Hybrid Deep Learning Framework for Skin Disease Classification Using an Enhanced Swin Transformer Network

Naveeth Babu[1], Ganesh Khekare [2],

[1] Postdoctoral Researcher, Computer Science and Engineering, Lincoln University College, Malaysia.

[2] Associate Professor, School of Computer Science & Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

pdf.naveeth@lincoln.edu.my, khekare.123@gmail.com

Abstract: Automated skin disease classification from dermoscopic and clinical images is challenging because lesions can exhibit substantial variation in color, texture, shape, illumination, scale, and background artifacts. This paper proposes a hybrid image-analysis framework centered on an Enhanced Swin Transformer Network (E-SwinTransNet) for multi-class skin disease classification The framework obtains labeled images from a benchmark dataset, applies median filtering for noise suppression and Contrast Limited Adaptive Histogram Equalization for local contrast enhancement, and then uses rotations, flips and controlled brightness variations to diversify the training set and reduce over fitting. The augmented images are transformed into visual tokens and processed by a hierarchical Swin Transformer with attention refinement and multi-scale feature fusion. A leakage-aware training protocol and comprehensive evaluation strategy are also specified to reach the desired output.

Keywords: skin disease classification, dermatology, deep learning, Swin Transformer, data augmentation, medical image analysis

1. Introduction

Skin disorders range from gentle lesions to clinically important cruel conditions and image-based assessment is increasingly supported by computer vision and deep learning. Visual appearance varies substantially between patients and imaging devices, while clinically meaningful differences may be concentrated in small regions such as pigment networks, borders, vascular structures and colour asymmetries. Deep Learning techniques demonstrated that automated skin-cancer image classification can reach dermatologist-level performance for selected tasks, establishing an important foundation for computer-aided dermatology [1]. Public datasets have accelerated reproducible research. HAM10000 contains 10,015 dermoscopic images covering seven common pigmented-lesion categories [2]. ISIC challenge datasets further established standardized tasks for lesion segmentation, attribute detection and disease classification [3]. However, class imbalance, acquisition variability, limited representation of rare categories and possible data leakage remain important methodological concerns [4], [5]. Transformers introduced attention as a mechanism for modeling relationships among tokens [6]. The Swin Transformer adapted this idea to vision by computing self-attention within local windows and shifting the windows between successive blocks, allowing cross-window information flow with more favorable computational scaling than full global attention [7]. This hierarchical representation is attractive for skin images because diagnostic cues occur at multiple scales.

2. Related work

2.1 Deep CNN-Based Classification

CNNs are effective at learning local patterns such as edges, colour transitions and texture. Landmark work by Esteva et al. showed that a deep neural network could classify selected skin-cancer categories at a level comparable with trained dermatologists [1]. Subsequent CNN studies have used transfer learning, regularization and augmentation. Their main limitations for multi-class skin-disease analysis include sensitivity to class imbalance and difficulty representing long-range relationships explicitly.

2.2 Datasets and Benchmarking

HAM10000 is a major benchmark with seven pigmented-lesion classes [2]. ISIC provides larger and evolving collections; the ISIC 2019 release contains 25,331 training images with diagnostic labels [10]. Multi-source datasets require careful splitting because images belonging to the same patient or lesion can otherwise appear in multiple partitions, producing overly optimistic estimates.

2.3 Vision Transformers and Swin Transformers

The Transformer architecture introduced self-attention as the central mechanism for token interaction [6]. Swin Transformer applies attention inside local windows and shifts the window partition between blocks, creating efficient cross-window communication while producing hierarchical feature maps [7]. Such multi-scale modelling is appropriate for lesions where both fine texture and global structure contribute to diagnosis.

2.4 Recent Hybrid Research

Sharma, Bhanodia and Bhushan (2026) proposed a hybrid CNN–Transformer model that combines local feature extraction with contextual Transformer representations [9]. A 2026 systematic review found a broader transition toward hybrid frameworks and multi-source feature fusion, but emphasized persistent issues in class imbalance, data leakage and external clinical validation [8]. These observations support the proposed E-SwinTransNet design.

3. Research Gap and Motivation

Existing skin disease classification systems often focus on a single modeling family, such as CNNs or standard transformers. CNN-based systems can learn detailed local texture but may not represent long-range relationships as effectively as attention-based architectures. Conversely, transformer models can capture broader dependencies but may require substantial data and can benefit from complementary local feature extraction.

Another challenge is the variability of skin images. Noise, uneven illumination, low contrast and limited examples can reduce model generalization. Although individual preprocessing and augmentation methods are common, there remains scope for an integrated framework that combines noise suppression, local contrast enhancement, controlled augmentation, hierarchical transformer learning and multi-scale feature extraction.

