Artificial Intelligence in Applied Sciences
Cutting-Edge Scientific Solutions

e-ISSN: 3108-4060
RESEARCH ARTICLE
Vol.2 / No.1 / 2026 / pp.32-36
https://doi.org/10.69882/adba.ai.2026015

Bridging the Gap Between Theoretical Performance and Clinical Utility in Multi-Class Skin Lesion Diagnosis

Furkan Sönmez[*,1] and **Fevzi Das**[α,2]
[*]Department of Computer Engineering, Igdir University, Igdir, Turkiye, [α]Department of Architecture and Town Planning, Igdir University, 76000 Igdir, Turkiye.

ABSTRACT The escalating global incidence of skin cancer necessitates the development of robust, objective, and automated diagnostic systems capable of augmenting clinical decision-making. This study presents a rigorous comparative analysis of four landmark Convolutional Neural Network (CNN) architectures, ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4, evaluated against the expansive and heterogeneous ISIC 2019 dataset. Comprising 25,331 high-resolution images across eight diagnostic categories, the dataset presents significant morphological challenges due to inherent visual ambiguity and class imbalance. Our findings reveal that EfficientNet-B5 achieves the highest predictive robustness with a peak accuracy of 0.8968 and an F1-score of 0.8458, leveraging its sophisticated compound scaling approach to capture subtle malignant markers. Concurrently, MobileNet-v3-Large demonstrated exceptional efficiency, yielding a nearly identical accuracy of 0.8965 with a minimal computational load of 0.4307 GFLOPs, making it a prime candidate for edge-computing applications. Despite its higher theoretical complexity, ResNet-101 provided the fastest real-world inference latency at 0.5032 ms, indicating superior hardware optimization. While these results underscore the transformative potential of deep learning in dermatology, misclassification patterns between melanoma and melanocytic nevi highlight persistent challenges in navigating fine-grained morphological boundaries. Ultimately, this research provides a holistic framework for selecting optimal architectural backbones based on specific clinical deployment constraints, bridging the gap between theoretical model performance and practical utility.

KEYWORDS
Skin lesion classification
Deep learning in dermatology
ISIC 2019 dataset
Computational efficiency
Model benchmarking

INTRODUCTION

Skin cancer represents a formidable global health crisis, with its prevalence reaching unprecedented levels over the last few decades (Gloster and Neal 2006; Armstrong and Kricker 1995; Leiter *et al.* 2020). Among the diverse spectrum of cutaneous malignancies, malignant melanoma is particularly notorious for its aggressive metastatic potential; however, it remains highly treatable when intercepted in its incipient stages. Despite the availability of advanced dermatoscopic techniques, the clinical diagnosis of skin lesions is fraught with challenges. The morphological overlap between benign and malignant pathologies, coupled with the subtle variations in pigment patterns and border irregularities, introduces a significant degree of intra-observer subjectivity. Consequently, there is an imperative clinical need for robust, objective, and automated diagnostic systems capable of augmenting a clinician's decision-making process (Madan *et al.* 2010).

The paradigm shift in medical image analysis has been primarily driven by the maturation of Deep Learning (DL), specifically through the evolution of Convolutional Neural Networks (CNNs) (Aruk *et al.* 2026; Cakmak and Pacal 2025a; Pacal and Cakmak 2025). These computational frameworks have demonstrated an extraordinary capacity to autonomously distill high-level hierarchical features from complex dermatological datasets, often identifying diagnostic biomarkers that elude manual visual inspection (Attallah 2024; Cakmak and Pacal 2025b; Cakmak 2025; Cakmak and Maman 2025). The transition from traditional hand-crafted feature engineering to end-to-end deep learning architecture has allowed for more nuanced classification across multi-class pathologies. Central to this progress is the availability of large-scale, annotated repositories such as the ISIC 2019 dataset, which provides a more diverse and challenging benchmark than its predecessors by incorporating a broader range of lesion categories and imaging conditions.

