AI Medicine
https://www.sciltp.com/journals/aim



Article

Improving Multilevel Classification and Characterization of Skin Cancer

Anvi Kumar [1], Meet Patel [2], Savya Vats [3], Arav Kumar [4] and Avimanyou Vatsa [2,*]

[1] Department of Computer Science, University of Texas, Austin, TX 78712, USA
[2] Department of Computer Science, Fairleigh Dickinson University, Teaneck, NJ 07666, USA
[3] Department of Computer Science, University of California, Los Angeles, CA 90095, USA
[4] Department of Computer Science, University of California, Berkeley, CA 94720, USA
* Correspondence: avatsa@fdu.edu

How To Cite: Kumar, A.; Patel, M.; Vats, S.; et al. Improving Multilevel Classification and Characterization of Skin Cancer. *AI Medicine* **2026**, *3*(2), 15. https://doi.org/10.53941/aim.2026.100015

Received: 21 August 2026
Revised: 15 November 2026
Accepted: 18 November 2026
Published: 25 November 2026

Abstract: In response to the growing number of skin cancer patients, this study helps identify seven different categories of skin cancer. Statistics from 2022 estimate that there were approximately 1.5 million people (Male—770,274 and Female—734,402) living with melanoma, with approximately 166,000 (∼11%) of those individuals under 50 years old in the USA. Additionally, the overall 5-year survival rate is as high as 94.7%, mainly due to advancements in immunotherapy and targeted therapies. Thus, the early detection of skin cancer and its effective treatment are complex processes. In this study, the seven different types of skin cancers are considered for classification. The Convolutional Neural Networks (CNNs) with different architectures play a vital role in identifying accurate patterns and classifying images. The CNN architectures (VGG16, and EfficientNetV2) require reconsideration for the multilabel classification of skin cancer. The diverse combination of hyperparameters, activation functions, and optimization methods for CNN architectures enables them to effectively and accurately identify patterns of skin cancer. The performance metrics of these three architectures—VGG16, Vision Transformers (ViTs), and EfficientNetV2—are measured and analyzed using seven performance metrics. As a result, VGG16 outperformed the others, achieving 97.55% accuracy and a 95.47% macro F1 score, calculated independently for each class.

Keywords: skin cancer; melanoma; classification; CNN; VGG16; ViT-B/16; EfficientNet; performance analysis

1. Introduction

The most common type of cancer is skin cancer, in which malignant cells form in the tissues of the skin. The most common types of skin cancer include squamous cell carcinoma, basal cell carcinoma, actinic keratosis, and melanoma. The high risk of squamous cell carcinoma and basal cell carcinoma, called nonmelanoma skin cancer (NMSC), is due to exposure to sunlight, particularly for people with fair and pale-yellow skin complexion [1]. The body parts (nose, ear, lower lip, or top of the hands, etc.) exposed to the sun lead to basal cell carcinoma and squamous cell carcinoma. In contrast, the chances of actinic keratosis are high if there are rough, red, pink, or brown scaly patches (flat or raised) on the skin. Additionally, the chances of getting most cancers, including skin cancer, increase as a person gets older [1].

Besides melanoma skin cancer, Squamous Cell Carcinoma (SCC) represents the majority of cancer-causing metastatic (stage 4 cancer, or cancer that has spread far into the body) deaths and diseases. The SCC can be treated and appropriately diagnosed with standard surgical procedures and histopathology. Newer technologies, such as dermoscopy (a non-invasive technique that magnifies the skin area 10 times compared to the naked human eye) and reflectance confocal microscopy, increase the likelihood of early detection and treatment of SCC. These diagnostic

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

techniques narrow the suspicious areas that are in critical need of attention. The most common signs of SCC in clinics are erythematous (redness of the skin caused by blood) scales or a slightly raised plaque on the skin. These types of SCC can go unnoticed by their carrier. Early detection of SCC prevents neoplastic (abnormal growth of cells) growth. High-frequency ultrasonography increased the correctness of diagnosis from 73% to 97% [2].

Actinic Keratosis (AK) is a cutaneous neoplasm that is the earliest manifestation of nonmelanoma skin cancer (NMSC). It is made of keratinocytes that develop from chronic UV exposure, and its presence increases the risk of both squamous cell carcinoma (SCC) and basal cell carcinoma (BCC). It is characterized by rough erythematous papules or plaques, usually asymptomatic. It can also appear with several other characteristics, which may increase the risk of progression to SCC or malignancy, including pigmentation, palpability, pain, pruritus, hyperkeratosis, inflammation, and ulceration. Additionally, AK can also be a cause of the formation of a cutaneous horn base. The extremities for these lesions pose an elevated risk of disease progression; lesions on the lips and ears carry an increased risk of progression, while lesions on the trunk and lower extremities also experience poor healing [3].

Basal cell carcinoma (BCC) is the most common type of skin neoplasm. It comes in six types: nodular, pigmented, cystic, sclerosing/morphea form, superficial, and nevoid. It usually appears on the face or the back of the hands. Metastasis is uncommon, as BCC typically grows slowly and spreads locally. Diagnosis is often made based on the appearance of the lesion, though confirmation generally requires an incisional or excisional biopsy. Low-risk lesions, typically measuring less than 1.5 cm in diameter and located in easily accessible areas, can be effectively managed with procedures such as cryotherapy, curettage, and electrodesiccation. Higher-risk or cosmetically sensitive cases may require radiation therapy or surgical excision to preserve function and appearance while ensuring a high cure rate [4]. Melanoma is less common than the other two types of skin cancer, but it invades nearby tissues and spreads to different parts of the body [1].

Melanoma is also an aggressive and life-threatening form of skin cancer where early and accurate diagnosis is critical for patient survival [5,6]. The standard for non-invasive examination is dermoscopy, a technique that enables the visualization of subsurface skin structures, significantly improving diagnostic accuracy over unaided observation [7,8]. However, the interpretation of dermoscopic images is subjective and dependent on clinician experience, which can lead to diagnostic inconsistencies [5]. To mitigate this, computer-aided diagnosis (CAD) systems leveraging deep learning, particularly convolutional neural networks (CNNs), have been developed. These systems have achieved performance comparable to, and in some cases exceeding, that of expert dermatologists in classifying lesions as either benign or malignant [9–11].