The motivation for E-SwinTransNet is therefore to combine the strengths of convolution and transformer representations. The model is designed to preserve fine lesion boundaries and texture while also learning relationships among spatially separated regions.

4. Proposed Methodology

The proposed research framework is illustrated as shown in the Fig 1. The first step is to obtaining skin disease images from the input dataset. The next step is preprocessing process in which the following procedures are been done.

4.1.1 Median filtering

It is a non-linear filtering technique used primarily in image and signal processing to remove noise while preserving edges better than a simple averaging (mean) filter. It is particularly effective at removing impulsive noise while preserving edges and fine details in an image.

Let I(x,y) be the intensity at pixel(x,y) and w be the window size

$$I_{\text{filtered}}(x,y)=\text{median}\{I(i,j)|(i,j)\in\text{window around }(x,y)\}$$

4.1.2 CLAHE

Contrast Limited Adaptive Histogram Equalization) is an advanced image augmentation technique designed to progress the contrast of images, especially in regions with varying lighting conditions. It is an addition of the traditional Histogram Equalization (HE) and Adaptive Histogram Equalization (AHE), addressing their limitations by limiting contrast amplification to prevent noise amplification. CLAHE can improve visibility of subtle pigment and texture differences under uneven illumination, but clip limit and tile-grid size should be tuned because excessive enhancement may amplify hairs, bubbles or background artefacts.

4.1.3 Data augmentation

This technique is used in machine learning, particularly in image processing to artificially increase the size and diversity of a training dataset by creating modified versions of existing data samples. This process helps improve the robustness, generalization and performance of models, especially when available data is limited.

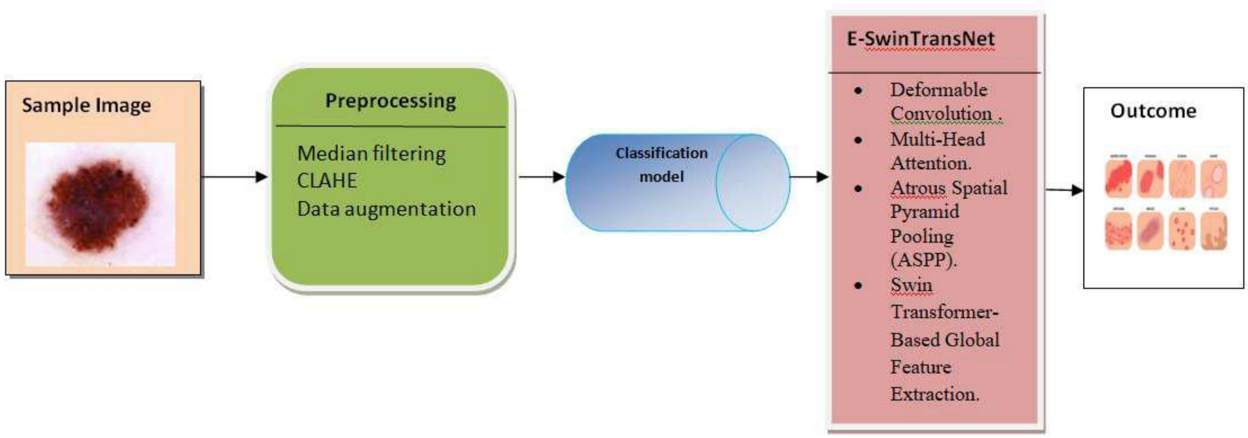

Figure 1. Skin disease classification using Enhanced Swin Transformer Network

These augmented images are used to feed into the proposed classification model known as Enhanced Swin Transformer Network (E-SwinTransNet).

4.2 Enhanced Swin Transformer Network (E-SwinTransNet)

4.2.1 Deformable Convolution

The traditional convolution layer is replaced by Deformable Convolution Network for better feature extraction and modeling of the complex shapes and boundaries of the skin lesions.

Let x be the input feature map.

y be the output feature map.

K be the convolution kernel size (e.g., 3x3).

p_k be the j-th position in the regular grid.

Δp_k be the learnable offset for position p_k.

The output at location p_0 is:

$$y(p_0) = \sum_{k=1}^{k} wk \cdot x(p_0 + p_k + \Delta p_k)$$

4.2.2 Multi-Head Attention

The model splits the image data into multiple parts. It analyzes color, shape, and texture separately. Then, it combines these results to spot skin disease or rashes with high accuracy. Different heads can focus on different aspects of the input, such as edges, textures, or semantic relationships. Multiple attention heads provide a richer, more nuanced understanding of the data.

