ADBA
Computers and Electronics in Medicine
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8609
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.42-47
https://doi.org/10.69882/adba.cem.2026015

Benchmarking State-of-the-Art Vision Transformer Architectures for the Automated Classification of Pigmented Skin Lesions

Md Saiful Islam [*,1], **André Chéagé Chamgoué** [β,2] **and Gurvinder Pal Dubb** [α,3]

*School of Engineering, Deakin University, Waurn Ponds Campus, Victoria, Australia, [β]Ngaoundere University, School of Geology and Mining Engineering, Meiganga, Cameroon, [α]Military Technological College, Department of Systems Engineering, Muscat, Sultanate of Oman.

ABSTRACT Skin cancer represents an escalating global public health challenge where early detection is paramount, potentially increasing five-year survival rates to 99%. While dermoscopy improves diagnostic sensitivity, its effectiveness often depends on clinician experience and is subject to inter-observer variability. To address these limitations, this study presents a rigorous comparative analysis of four state-of-the-art Vision Transformer (ViT) architectures, DeiT III-Base, Swin-Base, ViT-Base, and PiT-B, for the automated classification of pigmented skin lesions. We utilized the HAM10000 dataset (n=10,011) and implemented a stratified 70-15-15 split to ensure balanced training, validation, and testing phases. Images were resized to 224×224 pixels and normalized using ImageNet parameters, while transfer learning was employed to stabilize training and enhance generalization. Experimental results indicate that DeiT III-Base achieved superior diagnostic efficacy, reaching an accuracy of 92.04% and an F1-score of 85.44%. Furthermore, computational evaluation revealed that DeiT III-Base and ViT-Base offered highly efficient clinical throughput with sub-millisecond inference times (0.5674 ms and 0.5459 ms, respectively), whereas PiT-B exhibited the lowest computational workload (21.1067 GFLOPs). These findings underscore the viability of attention-based paradigms as robust real-time Computer-Aided Diagnosis (CAD) tools. Future research will explore the integration of multi-modal patient data and Explainable AI (XAI) to foster transparency and clinical trust.

KEYWORDS
Vision transformers (ViTs)
Skin cancer classification
HAM10000 Dataset
Computer-aided diagnosis (CAD)
DeiT III-Base

INTRODUCTION

Skin cancer constitutes a significant and escalating public health challenge globally, accounting for a substantial proportion of all diagnosed malignancies. With incidence rates rising steadily due to factors such as increased exposure to ultraviolet (UV) radiation, environmental changes, and aging populations, the burden on healthcare systems continues to grow. Among the various types, melanoma represents the most aggressive and lethal form, characterized by a high potential for metastasis (Gloster Jr and Neal 2006; Armstrong and Kricker 1995; Madan *et al.* 2010). However, prognosis is strongly correlated with the stage at diagnosis; early detection can increase the five-year survival rate to nearly 99%, whereas delayed diagnosis significantly diminishes treatment success and patient survival. Consequently, the development of rapid, accessible, and precise diagnostic mechanisms is not merely a clinical preference but a vital necessity to reduce mortality rates and improve patient outcomes (Gloster Jr and Brodland 1996).

Traditionally, the diagnosis of pigmented skin lesions relies on visual examination followed by dermoscopy, a non-invasive imaging technique that visualizes subsurface skin structures. While dermoscopy significantly enhances diagnostic sensitivity compared to naked-eye examination, its efficacy remains heavily dependent on the clinician's experience and training (Siegel *et al.* 2024; Jerant *et al.* 2000). The visual similarity between benign lesions (e.g., nevi, benign keratosis) and malignant tumors (e.g., melanoma, basal cell carcinoma) often leads to diagnostic ambiguity, resulting in either unnecessary biopsies or missed malignancies. Factors such as clinician fatigue, inter-observer variability, and the sheer volume of patients further complicate the manual diagnostic process. These limitations have necessitated the integration of Computer-Aided Diagnosis (CAD) systems to provide objective, consistent, and "second opinion" support for dermatologists.

In the last decade, the landscape of medical image analysis has been revolutionized by Deep Learning (DL), particularly Con-

Manuscript received: 5 September 2025,
Revised: 18 November 2025,
Accepted: 20 November 2025.

