Journal of Medical and Health Studies (JMHS)
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable Transformer-Based Skin Lesion Classification from Clinical Images

Ahmed Ali Linkon[1], Mostafizur Rahman Shakil[2], Shahriar Ahmed[3], Md Rashel Miah[4], Asif Hassan Malik[5]
[1]*Department of Computer Science, Westcliff University, Irvine, CA 92614, USA*
[2]*Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA*
[3]*School of Business, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010, USA*
[4]*Department of Business Administration, Westcliff University, Irvine, CA 92614, USA*
[5]*Department of Chemistry, York College, The City University of New York (CUNY), Jamaica, NY 11451, USA*
Corresponding Author: Asif Hassan Malik, **E-mail**: asifmalikbd@gmail.com

| ABSTRACT
Early and reliable detection of skin cancer is critical for reducing disease burden and improving patient outcomes, yet large-scale screening remains constrained by limited specialist availability and heterogeneous image acquisition conditions. This paper presents an efficient transformer-based framework for automated multiclass skin lesion classification, centered on an EfficientViT architecture designed to balance representational capacity and computational efficiency. The proposed approach is evaluated against lightweight transformer and CNN baselines, including DeiT-Tiny, Axial Attention Transformer, Swin Transformer-Tiny, and EfficientNetV2-S, using the PAD-UFES-20 dataset comprising 2,298 smartphone-acquired clinical images across six lesion categories. Experimental results show that EfficientViT achieves superior performance, reaching 99.40% accuracy and 99.78% PR-AUC, indicating robust discrimination under real-world acquisition variability. To enhance transparency and support clinical interpretability, Grad-CAM visual explanations are integrated to highlight lesion-relevant regions driving model predictions. Overall, the results demonstrate that EfficientViT provides an accurate and interpretable solution for practical skin lesion screening using consumer-grade images.

| KEYWORDS
Skin cancer detection, skin lesion classification, deep learning, Vision Transformer, Grad-CAM, PR-AUC, Explainable Artificial Intelligences.

| ARTICLE INFORMATION
ACCEPTED: 21 February 2026 **PUBLISHED:** 08 March 2026 **DOI:** 10.32996/jmhs.2026.7.5.7

1. Introduction

Skin cancer is among the most common malignancies worldwide and continues to impose a substantial public health and economic burden due to high incidence, recurrent clinical visits, and the costs associated with late-stage treatment. Delayed recognition of malignant lesions can lead to advanced disease progression, reduced survival, and increased healthcare expenditure, making early screening and accurate triage essential for improving outcomes. These challenges are amplified in low- and middle-income countries, where access to dermatologists, dermoscopic equipment, and timely diagnostic pathways is limited by geography, workforce shortages, and constrained resources. Scalable screening approaches that can operate with routine clinical photographs or smartphone-acquired images are therefore important for extending early detection beyond specialist centers[1].

Conventional diagnosis typically relies on clinician visual inspection followed by histopathological confirmation, which is time-consuming and may be influenced by subjective judgment and inter-observer variability. Reliable assessment is further complicated by lesion morphological diversity, variations in skin tone, differences in anatomical site, and inconsistent imaging conditions such as illumination changes, blur, and background clutter. These factors lead to pronounced intra-class variation and inter-class similarity across lesion categories, making robust multiclass recognition challenging in community screening and telemedicine settings. In practice, screening tools must not only achieve high accuracy but also behave consistently under

heterogeneous acquisition conditions and provide transparent outputs that support clinical interpretation and risk-aware triage[2][3].

Recent advances in artificial intelligence, particularly deep learning, have enabled automated analysis of skin lesion images at scale. Convolutional neural networks have been widely adopted for lesion classification because they learn discriminative features directly from data and can perform well under controlled benchmarks. However, CNNs often emphasize local textures and may underutilize broader contextual cues needed to separate visually similar lesions, especially when lesion boundaries are ambiguous or the surrounding skin context affects appearance. These limitations motivate architectures capable of modeling long-range dependencies while retaining fine-grained detail, particularly for smartphone-acquired images where acquisition noise and viewpoint variability can weaken purely local pattern recognition[4].

Recent studies collectively emphasize clinically oriented computer-aided diagnosis that couples performance with interpretability and deployment readiness across diverse modalities and conditions. Web application–based esophageal disease diagnosis illustrates an explicit design for low-resource clinical environments [13], while explainable stacking ensembles for cervical cancer diagnosis are reported both as conference evidence and as a journal-level formulation, reinforcing methodological continuity and translational intent [18,22]. Transformer-centered modeling has also been applied to oral cancer segmentation across binary and multiclass settings [14] and to prostate cancer classification in MRI via hybrid vision transformer designs [16]. For oncologic decision support, explainable deep stacking ensembles for brain tumor diagnosis [29], hierarchical Swin transformer ensembles for robust (including decentralized) breast cancer diagnosis [30], and attention-augmented hybrid transformers for efficient, explainable lung cancer diagnosis [31] collectively demonstrate the consolidation of ensemble–transformer paradigms. Beyond oncology, transfer learning–driven pipelines have been explored for lung cancer detection in smart healthcare [35], early leukemia diagnostics using image processing with transfer learning [24], and scalable pneumonia diagnosis from chest X-ray imagery [37], alongside ensemble deep learning for retinal disease recognition [40] and deep learning–based microorganism classification [41]. Complementary biomedical directions further broaden the scope toward precision wound healing with smart technologies [19], nanoconjugate strategies for glioblastoma therapy [20], and protein-degradation–oriented perspectives for reprogramming cancer immunity [21].

