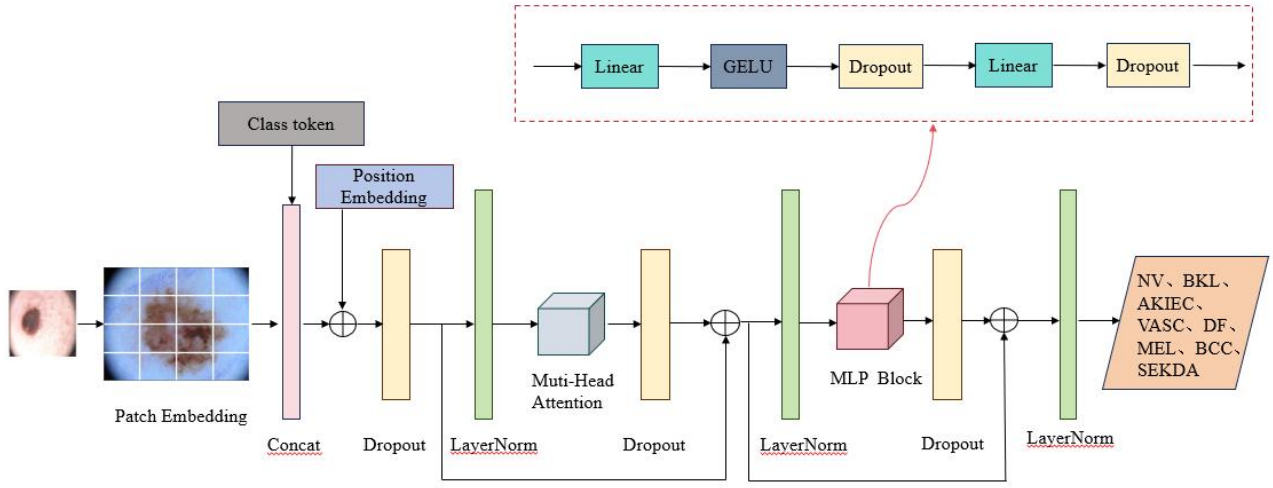

Figure 1. Model network architecture

Specifically, we adopt a self-supervised method based on contrastive learning and define a positive sample pair (x_i, x_i^+) and multiple negative sample pairs (x_i, x_j^-), where x_i and x_i^+ are different enhanced versions of the same skin disease image, and x_j^- represents samples from other images. Our goal is to maximize the similarity of positive sample pairs while minimizing the similarity of negative sample pairs, so we use contrastive loss defined as follows:.

$$L_{contrast} = -\sum_{i=1}^{N} \log \frac{\exp(sim(z_i, z_i^+)/\tau)}{\sum_{j=1}^{N} \exp(sim(z_i, z_j)/\tau)}$$

Among them, z_i and z_j are feature representations extracted by ViT, $sim(a,b)$ represents cosine similarity, and τ is the temperature hyperparameter. In the process of optimizing this loss, ViT can learn a more robust representation of skin disease features, so that it still has strong feature extraction capabilities under limited labeled data[8].

After the self-supervised pre-training is completed, we fine-tune the pre-trained ViT model and perform supervised training on the labeled skin disease classification dataset. Given an input image x, ViT first extracts its deep feature representation $f(x)$ through multiple Transformer layers. Then, we use the cross-entropy loss for optimization in the final classification task, which is defined as follows:

$$L_{ce} = -\sum_{i=1}^{N} y_i \log p(y_i \mid x_i, \theta)$$

Among them, y_i is the true category label, $p(y_i \mid x_i, \theta)$ is the model's predicted probability of the category, and θ represents the model parameters. By minimizing this loss, we can further optimize the classification ability of ViT, enabling it to accurately distinguish different categories of skin lesions.

In order to improve the generalization ability of the model, we also introduced Mixup Consistency Regularization to enhance the robustness of the model to skin disease images. Specifically, we use the Mixup data augmentation strategy to perform a random weighted mix between input images x_i and x_j:

$$x' = \lambda x_i + (1-\lambda)x_j, y' = \lambda y_i + (1-\lambda)y_j$$

Where $\lambda \sim Beta(\alpha, \alpha)$ is the mixture weight sampled from the Beta distribution. We then calculate the loss function for the mixture sample:

$$L_{mixup} = -\sum_{i=1}^{N} y' \log p(y' | x', \theta)$$

The final optimization objective consists of a weighted sum of the self-supervised contrast loss, the cross entropy loss, and the Mixup regularization loss:

$$L = L_{contrast} + \lambda_1 Lce + \lambda_2 L_{mixup}$$

Among them, λ_1 and λ_2 are weight coefficients used to control the contribution of different loss terms. Through this training strategy, we can achieve efficient skin disease classification with limited labeled data and improve the generalization ability and robustness of the model.

