International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-12, Issue-3, March 2026]

Neural Architecture Design for Edge-Based Dermatological Imaging

Bhaumik Mhatre[1*]; Vidhi Bhoir[2]; Jayshree Pawar[3]; Manthan Pardeshi[4]; Karishma Raut[5]

Department of CSE (AI & ML), Viva Institute of Technology, Mumbai University, India
*Corresponding Author

Abstract— *This review surveys neural network architectures proposed between 2020 and 2025 for dermatological image analysis on resource-constrained edge platforms, with an emphasis on minimizing inference latency and model footprint. The discussion covers lightweight convolutional neural network (CNN) families, such as MobileNet and ShuffleNet, as well as hybrid architectures that integrate convolutional feature extractors with attention or transformer-based modules to capture both local and global image context. Key engineering strategies, including depthwise and grouped convolutions, compact architectural blocks, and model compression techniques, are examined for their role in reducing parameter counts and accelerating on-device inference. The reviewed studies consistently demonstrate that carefully optimized models can achieve real-time performance with memory footprints in the order of a few megabytes, enabling deployment on smartphones and embedded vision systems. By prioritizing system-level performance metrics over clinical interpretation, this paper synthesizes engineering-driven approaches that support practical, low-latency skin image analysis under strict size, power, and compute constraints.*

Keywords— CNN architectures, edge computing, hybrid CNN–Transformer models, lightweight neural networks, model compression, on-device inference, transformer-based vision.

I. INTRODUCTION

Deep learning has become a central tool in medical image analysis, including applications in dermatology, particularly lesion classification and segmentation [1], [2]. While many high-performing models assume access to cloud or server-grade computing resources, this paradigm introduces latency, dependence on network connectivity, and potential data privacy concerns [3], [4]. Transmitting high-resolution skin images to remote servers can be impractical in real-time settings or in regions with limited connectivity, and it raises additional requirements for secure handling of sensitive visual data.

Edge computing offers an alternative by performing inference directly on or near the acquisition device, thereby reducing response time and retaining data locally [3], [4]. This shift, however, introduces stringent constraints on memory, computation, and energy consumption. Consumer-grade hardware such as smartphones, portable dermatoscopes, and embedded vision modules typically cannot support large-scale neural networks with tens of millions of parameters or high floating-point operation counts [4], [5]. As a result, recent research has emphasized architectural efficiency and deployment-aware design.

A wide range of compact CNN architectures, including SqueezeNet, MobileNet, ShuffleNet, and EfficientNet variants, have been widely adopted for dermatological imaging [6]–[9]. These models rely on depthwise separable convolutions, bottleneck structures, and channel reorganization to significantly reduce computational cost while maintaining competitive accuracy. In parallel, vision transformers and hybrid CNN–transformer architectures have been explored to introduce global context modeling within mobile-friendly design envelopes [10]–[13]. To further align models with edge constraints, compression techniques such as quantization, pruning, and knowledge distillation are routinely applied [14], [15].

This paper presents a focused review of engineering-oriented advances in neural architecture design for edge-based dermatological imaging between 2020 and 2025. Specifically, it examines (i) lightweight CNN models tailored for mobile and embedded deployment, (ii) transformer-based and hybrid architectures optimized for efficiency, (iii) model compression strategies that reduce size and inference cost, and (iv) evaluation metrics that reflect real-world deployment requirements, including latency, model complexity, and throughput. The goal is to synthesize practical design patterns and trade-offs relevant to building compact, responsive systems rather than to introduce novel algorithms or clinical claims.

II. LITERATURE SURVEY

TABLE 1
REPRESENTATIVE STUDIES ON EDGE-ORIENTED DEEP LEARNING FOR DERMATOLOGICAL IMAGING (2020–2025)

