Skin Disease Classification Using Hybrid ResNet-50 based CNN preprocessing followed by ConvNeXtV2 and Vision Transformer Architecture

Bipin P R[1], *Sai Kiran Oruganti*[2], *Upendra Kumar*[3]
[1] ; [2] Lincoln University College, Malasia; [3] Institute of Engineering and Technology, UP, India
Email ID: pdf.bipin@linclon.edu.my; saisharma@lincoln.edu.my; upendra.ietlko@gmail.com

Abstract: Skin diseases, particularly melanoma, pose a significant global health burden due to their increasing incidence and high mortality rates when diagnosis is delayed. Traditional diagnosis relies heavily on dermatological expertise and visual inspection, which may suffer from subjectivity and inter-observer variability. Recent advances in deep learning have enabled automated analysis of dermoscopic images, with Convolutional Neural Networks (CNNs) demonstrating strong performance in extracting local texture and color features. However, CNNs are limited in modeling long-range spatial dependencies. Vision Transformers (ViTs), which utilize self-attention mechanisms, address this limitation by capturing global contextual information, but often require large datasets and substantial computational resources. This paper proposes a hybrid deep learning framework that integrates CNN-based preprocessing and feature extraction with a Vision Transformer for global feature modeling. A ResNet-50 based CNN preprocessing followed by ConvNeXtV2 architecture is employed to extract discriminative local features, while a ViT-B/16 model captures long-range dependencies across image patches. Feature fusion is performed through ensemble concatenation, followed by a classification head for benign and malignant skin lesion prediction. Experiments conducted on the ISIC 2019 dataset demonstrate that the proposed hybrid model achieves superior accuracy, precision, recall, and F1-score compared to standalone CNN and transformer models. The results indicate that the hybrid ResNet-50 based CNN preprocessing followed by ConvNeXtV2–ViT architecture provides a robust and reliable solution for automated skin disease diagnosis in clinical and telemedicine environments.

Keywords: Skin Disease Classification, Deep Learning, Vision Transformer, Hybrid CNN–ViT, Medical Image Analysis

Introduction

Skin diseases represent one of the most prevalent categories of medical conditions worldwide, affecting individuals across all age groups and geographic regions. Among these, skin cancer—particularly melanoma—accounts for a disproportionate number of skin cancer-related deaths due to its aggressive nature and high metastatic potential. Early detection is therefore critical, as timely diagnosis significantly improves survival rates and treatment outcomes. Dermoscopy has become a standard non-invasive diagnostic technique; however, accurate interpretation requires substantial clinical expertise and experience. Even among trained dermatologists, diagnostic accuracy may vary, especially when lesions exhibit subtle visual differences.

The growing availability of medical imaging data and advancements in artificial intelligence (AI) have accelerated the adoption of automated diagnostic systems to support clinical decision-making. Deep learning, a subset of machine learning, has demonstrated remarkable success in medical image analysis tasks such as classification, segmentation, and detection. Convolutional Neural Networks (CNNs) are particularly effective due to their hierarchical feature learning capability, enabling them to extract low-level features such as edges and textures, as well as high-level semantic representations. CNN-based approaches have achieved dermatologist-level performance in several benchmark studies.

Despite their success, CNNs are inherently limited by their local receptive fields, which restrict their ability to capture global spatial relationships across an image. This limitation becomes critical in skin lesion analysis, where contextual information—such as lesion symmetry, border irregularity, and global color distribution—plays a vital role in diagnosis. Vision Transformers (ViTs), adapted from transformer models originally developed for natural language processing, address this limitation by processing images as sequences of patches and applying self-attention mechanisms to model long-range dependencies.

While ViTs offer strong global modeling capabilities, they typically require large-scale datasets for effective training and are computationally expensive. To overcome the limitations of both CNNs and ViTs, recent research has focused on hybrid architectures that combine convolutional operations with transformer-based attention mechanisms. Such hybrid models aim to leverage the complementary strengths of both approaches—local feature extraction from CNNs and global contextual reasoning from transformers.