Despite their success, these binary classification models have a fundamental limitation: they oversimplify the complex diagnostic process into a single output [12,13]. A dermatologist's assessment is not merely a benign-or-malignant decision; it involves the meticulous identification of multiple specific patterns, colors, and structures within a lesion. Key dermoscopic criteria, such as an atypical pigment network, blue-white veil, irregular streaks, or regression structures, are systematically evaluated to form a diagnosis [7]. Binary classifiers typically operate as "black boxes," providing a probability score without explaining which of these critical features led to the conclusion. This lack of transparency is a significant barrier to clinical adoption, as it prevents clinicians from verifying the model's reasoning against their own expert knowledge and can erode trust in the technology [14].

To address the shortcomings of binary classifiers, this study proposes a multilabel classification framework. Unlike traditional models that assign a single label, a multilabel approach trains an algorithm to simultaneously predict the presence of multiple predefined dermoscopic features in images. The purpose is to create a system that mimics the analytical methodology of a dermatologist, providing a rich and interpretable output that details the specific characteristics of a lesion.

2. Background

The diagnosis of melanoma has undergone progressive evolution, driven by advancements in both clinical techniques and computational analysis. A foundational element in modern melanoma diagnosis is dermoscopy, a non-invasive imaging modality that utilizes magnification and specialized optics to visualize subsurface skin structures invisible to the naked eye. Numerous studies and meta-analyses have confirmed that dermoscopy, when used by trained clinicians, significantly enhances diagnostic accuracy compared to unaided visual inspection, improving sensitivity for melanoma detection while also reducing the number of unnecessary biopsies of benign lesions [15]. The diagnostic process in dermoscopy is predicated on the systematic recognition of specific patterns and features. Methodologies such as the 7-point checklist, pattern analysis, and the ABCD rule codify this process, assigning diagnostic significance to features like atypical pigment networks, blue-white veils, irregular streaks, and regression structures [16].

The structured, feature-based nature of dermoscopic analysis has made it a fertile ground for the application of

Computer-Aided Diagnosis (CAD) systems. Early CAD systems relied on traditional image processing techniques for feature extraction (e.g., color, texture, and shape analysis) followed by classical machine learning classifiers, such as Support Vector Machines (SVMs) [17]. While these systems showed promise, their performance was often limited by the robustness of hand-engineered features. The advent of deep learning, particularly Convolutional Neural Networks (CNNs), marked a paradigm shift. CNNs obviate the need for manual feature engineering by learning hierarchical representations of features directly from pixel data. A landmark study demonstrated that a CNN could classify dermoscopic images for keratinocyte carcinomas and melanomas at a level of competence comparable to board-certified dermatologists [18]. Subsequent research has consistently validated the high performance of CNNs for the binary classification of melanoma [19].

Instead of predicting a single benign/malignant label, a multilabel framework is trained to simultaneously identify a multitude of diagnoses from a single image. This methodology mirrors the cognitive process of a dermatologist, providing a rich and structured output that is more diagnostically informative. Early work in this area demonstrated the feasibility of training a CNN to predict a small set of dermoscopic patterns. More recent studies have leveraged large, publicly available, and expertly annotated datasets to train models capable of identifying a wider array of features, effectively creating a "dermoscopic feature profile" for a given lesion [20]. This approach not only provides a basis for the final diagnosis but also quantifies the presence of specific warning signs, offering a more nuanced and clinically actionable assessment than a single malignancy score. By framing the diagnostic problem in a multilabel context, research is moving towards creating a more collaborative and transparent AI tool that can augment, rather than automate, clinical decision-making.

The EfficientNetV2 architecture is an advanced deep learning model that improves the original EfficientNet by being both faster and more accurate [21]. While the first version scaled network depth, width, and resolution in a fixed way, EfficientNetV2 uses a more flexible approach that adapts to the dataset and task - employing compound scaling to achieve an equilibrium among these three dimensions. It also introduces fused-MBConv layers and progressive learning, which helps the model train more quickly and with fewer computational resources. These improvements allow EfficientNetV2 to reach high accuracy on image classification tasks while keeping training and inference efficient. For medical imaging, such as dermoscopy, this balance of speed and performance makes EfficientNetV2 well suited for building practical clinical systems that can aid in early detection of skin cancer [21].

EfficientNetV2 improves on the original EfficientNet by introducing fused-MBConv layers and progressive learning. Unlike traditional CNNs, which often scale depth, width, and resolution independently, EfficientNetV2 employs compound scaling to achieve an equilibrium among these three dimensions. Through these changes, it achieves higher accuracy with faster training, making it much more viable for image classification [21]. Furthermore, recent studies in radiology that utilized EfficientNetV2 for skin lesion analysis, histopathology, and chest X-rays often report improvements over standard CNN baselines. Prior research has repeatedly demonstrated that EfficientNetV2 models pretrained on larger datasets and improved on through smaller medical data can reach high accuracy while maintaining training efficiency [22].

Moreover, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs, leveraging self-attention to model image patches as sequences of information. By understanding contextual relationships across the entire image, ViTs have achieved state-of-the-art performance in several benchmarks. This ability makes them an intriguing option for multi-label skin lesion classification, where overlapping visual features and inter-class similarities pose challenges for CNNs. We evaluate a ViT model on the HAM10000 dataset and compare its performance with VGG16 and EfficientNet architectures. Our results highlight the tradeoffs in F1 Score, Area Under the Receiver Operating Characteristic curve (AUROC), precision, and recall [20,23].

The Vision Transformer adapted the Transformer architecture from natural language processing to vision tasks [23]. Unlike CNNs, which rely on convolutions and pooling, ViTs divide images into fixed-size patches and treat them as tokens in a sequence. Through multi-head self-attention, ViTs capture long-range dependencies and holistic context, enabling robust performance in transfer learning applications.

In medical imaging, early studies have applied ViTs to radiology [24], histopathology [25], and dermatology [26], often reporting improvements over CNN baselines [27]. Researchers have shown that ViTs pretrained on large datasets and fine-tuned on smaller medical datasets can achieve high AUROC and F1 Scores, while also offering interpretability via attention maps. These findings support the inclusion of ViTs in this comparative study, alongside VGG16 and EfficientNet. Further, performance may be improved using a transfer learning process.