Vision transformers provide a strong alternative by leveraging self-attention to capture global relationships across the image, enabling joint learning of local lesion cues and global structural context. Motivated by these strengths and the need for efficient screening models, this work evaluates lightweight transformer-centric baselines, including DeiT-Tiny, Axial Attention Transformer, Swin Transformer-Tiny, and EfficientNetV2-S, and introduces EfficientViT as the proposed model for robust multiclass skin lesion classification. EfficientViT is designed to balance representational capacity with computational efficiency, making it suitable for practical screening pipelines where inference cost, stability, and generalization are important. To further improve trust and support error analysis, we integrate Grad-CAM visual explanations that highlight class-discriminative regions, enabling assessment of whether predictions are driven by lesion-relevant evidence rather than spurious background artifacts.

While prior studies have explored automated skin lesion recognition using deep learning, many are constrained by dataset-specific optimization, limited validation diversity, and insufficient interpretability analysis, which collectively restrict translation to real-world screening. This study addresses these gaps by benchmarking multiple lightweight models on the PAD-UFES-20 dataset, which contains smartphone-acquired clinical images and reflects practical acquisition variability. We adopt a unified pipeline that includes standardized preprocessing, multiclass training, and comprehensive evaluation using complementary metrics, alongside qualitative interpretability assessment using Grad-CAM to support transparent decision-making. Our key contributions are:

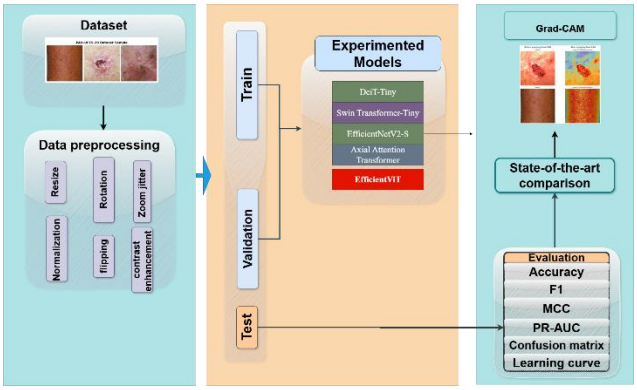

Fig. 1. Overview of the proposed methodology.

- Conducting a comparative study of lightweight transformer and efficient baselines (DeiT-Tiny, Axial Attention Transformer, Swin Transformer-Tiny, and EfficientNetV2-S) against the proposed EfficientViT for multiclass skin lesion classification.

- Developing an end-to-end experimental pipeline covering preprocessing, training, and evaluation on PAD-UFES-20 under realistic smartphone acquisition conditions.
- Incorporating Grad-CAM-based visual explanations to enhance model transparency and facilitate clinically meaningful interpretation and error analysis.
- Providing robustness-oriented analysis through multiple complementary metrics to support reliable screening-oriented performance reporting.

The rest of the paper is structured as follows: Section 2 presents related works on skin lesion classification and transformer-based models. Section 3 describes the dataset, preprocessing, and the proposed EfficientViT architecture. Section 4 reports the evaluation metrics and experimental results. Section 5 discusses findings and limitations, and Section 6 concludes the paper.

2. Related Works

Recent progress in automated skin cancer screening has been driven by deep learning models that learn discriminative lesion representations from dermoscopic and clinical imagery. CNN-based pipelines continue to be prevalent due to stable training and strong local feature extraction, their limited receptive field and reliance on local texture can reduce reliability when global context such as asymmetry or lesion skin contrast is diagnostically important. Consequently, recent work increasingly incorporates attention mechanisms and hybrid strategies to better model both local and global evidence.

Efficiency-oriented CNN refinements form one active line of research because practical screening tools often operate under compute constraints. Cheng et al. [5] introduced an enhanced MobileNet design that integrates a fused spatial–channel attention module to strengthen lesion localization cues while maintaining computational efficiency, reporting improved precision/recall/F1 on ISIC-2019. Their results highlight that lightweight attention can partially mitigate the locality limitation of standard CNN backbones, but robustness under dataset shift remains an open question when evaluation is confined to a single benchmark. Similarly, Abdullah et al. [6] presented a sequential CNN pipeline with ROI-focused preprocessing and reported 96.25% accuracy on HAM10000, demonstrating the benefit of structured preprocessing and staged feature learning. However, this type of dataset-specific optimization may inadvertently encode dataset artifacts, motivating evaluation protocols that explicitly test generalization across heterogeneous acquisition conditions and class distributions.

