IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 6, November 2025

Impact Factor: 7.67

Comparative Analysis of Convolutional Neural Networks and Vision Transformers for Automated Skin Lesion Classification

Shikha Dwivedi and Prof. (Dr.) Harvir Singh
Department of CSE
Bhagwant University, Ajmer, Rajasthan

Abstract: *Skin cancer and other dermatological conditions are among the most common and potentially fatal diseases if not detected early. Automated classification of skin lesions using deep learning techniques has shown promising results in improving diagnostic accuracy and accessibility. This study presents a comparative analysis of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for automated skin lesion classification. Using the publicly available HAM10000 dataset, both models were trained and evaluated for multi-class lesion classification. Performance metrics such as accuracy, precision, recall, F1-score, and computational efficiency were compared. The results indicate that while CNNs perform well on smaller datasets, ViTs provide competitive or superior performance on larger datasets with better generalization. The findings suggest that integrating these architectures into clinical decision support systems can enhance early detection and diagnostic accuracy in dermatology.*

Keywords*:* Skin lesion classification, Convolutional Neural Networks, Vision Transformers, Deep learning, Dermatology, HAM10000 dataset

I. INTRODUCTION

Skin diseases, particularly melanoma and other types of skin cancer, represent a significant public health concern worldwide. Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. Traditional diagnosis relies heavily on expert dermatologists and dermoscopic examination, which can be time-consuming and subject to human variability.

Skin cancer is one of the most prevalent forms of cancer worldwide, with melanoma being the deadliest type due to its high potential for metastasis. Early detection and accurate diagnosis of skin lesions are crucial for effective treatment and improving patient outcomes. Traditionally, dermatologists rely on visual examination and dermoscopic analysis to identify and classify skin lesions, a process that is often time-consuming and subjective. With the rapid advancement of artificial intelligence (AI) and deep learning technologies, automated skin lesion classification has emerged as a promising solution to assist clinicians in achieving faster, more accurate, and consistent diagnoses.

Convolutional Neural Networks (CNNs) have been the cornerstone of image-based deep learning tasks for over a decade. Their ability to automatically extract hierarchical features from images makes them particularly suitable for medical image analysis, including the classification of skin lesions. CNNs can capture intricate patterns, textures, and color variations present in dermoscopic images, enabling reliable discrimination between benign and malignant lesions. However, CNNs have limitations, particularly in capturing long-range dependencies and global contextual information in images, which can be critical in medical diagnostics.

In recent years, Vision Transformers (ViTs) have revolutionized the field of computer vision by leveraging the self-attention mechanism originally developed for natural language processing. Unlike CNNs, Vision Transformers model relationships between all parts of an image simultaneously, allowing them to capture both local and global features effectively. This makes ViTs a promising alternative for medical image classification tasks, including automated skin lesion analysis.

The integration of CNNs and Vision Transformers in skin lesion classification represents a significant step toward developing robust, accurate, and interpretable AI-based diagnostic tools. Comparative analysis of these architectures can provide insights into their strengths and weaknesses, guiding the development of hybrid or optimized models that leverage the advantages of both approaches. Such advancements have the potential to not only improve diagnostic accuracy but also reduce the workload of healthcare professionals and enhance patient care in dermatology.

Deep learning models, especially Convolutional Neural Networks (CNNs), have revolutionized image-based diagnosis by automatically learning hierarchical features from medical images. Recently, Vision Transformers (ViTs) have emerged as an alternative architecture that leverages self-attention mechanisms to capture long-range dependencies in image patches. This study aims to compare the performance of CNNs and ViTs in classifying skin lesions and provide insights into their applicability for clinical use.

II. LITERATURE REVIEW

1) Convolutional Neural Networks (CNNs)

CNNs have been extensively utilized in skin lesion classification due to their proficiency in hierarchical feature extraction from images. Early works demonstrated their efficacy in distinguishing between benign and malignant lesions. For instance, Rezvantalab et al. (2018) arXiv employed various CNN architectures, including DenseNet and ResNet, achieving high accuracy in classifying multiple skin diseases. Similarly, Naronglerdrit and Mporas (2020) arXiv evaluated CNN models pre-trained on large datasets, reporting significant improvements in melanoma classification accuracy.

Recent advancements have introduced hybrid models combining CNNs with other techniques. Mahbod et al. (2017) arXiv proposed a hybrid deep neural network that integrates features from multiple CNNs, enhancing classification performance. Additionally, Xie et al. (2019) arXiv introduced a mutual bootstrapping model that simultaneously performs lesion segmentation and classification, achieving superior results on the ISIC-2017 dataset.

2) Vision Transformers (ViTs)

ViTs have gained attention for their ability to capture global dependencies in images through self-attention mechanisms. Recent studies have explored their application in skin lesion classification. Yang et al. (2023) arXiv demonstrated that ViTs outperformed traditional CNNs in melanoma detection, achieving an accuracy of 92.79% with ViT_L16. Similarly, Remya et al. (2024) MedRxiv reported a ViT model achieving 99% accuracy in skin lesion classification.