Transfer learning has proven to be the best approach for medical image classification, as a simple yet powerful structure provides sufficient capabilities to achieve high results. While using images with 224×224 dimensions, CNN's VGG16 architecture performs well with the max pooling method. In addition to all these benefits, VGG16's simple architecture is also beneficial in terms of computational costs; it does not require $40 - 50$ layers to

compute at each training loop, but instead performs only 16 training loops and achieves significant results with less computational power.

In [28,29], a study conducted has improved the accuracy to 99% using ResNet50, while achieving 97.5% with MobileNetV2, 83% with InceptionV3, 86% with VGG16, and 87.3% with VGG19. Another research collaboration achieved accuracies of 91.12% for VGG16, 91.68% for VGG19, 92.51% for MobileNet, and 89.32% for ResNet-152 [28]. Moreover, VGG16 and ResNet50 are transfer learning models. Each model gave good accuracy. VGG16 achieved 93.45% accuracy, and ResNet50 gained 94.69% accuracy. One of the most notable points is that they observed 4 times the training speed in the VGG16 model compared to the ResNet50 [30]. In the study by Pande et al. [31], the accuracies were 93.03% for CNN, 95.53% for ResNet50, and 94.59% for VGG16.

3. Materials and Methods

The HAM10000 dataset [20,32] is one of the most widely used benchmarks for skin lesion classification, containing 10,015 dermoscopic images over seven diagnostic categories: Actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus-like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc). Its diversity and size have made it central in comparative studies of automated dermatology systems [22,33]. Notably, the dataset reflects real-world class imbalance, with nevi (NV) dominating rarer categories such as dermatofibroma (DF) or vascular lesions (VASC). The number of images in seven categories is illustrated in Figure 1.

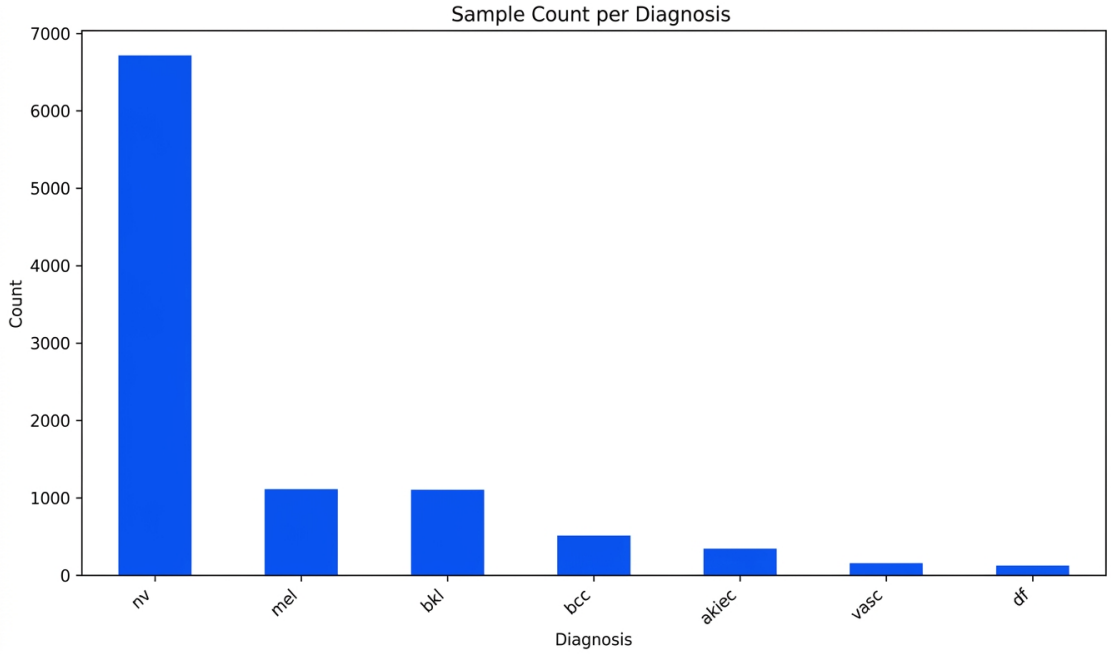


Figure 1. The data distribution of seven categories of images in the HAM10000 dataset.

The ViT model was trained and evaluated on the HAM10000 dataset, using a stratified split of 70/15/15 for training, validation, and testing. Labels were one-hot encoded for the seven diagnostic categories, and image preprocessing included resizing to 224×224 pixels, normalization to the range of $[-1, 1]$, and data augmentation with random flips, rotations, and color jitter to enhance generalization. The primary architecture was a Vision Transformer (ViT-B/16) implemented with the timm library and initialized from ImageNet [34]. To address class imbalance, it employed a weighted binary cross-entropy loss with class-specific positive weights proportional to the inverse frequency of each class. Training used AdamW optimization with weight decay (1×10^{-4}) and a cosine learning rate schedule. The evaluation metrics included accuracy, macro and micro F1 Scores, precision, recall, and AUROC. Because thresholding strongly affects multilabel performance, we tuned per-class thresholds on the validation set to maximize F1 and applied these thresholds at test time. This improved the precision-recall balance relative to the default 0.5 cutoff.

Another methodology for VGG16 is executed in a standard pipeline with various essential steps. The initial step begins with loading the "HAM10000 dataset," which comprises all images of 7 classes and a metadata file. Then, all images are converted to a 224×224 pixel dimension because the VGG16 model suits this dimension

for better results. Moreover, after changing the dimension, more data augmentation steps are done, including normalization to standardize the pixel value distribution [35].

The main advantage of using VGG16 is to leverage the transfer learning power and simplicity of VGG16 into this classification, which ultimately led to good results. The model checks CUDA availability, which checks for a compatible GPU for accelerated parallel computing, and prefers to train on what is available. The main parameters are the Cross-Entropy Loss as the objective function and the Adam optimizer, with a learning rate of 0.001 and a weight decay of 1×10^{-4}. All the images are divided into batches of 32 and trained for 500 epochs. For optimal time savings and to prevent unwanted session interruptions, model checkpoints were saved after every 10 epochs, ensuring that training would not have to start from scratch. Furthermore, a CSV file is logged for each epoch for subsequent analysis. The final step of the pipeline involves visualizing 500 training epochs, providing a comprehensive analysis.