Beyond single-architecture proposals, broader pipelines and surveys emphasize that reported gains are sensitive to the choice of preprocessing, label space, and evaluation setup. Kavithaa et al. [7] reviewed deep learning strategies for skin cancer detection and classification and reported an R-CNN–based result of 84.32% accuracy, underscoring the variability in outcomes across modeling paradigms and the importance of consistent metrics and validation design. These observations suggest that improvement claims should be interpreted alongside factors such as imbalance handling, augmentation realism, and lesion-type coverage, rather than accuracy alone.

Agro-environmental intelligence increasingly demonstrate how modern machine learning pipelines can deliver rapid, field-relevant decision support for crop health and management. A comprehensive synthesis of jute leaf disease detection highlights how classical machine learning and contemporary deep learning families jointly improve recognition performance under practical imaging constraints [10]. In parallel, deep models are being tailored for agricultural decision workflows, including automated weed species classification to support rice cultivation [15]. Disease-centric architectures continue to evolve toward transformer-driven and efficiency-aware designs, as shown in accelerated cotton leaf disease identification [17] and explainable transformer frameworks that extend leaf diagnostics to adjacent industrial inspection tasks such as fabric defect detection [33]. Complementing these efforts, explainability-enhanced vision transformer ensembles for mango leaf disease [27] and MaxViT-based solutions for soybean leaf/seed disease identification [28] further underscore the shift toward high-capacity yet interpretable architectures. Finally, web-based deployment for cucumber disease recognition reflects an emphasis on translational usability in resource-constrained settings [38].

Clinical translation and user-facing screening introduce additional constraints particularly explainability, calibration, and safety trade-offs. Liu et al. [1] proposed an XAI-enabled survival modeling approach that predicts future skin cancer risk directly from facial images, achieving a c-index of 0.72 and outperforming models relying on established risk factors. This work reinforces the importance of explainability for preventive and decision-support settings where actionable interpretation is required. In contrast, Manolakos et al. [8] evaluated an AI-assisted spectroscopy device and reported high sensitivity (97.04%) but low specificity (26.22%), illustrating that safety-driven designs can still impose substantial downstream burden through false positives if calibration and operating points are not carefully managed. Finally, Gaube et al. [3] found strong public preference for dermatologist screening over AI-enabled apps, with higher perceived trust and accuracy attributed to clinicians, indicating that model transparency and trust calibration are critical for adoption.

Contemporary work also demonstrates methodological transferability from perception-centric learning to security analytics, scientometrics, and language understanding. In enterprise cybersecurity, AI-powered threat detection has been positioned to enhance real-time response under modern operational constraints [11]. In scholarly ecosystems, edge-conditioned graph attention networks have been proposed for journal ranking in citation networks, highlighting the role of relational inductive bias and graph-conditioned parameterization for ranking and influence estimation [12]. Natural language processing continues to support large-scale inference from user-generated text, spanning sentiment extraction from online drug reviews [23],

comparative modeling for suicidal ideation detection using NLP with machine and deep learning methods [25], and multi-class sentiment classification for Bengali social media comments [26]. Human-centric sensing and recognition are also being advanced through ensemble transformers with post-hoc explanations for depression emotion and severity detection [34], deep learning for classroom activity classification [39], and multimodal perception via vision–audio fusion with hybrid and tensor fusion strategies [32]. Finally, explainable ensemble-based recognition has been extended to biodiversity and conservation contexts through rare medicinal plant recognition frameworks [36].

Overall, prior work often evaluates models within constrained datasets, provides limited insight into failure modes under imbalance and domain shift, and offers insufficient transparency to support calibrated decision-making. To address these limitations, our study emphasizes an XAI-centric framework that couples multi-scale representation learning with faithful visual rationales, integrates imbalance-aware optimization, and reports robustness-focused analysis (including class-wise errors and calibration-aware behavior) to improve reliability and user trust in practical screening scenarios.

3. Methodology

3.1 Data Description

Figure 2 presents the PAD-UFES-20 dataset [9], a public collection of clinical skin lesion photos taken with smartphone cameras in real outpatient settings. It contains 2,298 images from 1,641 lesions and 1,373 patients. The dataset includes six diagnosis groups: actinic keratosis, basal cell carcinoma, melanoma, nevus, squamous cell carcinoma, and seborrheic keratosis. Along with the images, it provides patient and lesion information such as age, sex, lesion location, skin type, and lesion size. Cancer cases are confirmed by biopsy. Because the photos were captured under different lighting and devices, image quality varies, and the classes are imbalanced, especially for melanoma. We therefore use patient-wise splitting and report balanced performance metrics.

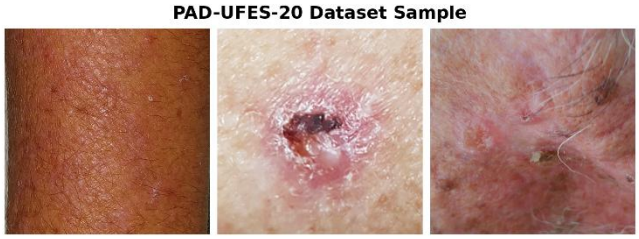

Fig. 2. Sample images from the PAD-UFES-20 dataset.