[1]Saiful.Islam@deakin.edu.au
[2]acchamgoue@gmail.com
[3]dubbgurvinderpal@gmail.com (**Corresponding author**)

volutional Neural Networks (CNNs) (ThangaPurni and Braveen 2025; Ozdemir and Pacal 2025; Çakmak and Pacal 2025; Cakmak and Maman 2025). Architectures such as ResNet, DenseNet, and EfficientNet have established themselves as the gold standard in skin lesion classification, demonstrating the ability to learn hierarchical feature representations directly from raw images (Karthik *et al.* 2024; Cakmak and Pacal 2025; Pacal and Cakmak 2025a,b). Despite their success, traditional CNNs primarily focus on local features due to their receptive field limitations, potentially missing long-range dependencies and global contextual information crucial for differentiating complex lesions. This limitation has paved the way for the adoption of Vision Transformers (ViTs) and hybrid architecture. Unlike CNNs, ViTs utilize self-attention mechanisms to capture global relationships across the entire image, offering a robust alternative for capturing fine-grained morphological details.

However, the rapid evolution of DL has precipitated a shift from pure convolutional networks to sophisticated attention-based paradigms. Consequently, there remains a critical need to rigorously evaluate how these distinct architectural strategies perform within the specific domain of skin cancer classification. In this study, we propose a comprehensive comparative analysis of advanced ViT architectures to identify the most effective mechanisms for skin lesion diagnosis. Rather than employing a broad spectrum of legacy models, our experimental framework rigorously benchmarks four distinct transformer-based methodologies that represent the state-of-the-art in attention mechanisms: the Data-efficient Image Transformer (DeiT III-Base), the standard ViT-Base, the hierarchical Swin Transformer (Swin-Base), and the Pooling-based Vision Transformer (PiT-B). Through this targeted evaluation, we aim to assess these architectures under unified conditions, providing critical insights into their generalization capabilities and clinical applicability for early skin cancer detection.

RELATED WORK

To overcome the locality bias of CNNs, researchers have increasingly adopted ViTs to model global context within dermoscopic images. Aruk *et al.* (2026) conducted a comprehensive comparative study evaluating 15 different CNNs against 15 ViT variants, including Swin and BeiT architectures, under identical training conditions. Their extensive analysis revealed that ViT models, particularly the Swin Transformer, consistently outperformed CNNs in classification accuracy, albeit at the cost of higher parameter counts and computational demands. Addressing the data-hungry nature of transformers, novel training paradigms have been introduced to improve robustness. Chaurasia *et al.* (2025) developed a multi-resolution model utilizing the DINOv2 self-supervised learning method to classify skin cancer subtypes from whole slide images (WSIs). By training on histological patches at various magnifications (10x to 400x), their model effectively captured multi-scale features, achieving an F1-score of 0.898 on external validation datasets.

Furthermore, specific architectural modifications have been proposed to tailor ViTs for the nuances of skin lesion analysis. Manju *et al.* (2025) proposed a preprocessing-optimized ViT model that integrates contrast enhancement and lesion segmentation directly into the workflow to remove artifacts before tokenization. This attention-enhanced model achieved an AUC-ROC score of 0.97, proving that feeding cleaner, segmented data into self-attention mechanisms significantly boosts diagnostic performance. Finally, the challenge of class imbalance in transformer training has been rigorously addressed. Sakib *et al.* (2025) introduced "LEVit," a framework that combines a hybrid ViT with extensive data augmentation and oversampling techniques to ensure uniform class distribution. Their approach not only achieved an F1 score of 98.11% on the ISIC 2019 dataset but also integrated Grad-CAM to generate class-specific heatmaps, ensuring the model's decisions were interpretable.

Building on the theme of architectural refinement for dermatological assessment, the literature has further evolved to address the synergy between localized features and global spatial reasoning. Pacal *et al.* (2024) advanced this structural capacity by proposing a Swin Transformer model that incorporates hybrid shifted window-based multi-head self-attention. Their methodology utilizes SwiGLU-based MLP layers to more effectively synchronize localized texture orientations with global spatial cues, thereby enhancing the network's sensitivity to minute malignant patterns that often elude standard convolutional filters. Beyond individual architectural optimizations, the focus has shifted toward improving decision robustness through multi-model integration. Bruno *et al.* (2025) expanded upon these gains by introducing a multi-scale attention and ensemble framework designed to aggregate features from diverse transformer-based learners. Their research demonstrates that combining multiple attention mechanisms effectively overcomes the inherent biases of single-architecture systems, leading to superior classification stability even in the presence of noisy or heterogeneous dermatoscopic data.