The EfficientNetV2 model was trained and evaluated on the HAM10000 dataset following a structured and optimized deep learning pipeline. The dataset was first loaded with its corresponding metadata and dermoscopic images, after which data augmentation techniques—such as random rotations, flips, and brightness adjustments—were applied to enhance model generalization and reduce overfitting. All images were resized to $300 \times 300 \times 3$ and normalized to the [0, 1] range to match the input requirements of EfficientNetV2-B0. A stratified split was employed to divide the dataset into training and validation subsets, ensuring proportional representation of all diagnostic classes. The base model architecture consisted of EfficientNetV2-B0 pretrained on ImageNet, which integrates convolutional and Squeeze-and-Excitation layers to capture both spatial and channel-wise feature dependencies efficiently. The pretrained backbone was fine-tuned on HAM10000, leveraging transfer learning to accelerate convergence and improve feature extraction quality. Training was performed for 500 epochs using GPU acceleration (CUDA) when available, with model checkpointing enabled to save the best-performing weights based on validation accuracy. The training loop incorporated key hyperparameter definitions, including learning rate scheduling, batch size, and optimizer configuration. Throughout training, losses and performance metrics were logged for real-time monitoring, and post-training visualization tools were employed to evaluate convergence trends and classification performance. This workflow ensured efficient utilization of computational resources while maximizing diagnostic accuracy across the skin lesion categories.

4. Results and Discussions

The ViT achieved a strong overall performance on HAM10000. On the held-out test set, AUROC reached 0.951, indicating excellent separability across lesion classes. Macro-F1 was 0.628 with micro-F1 of 0.804, reflecting robust classification across both common and rare categories. Accuracy reached 0.942. Significantly, precision-recall tradeoffs were improved by per-class threshold tuning, as macro precision increased from 0.618 to 0.639 while recall remained balanced at 0.626, as shown in Table 1.

Table 1. The performance metrics of ViT, VGG16, and EfficientNetV2 models (500 epochs).

Metrics / Models	Test Loss	Accuracy	F1 (Macro)	F1 (Micro)	Precision (Macro)	Recall (Macro)	AUROC
ViT-B/16	0.302	0.942	0.628	0.804	0.639	0.626	0.951
VGG16	0.176	0.9775	0.9547	0.9775	0.9595	0.9509	0.9965
EfficientNetV2	0.3251	0.9132	0.658	0.7812	0.672	0.645	0.935

Training dynamics (Figure 2) showed steady convergence, with validation F1 rising through epoch 500 before plateauing. The train-val F1 gap remained modest (< 0.1), suggesting limited overfitting. Loss curves similarly indicated stable optimization. Thus, these results suggest that ViTs can complement CNNs in dermatology by offering high discriminative ability and balanced precision-recall performance.

The following model is used to classify skin cancer lesions from the HAM10000 dataset using a VGG16 deep learning model and transfer learning. The model is trained for a total of 500 epochs, with adaptive data augmentation in effect, and a class-weighted loss function is used to handle the high-class imbalance of the dataset efficiently. The training yields very promising results, with a maximum validation accuracy of 97.75% achieved at epoch 500, along with a strong macro F1-score of 0.9547. At the termination of the training run, the model achieved a high validation accuracy level of 97.70%, substantiating that the adopted training approach yielded a very efficient and accurate classifier for dermatology images.

Based on the analysis of the training logs, the model performed well and generalized adequately to previously

unseen data. Overfitting was well-regulated, as evidenced by the slight disparity of only 0.20% between the final training accuracy (97.90%) and the final validation accuracy (97.70%). The effective generalization is likely a result of the simultaneous use of data augmentation, weight decay, and a ReduceLROnPlateau learning rate scheduler, which steered the model towards a stable convergence. The findings strongly support the notion that a fine-tuned VGG16 network is a robust and effective solution for this particular medical imaging classification task. The VGG16 outcome is listed in Table 1 and illustrated in Figure 3

