International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

LEVit-Skin: A balanced and interpretable transformer-CNN model for multi-class skin cancer diagnosis

Anamul Haque Sakib [1], Md Ismail Hossain Siddiqui [2], Sanjida Akter [3], Abdullah Al Sakib [4,*] and Mohammad Rasel Mahmud [5]

[1] *Department of Business Administration, International American University, Los Angeles, CA 90010, USA.*
[2] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*
[3] *Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh.*
[4] *Department of Information Technology, Westcliff University, Irvine, CA 92614, USA.*
[5] *Department of Management Information System, International American University, CA 90010, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1860-1873

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1166

Abstract

Skin cancer is a major cause of death, making early detection essential. This study presents LEVit, an explainable and class-balanced deep learning framework designed for multiclass skin lesion classification. LEVit combines a hybrid Vision Transformer (ViT) with a Convolutional Neural Network (CNN). We evaluated LEVit on two benchmark dermoscopic datasets: HAM10000, which consists of 10,015 images across 7 classes, and ISIC 2019, with 25,331 images spanning 8 classes. Both datasets have notable class imbalances. To address this issue, we applied advanced augmentation techniques to oversample minority classes, ensuring a uniform class distribution and enhancing the model's ability to generalize. LEVit effectively captures local lesion textures and global spatial relationships through its integrated self-attention and convolutional modules. We compared its performance against four state-of-the-art models: NASNet, SqueezeNet, SE-Net, and Xception, across four metrics: F1 Score, Specificity, Matthews Correlation Coefficient (MCC), and Precision-Recall Area Under the Curve (PR AUC). LEVit achieved outstanding results, with a F1 Score of 98.11% and a PR AUC of 98.57% on the ISIC 2019 dataset, and a F1 Score of 96.11% and a PR AUC of 96.62% on HAM10000. For interpretability, we utilized Grad-CAM to generate class-specific heatmaps, which highlight the key areas of lesions that influence the model's predictions. This work demonstrates that balanced training and a hybrid architecture can enhance both classification accuracy and interpretability in skin cancer diagnostics, effectively addressing the limitations of existing models and paving the way for reliable clinical applications.

Keywords: Skin cancer; Vision transformer; Deep learning; Explainable AI (XAI); Medical imaging.

1. Introduction

Skin cancer is one of the most common types of cancer worldwide, with more than 3 million non-melanoma cases and over 132,000 melanoma cases diagnosed each year, according to the World Health Organization [1]. In the United States, one in five people is expected to develop skin cancer by the age of 70, and more than two people die from skin cancer every hour [2], [3]. Melanoma, though less common, is the deadliest form of skin cancer due to its aggressive nature and high potential for metastasis. Early diagnosis is vital, as the five-year survival rate for localized melanoma is 90% [4]. However, this rate drops significantly to 27% once the cancer has spread to distant organs. However, early detection remains challenging due to subtle differences in the appearance of skin lesions and a heavy reliance on the expertise of dermatologists. Dermoscopic imaging has greatly improved dermatological diagnostics by providing magnified, high-

* Corresponding author: Abdullah Al Sakib

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

resolution views of skin lesions. However, the manual interpretation of these images is labor-intensive, highly subjective, and prone to variability between different observers [5], especially in low-resource clinical settings.

The development of automated skin lesion classification systems can significantly assist dermatologists by offering consistent, fast, and accurate second opinions. Among the various methods, CNNs has shown tremendous success in medical image analysis [6]. CNN-based architectures such as Xception, NASNet, and SE-Net have been employed for various image classification tasks and have demonstrated promising results [7]. Nonetheless, these models primarily extract local features, which limits their capacity to capture long-range dependencies or holistic lesion context. In contrast, ViTs offer superior global modeling ability but often require large-scale datasets and are computationally expensive, making them less practical in medical scenarios with limited labeled data.

One of the most significant challenges in dermoscopic image classification is class imbalance. Datasets like HAM10000 and ISIC 2019 contain a large number of nevus (NV) images, while rare classes such as vascular lesions (VASC) or dermatofibromas (DF) are underrepresented [8]. This imbalance biases models toward majority classes, reducing diagnostic sensitivity for clinically important minority classes. Additionally, most existing works provide limited model interpretability, which is critical in clinical settings where decisions must be explainable and verifiable [9].

To address these challenges, we propose a hybrid transformer-convolutional architecture based on the LEVit model. LEVit combines the convolutional inductive bias of CNNs with the global attention capabilities of ViTs. It enables the model to extract fine-grained local patterns and global contextual dependencies, making it highly suitable for skin lesion classification. Furthermore, we apply extensive data augmentation techniques to balance the dataset distributions and enhance generalization. We also employ Grad-CAM for explainability, offering pixel-level heatmaps that highlight the lesion regions most influential to the model's prediction, which supports clinical validation and trust. The key contributions are as follows:

- Addressed the class imbalance problem in dermoscopic datasets using advanced augmentation techniques, leading to improved model generalization across underrepresented classes.
- Proposed a novel skin cancer classification framework using the LEVit model that fuses convolutional and attention-based feature extraction to enhance lesion representation.
- Performed comparative and statistical performance analysis across multiple state-of-the-art models, demonstrating that the LEVit architecture achieves the highest scores on both HAM10000 and ISIC 2019 datasets.
- Integrated Grad-CAM-based explainability to visualize decision-making regions within lesions, enhancing clinical interpretability and trust in the model's outputs.

The rest of the paper is organized as follows: Section 2 provides an overview of related work and the limitations of existing skin lesion classification approaches. Section 3 describes the dataset characteristics, preprocessing techniques, and the proposed LEVit-based methodology. Section 4 presents experimental results, performance benchmarks, and statistical validation. Section 5 discusses the findings, implications, and outlines technical limitations and future directions. Finally, Section 6 concludes the study and highlights the practical relevance of our approach in dermatological diagnostics.

2. Related Works

In the past few years, transfer learning methods have become essential in numerous practical applications, covering areas like healthcare [10], [11], education [12], and industrial automation [13], [14]. Skin disease classification has been widely addressed using data-driven techniques that leverage both CNN and transformer architectures to capture local and global features of dermoscopic images. Rezaee et al. [15] introduced a hybrid model combining CNN branches with a transformer module via bi-directional feature fusion, achieving a 96.85% accuracy on the ISIC-2019 dataset, though without an explicit explainability mechanism. Similarly, Ahmad et al. [16] integrated DeepLabv3+ for segmentation with a ViT for classification, demonstrating impressive accuracies across multiple datasets; however, their work did not focus on extensive augmentation to address dataset imbalances. Rezaee and Zadeh [17] further advanced the field by proposing a bi-branch parallel framework that fused self-attention units and an optimized CNN backbone, which improved feature extraction yet lacked integrated XAI tools.