However, selecting an optimal architectural backbone for clinical deployment is not merely a question of peak accuracy; it requires a multifaceted evaluation of the trade-offs between predictive power and computational overhead. In modern medical environments, where real-time inference and integration into mobile or edge-computing platforms are increasingly vital, architec-

Manuscript received: 8 November 2025,
Revised: 16 December 2025,
Accepted: 16 January 2026.

[1] furkansonmez2024@gmail.com
[2] fevzi.das@igdir.edu.tr (**Corresponding author**).

tural efficiency is as quintessential as diagnostic sensitivity. While heavyweight models offer high-capacity feature extraction, lighter architecture provides the agility required for point-of-care applications. Systematic benchmarking across varying architectural paradigms is therefore essential to bridge the gap between theoretical model performance and practical clinical utility.

In this research, we conduct a rigorous comparative analysis of four landmark CNN architectures: ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4. By leveraging the ISIC 2019 dataset, our study investigates how distinct design philosophies, ranging from the residual learning of ResNet to the sophisticated compound scaling of EfficientNet, address the inherent complexities of multi-class skin lesion classification. Beyond traditional accuracy metrics, we scrutinize these models through the lens of computational complexity (GFLOPs) and parameter efficiency. This multifaceted evaluation ensures that the selected models are not only statistically robust but also optimized for the practical constraints of clinical environments, offering a balanced framework for high-stakes medical decision-making.

RELATED WORKS

The pursuit of high-precision diagnostic tools has led researchers to explore diverse architectural optimizations and learning strategies. In this context, Musthafa *et al.* (2024) demonstrated the efficacy of ensemble learning and model checkpoints, utilizing architectures like InceptionV3 and ResNet50 to achieve more stable and robust classification across various lesion types. Their work highlights the importance of strategic model saving and optimization to capture the most representative features during the training process. Building on the theme of architectural exploration, Rafeeque and Abini (2024) conducted a performance comparison between landmark models such as VGG16 and DenseNet, emphasizing how the depth and connectivity of these networks directly influence their ability to discern subtle malignant patterns in multi-class environments.

As models become more complex, the need for transparency in clinical decision-making has become paramount. Addressing this, Shah *et al.* (2024) introduced an explainable AI (XAI) framework that integrates Convolutional Neural Networks (CNNs) with Particle Swarm Optimization (PSO). By leveraging XAI, they provided a means to visualize the diagnostic markers driving the model's predictions, thereby bridging the gap between "black-box" algorithms and clinical interpretability. Similarly, Kumar *et al.* (2024) proposed SCCNet, a dedicated architecture for multi-class classification that provides estimated disease probabilities. Their approach moves beyond simple labels, offering a probabilistic output that aligns more closely with the nuanced nature of dermatological assessment.

The challenge of data variability and feature refinement has also been a focal point of recent research. . Surya *et al.* (2025) investigated the impact of rigorous image preprocessing and data augmentation on models like AlexNet and VGG, demonstrating that enhancing the quality of the input data is as critical as the choice of the network itself for melanoma detection. In a more specialized approach, Ali *et al.* (2024) presented a novel Fully Convolutional Encoder-Decoder Network (FCEDN) combined with SpaSA-based hyper-parameter optimization. Their work illustrates how adaptive CNN classification can be refined through automated optimization to handle the morphological diversity inherent in skin lesions. Finally, Sabir and Mehmood (2024) explored the boundaries of transfer learning, investigating how pretrained networks can be effectively fine-tuned for dermatological tasks. Their findings underscore that while deeper architectures offer significant representational power, the strategic application of knowledge from broader domains is essential for achieving high generalization in specialized medical datasets. Collectively, these studies establish a foundation for balancing predictive accuracy, computational feasibility, and clinical transparency.