Finally, ensuring that these sophisticated models are both trustworthy and clinically viable has become a primary objective. Dagnaw *et al.* (2024) addressed this by integrating ViTs with Explainable Artificial Intelligence (XAI) to bridge the gap between high-performance computing and clinical transparency. By providing dermatologists with interpretable saliency maps, their framework ensures that the diagnostic logic of the transformer is both visible and verifiable, fostering the necessary confidence for the adoption of automated screening tools in high-stakes medical environments.

MATERIALS AND METHODS

Dataset and Data Preprocessing

In this research, we worked with the HAM10000 dataset, which is essentially the gold standard collection used in the ISIC 2018 challenge (HAM 2025). It contains 10,011 dermoscopic photos covering seven different types of skin conditions, ranging from harmless moles to dangerous cancers like melanoma. If you look at Figure 1, you can see why this is such a difficult task: many of these lesions look incredibly similar to the untrained eye, making manual diagnosis a real challenge.

When setting up our experiments, we didn't just split the data randomly. We used a stratified split to make sure the balance of diseases remained consistent across our training and testing sets. We settled on a 70-15-15 distribution, which gave us 7,005 images to train our models, 1,498 to fine-tune them, and a final 1,508 to test how well they actually perform on "unseen" cases. You can see the exact breakdown of these numbers in Table 1.

To get the images ready for the AI, we resized everything to 224×224 pixels. We also normalized the colors based on ImageNet standards. This step is vital because it helps the model ignore background "noise" like lighting differences and instead focus strictly on the textures and patterns that actually matter for a correct diagnosis.

Vision Transformers (ViTs)

To address the inherent limitations of local receptive fields in convolutional networks, this study leverages ViTs to model long-range

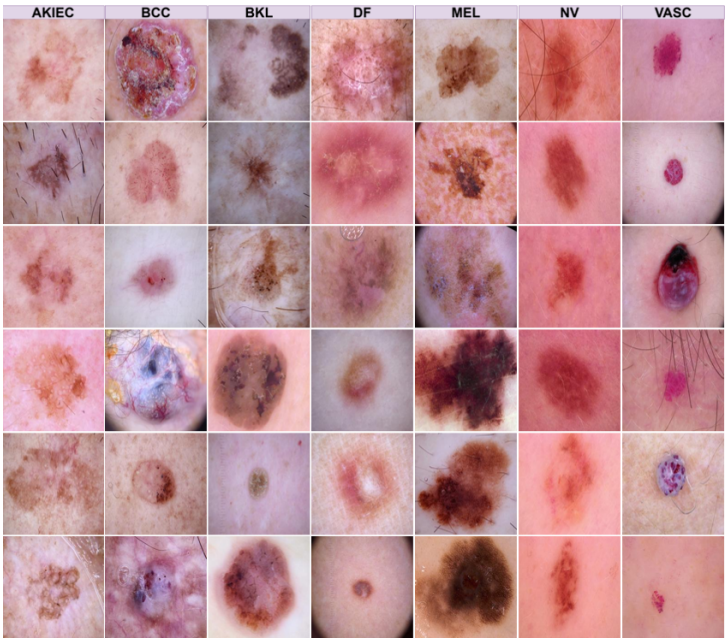

Figure 1 Representative dermoscopic samples from the HAM10000 dataset illustrating the seven skin lesion classes used in this study (AKIEC, BCC, BKL, DF, MEL, NV, and VASC).

Table 1 Distribution of the HAM10000 dataset across the seven diagnostic categories for training, validation, and testing subsets.

Class Name	Total	Train	Val	Test
BKL	1099	769	164	166
DF	115	80	17	18
VASC	142	99	21	22
AKIEC	323	226	48	49
MEL	1113	779	166	168
BCC	514	359	77	78
NV	6705	4693	1005	1007
Grand Total	10011	7005	1498	1508

dependencies and global semantic context, which are essential for distinguishing the subtle morphological variations in skin lesions. Our methodological framework rigorously benchmarks four distinct attention-based paradigms to evaluate their efficacy in dermoscopic analysis: the standard ViT-Base (Dosovitskiy *et al.* 2020), serving as a pure self-attention anchor; the DeiT III-Base (Touvron *et al.* 2022), selected for its superior data efficiency and refined training strategies suitable for medical datasets; the Swin Transformer (Swin-Base) (Liu *et al.* 2021), which introduces a hierarchical architecture with shifted windows to capture multi-scale features; and the Pooling-based Vision Transformer (PiT-B) (Ren *et al.* 2024), which incorporates spatial dimension reduction to bridge the gap between transformer flexibility and the structural abstraction of CNNs. By deploying this diverse set of architectures, we aim to dissect how specific structural innovations, ranging from hierarchical attention to pooling mechanisms, impact diagnostic precision in skin cancer classification.