3.2 Image Preprocessing

The PAD-UFES-20 images were processed under a unified pipeline to provide clean, model-ready inputs for training transformer- and CNN-based classifiers. Each image was resized to 224×224 and intensity-normalized to [0, 1]. To improve robustness to acquisition variability and mitigate class imbalance effects, augmentation included rotations (±15°), horizontal/vertical flips (p=0.5), zoom jitter (0.9–1.1), and brightness/contrast perturbations (0.8–1.2). Gaussian noise was additionally injected to emulate sensor noise and illumination fluctuations typical of smartphone capture. Samples were shuffled prior to training, labels were one-hot encoded for categorical objectives, and the dataset was split into 80% training, 15% validation, and 5% testing while maintaining class proportions across splits.

3.3 TL Models

1) *Deit-Tiny*: DeiT-Tiny is retained as a compact Vision Transformer baseline that captures global lesion context through self-attention while remaining computationally lighter than standard ViT variants. Following the ViT formulation, the input image is partitioned into non-overlapping patches and flattened into a token sequence, as defined in Equation (1). These patch tokens are then combined with positional embeddings and propagated through stacked transformer blocks, where the layer-wise representation update is performed according to Equation (2). This baseline provides a clear transformer-only reference for skin lesion classification under constrained latency requirements.

$$x_p = \text{Flatten}\big(\text{Patchify}(X)\big) \quad (1)$$

$$z_{l+1} = \text{MLP}\big(\text{MSA}\big(\text{LN}(z_l)\big) + z_l\big) + \text{MSA}\big(\text{LN}(z_l)\big) + z_l \quad (2)$$

2) Axial Attention Transformer: To better exploit directionally structured lesion patterns (e.g. elongated streaks, border contours, and texture anisotropy), we include an Axial Attention Transformer that computes attention independently along horizontal and vertical axes. The directional attention operation is formulated in Equation (3), which enables efficient long-range dependency modeling while reducing the quadratic cost associated with full 2D attention. By fusing the two axial attentions, this baseline improves orientation-robust feature learning for practical screening scenarios.

$$A_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \quad A_v = \text{Softmax}\left(\frac{Q_v K_v^T}{\sqrt{d_k}}\right) V_v \tag{3}$$

3) Swin Transformer-Tiny: Swin Transformer-Tiny is adopted as a strong hierarchical transformer baseline that scales efficiently to higher-resolution visual patterns. It performs self-attention within shifted local windows, improving computational efficiency while preserving fine-grained lesion cues needed for discriminating visually similar classes. Within each window, attention is computed using the window-based formulation in Equation (4), where relative positional bias helps maintain spatial relationships that are clinically relevant (e.g., boundary irregularity and local asymmetry).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right) V \tag{4}$$

4) *EfficientNetV2-S:* To represent an edge-friendly CNN baseline with strong accuracy–efficiency trade-offs, we employ EfficientNetV2-S. Its computational efficiency is largely driven by lightweight convolutional operations that decouple spatial filtering from channel mixing, which can be abstracted by the depthwise separable formulation in Equation (5). This baseline is particularly suitable for skin lesion screening where robust texture color extraction is required but compute and memory budgets may be constrained.

$$O = P * (D * I) \tag{5}$$

5) *EfficientViT:* We propose EfficientViT, an efficiency-oriented multi-scale transformer designed to jointly capture (i) fine local evidence (e.g., border sharpness, pigment networks, micro-texture) and (ii) coarse contextual structure (e.g., lesion geometry and surrounding skin context). The model processes multiple patch granularities using parallel transformer branches; for each scale sss, the corresponding branch generates an embedding as described in Equation (6). Multi-scale representations are then integrated via cross-scale attention fusion following Equation (7), enabling bidirectional information exchange between fine and coarse streams before classification. To address class imbalance common in real screening datasets, we optimize the network using focal loss as defined in Equation (8), which down-weights easy examples and focuses learning on harder (often minority-class) samples. Finally, to replace the earlier web-application emphasis, EfficientViT incorporates an explainability module that produces lesion-focused visual rationales (e.g., attention rollout and/or gradient-based saliency) to support transparency and trust in clinical usage, without modifying the mathematical formulation of the core learning objective. samples.

$$Z_i = \text{ViT}(S_i) \tag{6}$$

$$F = \text{Softmax}\left(\frac{Q_F K_F^T}{\sqrt{d_k}}\right) V_F \tag{7}$$

$$y = \text{Softmax}(W_o F + b_o) \tag{8}$$

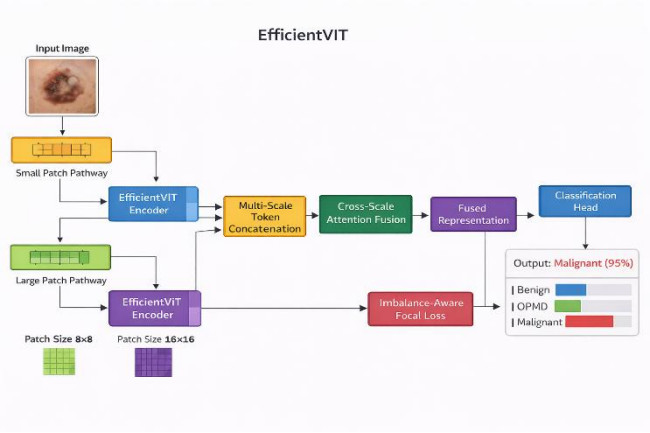

Fig. 3. Proposed EfficientViT model architecture.