Ensemble approaches such as that by Rahman et al. [7] have also been explored, where a weighted combination of multiple CNNs boosted recall scores significantly, though these methods often treat the classification process as a "black-box." More recently, Yang et al. [18] proposed a novel ViT model that applied class rebalancing and patch-based tokenization, achieving competitive performance with a 94.1% accuracy on HAM10000. Cai et al. [19] took a multimodal approach by fusing image features with clinical metadata through a mutual attention block, thereby enhancing the overall diagnostic accuracy.

Despite these advancements, a common shortcoming persists while many models achieve high accuracy, few integrate comprehensive data augmentation strategies to balance imbalanced datasets or incorporate explainability techniques. In our work, we address this gap by employing a ViT architecture augmented with advanced data augmentation methods to ensure dataset balance. Additionally, we embedded the Grad-CAM explainability tool into the pipeline, providing visual insights into the network's decision process. This integration not only enhances the transparency and interpretability of the classification outcomes but also builds clinician trust in automated diagnostic systems.

3. Materials and Methods

Figure 1 illustrates the complete workflow of our study. It begins with dermoscopic images from the HAM10000 and ISIC 2019 datasets, followed by image preprocessing (resizing and normalization), dataset balancing via augmentation and oversampling, and training using multiple deep learning models including NASNet, SqueezeNet, SE-Net, Xception, and the proposed LEVit. The performance is evaluated using metrics such as F1 Score, Specificity, MCC, PR AUC, confusion matrix, and learning curves, culminating in a state-of-the-art comparative analysis. We also integrated with Grad-CAM for improve diagnostic accuracy and model interpretability.

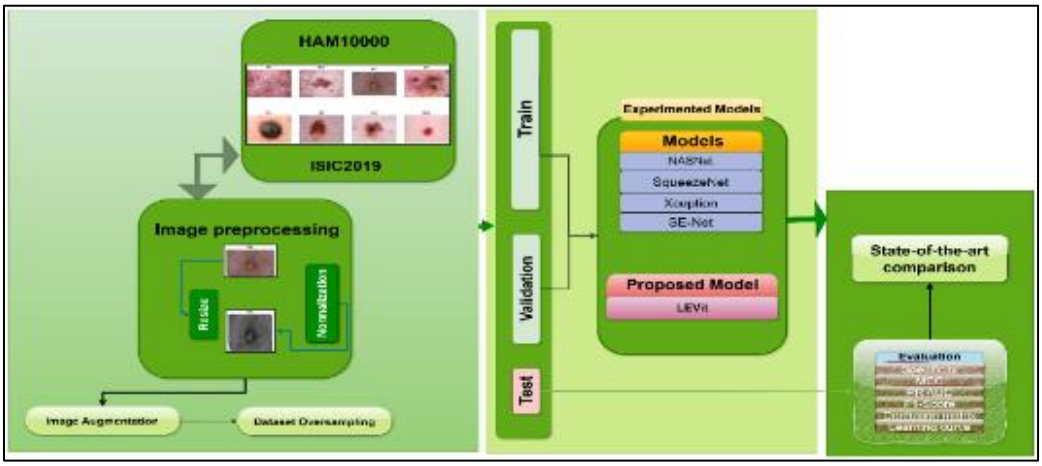

Figure 1 Overview of the proposed LEVit-based skin cancer classification framework

3.1. Data Description

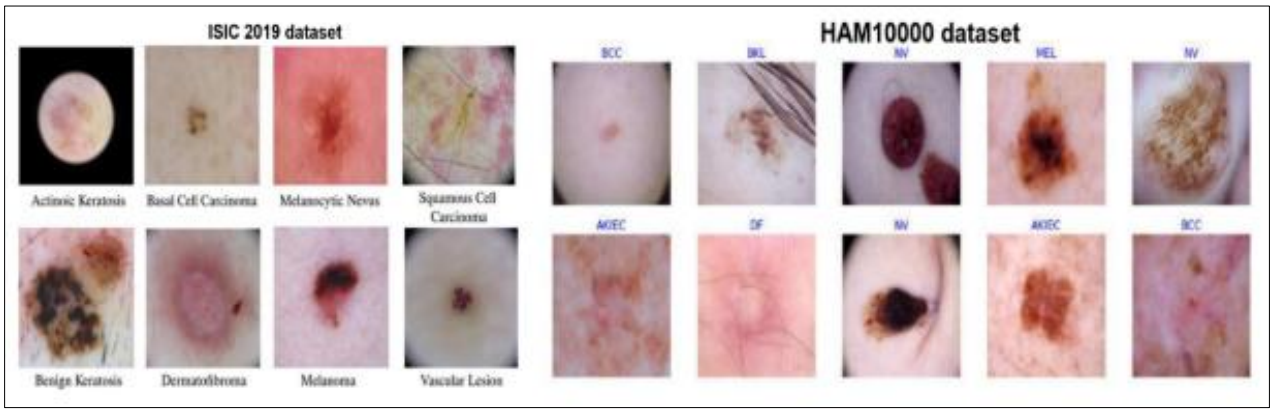

Figure 2 Samples from HAM10000 and ISIC 2019 dataset and each class

This study utilizes two widely recognized image datasets: HAM10000[20] and ISIC 2019[21]. Sample images from each of the both datasets are shown in Figure 2. The HAM10000 dataset contains 10,015 dermoscopic images covering seven classes of skin lesions AKIEC, BCC, BKL, DF, MEL, NV, and VASC providing a diverse representation of pigmented skin conditions. In contrast, the ISIC 2019 dataset comprises 25,331 images distributed across eight classes, including MEL, MN, BCC, AK, BK, DF, VASC, and SCC. Both datasets present significant challenges due to inherent class imbalances and high intra-class variability, underscoring the need for robust preprocessing and augmentation strategies to achieve reliable classification performance. The ISIC 2019 dataset complements HAM10000 by providing a larger, more varied

set of images, contributing to a comprehensive evaluation of our classification approach. Table 1 shows the class distribution for each dataset, split into 80% for training, 5% for validation, and 15% for testing.