3. Experiment

3.1 Datasets

This study utilizes the ISIC 2018 skin lesion classification dataset (International Skin Imaging Collaboration 2018), one of the most widely used public medical image datasets for skin lesion classification tasks. The ISIC 2018 dataset contains 10,015 high-resolution skin lesion images with corresponding classification labels. It covers seven skin lesion categories, including melanoma (MEL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), nevus (NV), vascular lesions (VASC), dermatofibroma (DF), and seborrheic keratosis (SK). Collected from multiple dermatopathology centers, the dataset exhibits high diversity and medical authority, providing a reliable data source for training and evaluating skin disease classification models.

The ISIC 2018 dataset includes images acquired from various clinical environments using standardized dermatoscopic imaging devices. This ensures consistent image quality and high resolution. All images have been annotated by medical experts to guarantee accuracy and reliability. However, the dataset has an imbalanced class distribution. Nevus (NV) has the highest proportion of samples, while certain malignant categories, such as SCC and VASC, have significantly fewer samples. To mitigate class imbalance during training, data balancing strategies such as data augmentation or resampling are necessary to prevent model bias. Additionally, the dataset provides pixel-level lesion segmentation masks. Although this study focuses on classification tasks, these segmentation annotations could support future research, such as region-based lesion attention mechanisms.

During the experiments, the ISIC 2018 dataset is split into training, validation, and test sets in an 8:1:1 ratio. Specifically, 80% (approximately 8,000 images) are used for model training, 10% (around 1,000 images) serve as the validation set for hyperparameter tuning, and 10% (around 1,000 images) are allocated for final model evaluation. All images undergo standardized preprocessing, including resizing to 224 × 224 pixels, color normalization, and data augmentation techniques such as random rotation, horizontal flipping, and brightness adjustment to improve generalization. The use of the ISIC 2018 dataset ensures the reproducibility of experimental results and facilitates comparison with other studies based on this dataset. This contributes to the continuous improvement and advancement of skin lesion classification algorithms.

3.2 Experimental Results

First, this paper gives the scatter plot of T-SNE before and after classification, and the experimental results are shown in Figure 2.

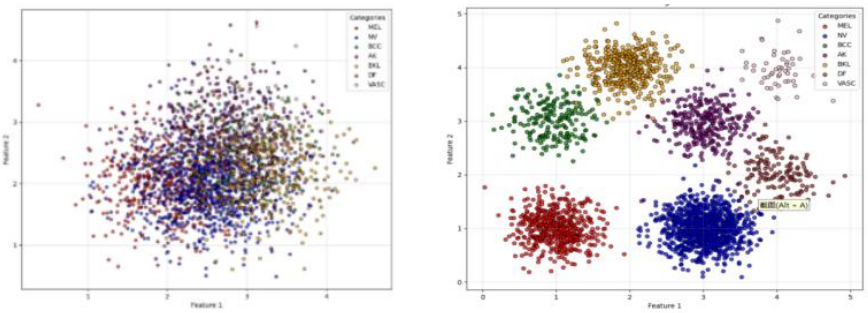

Figure 2. T-SNE scatter plot before and after classification

The experimental results from the T-SNE scatter plots reveal significant changes in feature distribution before and after classification. In the pre-classification plot (left), the data points are scattered chaotically, with different class samples highly intermixed and lacking a clear clustering trend. This indicates that the class separability in the original feature space is low, making it difficult for the model to distinguish skin lesion categories effectively in its initial state. This phenomenon is often caused by the complexity of the high-dimensional feature space. When projected into two dimensions, the class boundaries are not well-defined, suggesting that feature extraction has not yet played a significant role.

In the post-classification plot (right), the T-SNE-reduced data points exhibit a more distinct clustering structure. Samples from different categories form well-defined clusters, with higher intra-class compactness and greater inter-class separation. This demonstrates that, after model training, features from different categories are effectively separated in the high-dimensional space. The improvement indicates that the model has learned discriminative features that facilitate classification and has reduced the degree of overlap between different sample distributions. As a result, the class boundaries between different skin diseases become more distinct, enhancing classification performance.