3.4 Evaluation Metrics

For multiclass skin lesion classification, we report five complementary metrics. Accuracy measures the fraction of correctly predicted images over the full test set. Precision and recall characterize the trade-off between false alarms and missed cases, respectively, which is important when errors have different clinical consequences. The f1-score combines precision and recall into a single measure that penalizes extreme imbalance between them. Matthews correlation coefficient summarizes prediction quality using all entries of the confusion matrix and remains informative under class imbalance. Finally, PR-AUC summarizes precision–recall behavior across thresholds and is particularly sensitive to minority-class performance, where accuracy can be overly optimistic.

4. Result and Analysis

4.1 Performance Comparison of Experimental Models

Table I compares the proposed efficientvit with four baseline models on our dataset using accuracy, f1, mcc, and pr-auc. EfficientViT delivers the strongest overall performance, achieving 99.40 accuracy and 99.78 pr-auc, which suggests that its representation learning is highly effective for separating visually similar lesion categories under real-world smartphone variability. aa transformer also performs strongly, with high f1 (98.47) and the best mcc (97.80), indicating balanced predictions across classes and reduced sensitivity to class imbalance. deit-tiny remains a competitive compact transformer baseline (98.40 accuracy, 99.21 pr-auc), while swin transformer-tiny achieves high pr-auc (99.10) but slightly lower accuracy, consistent with its windowed attention favoring local texture patterns. efficientnetv2-s provides a solid cnn baseline with lower overall scores, supporting the advantage of transformer-style global context modeling for fine-grained lesion discrimination.

Table 1 Performance of experimental models.

Dataset	Model	Accu	F1	MCC	PR-AUC
PAD-UFES-20	EfficientViT	**99.40**	98.71	95.95	**99.78**
	EfficientNetV2-S	97.90	96.21	95.76	98.49
	AA Transformer	98.52	98.47	97.80	99.12
	Swin Transformer-Tiny	97.01	98.49	96.12	99.10
	DeiT-Tiny	98.40	97.90	97.50	99.21

4.2 Performance Validation

The confusion matrix in Fig. 4 demonstrates strong and well-balanced classification performance of EfficientViT on the PAD-UFES-20 dataset. Correct predictions dominate the main diagonal across all six lesion categories, indicating reliable class discrimination under real-world smartphone acquisition conditions. Melanoma and nevus are classified with particularly high consistency, while only limited confusion is observed among clinically and visually similar classes such as BCC, BK, AK, and SCC. These residual misclassifications are sparse and localized, reflecting expected ambiguity in borderline lesion appearances rather than systematic failure. Overall, the diagonal dominance confirms robust separability despite intra-class variability and class imbalance, suggesting that the model generalizes effectively to heterogeneous clinical imagery while maintaining stable minority-class recognition.

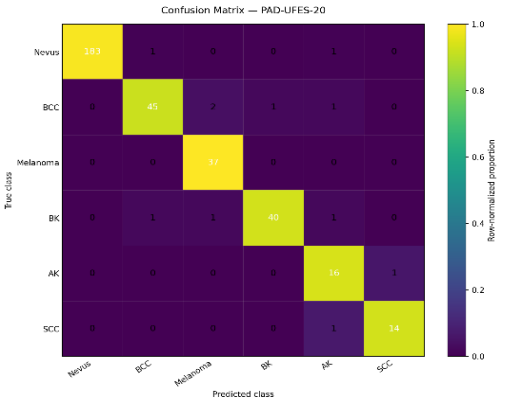

Fig. 4. Confusion matrix of the proposed model.

The learning curves in Figure 5 illustrate the training and validation behavior of EfficientViT on the PAD-UFES-20 dataset. Both loss and accuracy trajectories show smooth and consistent convergence, with the validation curves closely tracking the training

curves across epochs, indicating effective optimization and limited overfitting. The steady decrease in loss and the gradual rise in accuracy toward high final values suggest that the chosen learning rate and regularization strategy enable stable learning under class imbalance and image variability. Minor fluctuations in the validation curves are expected for a small test split and do not indicate instability, confirming reliable generalization on unseen samples.

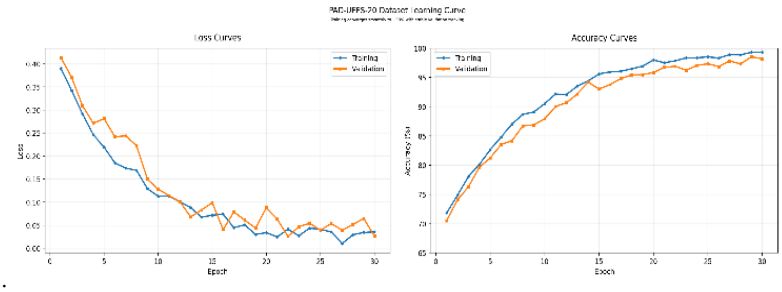

Fig. 5. Learning curve of the proposed model.