Transfer Learning

Training high-capacity DL architectures, particularly Vision Transformers, from scratch necessitates massive volumes of annotated data to overcome the lack of inherent inductive biases found in convolutional networks. Given the intrinsic scarcity of labeled medical imaging datasets compared to general-domain repositories, initializing models with random weights often leads to poor convergence and significant overfitting. To mitigate this challenge, we adopted a transfer learning strategy by leveraging models pre-trained on the ImageNet-1K dataset. This approach allows our network to inherit robust hierarchical feature representations, ranging from low-level edge detection to high-level semantic abstractions, learned from millions of natural images. By transferring this prior knowledge to the dermatological domain, we effectively bypass the computational burden of learning fundamental visual patterns de novo, thereby accelerating training stability and enhancing generalization performance on dermoscopic images (Ren *et al.* 2022).

In the implementation phase, the architectural adaptation involved replacing the original classification head, designed for the 1,000 distinct classes of ImageNet, with a task-specific Multi-Layer Perceptron (MLP) tailored to our seven diagnostic categories. We employed an end-to-end fine-tuning protocol rather than merely using the backbone as a fixed feature extractor. This allowed the pre-trained weights to be subtly adjusted via backpropagation, facilitating the domain adaptation process. By doing so, the models could recalibrate their attention mechanisms to focus on the specific, fine-grained morphological details pertinent to skin lesions, such as pigment networks and vascular structures, while retaining the robust generalization capabilities acquired during the pre-training phase (Bengio 2012).

Experimental Design and Training Protocol

To ensure the reproducibility and rigor of our comparative analysis, all DL architectures were implemented using the PyTorch framework on a high-performance workstation equipped with an NVIDIA GPU. The dataset was partitioned using a stratified sampling strategy to maintain the inherent class distribution across subsets, resulting in a split of approximately 70% for training (n=7,005), 15% for validation (n=1,498), and 15% for testing (n=1,508). Prior to feeding the networks, all images underwent standardization, including resizing to uniform dimensions compatible with the pre-trained transformer backbones (typically 224×224) and normalization using ImageNet mean and standard deviation parameters. This rigorous preprocessing pipeline was essential to stabilize the training dynamics and ensure that the attention mechanisms could effectively attend to lesion-specific features without being swayed by lighting or resolution inconsistencies.

The training phase was conducted using the Cross-Entropy Loss function to penalize classification errors across the seven diagnostic categories. To optimize the network weights, we employed the AdamW optimizer, widely recognized for its efficacy in training transformer models, initialized with an empirically tuned learning rate and weight decay to prevent overfitting. We utilized a dynamic learning rate scheduler (Cosine Annealing) to progressively reduce the learning rate, allowing the model to settle into sharper minima as training converged. The best-performing model weights were saved based on the validation loss metric to avoid the pitfalls of overfitting during extended epochs. Finally, the quantitative evaluation was performed on the unseen test set using standard metrics, including Accuracy, Precision, Recall, and F1-Score, to provide a holistic view of each model's diagnostic capability.

Performance Evaluation Metrics

To truly understand how these models hold up, we didn't just look at their accuracy. We used a multi-layered approach to evaluate them, looking at how well they handle different skin diseases and how they would actually run in a real-world clinic. We tracked standard scores like Accuracy, Precision, Recall, and the F1-Score to see how reliable the diagnoses are. However, for a model to be useful in a hospital, it also needs to be efficient. That's why we also measured "Params" to see how much memory the model takes up, "GFLOPs" to calculate the raw processing power required, and "Inference Time" to see how many milliseconds it takes for a doctor to get a result. You can see how all these factors compare for each model in Table 2. The formulas we used to calculate these results are shown above. Accuracy (Eq. 1) gives us the big picture of correct guesses, while Precision (Eq. 2) tells us how often the model is right when it flags a lesion as concerning. Recall (Eq. 3) is perhaps the most important for patients because it measures how many actual cancer cases the model caught without missing any. Finally, the F1-Score (Eq. 4) helps us find the sweet spot between being precise and being thorough, which is vital since some types of skin cancer in our dataset are much rarer than others.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