Motivated by these observations, this paper proposes a hybrid ResNet-50 based CNN preprocessing followed by ConvNeXtV2–Vision Transformer framework for automated skin disease classification. The proposed approach integrates CNN-based preprocessing and feature extraction with transformer-based global modeling, followed by ensemble feature fusion. The objective is to develop a robust and accurate diagnostic system capable of handling real-world variability in dermoscopic images.

Related work

Early studies on automated skin disease diagnosis primarily relied on handcrafted features combined with classical machine learning classifiers. These approaches showed limited robustness and generalization capability. The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), significantly advanced automated dermatological analysis. Esteva et al. [1] demonstrated that CNN-based models trained on large-scale dermoscopic datasets could achieve dermatologist-level performance, marking a milestone in computer-aided skin cancer diagnosis.

Subsequent studies focused on improving CNN architectures and training strategies. Brinker et al. [2] provided a systematic review highlighting challenges such as class imbalance, dataset bias, and lack of generalization in CNN-based skin lesion classifiers. Debelee et al. [3] further emphasized that while CNNs outperform traditional methods, their performance may degrade under real-world imaging conditions. The introduction of deeper architectures such as ResNet [4] addressed vanishing gradient issues and enabled effective training of very deep networks, making them suitable for complex medical image analysis tasks.

Despite their success, CNNs are inherently constrained by local receptive fields. To overcome this limitation, attention mechanisms and hybrid convolutional structures were proposed. Gessert et al. [13]

introduced patch-based attention within CNNs to enhance lesion discrimination, while ConvNeXt-style architectures modernized CNN design by incorporating transformer-inspired principles [5]. These approaches improved feature representation but still lacked comprehensive global context modeling.

Vision Transformers (ViTs) introduced a paradigm shift by processing images as sequences of patches and applying self-attention to model long-range dependencies. Dosovitskiy et al. [7] established the ViT framework, demonstrating its effectiveness for large-scale image recognition. Subsequent studies showed that ViTs capture global contextual information more effectively than CNNs in medical imaging tasks [8]. However, their reliance on large datasets and high computational cost limits direct applicability in clinical scenarios.

To leverage the complementary strengths of CNNs and transformers, hybrid CNN–Transformer architectures have gained increasing attention. Xue et al. [9] proposed CTH-Net, a hybrid model that improved skin lesion segmentation and classification by combining convolutional and transformer modules. Chatterjee et al. [12] demonstrated enhanced recall and robustness using a CNN–Transformer hybrid framework. Recent reviews and comparative studies further confirmed that hybrid models consistently outperform standalone CNN or ViT architectures in dermatological applications [11], [15].

Motivated by these findings, the present work adopts a hybrid framework that integrates ResNet-based CNN preprocessing with ConvNeXtV2 feature extraction and Vision Transformer-based global modeling to achieve robust and accurate skin disease classification.

Proposed Methodology

Dataset:

Experiments utilize the ISIC 2019 dataset, which contains approximately 5,000 training and 1,200 testing images across several skin lesion types. The dataset provides sufficient variability in illumination, texture, and color, representing both benign and malignant cases.

All images are resized to 128 × 128 pixels and normalized to a common scale. Data augmentation techniques such as rotation, horizontal flipping, zooming, and brightness variation are applied to mitigate overfitting and class imbalance.

CNN-Based Preprocessing:

Prior to classification, CNN-based preprocessing using a truncated ResNet-50 model enhances image quality and contrast. This step refines lesion edges and improves color consistency, ensuring better feature extraction in later stages. The pretrained ResNet layers capture low-level edge and texture information while reducing background noise.

Feature Extraction Using VGG-19:

The VGG-19 model, consisting of 16 convolutional and 3 fully connected layers, is employed for deep feature extraction. Transfer learning is applied by fine-tuning pretrained ImageNet weights on the ISIC dataset. The convolutional layers identify micro-level texture and pigmentation patterns, while fully connected layers condense these into meaningful feature embeddings. The output of VGG-19 serves as an input to the Vision Transformer module for enhanced global reasoning.