MATERIALS AND METHODS

Dataset and Data Preprocessing

The foundational pillar of this investigation is the ISIC 2019 dataset (r19 2019), which serves as an expansive and heterogeneous repository of 25,331 high-resolution dermatoscopic images. This corpus is uniquely characterized by its profound morphological diversity, spanning eight distinct diagnostic categories: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevus (NV), Squamous Cell Carcinoma (SCC), and Vascular lesions (VASC). As visualized in Figure 1, these pathologies exhibit significant visual ambiguity and intra-class variations, presenting a formidable challenge for automated diagnostic systems to distinguish between benign and malignant lesions with high precision.

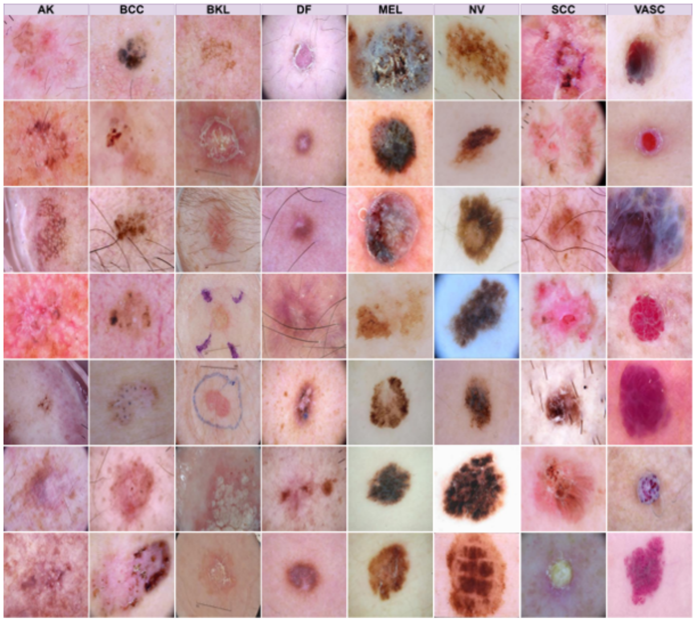

Figure 1 Representative dermatoscopic images from the ISIC 2019 dataset illustrating the morphological diversity and visual ambiguity across eight diagnostic categories

To ensure the statistical integrity of our comparative analysis and to promote robust model generalization, we implemented a rigorous stratified partitioning of the available data. The dataset was systematically divided into training (70%), validation (15%), and testing (15%) subsets, ensuring that the underlying distribution of lesion types remained consistent across all experimental phases. The precise numerical distribution of these samples across the eight diagnostic classes, culminating in a grand total of 25,331 images, is comprehensively documented in Table 1.

Table 1 Statistical distribution of the ISIC 2019 dataset across training, validation, and testing subsets for each of the eight skin lesion classes

Class Name	Total	Train	Val	Test
BKL	2624	1836	393	395
DF	239	167	35	37
VASC	253	177	37	39
AK	867	606	130	131
MEL	4522	3165	678	679
BCC	3323	2326	498	499
NV	12875	9012	1931	1932
SCC	628	439	94	95
GRAND TOTAL	25331	17728	3796	3807

Architectural Design Philosophies

To evaluate the trade-offs between predictive capacity and computational demand, our methodology benchmarks four landmark Convolutional Neural Network (CNN) architectures, each representing a distinct evolution in deep learning research: ResNet-101 (He *et al.* 2015), MobileNet-v3-Large (Howard *et al.* 2019), EfficientNet-B5 (Tan and Le 2019), and Inception-v4 (Szegedy *et al.* 2017). ResNet-101 leverages deep residual learning to facilitate the optimization of high-capacity networks by addressing the vanishing gradient problem. In contrast, MobileNet-v3-Large is specifically engineered for resource-constrained environments, utilizing depthwise separable convolutions to achieve high-speed inference with minimal parameter overhead. EfficientNet-B5 introduces a sophisticated compound scaling approach that simultaneously optimizes network depth, width, and resolution to maximize diagnostic sensitivity. Finally, Inception-v4 utilizes multi-scale convolutional modules to capture complex spatial hierarchies. Together, these models provide a comprehensive spectrum of design philosophies, ranging from the parameter-heavy residual blocks of ResNet to the agile, mobile-centric architecture of MobileNet.