Table 1 Class distribution after dataset splitting for experimental datasets

HAM10000 Distribution				
Class	Total	Train (80)	Validation (5)	Test (15)
AKIEC	327	261	16	50
BCC	514	411	25	78
BKL	1099	879	54	166
DF	115	92	5	18
MEL	1113	890	55	168
NV	6705	5364	335	1006
VASC	142	113	7	22
ISIC 2019 Distribution				
Class	Total	Train (80)	Validation (5)	Test (15)
MEL	4522	3617	226	679
NV	12875	10300	643	1932
BCC	3323	2658	166	499
AK	867	693	43	131
BK	2624	2099	131	394
DF	239	191	11	37
VASC	253	202	12	39
SCC	628	502	31	95

3.2. Data Preprocessing and Augmentation

Each dermoscopic image was resized to 224 × 224 pixels using bilinear interpolation to standardize input dimensions. After resizing, pixel intensities were normalized using Equation 1, where γ_k and β_k are learnable parameters, and ϵ is a small constant to prevent division by zero. This normalization ensures each channel has zero mean and unit variance, which accelerates and stabilizes ViT training.

$$y_{ijk} = \gamma_k \cdot \frac{x_{ijk} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + \beta_k \quad \text{for } i = 1, \dots, H;\ k = 1, \dots, W;\ k = 1, \dots, C \tag{1}$$

To further enhance generalization and reduce overfitting, various data augmentation techniques were applied (Table 2). These augmentations introduced variability by simulating different orientations, lighting conditions, and geometric distortions, enriching the training set and helping the model learn robust, discriminative features.

Table 2 Data augmentation strategies for enhanced model performance in skin cancer detection

Augmentation Technique	Description	Parameter Settings
Rotation	Simulates varied orientations	±30° range; multiple rotations
Horizontal/Vertical Flip	Mimics mirror images	Flip probability = 0.5
Random Crop & Translation	Captures different regions and scales	Crop ratio: 80–100%; shift ≤ 10%
Brightness & Contrast Adjust.	Simulates different lighting and exposure conditions	Factor range: 0.8–1.2
Shear & Zoom	Alters lesion geometry and scale	Shear: 0.0–0.2; Zoom: 0.8–1.2
Sigmoid Intensity Correction	Enhances features using non-linear pixel intensity adjustment	Empirically tuned sigmoid parameter

3.3. Data Balancing and Oversampling

To address the issue of class imbalance in the HAM10000 and ISIC 2019 datasets, we employed an oversampling technique. Minority classes, such as DF and VASC, were underrepresented compared to the dominant classes like NV and MN, which posed a risk of biased learning. We synthetically increased the sample count of the minority classes by augmenting their images using the methods described in Table 3. This process raised the sample sizes to 6,705 for HAM10000 and 12,875 for ISIC 2019. This balancing strategy ensured uniform representation across all class types, reducing model bias and enhancing feature learning for the less frequent classes [22]. Once the datasets were balanced, we divided them into training (80%), validation (5%), and testing (15%) subsets. When combined with the ViT and Grad-CAM explainability, this pipeline significantly contributed to the reliability and high performance of the proposed classification framework.

3.4. Experimental Models

In this study, we evaluated a set of state-of-the-art deep learning models alongside our proposed model, LEVit, to benchmark and enhance skin cancer classification.

3.4.1. NASNet

It is a CNN architecture discovered through neural architecture search. Its modular design—with repetitive normal and reduction cells—allows NASNet to learn robust feature hierarchies with minimal human intervention. Its strong performance on large-scale image recognition tasks [23], [24] makes it a valuable baseline for skin lesion analysis, where subtle texture variations must be accurately captured.

Table 3 Balanced data distribution for both datasets via oversampling techniques

	HAM10000 dataset balancing				
Class	Actual	After Augmentation	Training	Validation	Testing
AKIEC	327	6,705	5,364	335	1,006
BCC	514	6,705	5,364	335	1,006
BKL	1,099	6,705	5,364	335	1,006
DF	115	6,705	5,364	335	1,006
MEL	1,113	6,705	5,364	335	1,006
NV	6,705	6,705	5,364	335	1,006
VASC	142	6,705	5,364	335	1,006
	ISIC 2019 dataset balancing				
MEL	4,522	12,875	10,300	644	1,931

MN	12,875	12,875	10,300	644	1,931
AK	867	12,875	10,300	644	1,931
BCC	3,323	12,875	10,300	644	1,931
DF	239	12,875	10,300	644	1,931
VASC	253	12,875	10,300	644	1,931
SCC	628	12,875	10,300	644	1,931
BK	2,624	12,875	10,300	644	1,931

3.4.2. SqueezeNet

SqueezeNet is celebrated for its lightweight architecture, leveraging "fire modules" that combine squeeze layers (using 1×1convolutions) with expand layers (using both 1×1 and 3×3 convolutions). This design dramatically reduces the model's parameter count while maintaining competitive accuracy [25]. SqueezeNet is particularly useful in clinical contexts where computational resources may be limited, and rapid inference is needed.

3.4.3. Xception

Xception extends the idea of depthwise separable convolutions, decoupling spatial feature extraction from channel-wise processing. This results in efficient architecture that minimizes redundancy while capturing fine-grained, discriminative features [26]. Xception's ability to delineate subtle differences in lesion morphology and texture is critical for distinguishing between similar skin cancer types.

3.4.4. SE-Net

Squeeze-and-Excitation Networks introduces a channel attention mechanism that selectively emphasizes important features. By performing a squeeze operation through global average pooling followed by an excitation step, SE-Net re-calibrates the importance of each feature map. This targeted re-weighting is particularly beneficial in medical imaging, where certain features—such as color and border irregularities—are more indicative of malignancy than others [27].

3.4.5. Proposed LEVit

Our proposed LEVit model integrates the strengths of ViTs with enhanced local feature extraction to address the challenges of skin lesion variability. LEVit leverages self-attention mechanisms as shown in Figure 3, it captures both global contextual relationships and fine-grained details, which is crucial for accurate skin cancer classification. The model's architecture is defined by a series of complex operations. The input image $X \in R^{H\times W\times C}$is divided into N non-overlapping patches. Each patch x_i is flattened and projected into an embedding shown in Equation 2. In Equation 3, the output of the multi-head attention is passed through a position-wise feed-forward network which allows the model to learn complex non-linear transformations of the features.