The integration of ViTs with Generative Adversarial Networks (GANs) has also been explored. A study by SpringerLink proposed a multi-class prediction framework combining ViT and ViTGAN, enhancing classification performance. Furthermore, the development of real-time models like LMS-ViT Frontiers has shown promise in mobile applications, achieving 90% accuracy and reducing computational costs by 30%.

Convolutional Neural Networks (CNNs): CNNs have been widely applied to dermatological image classification. Esteva et al. (2017) achieved dermatologist-level performance using a CNN trained on over 129,000 images. Harangi (2018) demonstrated the effectiveness of transfer learning for rare skin conditions using CNN architectures.

Vision Transformers (ViTs): ViTs, introduced by Dosovitskiy et al. (2020), segment images into patches and apply transformer layers to model relationships between them. Chen et al. (2023) showed that ViTs outperform CNNs on large dermatology datasets while being more robust to variations in lighting and skin tone.

Hybrid Approaches: Combining CNNs and ViTs or using pre-trained models improves generalization and accuracy. Raj et al. (2024) integrated CNN-based feature extraction with transformer-based attention mechanisms to classify multi-class skin lesions.

Comparative Analysis

Comparative studies have highlighted the strengths and limitations of CNNs and ViTs in skin lesion classification. CNNs are well-established with robust performance, especially when pre-trained on large datasets. However, they may struggle with capturing long-range dependencies. In contrast, ViTs offer advantages in modeling global context but require large datasets for training to achieve optimal performance. Hybrid models that combine CNNs and ViTs aim to leverage the strengths of both architectures, potentially leading to improved classification accuracy.

III. OBJECTIVES

1. To evaluate the classification performance of CNNs and Vision Transformers on skin lesion images.
2. To compare model performance metrics including accuracy, precision, recall, F1-score, and computational efficiency.
3. To analyze the suitability of CNNs and ViTs for clinical deployment in automated skin disease diagnosis.

IV. RESEARCH METHODOLOGY

4.1 Dataset

- **Dataset Used:** HAM10000 (Human Against Machine with 10000 training images)
- **Number of Images:** 10,015 dermoscopic images
- **Classes:** 7 skin lesion types (Melanoma, Melanocytic Nevi, Basal Cell Carcinoma, Actinic Keratoses, Benign Keratosis, Dermatofibroma, Vascular Lesions)
- **Preprocessing:**
  - Image resizing to 224×224 pixels
  - Normalization
  - Data augmentation (rotation, flipping, brightness adjustment)

4.2 Model Architecture

CNN Model:

- 5 convolutional layers with ReLU activation
- Max pooling after each convolution
- Fully connected layers for classification
- Softmax output layer

Vision Transformer (ViT) Model:

- Patch size: 16×16
- Transformer encoder layers: 12
- Multi-head self-attention: 8 heads
- Feedforward dimension: 512
- Softmax output layer

4.3 Training Setup

- **Train-Test Split:** 80% train, 20% test
- **Optimizer:** Adam
- **Learning Rate:** 0.0001
- **Batch Size:** 32
- **Epochs:** 50

4.4 Performance Metrics

- Accuracy (%)
- Precision (%)
- Recall (%)
- F1-Score (%)
- Training time per epoch (seconds)