RESULTS

Quantitative Performance and Comparative Analysis

The empirical evaluation of the four benchmarked ViT architectures reveals distinct trade-offs between diagnostic precision and computational efficiency. As summarized in Table 2, the DeiT III-Base model achieved the highest overall performance, reaching an accuracy of 92.04% and an F1-score of 85.44%. This superior performance suggests that the data-efficient training strategies and refined attention mechanisms of DeiT III are particularly effective for capturing the fine-grained morphological features necessary for skin lesion classification. Swin-Base followed closely with an accuracy of 91.64%, demonstrating the benefit of hierarchical feature extraction in dermoscopic analysis.

In terms of computational complexity, PiT-B exhibited the highest efficiency regarding memory footprint and processing workload, utilizing only 72.75M parameters and 21.1067 GFLOPs. However, this reduction in GFLOPs did not translate to the fastest execution; PiT-B recorded the highest Inference Time (1.5418 ms), likely due to its unique pooling-based architecture which may hinder parallelization compared to the standard transformer blocks. Conversely, ViT-Base and DeiT III-Base provided the most rapid predictions with inference times of 0.5459 ms and 0.5674 ms, respectively, making them the most suitable candidates for high-throughput clinical screening.

Analysis of Model Predictions and Error Patterns

To dissect the model's decision-making process, we analyzed the confusion matrix for the top-performing DeiT III-Base model, as shown in Figure 2. The diagonal elements indicate high sensitivity for the most prevalent class, Melanocytic Nevi (NV), with 977 correct predictions. However, notable confusion exists between Actinic Keratoses (AKIEC) and Benign Keratosis (BKL), as well as between Melanoma (MEL) and NV. This reflects the inherent "visual mimicry" described in clinical literature, where benign and malignant lesions share overlapping pigment patterns.

A qualitative assessment of these results is provided in Figure 3, which visualizes both successful and erroneous predictions alongside their confidence scores. True predictions often correspond to lesions with clear, well-defined diagnostic structures, such as the distinct vascular patterns in VASC or the characteristic symmetry in NV. In contrast, misclassified cases, such as a MEL predicted as AKIEC with 49.3% confidence, often involve ambiguous textures or peripheral artifacts that challenge the self-attention mechanism.

Table 2 Performance comparison of the benchmarked ViT architectures on the HAM10000 test set, evaluated by predictive metrics and computational efficiency.

Models	Accuracy	Precision	Recall	F1 Score	Params (M)	Gflops	Inference Time (Ms)
DeiT III-Base	0.9204	0.8794	0.8348	0.8544	85.82	33.6955	0.5674
ViT-Base	0.8952	0.8508	0.7704	0.8060	85.80	33.6955	0.5459
Swin-Base	0.9164	0.8703	0.8356	0.8508	86.75	30.3375	0.7927
PiT-B	0.9072	0.8677	0.8599	0.8606	72.75	21.1067	1.5418

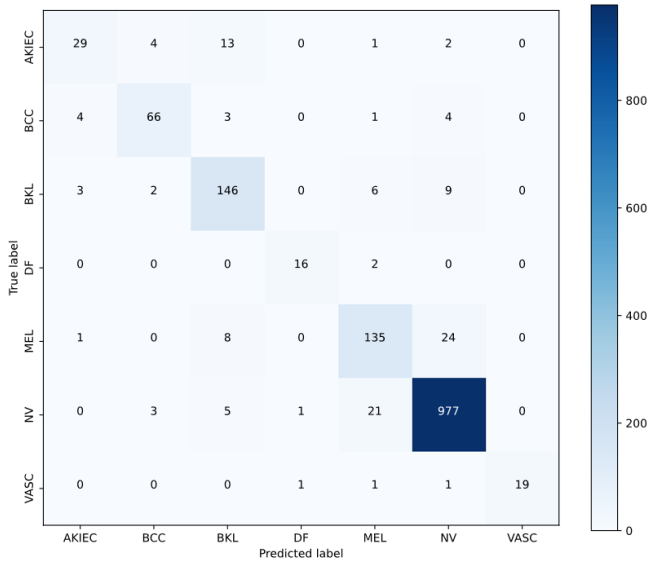

Figure 2 Confusion matrix for the top-performing DeiT III-Base model, illustrating class-specific diagnostic performance and inter-class misclassification patterns.