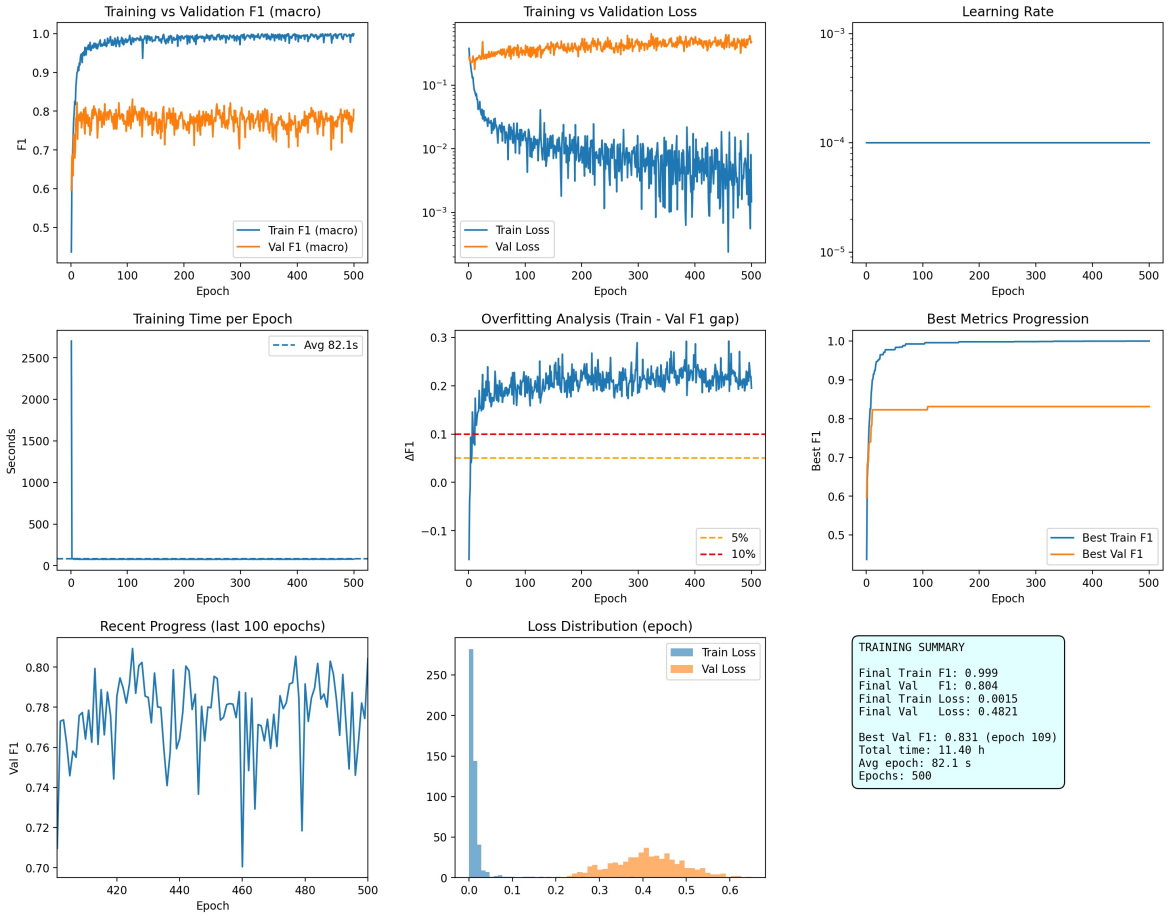


Figure 2. The training dynamics and performance metric of the ViT model.

The training progression of the EfficientNetV2-B3 model on the HAM10000 dataset is clearly illustrated in the graphs and summary metrics. As depicted in the validation accuracy plot, the model exhibited rapid gains during the initial epochs, with accuracy plateauing around 90–91% after 500 epochs. This suggests that EfficientNetV2 effectively captured relevant dermatological features early in training, with subsequent epochs primarily refining performance rather than producing substantial increases. The final recorded validation accuracy was 91.32%, accompanied by a validation loss of 0.3251, indicating a well-balanced trade-off between precision and generalization.

The model was trained using the AdamW optimizer, which facilitated stable convergence and mitigated overfitting through decoupled weight decay regularization. The optimizer was configured with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and a weight decay of 0.01. A cosine annealing learning rate schedule with warm restarts was employed to progressively reduce the learning rate over time, enabling the model to escape shallow local minima during the fine-tuning phase. This setup promoted smoother training dynamics and reduced overfitting to underrepresented classes, particularly in conjunction with the class-weighted loss function. The relatively conservative learning rate combined with moderate weight decay yielded slower but more stable convergence, which accounts for the steady improvement and minor oscillations observed in the loss curves.

The loss trajectories further corroborate this interpretation: both training and validation losses decreased sharply within the first 20 epochs before plateauing, a pattern characteristic of well-regularized convergence. Minor fluctuations in the later epochs are consistent with the cosine learning rate adjustments and the fine-tuning process. Overall, the EfficientNetV2-B0 model demonstrated efficient and stable training behavior over 100 epochs,

completing in 11 hours and 42 minutes. Although it did not surpass VGG16 in raw accuracy, the results and training curves highlight EfficientNetV2's robustness, stability, and computational efficiency in processing high-resolution medical imagery. Its outcome is listed in Table 1 and illustrated in Figure 4

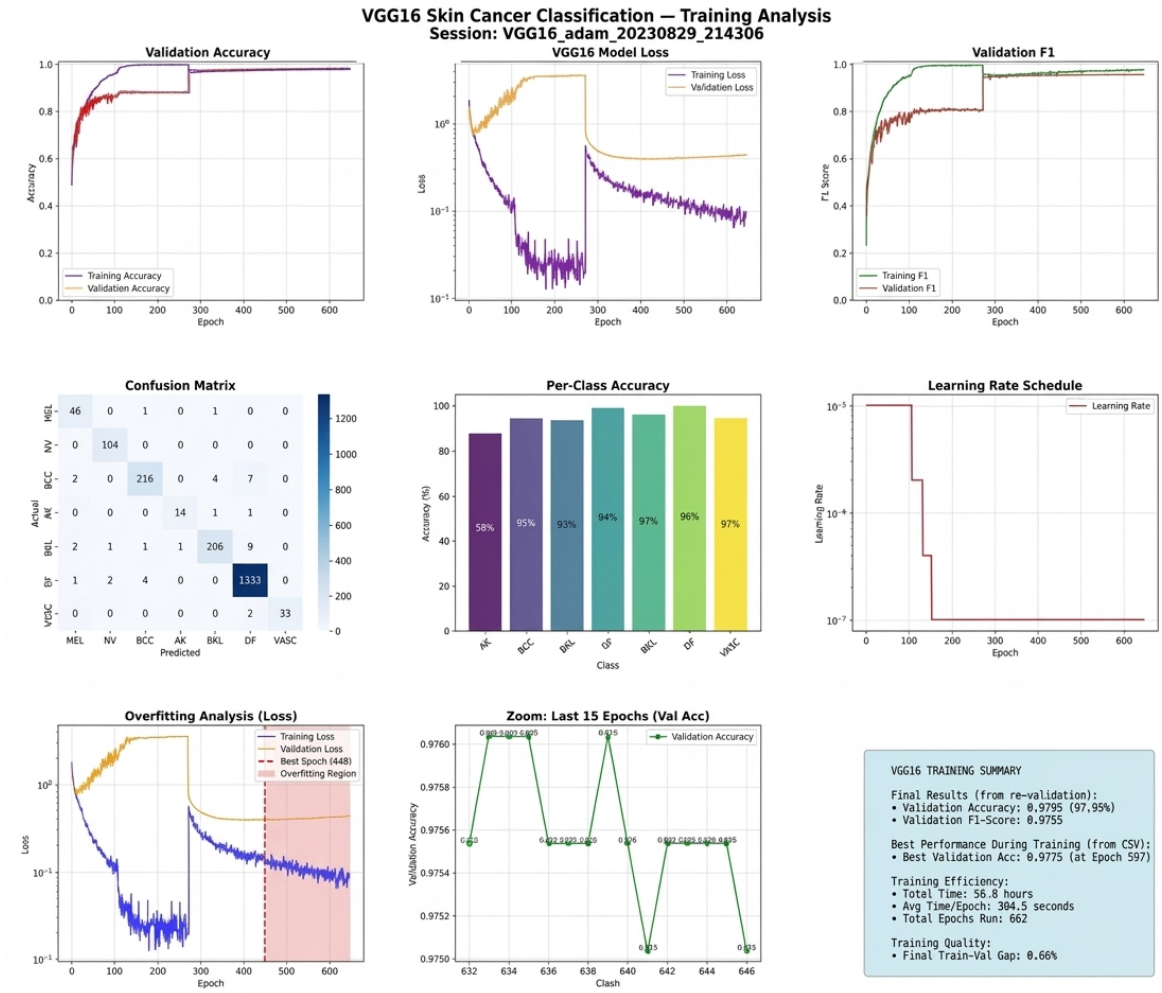


Figure 3. The training dynamics and performance metric of the VGG16 model.