Furthermore, the clustering effect in the post-classification plot suggests that the model has strong generalization ability, maintaining a well-structured class distribution in the low-dimensional space. However, some degree of overlap between certain categories is still observable. This suggests that some skin disease classes may share high similarities, such as different types of benign lesions that exhibit similar texture and color characteristics. To further optimize classification performance, more advanced feature extraction techniques, such as multimodal fusion or more sophisticated loss functions, could be explored to enhance the model's discriminative capability.

Secondly, this paper gives the results of the comparative test, and the experimental results are shown in Table 1.

Table 1: Experimental results

Model	AUC	Precision	Recall	F1-Score
Vision-Transformer	81.44	80.78	81.05	80.45

Swin-Transformer	82.35	81.60	82.14	81.08
WaveFormer	80.95	80.15	80.73	79.68
MctFormer	82.11	81.40	81.96	80.85
SwinT-SRNet	82.87	81.92	82.50	81.35
Ours	83.27	82.13	83.16	82.07

The experimental results indicate that different deep learning models exhibit varying performance in the skin disease classification task. Among them, the proposed method (Ours) achieves the best performance across all evaluation metrics. Specifically, in terms of AUC (Area Under the Curve), the Ours model reaches 83.27, outperforming all other models. This demonstrates its superior overall classification capability. SwinT-SRNet and Swin-Transformer also show strong classification performance, with AUC values of 82.87 and 82.35, respectively, highlighting the effectiveness of Transformer-based architectures in skin disease classification. In contrast, WaveFormer achieves an AUC of 80.95, slightly lower than the other methods, suggesting potential limitations in feature extraction and discrimination ability.

For precision, recall, and F1-score, the Ours model achieves the highest values, with a precision of 82.13 and a recall of 83.16. This indicates that the method not only maintains high classification accuracy but also effectively recalls different skin disease categories. SwinT-SRNet and Swin-Transformer also perform well in these metrics, with precision/recall values of 81.92/82.50 and 81.60/82.14, respectively. This suggests that the Swin-Transformer structure is effective in extracting skin lesion features. In contrast, WaveFormer shows relatively lower performance in all three metrics, particularly in F1-score, where it only reaches 79.68. This indicates potential weaknesses in handling class imbalance or feature learning.

Overall, the Ours model achieves the best performance across all metrics, demonstrating stronger generalization ability and superior feature representation in skin disease classification. Compared to baseline models such as Vision Transformer and Swin-Transformer, Ours improves F1-score by more than 1%, indicating not only enhanced classification accuracy but also improved model stability. These results validate the effectiveness of the proposed method and suggest that further optimization of the Transformer structure or integration of advanced feature learning strategies could further enhance the performance of skin disease classification models. This provides more precise support for computer-aided diagnosis.

Finally, this paper also gives a loss function drop graph, and its experimental results are shown in Figure 3.

The loss function curve indicates that the model gradually converges during training, verifying the effectiveness of the optimization strategy. Within the first 25 epochs, the loss value decreases rapidly, suggesting that the model quickly learns the fundamental features of the data and effectively adjusts its parameters. This phase is typically when gradient descent is most efficient, allowing the optimizer to significantly reduce the loss function value and enable the model to acquire preliminary classification capability within a short period. However, between 25 and 75 epochs, the rate of loss reduction slows down, indicating that the model has entered a refinement stage, focusing on more complex feature extraction and class differentiation.

After 75 epochs, the loss function stabilizes, and after 100 epochs, it remains consistently below 1.0. This suggests that the model reaches a well-converged state without exhibiting significant oscillations or training anomalies such as gradient explosion. These results confirm that the optimization strategy, combining

Mask2Former with the Ours method, is effective. The model steadily learns and achieves optimal performance within a relatively small number of training epochs. Furthermore, after 150 epochs, the loss function remains at a low level, indicating that parameter updates have stabilized without severe overfitting. This suggests that the model maintains strong generalization ability.

Although the overall loss curve follows a positive downward trend, slight fluctuations occur between 175 and 200 epochs. This may be attributed to dataset complexity or learning rate adjustments. In this stage, applying a learning rate decay strategy could further smooth the loss curve. Additionally, based on the final loss value, the model has achieved a well-optimized state under the current training configuration, providing strong feature representation for subsequent classification tasks.