Author (s) & Year	Model / Architecture	Model Category	Task	Dataset(s)	Key Performance Metrics	Model Size / FLOPs	Edge Deployment / Platform
Baig et al., 2023 [6]	Light-Dermo (ShuffleNet + SE)	CNN	Classification	ISIC	Accuracy, Sensitivity, Specificity, F1	~2–3M params	Smartphone (TensorFlow Lite)
Cheng et al., 2024 [7]	Enhanced MobileNet + Attention	CNN	Classification	ISIC 2019	Accuracy, Precision, Recall, F1	~3–4M params	Mobile / Embedded
Yu et al., 2025 [5]	MedLiteNet (MobileNetV2 + Transformer Tokens)	Hybrid	Segmentation	ISIC 2018	Dice, IoU	<3.3M params, ~7 GFLOPs	GPU prototype, mobile-oriented
Jiao et al., 2025 [10]	HCViT-Net	Hybrid	Segmentation	ISIC 2017/2018	mIoU, Dice	~5.76M params, ~7.5 GFLOPs	Embedded GPU / Mobile
Tang et al., 2024 [19]	SkinSwinViT	Transformer	Classification	ISIC	Accuracy, AUC, F1	~31M params	High-end Mobile / GPU
He et al., 2025 [20]	RFCS-Net (MobileViT based)	Hybrid	Classification	ISIC, HAM10000	Accuracy, F1, Latency	+0.1M params, +0.02 GFLOPs	Smartphone / Edge GPU
Valova et al., 2023 [2]	MobileNet App	CNN	Classification	Clinical Images	Accuracy, Latency	~3–5M params	Smartphone (CoreML/TFLite)
Tjong et al., 2025 [13]	ResNet/MobileNet (Jetson)	CNN	Classification	PH2	Accuracy, FPS	~4–25M params	NVIDIA Jetson Nano
Winata et al., 2025 [12]	MTAKD Student Model	CNN (Compressed)	Classification	ISIC	AUC, Latency	~50× smaller than teacher	Smartphone (TFLite)
Córdova Cárdenas et al., 2025 [22]	Quantized CNN/ViT Framework	CNN / Hybrid	Classification	Mixed	Latency, Energy, Accuracy	INT8, MB-scale models	Embedded Systems

2.1 CNN-Based Lightweight Architectures

Early edge-oriented dermatology systems primarily employed compact CNNs designed to minimize parameters and floating-point operations while preserving discriminative capability. Architectures such as SqueezeNet, MobileNet, and ShuffleNet introduced depthwise separable convolutions, group convolutions, and channel shuffling to reduce computation relative to conventional convolutional pipelines [8]. These design principles have been widely adopted and extended in dermatological imaging.

Baig et al. proposed **Light-Dermo**, which augments a ShuffleNet backbone with squeeze-and-excitation (SE) modules to emphasize informative channels while maintaining a low parameter budget [6]. The integration of lightweight attention mechanisms enabled improved feature representation without significantly increasing model complexity. Similarly, Cheng et al. introduced an enhanced MobileNet architecture that fuses spatial and channel attention to improve lesion classification performance while retaining the efficiency of a depthwise convolutional backbone [7].

Custom compact CNNs have also been reported. Nawaz et al. demonstrated that shallow, task-specific CNNs can outperform large pretrained models when optimized for dermoscopic texture and color patterns [8]. EfficientNet variants, which apply compound scaling across network depth, width, and resolution, have likewise been adapted to dermatology tasks, providing a structured approach to balancing performance and computational cost [18]. Across these works, the dominant pattern is the use of lightweight convolutional blocks combined with selective attention or multi-scale feature aggregation to maintain accuracy under tight resource constraints. Attention-augmented lightweight CNNs such as MedNet further demonstrate that selective feature recalibration can improve performance under tight parameter budgets [17].

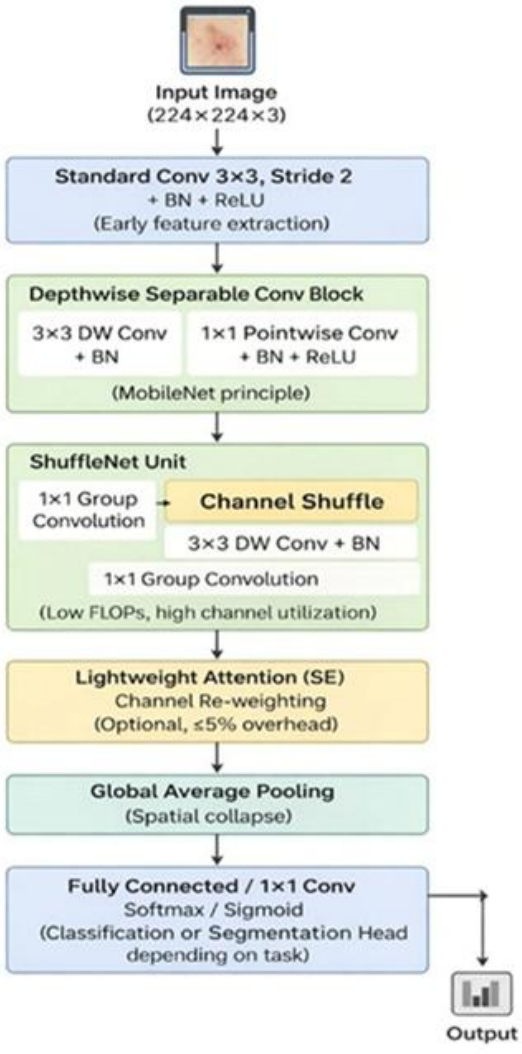

FIGURE 1: Generalized CNN-Based Lightweight Architecture for Edge-Based Dermatological Image Analysis

(Note: The figure is a conceptual synthesis of commonly adopted lightweight convolutional design patterns, including depthwise separable convolutions, group convolution with channel shuffling, and lightweight channel attention mechanisms, as reported in prior dermatological and edge-oriented studies. The diagram is redrawn for illustrative purposes to highlight representative architectural components used in resource-constrained deployments [6]–[8], [18].)