Data Augmentation Strategy

A critical challenge inherent in dermatological imaging is the pronounced class imbalance, particularly the dominance of Melanocytic Nevus (NV) samples. To mitigate this and enhance the generalization capabilities of the models, we implemented an extensive data augmentation strategy. This pipeline included geometric transformations, such as random rotations, spatial scaling, and horizontal/vertical flipping, designed to simulate the varied orientations and perspectives encountered in clinical dermoscopy. These techniques ensure that the networks learn to distill invariant diagnostic features rather than memorizing dataset-specific noise, thereby bridging the gap between theoretical performance and practical clinical utility (Wang *et al.* 2024).

Performance Evaluation Metrics

To rigorously assess the diagnostic efficacy of the benchmarked architectures, we utilize a multifaceted suite of performance metrics that account for both statistical robustness and clinical applicability. The primary metric, Accuracy, provides an overall measure of the model's ability to correctly classify lesions across all eight diagnostic categories as defined in (1). However, given the high-stakes nature of dermatological screening, we extend our evaluation to include Precision and Recall to specifically measure the fidelity of malignancy detection and the model's sensitivity to true positive cases, as formulated in (2) and (3) respectively. To reconcile the potential trade-offs between these two metrics, especially in the context of the inherent class imbalance found within the ISIC 2019 dataset, the F1-Score is employed as a harmonic mean (4), ensuring a balanced representation of diagnostic performance across minority and majority classes. Beyond these conventional accuracy-based metrics, our analysis incorporates critical computational parameters such as model complexity (Params), floating-point operations (GFLOPs), and real-world inference latency (ms) to determine the practical feasibility of deploying these models within resource-constrained or real-time medical environments.

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

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

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

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

RESULTS

Quantitative Performance and Comparative Analysis

The empirical evaluation of the four benchmarked CNN architectures reveals a compelling spectrum of performance, highlighting the trade-offs between depth-wise feature extraction and computational agility. As detailed in Table 2.

EfficientNet-B5 emerged as the superior model in terms of predictive robustness, achieving a peak Accuracy of 0.8968 and a leading F1-Score of 0.8458. These results underscore the efficacy of the compound scaling method in distilling high-level diagnostic markers from the heterogeneous ISIC 2019 dataset. Interestingly, MobileNet-v3-Large demonstrated nearly identical accuracy (0.8965) and a highly competitive F1-Score (0.8383), despite maintaining a significantly smaller parameter footprint of only 4.21M compared to the 28.36M of EfficientNet-B5.

The comparative analysis further illustrates the impact of architectural design on computational overhead. While ResNet-101 and Inception-v4 provided respectable performance with accuracies of 0.8521 and 0.8697 respectively, they exhibited considerably higher computational costs, with ResNet-101 reaching 15.7288 GFLOPs. From a clinical deployment perspective, MobileNet-v3-Large presented a highly favorable profile, achieving the lowest GFLOP count (0.4307) while maintaining high diagnostic sensitivity. However, it is noteworthy that ResNet-101 offered the fastest real-world inference time at 0.5032 ms, suggesting that its residual structure remains highly optimized for parallel processing on modern GPU hardware.