4.2.3 Atrous Spatial Pyramid Pooling (ASPP)

This block is added to the proposed model to extract contextual information from multiple scales. This lets a network capture both fine local details and broad global context without losing resolution. It enriches feature representations, leading to more accurate segmentation and scene parsing. The outputs of all branches are concatenated along the channel dimension. A final 1x1 convolution fuses these multi-scale features into a rich representation.

4.2.4 Swin Transformer-Based Global Feature Extraction

The Swin Transformer combines hierarchical feature learning with window-based self-attention to efficiently process high-resolution images. In this step the following techniques are been implemented to classify the accurate skin disease.

4.2.4.1 Patch Partition:

The input image is divided into fixed-size patches, and each patch is converted into a feature embedding. This approach enables the model to capture local features efficiently and is a core component of the Swin Transformer architecture, which employs hierarchical and window-based self-attention mechanisms.

4.2.4.2 Window-Based Self-Attention:

Self-attention is computed within local windows, allowing the model to capture local features efficiently. Instead of computing self-attention globally across the entire image, it restricts attention computation to small, non-overlapping or overlapping windows, enabling scalable and effective modeling of local relationships.

4.2.4.3 Shifted Window Attention:

Window positions are shifted between layers, enabling information exchange across neighboring regions and improving global feature learning. It enables cross-window communication and improves modeling of long-range dependencies while maintaining computational efficiency.

4.2.4.4 Hierarchical Feature Learning:

The model progressively merges patches and learns multi-scale feature representations across multiple stages. It involves progressively capturing features at multiple levels of abstraction, from low-level details to high-level semantic information, by building a multi-layered, multi-scale representation of the input data.

5. Conclusion

This research paper presented a hybrid deep learning framework for skin disease classification using an Enhanced Swin Transformer Network. The pipeline starts with skin disease image acquisition and applies median filtering for noise reduction, CLAHE for local contrast enhancement, and controlled data augmentation through rotations, flips and brightness variation. The resulting images are processed by E-SwinTransNet, which combines convolution local feature learning, shifted-window self-attention and multi-scale contextual representation. The framework is designed to address the complementary requirements of fine-grained lesion analysis and broader spatial reasoning. A rigorous patient-level evaluation, class-wise metrics and ablation studies are essential for validating the proposed approach. The framework provides a research foundation for accurate and potentially explainable automated skin disease classification.

6. References

1. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." Nature, Volume 542, pages 115–118, 2017. https://doi.org/10.1038/nature21056.
2. Philipp Tschandl, Cliff Rosendahl, Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." Scientific Data, Volume 5, Article 180161, 2018. https://doi.org/10.1038/sdata.2018.161.
3. Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern. "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)." MICCAI Workshop Proceedings, 2018. https://doi.org/10.48550/arXiv.1902.03368.
4. Zhenyu Yan, Xue Wang, Ying Li, et al. "Skin Cancer Classification With Deep Learning: A Systematic Review." Frontiers in Oncology, Volume 12, Article 893972, 2022. https://doi.org/10.3389/fonc.2022.893972.
5. Mohammed A. Kassem, Khalid M.Hosny, Robertas Damasevicius , "Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review." Diagnostics, Volume 11, 2021. https://doi.org/10.3390/diagnostics11081390.
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. "Attention Is All You Need." Advances in Neural Information Processing Systems, Volume 30, pages 5998 – 6008, 2017. https://doi.org/10.48550/arXiv.1706.03762.
7. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021. https://doi.org/ 10.1109/ICCV48922.2021.00986.
8. Neven Saleh, Abdel Rahman Shaaban, Ahmed M. Salaheldin. "A comprehensive review of deep learning applications in the segmentation and classification of skin cancer." Biomedical Physics & Engineering Express, Volume 12, Issue 3, 2026. https://doi.org/10.1088/2057-1976/ae74d6.
9. Arpit Kumar Sharma, Praveen Kumar Bhanodia, Megha Bhushan. "A hybrid deep learning approach for skin lesion classification using convolutional neural networks and transformers." Network Modeling Analysis in Health Informatics and Bioinformatics, Volume 15, Article 119, 2026. https://doi.org/10.1007/s13721-026-00794-4.
10. Abdulmateen Adebiyi, Nader Abdalnabi, Emily Hoffman Smith, Jesse Hirner, Eduardo J. Simoes, Mirna Becevic, Praveen Rao. "Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 and ISIC 2017 Datasets." medRxiv, 2025. https://doi.org/10.1101/2024.05.30.24308213.