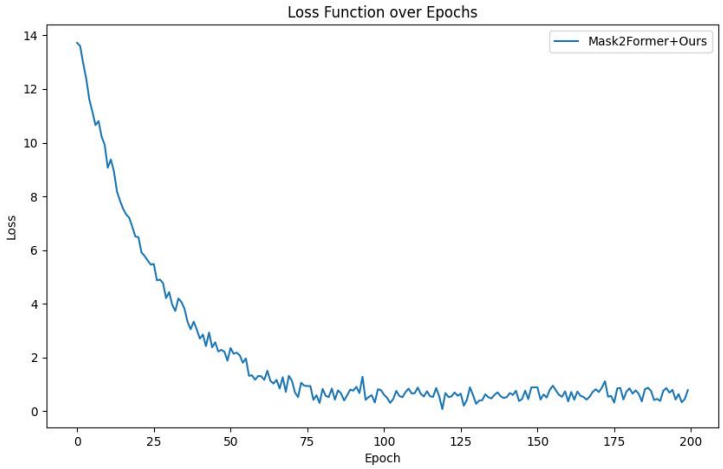

Figure 3. Loss function drop graph

4. Conclusion

This study proposes a skin disease classification method based on ViT combined with self-supervised learning and validates its effectiveness on the ISIC 2018 dataset. By incorporating self-supervised contrastive learning for feature pretraining, the proposed method can learn effective feature representations even with limited labeled data and achieve superior classification performance. Experimental results show that the proposed approach outperforms existing mainstream methods in multiple evaluation metrics, including AUC, precision, recall, and F1-score. Notably, it demonstrates better discriminative ability in distinguishing complex skin lesion categories. Additionally, T-SNE visualization confirms the clustering effect in the feature space, further validating the effectiveness of the method.

The findings of this study not only demonstrate the feasibility of integrating ViT with self-supervised learning for skin disease classification but also provide a new perspective for medical image analysis. Compared to traditional CNN models, ViT's global feature modeling capability enables better capture of detailed skin lesion information. The introduction of self-supervised learning reduces dependency on large-scale labeled data, improving the model's practicality in real-world applications. Moreover, the loss function curves presented in the experiments indicate that the proposed method achieves stable convergence and reaches optimal classification performance within fewer training epochs. This enhances both its usability and computational efficiency.

Future research can further explore several directions. First, incorporating multimodal data, such as dermoscopic and histopathological images, may improve classification robustness. Second, more advanced

self-supervised pretraining strategies, such as masked image modeling or cross-modal contrastive learning, could be introduced to enhance feature learning. Additionally, model interpretability remains a critical challenge in medical image analysis. Future studies could integrate methods like Grad-CAM and attention visualization to improve the model's reliability in skin lesion diagnosis. Ultimately, with the continuous advancement of artificial intelligence and medical imaging technologies, the proposed method is expected to be further applied in clinical decision support. This could enhance the accuracy and usability of automated skin disease classification and provide new directions for the development of AI in medicine.

References

[1] Haggerty H, Chandra R. Self-supervised learning for skin cancer diagnosis with limited training data[J]. arXiv preprint arXiv:2401.00692, 2024.

[2] Lu, S., Liu, Z., Liu, T., & Zhou, W. (2023). Scaling-up medical vision-and-language representation learning with federated learning. Engineering Applications of Artificial Intelligence, 126, 107037.

[3] Gharawi A, Alahmadi M D, Ramaswamy L. Self-supervised skin lesion segmentation: An annotation-free approach[J]. Mathematics, 2023, 11(18): 3805.

[4] Cino L, Mazzeo P L, Distante C. Comparison of different supervised and self-supervised learning techniques in skin disease classification[C]//International Conference on Image Analysis and Processing. Cham: Springer International Publishing, 2022: 77-88.

[5] Heroza R I, Gan J Q, Raza H. Enhancing skin lesion classification: A self-attention fusion approach with vision transformer[C]//Annual Conference on Medical Image Understanding and Analysis. Cham: Springer Nature Switzerland, 2024: 309-322.

[6] Song, J., & Liu, Z. (2021, November). Comparison of Norm-Based Feature Selection Methods on Biological Omics Data. In Proceedings of the 5th International Conference on Advances in Image Processing (pp. 109-112).

[7] Özbay E, Özbay F A, Gharehchopogh F S. Kidney tumor classification on ct images using self-supervised learning[J]. Computers in Biology and Medicine, 2024, 176: 108554.

[8] Yeh K, Jabal M S, Gupta V, et al. Transformer-Based Self-Supervised Learning for Histopathological Classification of Ischemic Stroke Clot Origin[J]. arXiv preprint arXiv:2405.00908, 2024.