To ensure stable gradient flow and deeper network training, residual connections and layer normalization are applied throughout (shown in Equations 4-5) allowing the model to efficiently combine initial features with their refined representations. LEVit was chosen as our proposed model because its transformer-based self-attention mechanism addresses the high intra-class variability present in medical images [28]. By explicitly modeling long-range dependencies and capturing both global context and local details, LEVit provides superior discriminative power crucial for challenging skin cancer classification tasks. Additionally, the integration of residual learning and normalization schemes further stabilizes training, ensuring that the model generalizes well across diverse datasets [29].

$$zi = \text{Linear}\big(\text{Flatten}(x_i)\big) + p_i, \quad i = 1, \dots, N, \tag{2}$$

$$\text{Attention}_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \tag{3}$$

$$\text{FFN}(z') = \text{GELU}(z' W_1 + b_1) W_2 + b_2, \tag{4}$$

$$z'' = \text{LayerNorm}\big(z + \text{FFN}(z')\big), \tag{5}$$

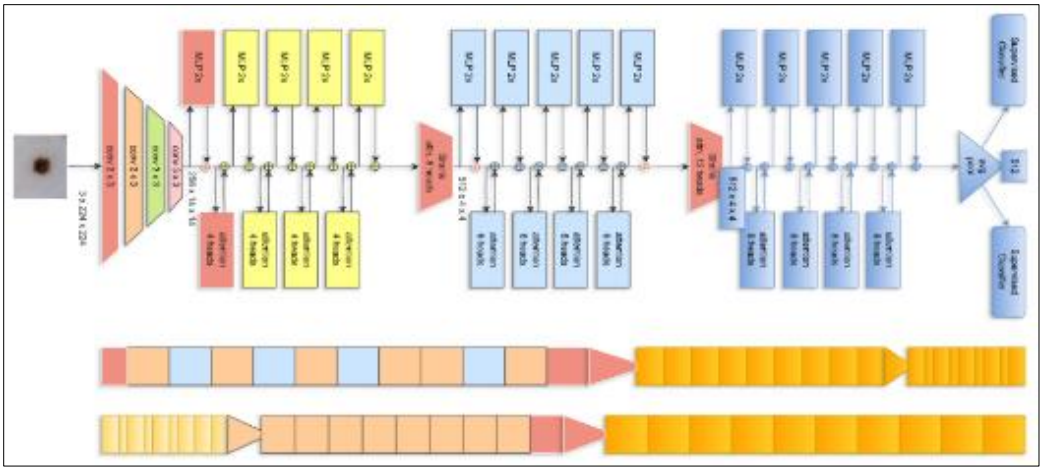

Figure 3 Architecture of the proposed LEViT model

3.6. Evaluation Metrics

To thoroughly evaluate the performance of the proposed and baseline models for skin cancer classification, we used four widely accepted evaluation metrics: F1 Score, Specificity, MCC, and PR AUC. The F1 Score is the harmonic mean of precision and recall, and it balances false positives and false negatives. This balance is crucial in medical imaging, where both types of errors can have serious consequences. Specificity measures the proportion of true negatives correctly identified, which helps to minimize false positives and unnecessary clinical interventions. The MCC provides a balanced evaluation by considering all components of the confusion matrix, making it a more dependable metric than accuracy in scenarios with class imbalances. Lastly, the PR AUC illustrates the trade-off between precision and recall across various thresholds, offering a more informative assessment than ROC AUC for imbalanced datasets. High values across these metrics indicate a model's robustness and its potential for clinical application. To ensure unbiased evaluation, stratified K-fold cross-validation was employed, with K set to 10. This approach maintains the class distribution across all folds, ensuring each fold is a good representative of the overall dataset.

3.7. Training Parameters

Table 4 outlines the key configurations used to train the LEVit model for skin cancer classification. Images were resized to 224×224 and processed in batches of 32 over 30 epochs. An AdamW optimizer with a learning rate of 1e-4 and weight decay of 0.01 was employed, along with a cosine annealing scheduler for gradual learning rate reduction. The model used a patch size of 16×16 for tokenization, 12 transformer layers, and 8 attention heads. Dropout was set to 0.1 to prevent overfitting, and categorical cross-entropy was used as the loss function for multi-class classification. These settings ensured training stability, generalization, and effective feature representation across diverse lesion types.

4. Results and Discussion

4.1. Performance Comparison

The results summarized in the Table 5, provide a comprehensive performance evaluation of our experimental deep learning models across the HAM1000 and ISIC 2019 datasets, both before and after the application of augmentation strategies. The "Before Augmentation" results establish the baseline performance of the models when trained on the original, unaugmented data, whereas the "After Augmentation" results illustrate the impact of advanced augmentation techniques designed to enrich the training set by introducing additional variability.

Across both datasets, the observed improvements in all performance metrics following augmentation are significant, particularly in PR AUC, which is critical in assessing the precision-recall balance of diagnostic models. The proposed LEVit model, leveraging a hybrid transformer-based architecture, consistently achieves higher or comparable performance relative to the other models; for instance, its F1 score improves markedly from 92.10 ± 0.9 to 96.11 ± 0.3 on the HAM1000 dataset and reaches 98.11 ± 0.5 on the ISIC 2019 dataset after augmentation. This trend underscores the model's ability to capture both global contextual features and subtle local details in dermoscopic images. Furthermore, the consistency of performance across both datasets—as evidenced by similar trends in the improvement of metrics—and the low standard deviations in post-augmentation scenarios highlight the enhanced reliability of the models under augmented training conditions. Overall, the data demonstrates that image augmentation is pivotal for

addressing class imbalance and improving model generalizability in skin lesion classification, with the proposed LEVit model emerging as a robust candidate for clinical application.