Analysis of Model Predictions and Error Patterns

To gain deeper insight into the classification behavior of our most accurate model, we scrutinized the EfficientNet-B5 predictions through the confusion matrix presented in Figure 2. The model exhibited exceptional proficiency in identifying NV, correctly classifying 1,866 out of 1,932 test samples, which is vital given the

Table 2 Comparative performance analysis of ResNet-101, MobileNet-v3-Large, EfficientNet-B5, and Inception-v4 based on diagnostic metrics and computational complexity

Models	Accuracy	Precision	Recall	F1 Score	Params (M)	GFLOPs	Inference Time (ms)
ResNet-101	0.8521	0.8042	0.7490	0.7729	42.52	15.7288	0.5032
MobileNet-v3-Large	0.8965	0.8774	0.8093	0.8383	4.21	0.4307	1.9880
EfficientNet-B5	0.8968	0.8805	0.8166	0.8458	28.36	4.6551	3.4748
Inception-v4	0.8697	0.8333	0.7799	0.8046	41.16	12.2450	0.5516

prevalence of this category in clinical practice. Similarly, BCC and MEL showed high true positive rates, with 465 and 545 correct identifications respectively.

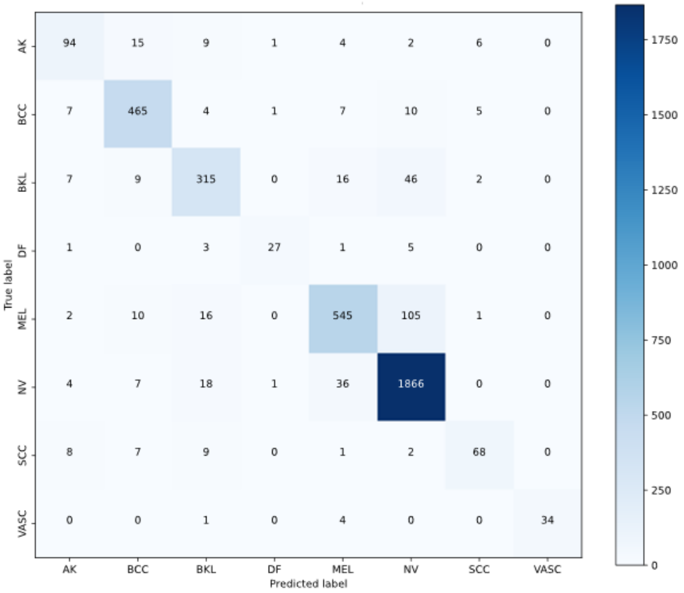

Figure 2 Confusion matrix of the EfficientNet-B5 model on the ISIC 2019 test set, highlighting classification performance and common morphological overlaps

Despite these strengths, the matrix reveals critical areas of morphological overlap that contribute to misclassification. A notable percentage of MEL cases were incorrectly labeled as NV, a common diagnostic pitfall in dermatology due to the subtle pigmentary transitions between benign and malignant melanocytic lesions. Furthermore, AK and SCC exhibited mutual confusion, likely stemming from their shared epithelial origin and similar scaling textures observed in dermatoscopic imagery.

A qualitative assessment of the classification outcomes, as illustrated in Figure 3, provides a visual context for both the successes and the limitations of the automated pipeline. The "True Predictions" panels demonstrate the model's ability to successfully capture the hallmark features of various pathologies, such as the distinct vascular structures in VASC and the characteristic pigment networks in MEL. Conversely, the "Misclassified Predictions" highlight the profound visual ambiguity inherent in skin lesion analysis. In several instances, lesions with atypical presentations, such as Melanoma mimicking the symmetry of a Benign Keratosis, led to false negative results. These visual findings emphasize that while CNNs possess remarkable feature-extraction capabilities, the clinical reliability of such systems is highly dependent on the model's ability to navigate the fine-grained morphological boundaries that often blur the distinction between benignity and malignancy.