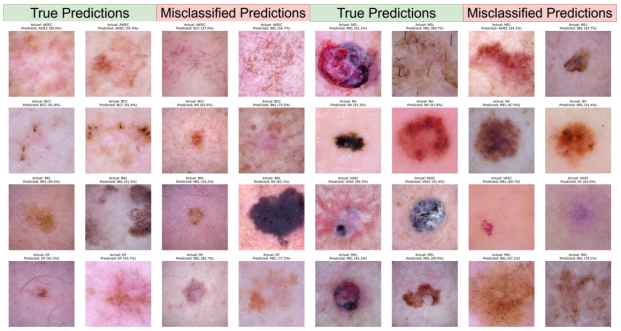

Figure 3 Qualitative analysis of model predictions showing representative examples of successful classifications and erroneous predictions with their corresponding confidence scores.

DISCUSSION

Interpretation of Key Findings

The results of this study underscore the effectiveness of ViTs in overcoming the locality bias of traditional CNNs. The top performance of DeiT III-Base and Swin-Base confirms that modeling global contextual relationships is vital for differentiating complex lesions. Specifically, the self-attention mechanism allows these models to attend to subtle, long-range dependencies across the lesion surface, which are often missed by the localized receptive fields of standard convolutional filters. Furthermore, the use of Transfer Learning from ImageNet-1K was essential in stabilizing training and achieving high accuracy despite the limited size of medical datasets compared to natural image repositories.

Clinical Implications, Limitations, and Future Directions

From a clinical perspective, the high accuracy and sub-millisecond inference times of models like DeiT III-Base suggest their viability as real-time CAD tools. Such systems could provide a critical "second opinion," potentially reducing the rate of unnecessary biopsies for benign lesions while ensuring that early-stage melanomas are not overlooked. However, this study has limitations. The HAM10000 dataset is significantly imbalanced, with NV accounting for over 60% of the samples. As seen in the confusion matrix (Figure 2), this imbalance can bias models toward predicting the majority class. Future research should focus on incorporating multi-modal data, such as patient age, anatomical location, and clinical history, to further refine diagnostic precision. Additionally, exploring Explainable AI (XAI) techniques, beyond the visual samples in Figure 3, will be crucial for building clinician trust and ensuring the transparency of DL models in high-stakes medical environments.

CONCLUSION

This study demonstrates the efficacy of advanced ViT architectures in the automated classification of pigmented skin lesions using the HAM10000 dataset. By leveraging global self-attention mechanisms to model long-range dependencies, our framework successfully overcame the locality limitations of traditional CNNs, with the DeiT III-Base architecture emerging as the superior model, achieving a peak accuracy of 92.04% and an F1-score of 85.44%. The comparative analysis revealed that while hierarchical structures like Swin-Base offer competitive diagnostic precision, the data-efficient strategies of DeiT III provide an optimal balance between predictive sensitivity and computational throughput, characterized by sub-millisecond inference times suitable for real-time clinical screening. Furthermore, the integration of transfer learning from large-scale natural image repositories was essential for stabilizing training and achieving high performance despite the inherent class imbalances within the dermoscopic data. Ultimately, these findings underscore the potential of transformer-based paradigms as robust CAD tools in dermatology. Future research will focus on the integration of multi-modal data, such as patient history and anatomical location, alongside XAI techniques to ensure the transparency and clinical reliability of DL deployments in high-stakes healthcare environments.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The dataset analyzed for this study is the public dataset, which is available on Kaggle: https://www.kaggle.com/datasets/surajghuwalewala/ham1000-segmentation-and-classification?select=GroundTruth.csv

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

2025 Skin cancer: Ham10000 dataset.

Armstrong, B. K. and A. Kricker, 1995 Skin cancer. Dermatologic Clinics **13**: 583–594.

Aruk, I., I. Pacal, and A. N. Toprak, 2026 A comprehensive comparison of convolutional neural network and visual transformer models on skin cancer classification. Computational Biology and Chemistry **120**.

Bengio, Y., 2012 Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 29th International Conference on Machine Learning*, volume 27, pp. 17–36.

Bruno, A., A. Artesani, P. L. Mazzeo, F. Janan, G. Yang, *et al.*, 2025 Boosting skin cancer classification: A multi-scale attention and ensemble approach with vision transformers. Sensors **25**: 2479.