4.3 Model Transparency

Figure 6 illustrates the qualitative interpretability of EfficientViT using Grad-CAM visualizations on representative test samples from the PAD-UFES-20 dataset. The left column shows the original input images, while the right column presents the corresponding Grad-CAM overlays highlighting image regions that most strongly influence the model's predictions. In the first example, the activation map concentrates on the lesion core and irregular boundaries, indicating that the model attends to clinically relevant structures such as pigmentation variation and lesion asymmetry. In the second example, Grad-CAM emphasizes diffuse abnormal regions rather than background skin texture, suggesting that the model focuses on subtle visual cues associated with inflammatory or keratinized patterns. The localized and coherent activation patterns across both cases demonstrate that EfficientViT bases its predictions on meaningful lesion regions rather than spurious background artifacts. This behavior supports improved transparency and provides qualitative evidence that the model's decisions align with clinically interpretable visual features.

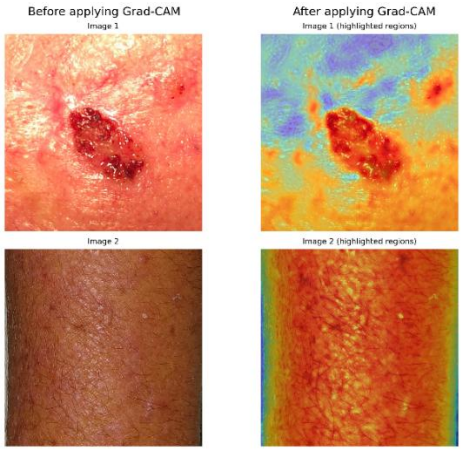

Fig. 6 Grad-CAM heatmaps for skin lesion classification.

4.4 State-of-The-Art Comparison

TABLE II summarizes representative recent studies on skin lesion classification across datasets with different sizes and acquisition conditions. Prior attention-based CNN and hybrid fusion approaches report strong accuracy, but variations in dataset curation, class taxonomy, and validation strategy reduce direct comparability of the reported numbers. In this context, EfficientViT achieves the best result on PAD-UFES-20 (99.40%) while using a compact transformer-style design aimed at efficient feature learning from smartphone-acquired clinical images. The table also indicates that high accuracy is attainable on larger benchmarks using hybrid backbones, whereas results on smaller or differently curated datasets may be inflated or unstable due to limited diversity. Overall, the comparison supports EfficientViT as a competitive and robust choice under realistic PAD-UFES-20 conditions.

Table 2 Performance comparison with previous studies.

Model	Data sample size	Result (%)
MobileNet-MFS (attention-enhanced MobileNetV3)[5]	2513	87.3
SNC_Net (HC features + InceptionV3 fusion)[4]	5332	97.81
MedFusionNet (ConvNeXt + ViT fusion)[3]	10015	98.8
RvXmBlendNet (multi-architecture hybrid)[6]	658	98.26
EfficientViT (Our Proposed)	**2298**	**99.40**

5. Discussion

This study demonstrates that an efficient transformer-based architecture can achieve highly accurate multiclass skin lesion classification on PAD-UFES-20 under smartphone-driven acquisition variability. The strong performance of EfficientViT suggests that compact attention-based representations effectively capture both fine texture cues and global lesion structure needed to separate visually similar classes. The confusion matrix and learning curves further indicate stable optimization and consistent generalization with limited overfitting. However, the evaluation remains restricted to a single public dataset with a small held-out test split, which may not fully reflect clinical diversity. Future work should validate across external cohorts with broader variation in skin tone, devices, and lesion subtypes, and strengthen interpretability through systematic Grad-CAM analysis and quantitative faithfulness checks.

6. Conclusion

This paper introduces an efficient transformer-based framework for multiclass skin lesion classification using EfficientViT, with Grad-CAM incorporated to support transparent decision-making. Evaluations on PAD-UFES-20 show that EfficientViT achieves the strongest overall performance across accuracy, F1-score, MCC, and PR-AUC, indicating reliable discrimination under heterogeneous smartphone-acquired clinical images. Grad-CAM overlays consistently emphasize lesion-relevant regions, providing practical interpretability for error analysis and trust calibration. These findings support EfficientViT as a strong candidate for accurate and interpretable skin lesion screening, motivating further external validation toward real-world clinical adoption.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Liu X, Sangers TE, Nijsten T, et al. Articles Predicting skin cancer risk from facial images with an explainable arti fi cial intelligence (XAI) based approach : a proof-of-concept study. EClinicalMedicine. 2024;71:102550. doi:10.1016/j.eclinm.2024.102550

[2] Brancaccio G, Balato A, Malvehy J, Puig S, Argenziano G, Kittler H. Artificial Intelligence in Skin Cancer Diagnosis : A Reality Check. Journal of Investigative Dermatology. 2024;144(3):492-499. doi:10.1016/j.jid.2023.10.004

[3] Gaube S, Biebl I, Karin M, Engelmann M, Kleine A kathrin, Lermer E. Social Science & Medicine Comparing preferences for skin cancer screening : AI-enabled app vs dermatologist. Soc Sci Med. 2024;349(March):116871. doi:10.1016/j.socscimed.2024.116871