Table 4 Hyperparameter settings used for training the LEVit Model on both datasets

Hyperparameter	**Value**
Input Image Size	224 × 224
Optimizer	AdamW
Learning Rate	1.00E-04
Batch Size	32
Epochs	30
Weight Decay	0.01
Learning Rate Scheduler	Cosine Annealing
Loss Function	Categorical Cross-Entropy
Patch Size	16 × 16
Dropout Rate	0.1
Number of Heads	8
Transformer Layers	12

Table 4 Performance comparison for experimental classifiers on both dataset before and after augmentation

Ham10000 Dataset Results Before Augmentation				
Model	F1	Specificity	MCC	PR AUC
NASNet	93.21 ± 0.4	92.45 ± 0.9	88.69 ± 1.7	90.12 ± 0.9
SqueezeNet	91.88 ± 0.5	90.77 ± 1.9	86.55 ± 1.8	88.90 ± 1.1
SE-Net	89.75 ± 1.6	88.12 ± 0.8	83.72 ± 1.9	86.43 ± 1.8
Xception	92.40 ± 1.4	91.29 ± 1.7	87.44 ± 1.4	89.15 ± 1.0
LEVit	92.10 ± 0.9	92.26 ± 0.4	89.21 ± 0.9	91.64 ± 0.8
Ham10000 Dataset Results After Augmentation				
NASNet	97.01 ± 1.3	96.25 ± 1.1	94.13 ± 1.1	96.03 ± 1.0
SqueezeNet	96.28 ± 0.9	95.40 ± 0.5	93.26 ± 0.7	95.12 ± 0.7
Xception	95.32 ± 1.4	93.18 ± 1.5	91.08 ± 1.5	94.85 ± 1.1
SE-Net	95.75 ± 0.4	94.86 ± 0.6	92.58 ± 0.8	94.02 ± 0.8
LEVit	96.11 ± 0.3	96.29 ± 0.8	95.51 ± 0.9	96.62 ± 0.6
ISIC 2019 Dataset Results Before Augmentation				
NASNet	93.21 ± 0.6	93.45 ± 0.7	90.69 ± 1.4	91.12 ± 0.7
SqueezeNet	89.88 ± 1.5	90.53 ± 1.5	87.55 ± 1.9	88.90 ± 1.3
SE-Net	88.75 ± 1.2	87.12 ± 0.9	84.72 ± 1.8	85.43 ± 1.5
Xception	95.63 ± 1.1	96.01 ± 0.9	94.44 ± 1.1	95.15 ± 0.9
LEVit	96.02 ± 1.0	96.33 ± 0.6	92.21 ± 0.8	97.64 ± 0.6

ISIC 2019 Dataset Results After Augmentation				
NASNet	94.93 ± 1.1	95.08 ± 1.0	93.09 ± 1.4	95.12 ± 1.0
SqueezeNet	91.97 ± 0.8	92.40 ± 0.6	91.26 ± 1.2	92.40 ± 0.9
Xception	97.32 ± 0.4	96.95 ± 0.9	92.08 ± 1.1	97.09 ± 0.8
SE-Net	96.83 ± 0.7	96.65 ± 0.9	94.58 ± 0.8	96.81 ± 0.7
LEVit	98.11 ± 0.5	98.29 ± 0.7	97.19 ± 0.6	98.57 ± 0.2

4.2. Performance Validation

Figure 4 displays the confusion matrices for the HAM10000 and ISIC 2019 datasets, highlighting the class-wise performance of the LEVit model. In both cases, the matrices exhibit strong diagonal dominance, confirming the model's capacity to correctly distinguish between multiple lesion types. For HAM10000, NV, BCC, and MEL classes show minimal off-diagonal errors, while minor misclassifications occur primarily between visually similar classes such as BKL vs. AKIEC or MEL. For ISIC 2019, the model maintains high fidelity across major classes such as BCC, MN, and BK, with limited confusion observed between AK and SCC—two categories with overlapping dermoscopic patterns. The sparse off-diagonal entries reinforce the model's low error dispersion, aligning with the high F1 and MCC scores reported. These matrices validate the model's robustness in handling intra-class variability and inter-class ambiguity, particularly under class-balanced and augmented training conditions.

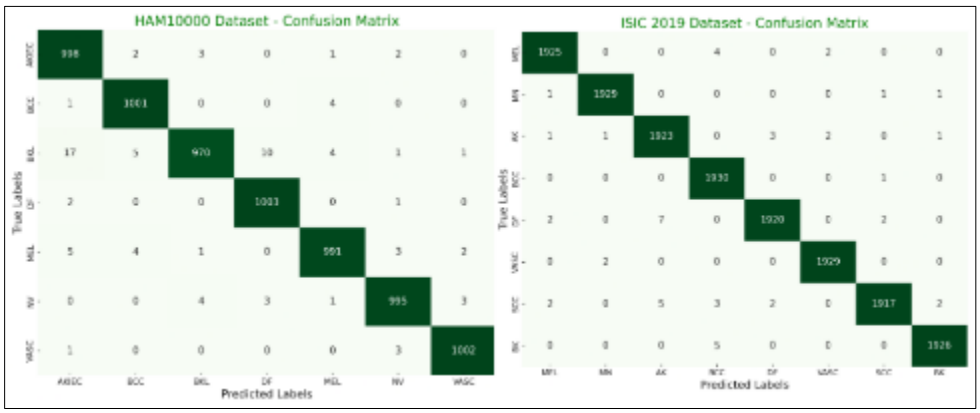

Figure 4 Confusion matrices of the LEVit model on both datasets

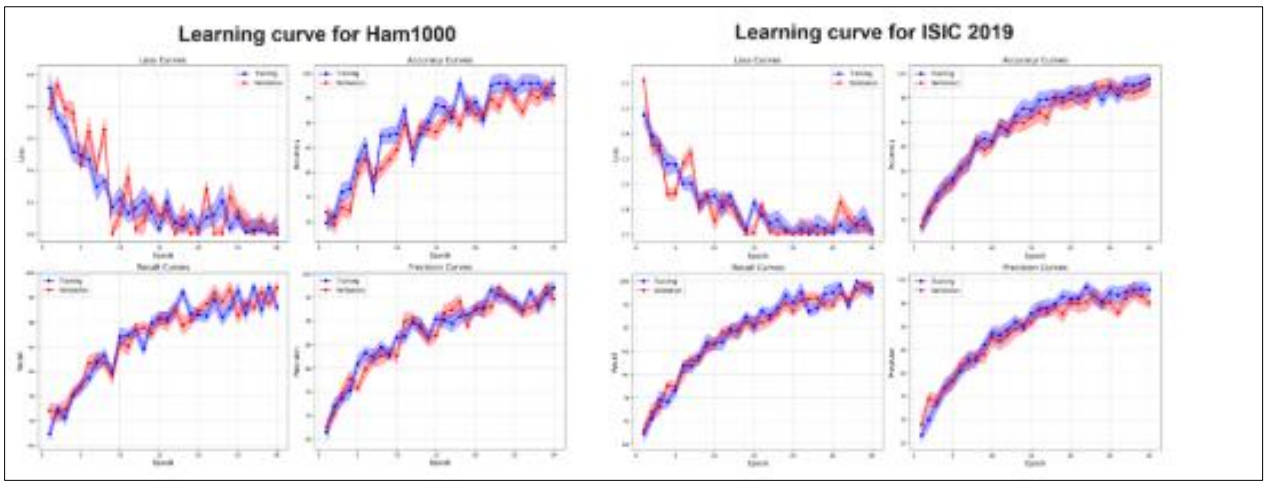

Figure 5 Learning curves of the LEVit model trained on both datasets

For the HAM10000 dataset, the loss curves show a sharp decline during the initial epochs followed by stable minimization, while the corresponding accuracy, recall, and precision curves exhibit a consistent upward trajectory. Although minor oscillations are observed, the convergence patterns between training and validation curves remain