The evaluations of the three deep learning architectures on the HAM10000 skin lesion classification dataset revealed that training duration had a significant impact on performance, alongside architectural design. VGG16 achieved the best results with 97.55% accuracy, 95.47% macro F1, and an exceptional 99.65% AUROC after 500 epochs of training with standard cross-entropy loss and Adam optimization on 224×224 images.

The EfficientNetV2 model delivered moderate performance—91.32% accuracy and 0.658 macro F1 after 500 epochs—at 300×300 input resolution, while ViT-B/16 showed 94.2% accuracy and 0.628 macro F1 after only 500 epochs of training, significantly fewer than the other models. The gaps between the macro and micro F1 scores reveal how each model handles the class imbalance. The VGG16 maintained tight alignment (0.9547 vs. 0.9775), indicating balanced performance across all seven diagnostic categories, while both ViT (0.628 vs. 0.804) and EfficientNetV2 (0.658 vs. 0.781) showed substantial discrepancies, favoring the majority classes despite balancing strategies.

The dramatic similarity in training duration- 500 epochs for VGG16, 500 epochs for EfficientNetV2, and 500 for ViT makes the direct architectural comparison easy. ViT especially requires more training, as its low-test loss relative to its accuracy suggests early-stage learning. VGG16's strong performance with its simpler architecture demonstrates the effectiveness of extended training on CNNs for medical imaging. At the same time, EfficientNetV2's intermediate results at 100 epochs suggest its efficient design benefits from moderate training duration.

Current results favor VGG16, though the $33\times$ difference in training epochs between ViT and VGG16 prevents definitive conclusions about architectural superiority. The performance metrics of three models are demonstrated in Figure 5

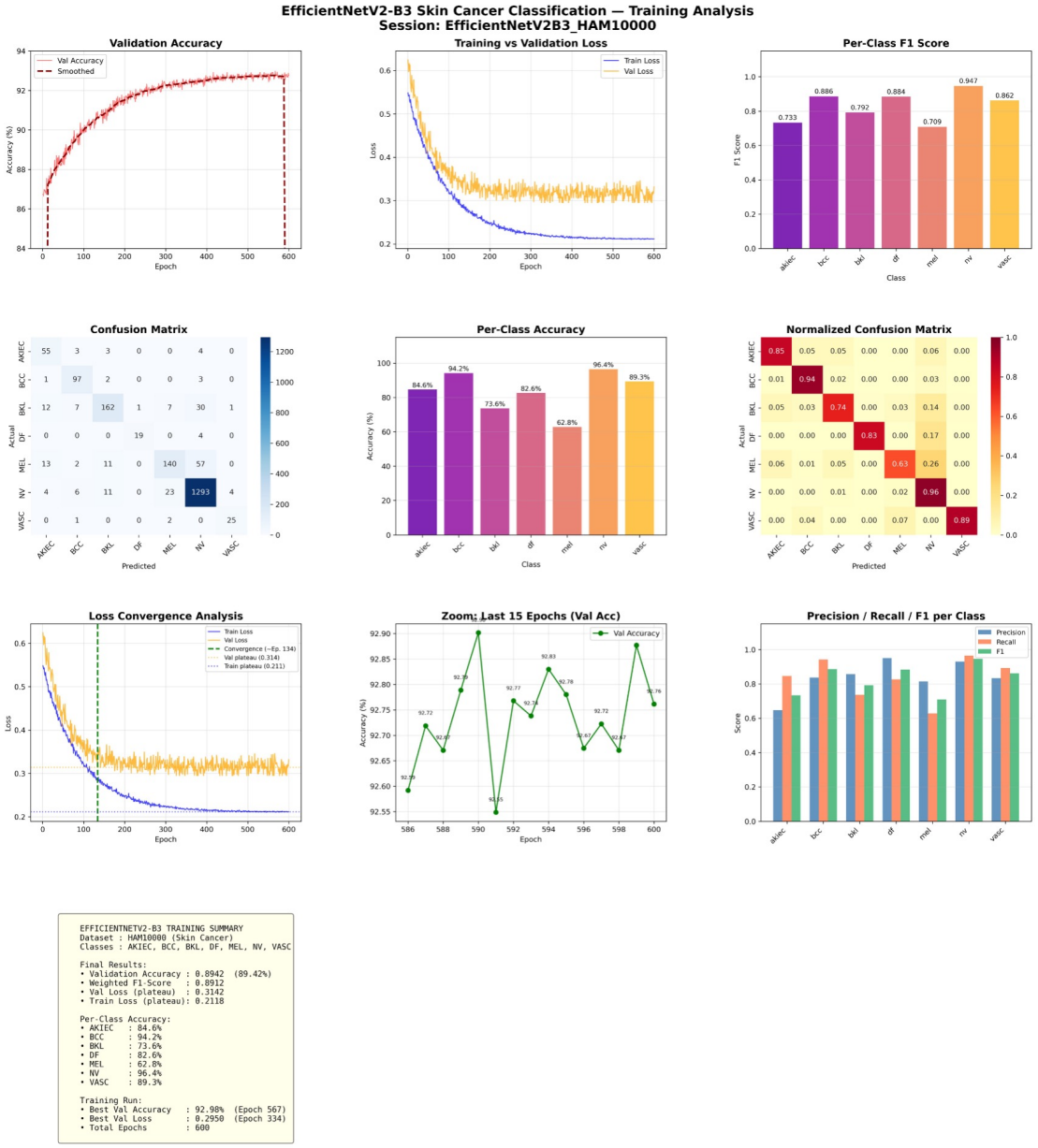


Figure 4. The training dynamics and performance metric of the EfficientNetV2-B3 model.