[4] Pham TC, Luong CM, Visani M, Hoang VD. Deep CNN and Data Augmentation for Skin Lesion Classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018;10752 LNAI:573–82. doi:10.1007/978-3-319-75420-8_54

[5] Nasreen G, Haneef K, Tamoor M, Irshad A. Review: a comparative study of state-of-the-art skin image segmentation techniques with CNN. Multimedia Tools and Applications 2022 82:7. 2022 Sep 10;82(7):10921–42. doi:10.1007/s11042-022-13756-5

[6] Anand V, Gupta S, Koundal D, Singh K. Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images. Expert Syst Appl. 2023 Mar 1;213(7):119230. doi:10.1016/j.eswa.2022.119230

[7] Kavitha C, Priyanka S, Priyanka S, Praveen M, Kusuma V. ScienceDirect ScienceDirect Skin Skin Cancer Cancer Detection Detection and and Classification Classification using using Deep Deep Learning Learning Techniques Techniques. Procedia Comput Sci. 2024;235(2023):2793-2802. doi:10.1016/j.procs.2024.04.264

[8] Manolakos D, Patrick G, Geisse JK, Rabinovitz H. Use of an elastic-scattering spectroscopy and artificial intelligence device in the assessment of lesions suggestive of skin cancer: A comparative effectiveness study. JAAD Int. 2023;14:52-58. doi:10.1016/j.jdin.2023.08.019

[9] Tumpa PP, Kabir MA. A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN). Clinical eHealth. 2023 Dec 1;6:76–84. doi:10.1016/j.sintl.2021.100128

[10] Haque R, Khan M, Pranto MN, et al. Data-Centric Approach for Leather Quality Control Using Advanced Vision Transformer Models. Proceedings - International Conference on Next Generation Communication and Information Processing, INCIP 2025. Published online 2025:200-205. doi:10.1109/INCIP64058.2025.11019741

[11] Sultana S, Rahman MM, Hossain MS, Gony MdN, Rafy A. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments. World Journal of Advanced Engineering Technology and Sciences. 2022 Aug 30;6(2):136–46. doi:10.30574/wjaets.2022.6.2.0079

[12] Abid SM, Xiaoping Q, Islam MM, Islam MA, Rahman MM, Alam ARM. Edge-Conditioned GAT for Journal Ranking in Citation Networks. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2025. 2025;1612–9. doi:10.1109/AINIT65432.2025.11035687

[13] Al Masum A, Limon ZH, Islam MA, Rahman MS, Khan M, Afridi SS, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. 2024;153–8. doi:10.1109/BECITHCON64160.2024.10962580

[14] Hossain A, Sakib A, Pranta ASUK, Debnath J, Tarafder MTR, Islam S, et al. Transformer-Based Ensemble Model for Binary and Multiclass Oral Cancer Segmentation. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11012921

[15] Rahman H, Khan MA, Khan S, Limon ZH, Siddiqui MIH, Chakroborty SK, et al. Automated Weed Species Classification in Rice Cultivation Using Deep Learning. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11014047

[16] Debnath J, Bin Mohiuddin A, Pranta ASUK, Sakib A, Hossain A, Shanto MM, et al. Hybrid Vision Transformer Model for Accurate Prostate Cancer Classification in MRI Images. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11013952

[17] Rahman MM, Hossain MS, Dhakal K, Poudel R, Islam MM, Ahmed MR, et al. A Novel Transformer Model for Accelerated and Efficient Cotton Leaf Disease Identification. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11172151

[18] Bin Mohiuddin A, Rahman MM, Gony MN, Shuvra SMK, Rafy A, Ahmed MR, et al. Accelerated and Accurate Cervical Cancer Diagnosis Using a Novel Stacking Ensemble Method with Explainable AI. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11171850

[19] Malik AH, Rahman S. Toward precision wound healing: Integrating regenerative therapies and smart technologies. International Journal of Science and Research Archive. 2025 Sep 30;16(3):244–57. doi:10.30574/ijsra.2025.16.3.2492

[20] Malik AH, Rahman S. Hybrid Temozolomide Nanoconjugates: A polymer–drug strategy for enhanced stability and glioblastoma therapy. International Journal of Science and Research Archive. 2025 Sep 30;16(3):258–68. doi:10.30574/ijsra.2025.16.3.2493

[21] Malik AH, Rahman S. Molecular erasers: Reprogramming cancer immunity through protein degradation. World Journal of Advanced Engineering Technology and Sciences. 2025 Sep 30;16(3):277–91. doi:10.30574/wjaets.2025.16.3.1335

[22] Siddiqui MIH, Khan S, Limon ZH, Rahman H, Khan MA, Al Sakib A, et al. Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI. Inform Med Unlocked. 2025 Jan 1;56(2):101657. doi:10.1016/j.imu.2025.101657

[23] Haque R, Laskar SH, Khushbu KG, Hasan MJ, Uddin J. Data-Driven Solution to Identify Sentiments from Online Drug Reviews. Computers 2023, Vol 12,. 2023 Apr 21;12(4). doi:10.3390/computers12040087