closely aligned, indicating effective learning without significant overfitting. The model maintains stable generalization throughout training, as evidenced by the narrow gap between training and validation scores across all metrics. On the ISIC 2019 dataset, the LEVit model demonstrates even smoother and faster convergence. The loss decreases sharply within the first few epochs and plateaus with minimal fluctuation, while accuracy, recall, and precision curves exhibit strong monotonic improvement. The marginal difference between training and validation curves further confirms that the model is capable of capturing the underlying data distribution effectively, benefiting from the dataset's size and class diversity. Figure 5 illustrates the learning dynamics of the LEVit model over 30 training epochs on the HAM10000 (left) and ISIC 2019 (right) datasets, evaluated using loss, accuracy, recall, and precision metrics for both training and validation sets.

4.3. Model's Transparency

Figure 6 shows Grad-CAM visualizations that illustrate the decision-making process of the skin lesion classification model for both datasets. Panel (a) features seven classes from the HAM10000 dataset: NV (Melanocytic Nevus), AKIEC (Actinic Keratoses), VASC (Vascular Lesion), BCC (Basal Cell Carcinoma), BKL (Benign Keratosis), DF (Dermatofibroma), and MEL (Melanoma). Each lesion image is paired with a Grad-CAM heatmap, highlighting key regions that influenced the model's predictions. The model tends to focus on the center of lesions or their texture irregularities, which are important for accurate diagnosis. Panel (b) covers eight classes from the ISIC 2019 dataset, including NV, AK, VASC, BCC, BK, DF, MEL, and SCC (Squamous Cell Carcinoma). The Grad-CAM maps indicate where the network pays attention for each lesion image. These heatmaps are generally centered on the lesions, showing that the model recognizes critical features like shape irregularities and pigmentation differences. For complex or malignant lesions such as MEL and SCC, the attention maps cover larger areas, reflecting the challenging visual characteristics of these cases.

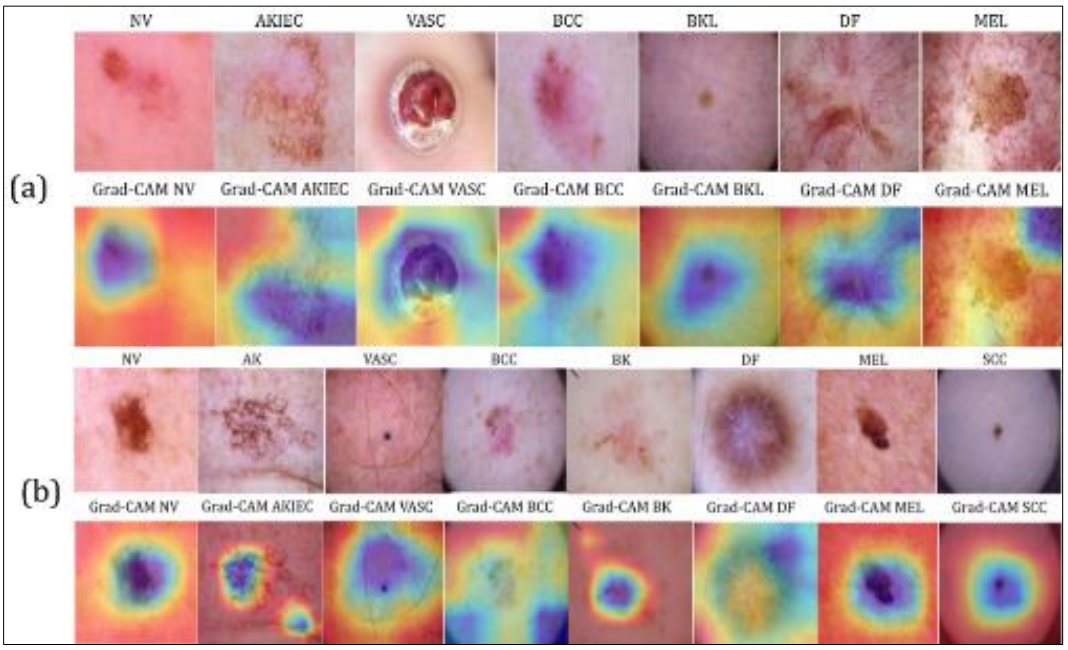

Figure 6 Grad-CAM visualization for both (a) HAM10000 and (b) ISIC 2019 datasets

4.4. State-of-the-Art Comparison

Table 6 presents a comparative analysis of recent state-of-the-art approaches for skin cancer classification using dermoscopic image datasets. Most prior studies achieved high performance using a variety of CNN and transformer-based models; however, several limitations persist, particularly regarding dataset imbalance and generalizability. For instance, Rezaee[15] et al. utilized ISIC-2019 without balancing and reported an accuracy of 96.85%, while Rezaee & Zadeh achieved up to 97.48% across ISIC-2019 and PH2 using a bi-branch transformer-CNN architecture. Ahmad et al. evaluated their model across multiple datasets, including ISIC-16 to ISIC-20 and HAM10000, but without balancing strategies, resulting in variable performance, with accuracy dropping to 93.47%.