2.2 Transformer and Hybrid Models

Vision transformers (ViTs) and hybrid CNN–transformer architectures have gained attention for their ability to capture long-range dependencies through self-attention. While standard ViTs are often computationally demanding, recent work has focused on mobile-oriented variants that integrate transformer blocks within compact CNN frameworks.

MobileViT combines convolutional layers for local feature extraction with lightweight transformer modules to model global context, achieving favorable accuracy–efficiency trade-offs on image classification benchmarks [19]. Building on this concept, Pan et al. introduced **EdgeViT**, which employs a local–global attention structure to maintain low latency on mobile platforms while retaining competitive performance [20].

In dermatological imaging, hybrid models have been applied to both classification and segmentation. Yu et al. proposed **MedLiteNet**, which uses a MobileNet-based encoder and multi-scale attention tokens to achieve efficient lesion segmentation with a small parameter footprint [5]. Jiao et al. introduced **HCViT-Net**, integrating multi-scale convolutional features with query-based transformer modules to improve segmentation performance while controlling model complexity [10]. For classification, Tang et al. presented **SkinSwinViT**, which adapts a Swin Transformer backbone for skin lesion analysis, demonstrating that transformer-based designs can rival CNNs when appropriately scaled and trained [19].

These studies indicate that hybrid architectures can enhance representational capacity by combining local convolutional features with global attention. However, they also highlight the need for careful architectural design and compression to ensure feasibility on edge hardware. Additional hybrid CNN–transformer designs using focal loss and compact attention mechanisms have also been explored for skin lesion classification [23].

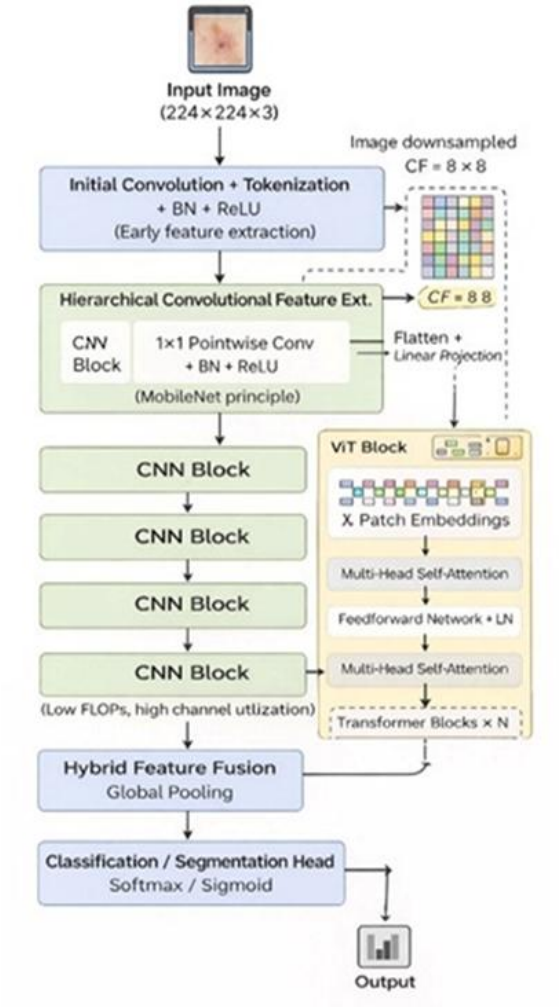

FIGURE 2: Generalized Hybrid CNN–Transformer Architecture for Edge-Based Dermatological Image Analysis

2.3 Edge Deployment Considerations

A recurring theme in the literature is the practical deployment of dermatology models on mobile and embedded platforms. Valova et al. developed a smartphone application based on MobileNet to demonstrate real-time, on-device skin condition analysis without reliance on cloud connectivity [2]. Tjong et al. evaluated transfer-learned CNNs on an NVIDIA Jetson Nano, illustrating the feasibility of GPU-accelerated inference in compact embedded systems [13].

Several works emphasize the conversion of trained models into mobile-friendly formats such as TensorFlow Lite or CoreML to enable efficient execution on smartphones. Pirahandeh demonstrated a Raspberry Pi–based prescreening system for rural deployment, highlighting the importance of low-cost hardware compatibility [3]. These studies consistently report reductions in latency and improved data privacy when inference is performed locally rather than via remote servers.

2.4 Model Compression Techniques

To further reduce computational and memory demands, many studies apply model compression. **Quantization** techniques map 32-bit floating-point parameters to lower-precision representations, often 8-bit integers, significantly reducing memory usage and accelerating inference on supported hardware [22]. **Pruning** methods remove redundant weights or channels, yielding smaller dense networks with minimal performance degradation.