[24] Haque R, Al Sakib A, Hossain MF, Islam F, Ibne Aziz F, Ahmed MR, et al. Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning. BioMedInformatics 2024, Vol 4, Pages 966-991. 2024 Apr 1;4(2):966–91. doi:10.3390/biomedinformatics4020054

[25] Haque R, Islam N, Islam M, Ahsan MM. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies 2022, Vol 10,. 2022 Apr 29;10(3). doi:10.3390/technologies10030057

[26] Haque R, Islam N, Tasneem M, Das AK. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering. 2023 Jun 1;4:21–35. doi:10.1016/j.ijcce.2023.01.001

[27] Noman A Al, Hossain A, Sakib A, Debnath J, Fardin H, Sakib A Al, et al. ViX-MangoEFormer: An Enhanced Vision Transformer–EfficientFormer and Stacking Ensemble Approach for Mango Leaf Disease Recognition with Explainable Artificial Intelligence. Computers 2025, Vol 14,. 2025 May 2;14(5). doi:10.3390/computers14050171

[28] Pranta ASUK, Fardin H, Debnath J, Hossain A, Sakib AH, Ahmed MR, et al. A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. Computers 2025, Vol 14,. 2025 May 18;14(5). doi:10.3390/computers14050197

[29] Haque R, Khan MA, Rahman H, Khan S, Siddiqui MIH, Limon ZH, et al. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Comput Biol Med. 2025 Jun 1;191:110166. doi:10.1016/j.compbiomed.2025.110166 PubMed PMID: 40249992.

[30] Ahmed MR, Rahman H, Limon ZH, Siddiqui MIH, Khan MA, Pranta ASUK, et al. Hierarchical Swin Transformer Ensemble with Explainable AI for Robust and Decentralized Breast Cancer Diagnosis. Bioengineering 2025, Vol 12,. 2025 Jun 13;12(6). doi:10.3390/bioengineering12060651

[31] Debnath J, Uddin Khondakar Pranta AS, Hossain A, Sakib A, Rahman H, Haque R, et al. LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis. Inform Med Unlocked. 2025 Jan 1;57(1):101669. doi:10.1016/j.imu.2025.101669

[32] Ahmed MR, Haque R, Rahman SMA, Reza AW, Siddique N, Wang H. Vision-audio multimodal object recognition using hybrid and tensor fusion techniques. Information Fusion. 2026 Feb 1;126(1):103667. doi:10.1016/j.inffus.2025.103667

[33] Rahman Swapno SMM, Sakib A, Uddin Khondakar Pranta AS, Hossain A, Debnath J, Al Noman A, et al. Explainable transformer framework for fast cotton leaf diagnostics and fabric defect detection. iScience. 2026 Feb 20;29(2):114411. doi:10.1016/j.isci.2025.114411

[34] Islam S, Haque R, Khan MA, Mohiuddin A Bin, Hossain Siddiqui MI, Limon ZH, et al. Ensemble Transformer with Post-hoc Explanations for Depression Emotion and Severity Detection. iScience. 2026 Feb 20;29(2):114605. doi:10.1016/j.isci.2025.114605

[35] Haque R, Sultana S, Rafy A, Babul Islam M, Arafat MA, Bhattacharya P, et al. A Transfer Learning-Based Computer-Aided Lung Cancer Detection System in Smart Healthcare. IET Conference Proceedings. 2024;2024(37):594–601. doi:10.1049/icp.2025.0858

[36] Khan S, Rahman H, Hossain Siddiqui MI, Hossain Limon Z, Khan MA, Haque R, et al. Ensemble-Based Explainable Approach for Rare Medicinal Plant Recognition and Conservation. 2025 10th International Conference on Information and Network Technologies, ICINT 2025. 2025;88–93. doi:10.1109/ICINT65528.2025.11030872

[37] Haque R, Mamun MA Al, Ratul MH, Aziz A, Mittra T. A Machine Learning Based Approach to Analyze Food Reviews from Bengali Text. 12th International Conference on Electrical and Computer Engineering, ICECE 2022. Published online 2022:80-83. doi:10.1109/ICECE57408.2022.10088971

[38] Nobel SMN, Swapno SMMR, Islam MR, Safran M, Alfarhood S, Mridha MF. A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. Scientific Reports 2024 14:1. 2024;14(1):1-25. doi:10.1038/s41598-024-64987-5

[39] Khushubu KG, Masum A Al, Rahman MH, et al. TransUNetB: An advanced Transformer–UNet framework for efficient and explainable brain tumor segmentation. Inform Med Unlocked. 2025;59(10):101706. doi:10.1016/j.imu.2025.101706

[40] Al Masum A, Limon ZH, Islam MA, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. Published online 2024:153-158. doi:10.1109/BECITHCON64160.2024.10962580

[41] Rahman S, Parameshachari BD, Haque R, Masfequier Rahman Swapno SM, Babul Islam M, Nobel SN. Deep Learning-Based Left Ventricular Ejection Fraction Estimation from Echocardiographic Videos. 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT 2023. Published online 2023. doi:10.1109/EASCT59475.2023.10392607