Table 5 Comparative analysis of state-of-the-art methods for skin cancer classification

Reference	Dataset Name	Dataset Balancing	Result (%)
Rezaee et al.[15]	ISIC-2019	No	96.85
Ahmad et al.[16]	ISIC-19, HAM10000	No	93.47
Zadeh et al. [17]	ISIC-2019, PH2	No	96.87–97.48
Damaševičius[30]	HAM10000, ISIC-2018, ISIC-2019, PH2	No	93.47–98.98
Rahman et al.[7]	HAM10000, ISIC-2019	Yes	94
Xin et al.[31]	HAM10000, Clinical dermoscopy dataset	No	94.1–94.3
Yang et al.[18]	HAM10000, Edinburgh DERMOFIT	Yes	94.1
Ours (LEViT)	HAM10000 and ISIC-2019	Yes	96.62, 98.57

Maqsood and Damaševičius[30] employed a deep feature fusion and selection framework across four datasets and achieved up to 98.98%, though their performance on ISIC-2019 was only 93.47%, again without addressing class imbalance. Rahman et al.[7] introduced a weighted ensemble method with explicit dataset balancing and achieved 94% accuracy. Similarly, Yang et al. applied balancing techniques and obtained 94.1% on HAM10000. Xin et al.[31] proposed a transformer model without dataset balancing, yielding results between 94.1% and 94.3%.

In contrast, our proposed approach applied comprehensive dataset balancing via advanced augmentation strategies to address class imbalance in both HAM10000 and ISIC-2019. This led to substantial improvements, achieving 96.62% accuracy on HAM10000 and 98.57% on ISIC-2019. These results outperform most existing models under similar conditions, demonstrating LEVit's superior generalization and feature extraction capabilities. Furthermore, the integration of XAI through Grad-CAM and robust evaluation across diverse metrics underscores the clinical reliability of our method. The results confirm that careful dataset handling, in combination with hybrid attention architecture, is critical for advancing automated skin cancer diagnosis.

5. Discussion

Our proposed model consistently outperformed competing architectures due to its hybrid design, which integrates convolutional priors with transformer-based global self-attention. This architectural synergy allows LEVit to simultaneously model local texture patterns and long-range dependencies, both of which are critical in dermoscopic image analysis where lesions exhibit high intra-class variability and inter-class visual similarity. In contrast to traditional CNNs that are limited in global contextual modeling, and pure transformers that require large-scale datasets for stable convergence, LEVit maintains a favorable trade-off between expressiveness and data efficiency. These capabilities translated into superior F1 scores, MCC, and PR AUC values across both the HAM10000 and ISIC 2019 datasets, particularly following dataset balancing via augmentation. Beyond raw performance, the broader implications of this work are significant. From an economic standpoint, automated lesion classification using LEVit could reduce dependency on specialist consultations, lower diagnostic costs, and enable scalable screening in high-volume or resource-limited settings. Socially, such deployment can facilitate early cancer detection in underserved populations, thereby reducing diagnostic latency and improving prognostic outcomes.

However, the approach is not without limitations. LEVit's transformer blocks incur notable computational overhead, which may hinder real-time deployment on mobile or edge devices unless quantization or pruning is introduced. Furthermore, while we employed Grad-CAM for visual interpretability, its current resolution may not always provide clinically actionable insight into specific lesion subregions or boundaries. Future work could benefit from integrating higher-resolution interpretability frameworks such as Score-CAM++ or Layer-CAM, which offer finer localization. To further optimize the model, we suggest exploring hierarchical multi-scale attention mechanisms to better capture structural lesion features at different granularities. Additionally, incorporating patch-wise token pruning or lightweight self-distillation could reduce inference time while maintaining accuracy. Finally, extending LEVit into a multimodal framework that jointly processes dermoscopic images and structured metadata (e.g., age, lesion site, risk factors) could offer improved diagnostic performance and deeper clinical relevance.

6. Conclusion

This study proposed a LEVit-based deep learning framework for automated skin cancer classification, evaluated on two benchmark datasets: HAM10000 and ISIC 2019. The model demonstrated superior performance across multiple evaluation metrics outperforming well-established architectures such as NASNet, Squeeze Net, SE-Net, and Xception. These results validate the effectiveness of combining local convolutional priors with global attention mechanisms for robust visual pattern recognition in complex medical imaging tasks. The implications of these findings extend beyond academic significance. In practical terms, the proposed system could enhance clinical diagnostic workflows by enabling faster, more consistent lesion assessment. Its ability to generalize across datasets suggests strong potential for integration into digital dermoscopy tools, improving diagnostic reach in both centralized hospitals and decentralized teledermatology platforms. Nevertheless, challenges remain. The model's reliance on dermoscopic inputs alone limits its contextual understanding, especially in borderline or rare cases. In addition, while augmentation addressed class imbalance, real-world data heterogeneity (e.g., varying skin tones, lighting conditions) still poses generalization challenges. Future iterations could integrate domain adaptation techniques to enhance cross-domain robustness or fuse multimodal clinical metadata to contextualize image-based predictions. When LEVit properly tuned and supported by data augmentation and interpretability tools like Grad-CAM, can serve as a high-performing, explainable model for multi-class skin cancer classification. Its hybrid attention-convolution architecture offers a viable path forward for designing intelligent diagnostic systems that are not only accurate but also transparent and scalable for real-world deployment.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] S. Hao et al., "ConvNeXt-ST-AFF: A Novel Skin Disease Classification Model Based on Fusion of ConvNeXt and Swin Transformer," IEEE Access, vol. 11, pp. 117460–117473, 2023, doi: 10.1109/ACCESS.2023.3324042.

[2] I. Kousis, I. Perikos, I. Hatzilygeroudis, and M. Virvou, "Deep Learning Methods for Accurate Skin Cancer Recognition and Mobile Application," Electronics 2022, Vol. 11, Page 1294, vol. 11, no. 9, p. 1294, Apr. 2022, doi: 10.3390/ELECTRONICS11091294.

[3] L. Nanz, U. Keim, A. Katalinic, T. Meyer, C. Garbe, and U. Leiter, "Epidemiology of Keratinocyte Skin Cancer with a Focus on Cutaneous Squamous Cell Carcinoma," Cancers 2024, Vol. 16, Page 606, vol. 16, no. 3, p. 606, Jan. 2024, doi: 10.3390/CANCERS16030606.

[4] G. Rajput, S. Agrawal, G. Raut, and S. K. Vishvakarma, "An accurate and noninvasive skin cancer screening based on imaging technique," Int J Imaging Syst Technol, vol. 32, no. 1, pp. 354–368, Jan. 2022, doi: 10.1002/IMA.22616.

[5] R. Haque et al., "Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning," BioMedInformatics 2024, Vol. 4, Pages 966-991, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[6] R. Haque et al., "A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis," 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 255–262, Sep. 2024, doi: 10.1109/IC3I61595.2024.10829132.