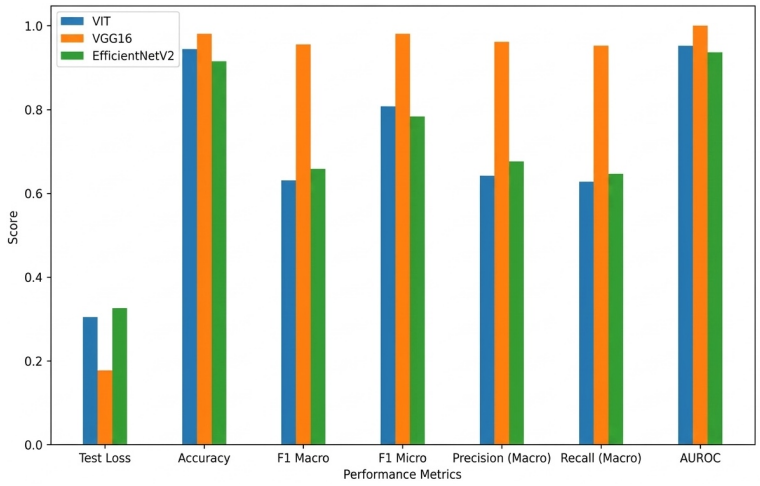


Figure 5. Performance metrics of ViT-B/16, VGG16, and EfficientNetV2 models.

5. Conclusions

This study presents the methodology for developing a multilabel deep learning framework, the results of its empirical validation, and a discussion of its implications for melanoma diagnosis. The overall contribution to the literature is a successful shift from opaque, binary classification systems to an interpretable, feature-based model that emulates a clinician's diagnostic reasoning. The primary strength of our approach, as demonstrated by the results, is its high fidelity in simultaneously identifying clinically significant dermoscopic features from a single image, thereby enhancing diagnostic transparency and providing a richer output than a simple malignancy score. However, a key weakness discovered during our analysis is the model's inherent dependence on the training dataset, which introduces potential biases related to underrepresented skin tones or rare lesion types and may limit its real-world generalizability. What is currently missing is a validation of this model in a true clinical setting and its integration with non-imaging patient data. To build upon this study, the immediate next steps must be twofold: first, to conduct rigorous validation on external, multi-source, and demographically diverse datasets to address the weaknesses of fairness and robustness explicitly. Second, future research should address the missing context by developing multimodal architectures that integrate our image feature analysis with structured clinical data from Electronic Health Records, such as patient history and lesion evolution. The ultimate progression from this work would be the implementation of this enhanced model in a prospective clinical trial to measure its impact on diagnostic accuracy and patient outcomes definitively. Overall, among all the models evaluated, VGG16 achieved the highest performance across nearly all metrics, displaying superior accuracy, F1 scores, and AUROC compared to both ViT-B/16 and EfficientNetV2. While EfficientNetV2 exhibited competitive F1 and precision values, its overall performance trailed behind VGG16, indicating that traditional CNN architectures can still outperform more recent transformer-based and hybrid models on the HAM10000 dataset. In future, the classification and prediction may be done using different ViT models, Vision Language Model (VLM), and Agentic AI - image to text prediction.

Author Contributions

A.K.(Anvi Kumar), M.P., S.V., A.K.(Arav Kumar) and A.V.: conceptualization, methodology, software, data curation, writing—original draft preparation, visualization, investigation, validation, writing—reviewing and editing; A.V.: supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets for this study can be found on www.kaggle.com. The urls are cited in the references.