Cakmak, Y. and A. Maman, 2025 Deep learning for early diagnosis of lung cancer. Computational Systems and Artificial Intelligence **1**: 20–25.

Cakmak, Y. and I. Pacal, 2025 Comparative analysis of transformer architectures for brain tumor classification. Exploratory Medicine **6**.

Çakmak, Y. and N. Pacal, 2025 Deep learning for automated breast cancer detection in ultrasound: A comparative study of four cnn architectures. Artificial Intelligence in Applied Sciences **1**: 13–19.

Chaurasia, A. K., P. W. Toohey, H. C. Harris, and A. W. Hewitt, 2025 Multi-resolution vision transformer model for histopathological skin cancer subtype classification using whole slide images. Computers in Biology and Medicine **196**.

Dagnaw, G. H., M. El Mouhtadi, and M. Mustapha, 2024 Skin cancer classification using vision transformers and explainable artificial intelligence. Journal of Medical Artificial Intelligence **7**.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, *et al.*, 2020 An image is worth 16x16 words: Transformers for image recognition at scale.

Gloster Jr, H. M. and D. G. Brodland, 1996 The epidemiology of skin cancer. Dermatologic Surgery **22**: 217–226.

Gloster Jr, H. M. and K. Neal, 2006 Skin cancer in skin of color. Journal of the American Academy of Dermatology **55**: 741–760.

Jerant, A. F., J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, 2000 Early detection and treatment of skin cancer. American family physician **62**: 357–368.

Karthik, R., R. Menaka, S. Atre, J. Cho, and S. V. Easwaramoorthy, 2024 A hybrid deep learning approach for skin cancer classification using swin transformer and dense group shuffle non-local attention network. IEEE Access **12**: 158040–158051.

Liu, Z., Y. Lin, Y. Cao, *et al.*, 2021 Swin transformer: Hierarchical vision transformer using shifted windows.

Madan, V., J. T. Lear, and R.-M. Szeimies, 2010 Non-melanoma skin cancer. The lancet **375**: 673–685.

Manju, V. N., D. S. Dayana, N. Patwari, K. P. B. Madavi, and K. K. Sowjanya, 2025 Attention-enhanced vision transformer model for precise skin cancer detection. In *Proceedings of the 2025 International Conference on Emerging Technologies in Computing and Communication (ETCC)*, IEEE.

Ozdemir, B. and I. Pacal, 2025 An innovative deep learning framework for skin cancer detection employing convnextv2 and focal self-attention mechanisms. Results in Engineering **25**: 103692.

Pacal, I., M. Alaftekin, and F. D. Zengul, 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**: 3174–3192.

Pacal, I. and Y. Cakmak, 2025a A comparative analysis of u-net-based architectures for robust segmentation of bladder cancer lesions in magnetic resonance imaging. Eurasian Journal of Medicine and Oncology **9**: 268–283.

Pacal, I. and Y. Cakmak, 2025b *Diagnostic Analysis of Various Cancer Types with Artificial Intelligence*. Duvar Yayınları.

Ren, H., J. Guo, S. Cheng, and Y. Li, 2024 Pooling-based visual transformer with low complexity attention hashing for image retrieval. Expert Systems with Applications **241**: 122745.

Ren, Z., H. Zhang, T. Huang, *et al.*, 2022 Deep learning (cnn) and transfer learning: A review. Journal of Physics: Conference Series **2273**: 012029.

Sakib, A. H., M. I. H. Siddiqui, S. Akter, A. Al Sakib, and M. R. Mahmud, 2025 Levit-skin: A balanced and interpretable transformer-cnn model for multi-class skin cancer diagnosis. International Journal of Science and Research Archive **15**: 1860–1873.

Siegel, R. L., A. N. Giaquinto, and A. Jemal, 2024 Cancer statistics, 2024. CA: a cancer journal for clinicians **74**: 12–49.

ThangaPurni, J. and M. Braveen, 2025 Unified arp-vit-cnn system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions. Array p. 100515.

Touvron, H., M. Cord, and H. Jégou, 2022 Deit iii: Revenge of the vit.

How to cite this article: Islam, M. S., Chamgoué, A. C. and Dubb, G. P. Benchmarking State-of-the-Art Vision Transformer Architectures for the Automated Classification of Pigmented Skin Lesions. *Computers and Electronics in Medicine*, 3(1), 42-47, 2026.