[7] Z. Rahman, Md. S. Hossain, Md. R. Islam, Md. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin lesion classification based on ensemble learning," Inform Med Unlocked, vol. 25, p. 100659, 2021, doi: 10.1016/j.imu.2021.100659.

[8] E. H. Houssein, D. A. Abdelkareem, G. Hu, M. A. Hameed, I. A. Ibrahim, and M. Younan, "An effective multiclass skin cancer classification approach based on deep convolutional neural network," Cluster Comput, vol. 27, no. 9, pp. 12799–12819, Dec. 2024, doi: 10.1007/S10586-024-04540-1/TABLES/8.

[9] U. A. Lyakhova and P. A. Lyakhov, "Systematic review of approaches to detection and classification of skin cancer using artificial intelligence: Development and prospects," Comput Biol Med, vol. 178, p. 108742, Aug. 2024, doi: 10.1016/J.COMPBIOMED.2024.108742.

[10] M. Sohaib, M. J. Hasan, and Z. Zheng, "A multichannel analysis of imbalanced computed tomography data for lung cancer classification," Meas Sci Technol, vol. 35, no. 8, p. 085401, May 2024, doi: 10.1088/1361-6501/AD437F.

[11] S. Ahmmed et al., "Enhancing Brain Tumor Classification with Transfer Learning across Multiple Classes: An In-Depth Analysis," BioMedInformatics 2023, Vol. 3, Pages 1124-1144, vol. 3, no. 4, pp. 1124–1144, Dec. 2023, doi: 10.3390/BIOMEDINFORMATICS3040068.

[12] R. Haque, P. B.D, M. K. Hasan, A. H. Sakib, A. U. Rahman, and M. B. Islam, "Scientific Article Classification: Harnessing Hybrid Deep Learning Models for Knowledge Discovery," 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), pp. 1–7, Nov. 2023, doi: 10.1109/AIKIIE60097.2023.10389945.

[13] M. Sohaib, M. J. Hasan, M. A. Shah, and Z. Zheng, "A robust self-supervised approach for fine-grained crack detection in concrete structures," Scientific Reports 2024 14:1, vol. 14, no. 1, pp. 1–20, Jun. 2024, doi: 10.1038/s41598-024-63575-x.

[14] A. Al-Sakib et al., "Robust Phishing URL Classification Using FastText Character Embeddings and Hybrid Deep Learning," 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 53–58, 2024, doi: 10.1109/RAAICON64172.2024.10928513.

[15] K. Rezaee, M. R. Khosravi, L. Qi, and M. Abbasi, "SkinNet: A Hybrid Convolutional Learning Approach and Transformer Module Through Bi-directional Feature Fusion," in 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), IEEE, Jun. 2022, pp. 1–6. doi: 10.1109/IC3SIS54991.2022.9885591.

[16] I. Ahmad, J. Amin, M. IkramUllah Lali, F. Abbas, and M. Imran Sharif, "A novel Deeplabv3+ and vision-based transformer model for segmentation and classification of skin lesions," Biomed Signal Process Control, vol. 92, p. 106084, Jun. 2024, doi: 10.1016/j.bspc.2024.106084.

[17] K. Rezaee and H. G. Zadeh, "Self-attention transformer unit-based deep learning framework for skin lesions classification in smart healthcare," Discover Applied Sciences, vol. 6, no. 1, p. 3, Jan. 2024, doi: 10.1007/s42452-024-05655-1.

[18] G. Yang, S. Luo, and P. Greer, "A Novel Vision Transformer Model for Skin Cancer Classification," Neural Process Lett, vol. 55, no. 7, pp. 9335–9351, Dec. 2023, doi: 10.1007/s11063-023-11204-5.

[19] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," Vis Comput, vol. 39, no. 7, pp. 2781–2793, Jul. 2023, doi: 10.1007/s00371-022-02492-4.

[20] P. Tschandl, C. Rosendahl, and H. Kittler, "Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci Data, vol. 5, Aug. 2018, doi: 10.1038/sdata.2018.161.

[21] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," Med Image Anal, vol. 75, Jan. 2022, doi: 10.1016/j.media.2021.102305.

[22] Md. R. Ahmed et al., "Towards Automated Detection of Tomato Leaf Diseases," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), pp. 387–392, May 2024, doi: 10.1109/ICEEICT62016.2024.10534559.

[23] M. S. Rahman et al., "Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models," 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 59–64, 2024, doi: 10.1109/RAAICON64172.2024.10928353.

[24] A. Altamimi, F. Alrowais, H. Karamti, M. Umer, L. Cascone, and I. Ashraf, "An improved skin lesion detection solution using multi-step preprocessing features and NASNet transfer learning model," Image Vis Comput, vol. 144, p. 104969, Apr. 2024, doi: 10.1016/J.IMAVIS.2024.104969.

[25] A. Al-Sakib, F. Islam, R. Haque, M. B. Islam, A. Siddiqua, and M. M. Rahman, "Classroom Activity Classification with Deep Learning," 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, 2024, doi: 10.1109/ICICACS60521.2024.10498187.

[26] M. D. Hosen et al., "Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[27] J. C. Ni, S. H. Lee, Y. C. Shen, and C. S. Yang, "Improved U-Net based on ResNet and SE-Net with dual attention mechanism for glottis semantic segmentation," Med Eng Phys, vol. 136, p. 104298, Feb. 2025, doi: 10.1016/J.MEDENGPHY.2025.104298.

[28] A. R. W. Sait, "A LeViT–EfficientNet-Based Feature Fusion Technique for Alzheimer's Disease Diagnosis," Applied Sciences 2024, Vol. 14, Page 3879, vol. 14, no. 9, p. 3879, Apr. 2024, doi: 10.3390/APP14093879.

[29] N. A. Aljarallah, A. K. Dutta, and A. R. W. Sait, "Image classification-driven speech disorder detection using deep learning technique," SLAS Technol, vol. 32, p. 100261, Jun. 2025, doi: 10.1016/J.SLAST.2025.100261.

[30] S. Maqsood and R. Damaševičius, "Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare," Neural Networks, vol. 160, pp. 238–258, Mar. 2023, doi: 10.1016/j.neunet.2023.01.022.

[31] C. Xin et al., "An improved transformer network for skin cancer classification," Comput Biol Med, vol. 149, p. 105939, Oct. 2022, doi: 10.1016/j.compbiomed.2022.105939.