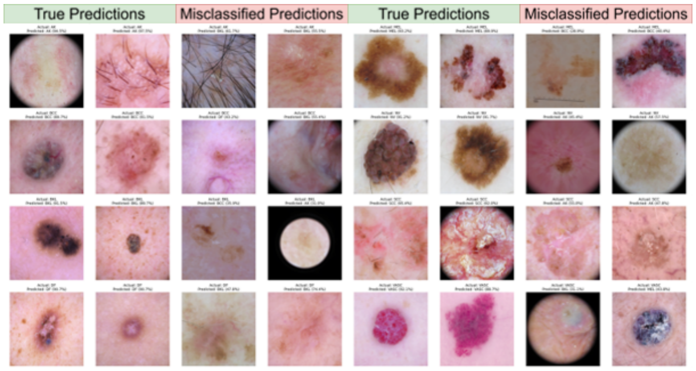

Figure 3 Qualitative assessment of classification results showcasing successful true predictions and representative misclassified samples with associated confidence levels

DISCUSSION

Interpretation of Key Findings

Our comparative analysis underscores a pivotal trade-off between predictive robustness and computational efficiency. EfficientNet-B5 emerged as the most accurate architecture, achieving a peak Accuracy of 0.8968 and a leading F1-Score of 0.8458. This superiority is largely attributed to its compound scaling philosophy, which systematically optimizes network depth, width, and resolution to capture the subtle morphological hallmarks of malignancy. However, the performance of MobileNet-v3-Large is arguably more noteworthy from a practical standpoint; it achieved a nearly identical Accuracy of 0.8965 while utilizing only 4.21M parameters and the lowest recorded computational load of 0.4307 GFLOPs.

An intriguing observation from our results is the discrepancy between theoretical complexity (GFLOPs) and real-world Inference Time. While ResNet-101 exhibited significantly higher GFLOPs than the mobile-centric models, it yielded the fastest inference latency at 0.5032 ms. This suggests that the classic residual structure remains exceptionally well-optimized for the parallel processing capabilities of modern GPU hardware. These findings imply that the "best" model is highly dependent on the target deployment environment, whether it be a high-throughput clinical workstation or a resource-constrained edge device.

Clinical Implications and Diagnostic Challenges

The error patterns identified through our confusion matrix analysis in Figure 2 reflect the same diagnostic pitfalls encountered by board-certified dermatologists. The frequent misclassification of MEL as NV highlights the critical challenge posed by subtle pigmentary transitions and "look-alike" lesions. Similarly, the mutual confusion between AK and SCC suggests that the shared epithelial origins of these pathologies lead to overlapping visual textures that remain difficult for even sophisticated CNNs to resolve.

Qualitative assessments further reveal that while models successfully identify hallmark features like vascular structures, they remain susceptible to atypical presentations. As seen in Figure 3, instances of malignancy mimicking benign symmetry can lead to dangerous false negatives. These results emphasize that while automated tools provide powerful decision support, their clinical reliability is fundamentally limited by the fine-grained morphological boundaries of skin cancer.

Limitations and Future Directions

Despite the high performance achieved, this study faces limitations inherent in the ISIC 2019 dataset, primarily the significant class imbalance where NV accounts for over half of the samples. While our data augmentation strategies mitigated this effect, future work should explore more advanced synthetic data generation or cost-sensitive learning to further improve minority class detection. Additionally, moving beyond purely visual data to incorporate patient metadata, such as age, anatomical site, and clinical history, could provide the contextual cues necessary to resolve the visual ambiguities identified in this research.

CONCLUSION

This research has established a comprehensive benchmarking framework for multi-class skin lesion classification by evaluating the intricate trade-offs between architectural design philosophies and diagnostic efficacy. Our investigation underscores that while EfficientNet-B5 provides the most robust predictive power, the agile architecture of MobileNet-v3-Large offers a remarkably viable alternative for resource-constrained point-of-care applications, maintaining high sensitivity with a significantly reduced parameter footprint. Furthermore, the observation that ResNet-101 delivers the fastest inference latency despite higher GFLOPs highlights that real-world clinical utility is as much a product of hardware-specific optimization as it is of algorithmic complexity. Although the deep learning models evaluated here effectively navigate profound visual ambiguities, the persistent confusion between malignant melanoma and benign nevi remains a significant diagnostic hurdle that mirrors clinician-level difficulties. To transcend these limitations, future research must move beyond purely visual data to integrate patient metadata and explore advanced learning strategies to address the inherent class imbalances found in large-scale dermatological repositories. In summary, this study provides the quantitative evidence and qualitative insights necessary to guide the selection of deep learning backbones for the next generation of automated dermatological screening tools.