Acknowledgments

We gratefully acknowledge the support of Olsen College of Engineering and Science, Fairleigh Dickinson University, Teaneck, NJ 07666, USA.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Given the role as Associate Editor, Avimanyou Vatsa had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. NCI. Skin Cancer (Including Melanoma)—Patient Version. Available online: https://www.cancer.gov/types/skin (accessed on 15 January 2025).
2. Combalia, A.; Carrera, C. Squamous Cell Carcinoma: An Update on Diagnosis and Treatment. *Dermatol. Pract. Concept.* **2020**, *10*, e2020066. https://doi.org/10.5826/dpc.1003a66.
3. Rigel, D.S.; Stein Gold, L.F. The Importance of Early Diagnosis and Treatment of Actinic Keratosis. *J. Am. Acad. Dermatol.* **2013**, *68*, S20–S27. https://doi.org/10.1016/j.jaad.2012.10.001.
4. Baba, P.U.F.; Hassan, A.U.; Khurshid, J.; et al. Basal Cell Carcinoma: Diagnosis, Management and Prevention. *J. Mol. Pathol.* **2024**, *5*, 153–170. https://doi.org/10.3390/jmp5020010.
5. Vipin, V.; Nath, M.K.; Sreejith, V.; et al. Detection of Melanoma Using Deep Learning Techniques: A Review. In Proceedings of the 2021 International Conference on Communication, Control and Information Sciences (ICCISc), Idukki, India, 16–18 June 2021; pp. 1–8. https://doi.org/10.1109/iccisc52257.2021.9484861.
6. Salam, A.; N, A. Revolutionizing Dermatology: The Role of Artificial Intelligence in Clinical Practice. *IP Indian J. Clin. Exp. Dermatol.* **2024**, *10*, 107–112. https://doi.org/10.18231/j.ijced.2024.021.
7. Holmes, G.A.; Vassantachart, J.M.; Limone, B.A.; et al. Using Dermoscopy to Identify Melanoma and Improve Diagnostic Discrimination. *Fed. Pract.* **2024**, *35* (Suppl. 4), S39–S45.
8. Massone, C.; Hofman-Wellenhof, R.; Chiodi, S.; et al. Dermoscopic Criteria, Histopathological Correlates and Genetic Findings of Thin Melanoma on Non-Volar Skin. *Genes* **2021**, *12*, 1288. https://doi.org/10.3390/genes12081288.
9. Omiye, J.A.; Gui, H.; Daneshjou, R.; et al. Principles, Applications, and Future of Artificial Intelligence in Dermatology. *Front. Med.* **2023**, *10*, 1278232. https://doi.org/10.3389/fmed.2023.1278232.
10. Ye, Z.; Zhang, D.; Zhao, Y.; et al. Deep Learning Algorithms for Melanoma Detection Using Dermoscopic Images: A Systematic Review and Meta-Analysis. *Artif. Intell. Med.* **2024**, *155*, 102934. https://doi.org/10.1016/j.artmed.2024.102934.
11. Kreouzi, M.; Theodorakis, N.; Feretzakis, G.; et al. Deep Learning for Melanoma Detection: A Deep Learning Approach to Differentiating Malignant Melanoma from Benign Melanocytic Nevi. *Cancers* **2025**, *17*, 28. https://doi.org/10.3390/cancers17010028.
12. Ghanem, M.; Ghaith, A.K.; El-Hajj, V.G.; et al. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain Sci.* **2023**, *13*, 1723. https://doi.org/10.3390/brainsci13121723.
13. Loh, D.R.; Hill, E.D.; Liu, N.; et al. Limitations of Binary Classification for Long-Horizon Diagnosis Prediction and Advantages of a Discrete-Time Time-to-Event Approach: Empirical Analysis. *JMIR AI* **2025**, *4*, e62985. https://doi.org/10.2196/62985.
14. Radhika, V.; Chandana, B.S. MSCDNet-Based Multi-Class Classification of Skin Cancer Using Dermoscopy Images. *PeerJ Comput. Sci.* **2023**, *9*, e1520. https://doi.org/10.7717/peerj-cs.1520.
15. Vestergaard, M.E.; Macaskill, P.; Holt, P.E.; et al. Dermoscopy Compared with Naked Eye Examination for the Diagnosis of Primary Melanoma: A Meta-Analysis of Studies Performed in a Clinical Setting. *Br. J. Dermatol.* **2008**, *159*, 669–676. https://doi.org/10.1111/j.1365-2133.2008.08713.x.
16. Argenziano, G.; Fabbrocini, G.; Carli, P.; et al. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis. *Arch. Dermatol.* **1998**, *134*, 1563–1570. https://doi.org/10.1001/archderm.134.12.1563.
17. Celebi, M.E.; Kingravi, H.A.; Uddin, B.; et al. A Methodological Approach to the Classification of Dermoscopy Images. *Comput. Med. Imaging Graph.* **2007**, *31*, 362–373. https://doi.org/10.1016/j.compmedimag.2007.01.003.
18. Esteva, A.; Kuprel, B.; Novoa, R.A.; et al. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. https://doi.org/10.1038/nature21056. Correction to *Nature* **2017**, *546*, 686. https://doi.org/10.1038/nature22985.
19. Brinker, T.J.; Hekler, A.; Enk, A.H.; et al. A Convolutional Neural Network Trained with Dermoscopic Images Performed on Par with 145 Dermatologists in a Clinical Melanoma Image Classification Task. *Eur. J. Cancer* **2019**, *111*, 148–154. https://doi.org/10.1016/j.ejca.2019.02.005.
20. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161. https://doi.org/10.1038/sdata.2018.161.
21. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. *arXiv* **2021**, arXiv:2104.00298.
22. Vatsa, A.; Kumar, A.; Vats, S.; et al. Comparing the Performance of Classification Algorithms for Melanoma Skin Cancer. In Proceedings of the 2023 IEEE Integrated STEM Education Conference (ISEC), Laurel, MD, USA, 11 March 2023; pp. 375–380. https://doi.org/10.1109/isec57711.2023.10402205.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
24. Hatamizadeh, A.; Tang, Y.; Nath, V.; et al. UNETR: Transformers for 3D Medical Image Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584. https://doi.org/10.1109/wacv51458.2022.00181.

25. Chen, R.J.; Chen, C.; Li, Y.; et al. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16144–16155. https://doi.org/10.1109/cvpr52688.2022.01567.
26. Yolcu Oztel, G. Vision Transformer and CNN-Based Skin Lesion Analysis: Classification of Monkeypox. *Multimed. Tools Appl.* **2024**, *83*, 71909–71923. https://doi.org/10.1007/s11042-024-19757-w.
27. Li, J.; Chen, J.; Tang, Y.; et al. Transforming Medical Imaging with Transformers? A Comparative Review of Key Properties, Current Progresses, and Future Perspectives. *Med. Image Anal.* **2023**, *85*, 102762. https://doi.org/10.1016/j.media.2023.102762.
28. Tahir, M.; Naeem, A.; Malik, H.; et al. DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images. *Cancers* **2023**, *15*, 2179. https://doi.org/10.3390/cancers15072179.
29. Mahmud, M.Z.; Reza, M.S.; Alve, S.R.; et al. Advance Transfer Learning Approach for Identification of Multi-class Skin Disease with LIME Explainable AI Technique. In Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 20–22 December 2024; pp. 109–114. https://doi.org/10.1109/iccit64611.2024.11021908.
30. Mahmud, A.; Ajami, H.; Hoque, M.S.U.; et al. Comparison of Performance of Two Deep Learning Models for Classification of Skin Lesions Using Image Resampling Technique for Data Augmentation. In Proceedings of the Applications of Digital Image Processing XLVII, San Diego, CA, USA, 18–23 August 2024; Volume 13137, pp. 1313705. https://doi.org/10.1117/12.3027916.
31. Pande, Y.; Wagre, S.S.; Alone, S.; et al. Skin Lesion Classification Using Deep Learning. *Int. J. Res. Appl. Sci. Eng. Technol.* **2025**, *13*, 2010–2015. https://doi.org/10.22214/ijraset.2025.72450.
32. SIIM-ISIC Melanoma Classification. Available online: https://www.kaggle.com/c/siim-isic-melanoma-classification/ (accessed on 15 January 2021).
33. Vatsa, A.K. SDFS: A Standardization Technique for Nonparametric Analysis. *IJonEST* **2021**, *3*, 30–43. https://doi.org/10.46328/ijonest.34.
34. PyTorch Models and Pre-Trained Weights. Available online: https://docs.pytorch.org/vision/main/models.html (accessed on 15 July 2021).
35. Kumar, A.; Vatsa, A. Untangling Classification Methods for Melanoma Skin Cancer. *Front. Big Data* **2022**, *5*, 848614. https://doi.org/10.3389/fdata.2022.848614.