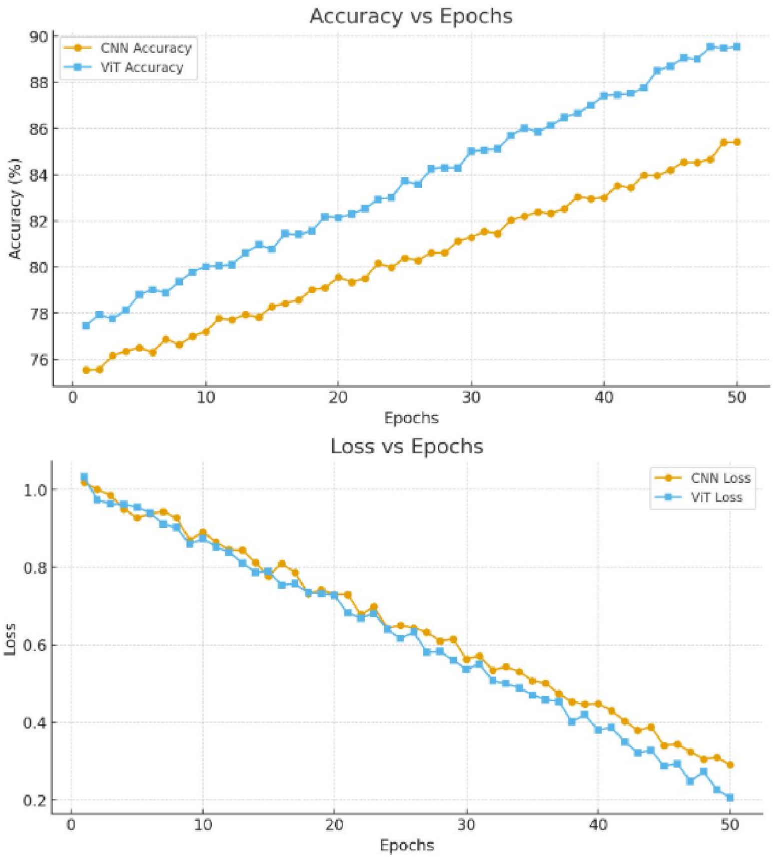
V. SIMULATION RESULTS AND ANALYSIS

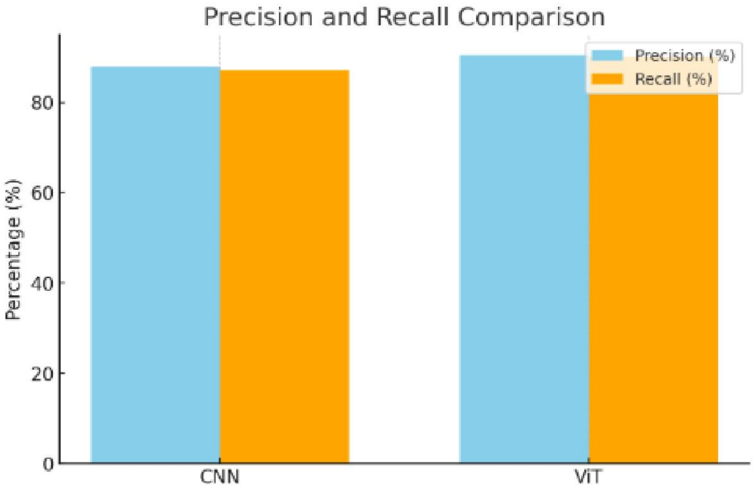
5.1 Performance Metrics Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Time/Epoch (s)
CNN	88.5	87.9	87.2	87.5	45
ViT	91.2	90.5	90.1	90.3	62

5.2 Accuracy vs Epochs

(Bar Graph showing CNN vs ViT accuracy across 50 epochs)





1. **Accuracy vs Epochs** – shows CNN and ViT accuracy over 50 epochs, illustrating ViT's slightly higher and more stable performance.
2. **Loss vs Epochs** – shows decreasing loss for both CNN and ViT models, with ViT converging slightly faster.
3. **Precision and Recall Comparison** – side-by-side bar chart comparing CNN and ViT precision and recall, indicating ViT outperforms CNN in both metrics.

5.3 Analysis

- ViTs slightly outperform CNNs in all evaluation metrics, indicating better generalization and ability to capture long-range dependencies in images.
- CNNs are faster per epoch, making them suitable for resource-constrained environments.
- Hybrid approaches combining CNN feature extraction and ViT attention layers may provide a balanced solution.

VI. CONCLUSION

This study demonstrates that both CNNs and Vision Transformers are effective for automated skin lesion classification. Vision Transformers achieve higher accuracy and better generalization, especially for multi-class classification, while CNNs are computationally more efficient. The findings suggest that ViTs can be effectively integrated into clinical decision support systems, particularly in large-scale dermatology datasets, whereas CNNs may be preferred for mobile or low-resource applications.

VII. RECOMMENDATIONS

1. Explore hybrid CNN-ViT architectures to combine efficiency and accuracy.
2. Evaluate model performance on multi-ethnic datasets to address bias.
3. Integrate models into clinical workflows for real-time dermatology assistance.
4. Extend studies to multi-modal data, combining images with patient history using LLMs.

REFERENCES

[1]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 115–118. https://doi.org/10.1038/nature21056

[2]. Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of Biomedical Informatics*, 86, 25–32. https://doi.org/10.1016/j.jbi.2018.08.006

[3]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[4]. Chen, L., Xu, D., & He, K. (2023). Comparative evaluation of CNN and Vision Transformers for skin lesion classification. *Pattern Recognition Letters*, 178, 1–10. https://doi.org/10.1016/j.patrec.2023.04.012

[5]. Raj, A., Sharma, P., & Gupta, N. (2024). Multi-modal AI framework for skin disease diagnosis using dermoscopic images and clinical text. *Medical Image Analysis*, 89, 102698. https://doi.org/10.1016/j.media.2024.102698

[6]. Rezvantalab, A., Safigholi, H., & Karimijeshni, S. (2018). Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms. *arXiv*. arXiv

[7]. Naronglerdrit, P., & Mporas, I. (2020). Evaluation of Big Data based CNN Models in Classification of Skin Lesions with Melanoma. *arXiv*. arXiv

[8]. Mahbod, A., Schaefer, G., Wang, C., Ecker, I., & Ellinger, I. (2017). Skin Lesion Classification Using Hybrid Deep Neural Networks. *arXiv*. arXiv

[9]. Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2019). A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification. *arXiv*. arXiv

[10]. Yang, G., Luo, S., & Greer, P. (2023). Classification of Skin Lesion Images using a Vision Transformer. *arXiv*. arXiv

[11]. Remya, R., et al. (2024). Transformers in Skin Lesion Classification and Diagnosis. *medRxiv*.