Acknowledgments

This study was supported by the Turkish Health Institutes Presidency (TÜSEB) within the scope of the 2025-A1-01 Call, Project No. 44675. The authors gratefully acknowledge TÜSEB for its financial support and scientific contribution to this research. The experimental computations were carried out using the computing infrastructure of the Artificial Intelligence and Big Data Application and Research Center, Igdir University, which provided essential resources for model training and evaluation.

Ethical standard

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

Availability of data and material

The dataset analyzed for this study is the public dataset, which is available on Kaggle: https://www.kaggle.com/datasets/salviohexia/isic-2019-skin-lesion-images-for-classification?resource=download.

Conflicts of interest

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

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used artificial intelligence tools in order to improve the readability and language quality of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

LITERATURE CITED

2019 Isic 2019 skin lesion images for classification. Kaggle Dataset, Accessed December 23, 2025.

Ali, R., A. Manikandan, R. Lei, and J. Xu, 2024 A novel spasa based hyper-parameter optimized fcedn with adaptive cnn classification for skin cancer detection. Scientific Reports **14**.

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

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

Attallah, O., 2024 Skin-cad: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level cnns features and transfer learning. Computers in Biology and Medicine **178**.

Cakmak, Y., 2025 Machine learning approaches for enhanced diagnosis of hematological disorders. Computational Systems and Artificial Intelligence **1**: 8–14.

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

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

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

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

He, K., X. Zhang, S. Ren, and J. Sun, 2015 Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Howard, A., M. Sandler, G. Chu, *et al.*, 2019 Searching for mobilenetv3. arXiv preprint .

Kumar, V. A., C. Chandana, G. Supraja, *et al.*, 2024 Sccnet: Skin cancer detection and multi-class classification using deep cnn model with estimated disease probabilities. SN Computer Science **5**.

Leiter, U., U. Keim, and C. Garbe, 2020 Epidemiology of skin cancer: Update 2019. Advances in Experimental Medicine and Biology **1268**: 123–139.

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

Musthafa, M. M., M. T R, V. K. V, and S. Guluwadi, 2024 Enhanced skin cancer diagnosis using optimized cnn architecture and checkpoints for automated dermatological lesion classification. BMC Medical Imaging **24**.

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

Rafeeque, S. and M. A. Abini, 2024 Performance comparison of skin cancer detection and classification using cnn. In *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*.

Sabir, R. and T. Mehmood, 2024 Classification of melanoma skin cancer based on image data set using different neural networks. Scientific Reports **14**.

Shah, S. A. H., S. T. H. Shah, R. Khaled, *et al.*, 2024 Explainable ai-based skin cancer detection using cnn, particle swarm optimization and machine learning. Journal of Imaging **10**: 332.

Surya, V., B. H. Gollavilli, U. H. Nagarajan, *et al.*, 2025 Cloud-based cnn for automated skin cancer detection and classification in healthcare. International Journal of Science and Engineering Applications **14**: 40–45.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi, 2017 Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence .

Tan, M. and Q. V. Le, 2019 Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint .

Wang, Z., P. Wang, K. Liu, *et al.*, 2024 A comprehensive survey on data augmentation. arXiv preprint .

How to cite this article: Sönmez, F. B. and Das, F. Bridging the Gap Between Theoretical Performance and Clinical Utility in Multi-Class Skin Lesion Diagnosis. *Artificial Intelligence in Applied Sciences*, 2(1),32-36, 2026.