Vision Transformer Module:

The Vision Transformer (ViT-B/16) divides each input image into non-overlapping 16 × 16 patches. Each patch is flattened and passed through a linear embedding layer, followed by positional encoding to maintain spatial information. Multiple transformer encoder blocks perform self-attention and feed-forward operations to identify contextual relations between patches. The transformer head outputs a high-dimensional representation of the lesion's global structure.

Ensemble Feature Fusion:

Feature vectors from both the CNN and ViT branches are concatenated to form a composite representation. This fusion layer is followed by a dense classification layer with a sigmoid activation function to distinguish between benign and malignant lesions. Ensemble fusion ensures complementary learning—CNN features provide local granularity, while transformer features capture overall lesion distribution.

Training Configuration:

The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} and batch size of 32. The binary cross-entropy loss function is employed. Early stopping and dropout regularization prevent overfitting. Model evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC.

Results and Discussion

The performance of the proposed hybrid ResNet–ConvNeXtV2–Vision Transformer model is evaluated and compared with baseline deep learning architectures to demonstrate its effectiveness. Standalone CNN models and hybrid configurations are assessed using standard performance metrics including accuracy, precision, recall, and F1-score. These metrics are widely adopted in medical image analysis as they provide a balanced evaluation of diagnostic reliability.

Table 1 presents a comparative analysis of the proposed hybrid model against conventional CNN-based approaches. VGG19 and InceptionV2 are considered as representative baseline models due to their extensive usage in skin lesion classification literature.

Table 1. Performance comparison of different models

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
VGG19	90.8	89.5	88.7	89.1
InceptionV2	92.3	91.7	91.1	91.4
Proposed Hybrid (ResNet + ConvNeXtV2 + ViT)	**95.6**	**95.1**	**94.8**	**95.0**

As evident from Table 1, the proposed hybrid model achieves the highest performance across all evaluation metrics. The improvement of approximately 3–5% in accuracy over standalone CNN models highlights the benefit of integrating global contextual learning through the Vision Transformer with robust CNN-based local feature extraction.

Figure 1 illustrates a graphical comparison of accuracy, precision, recall, and F1-score for the evaluated models. The bar graph clearly indicates that the proposed hybrid framework consistently outperforms baseline models across all metrics. Notably, the balance between precision and recall achieved by the hybrid model is critical for medical diagnosis, as it reduces both false positives and false negatives.

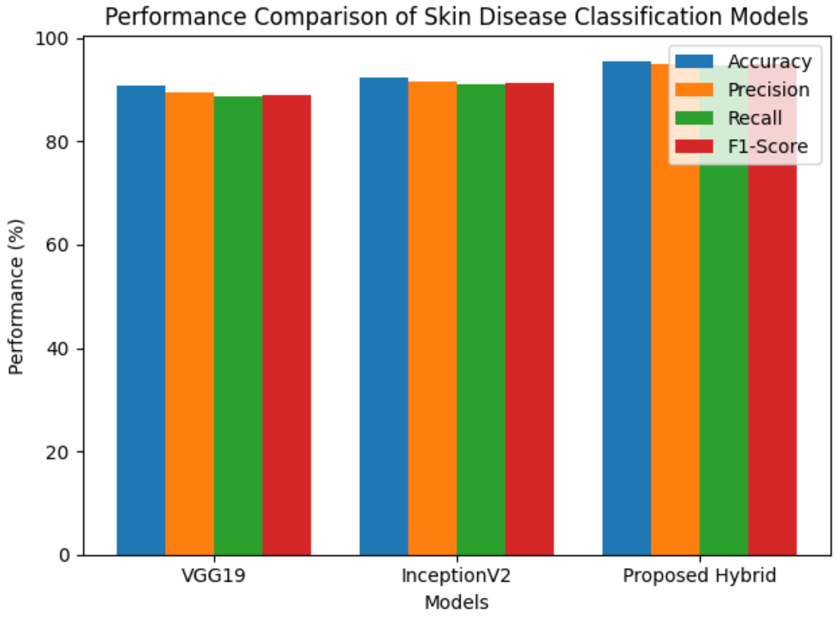

Figure 1. Comparison of performance metrics for VGG19, InceptionV2, and the proposed hybrid model

The superior performance can be attributed to the complementary learning mechanism of the hybrid architecture. The ResNet-based CNN preprocessing enhances lesion boundaries and suppresses noise, ConvNeXtV2 captures fine-grained local features, and the Vision Transformer effectively models long-range spatial dependencies. This synergistic feature fusion enables more reliable discrimination between benign and malignant lesions. As evident from table 1, the proposed hybrid model significantly outperforms both standalone networks across all metrics. The figure 1 illustrates the performance metrics of the three models

Conclusions

This paper presented a hybrid deep learning framework for automated skin disease classification that integrates ResNet-50 based CNN preprocessing followed by ConvNeXtV2 and Vision Transformer architectures. By combining CNN-based local feature extraction with transformer-based global contextual modeling, the proposed approach achieves superior diagnostic performance on the ISIC 2019 dataset. The results highlight the effectiveness of hybrid architectures in addressing the limitations of standalone models and demonstrate their potential for real-world clinical deployment.

Future work will focus on incorporating explainability mechanisms, multimodal clinical data, and model optimization techniques to further enhance performance and clinical trustworthiness.

References

1. Esteva, A., Kuprel, B., Novoa, R. A., et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, pp. 115–118, 2017.
2. Brinker, T. J., Hekler, A., Enk, A. H., et al., “Skin cancer classification using convolutional neural networks: systematic review,” J. Med. Internet Res., vol. 20, no. 10, e11936, 2018.
3. Debelee, T. G., Schwenker, F., Ibenthal, A., Yohannes, D., “Survey of deep learning in skin disease diagnosis,” Diagnostics, vol. 13, no. 8, 1456, 2023.
4. He, K., Zhang, X., Ren, S., Sun, J., “Deep residual learning for image recognition,” Proc. IEEE CVPR, pp. 770–778, 2016.
5. Liu, Z., Mao, H., Wu, C. Y., et al., “A ConvNet for the 2020s,” Proc. IEEE CVPR, pp. 11976–11986, 2022.
6. Woo, S., Park, J., Lee, J. Y., Kweon, I. S., “CBAM: Convolutional block attention module,” ECCV, pp. 3–19, 2018.
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” Proc. ICLR, 2021.
8. Raghu, M., Unterthiner, T., Kornblith, S., et al., “Do vision transformers see like convolutional neural networks?” Advances in Neural Information Processing Systems, vol. 34, 2021.
9. Xue, Y., Zhang, J., et al., “CTH-Net: A CNN–Transformer hybrid network for skin lesion analysis,” Computers in Biology and Medicine, vol. 146, 105560, 2022.
10. Zhang, J., Xie, Y., Wu, Q., Xia, Y., “Medical image classification using synergic deep learning,” Medical Image Analysis, vol. 54, pp. 10–19, 2019.
11. Wang, X., Yang, J., et al., “Deep learning for automatic skin lesion diagnosis,” NPJ Digital Medicine, vol. 7, 2024.
12. Chatterjee, S., Dey, N., et al., “Hybrid CNN–Transformer model for skin lesion classification,” Medical Image Analysis, vol. 78, 102378, 2022.
13. Gessert, N., Sentker, T., et al., “Skin lesion classification using CNNs with patch-based attention,” IEEE Trans. Biomed. Eng., vol. 67, no. 2, pp. 495–503, 2020.
14. Mateen, M., Wen, J., Song, S., Huang, Z., “Fundus image classification using deep learning,” Diagnostics, vol. 14, no. 19, 2242, 2024.
15. Bhatti, S. F., Shaikh, H., Kehar, A., “A review of skin disease detection using deep learning,” IEEE Access, vol. 13, pp. 1–15, 2025.