Knowledge distillation has been particularly influential in dermatology applications. Winata et al. proposed a multi-teacher agreement framework in which a compact student network learns from multiple large teacher models, achieving substantial reductions in model size and inference time while maintaining classification performance [12]. Xu et al. demonstrated that combining pruning and quantization on standard CNN backbones can preserve accuracy while producing highly compact models suitable for deployment [9]. Other studies have proposed computationally efficient CNN architectures that explicitly balance accuracy and resource usage through architectural simplification and optimized training strategies [24], [25].

2.5 Evaluation Metrics Used in Prior Work

In addition to conventional predictive metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), edge-focused studies increasingly report system-level measures. These include parameter count, floating-point operations (FLOPs), inference latency, throughput, memory footprint, and, in some cases, energy consumption. For segmentation tasks, overlap-based metrics such as the Dice coefficient and Intersection-over-Union (IoU) remain standard [10].

The combined reporting of predictive and efficiency metrics reflects a growing recognition that deployment feasibility is as critical as model accuracy in edge-based dermatological imaging.

III. SUMMARY OF FINDINGS

The reviewed literature demonstrates a strong and consistent emphasis on efficiency-driven neural architecture design. Lightweight CNNs remain the dominant foundation for edge-based dermatological systems due to their favorable balance between accuracy and computational cost. Depthwise separable convolutions, inverted residual blocks, and compact attention mechanisms are widely used to extract discriminative features while minimizing parameters and operations.

Transformer-based and hybrid models introduce global context modeling that can enhance classification and segmentation performance, particularly in visually complex lesion patterns. However, their higher computational overhead necessitates careful scaling and frequent use of compression techniques to meet edge constraints. Hybrid CNN–transformer architectures often achieve superior representational capacity, but they typically rely on distillation, pruning, or quantization to remain feasible on mobile and embedded platforms.

Model compression emerges as a unifying strategy across architectural families. Quantization and knowledge distillation, in particular, are frequently applied to reduce memory footprint and inference time with limited impact on predictive performance. The literature also reflects a shift toward deployment-aware evaluation, where metrics such as latency, model size, and throughput are reported alongside traditional accuracy measures.

Overall, the dominant engineering trend is the co-optimization of architectural efficiency and deployment feasibility. By integrating compact design principles, selective attention mechanisms, and systematic compression, recent systems demonstrate that real-time, on-device dermatological image analysis is achievable within strict size and power budgets.

IV. RESEARCH GAPS

A primary limitation identified across the literature is the **lack of unified optimization frameworks** that simultaneously address accuracy, latency, and power consumption. Most approaches prioritize predictive performance during model design and subsequently apply compression or acceleration techniques to improve deployment efficiency [9], [22]. This sequential optimization strategy can result in suboptimal trade-offs, as improvements in inference speed or memory usage may lead to increased energy consumption or reduced numerical stability on low-precision hardware. The absence of systematic multi-objective design methodologies limits the ability to consistently meet the tight operational envelopes required by edge-based dermatological systems.

Another gap is the **limited extent of evaluation on real-world edge hardware**. Many studies report latency, throughput, or energy metrics derived from desktop-class GPUs or simulated mobile environments rather than direct measurements on consumer-grade devices, microcontrollers, or embedded accelerators [13], [22]. As a result, performance claims may not generalize across heterogeneous hardware platforms. The lack of standardized benchmarking protocols and reference hardware configurations further complicates reproducibility and fair comparison across studies.

Dependence on curated and controlled datasets remains a significant concern for deployment robustness. Training and evaluation are often conducted on dermoscopic or clinical datasets captured under standardized imaging conditions, whereas edge-based systems must operate in unconstrained environments with variable lighting, background clutter, and acquisition angles [5]. This discrepancy limits understanding of model generalization and may mask biases related to data homogeneity, which can become evident only after real-world deployment.

The **scalability and maintainability of hybrid CNN–transformer architectures** also present unresolved challenges. While these models demonstrate improved representational capacity, they typically involve multi-stage pipelines and higher architectural complexity than pure CNN counterparts [10], [15], [19]. This complexity increases engineering overhead when adapting models to new hardware platforms or updating them to accommodate evolving deployment requirements, raising concerns regarding long-term maintainability in production environments.

Finally, **explainability under edge constraints** remains largely unexplored. Although several hybrid and attention-based models employ post-hoc visualization techniques during offline evaluation, such methods are rarely integrated into on-device inference pipelines due to their computational overhead [11], [22]. The absence of lightweight, deployment-compatible interpretability mechanisms limits transparency and complicates debugging and validation in real-world edge deployments.

V. CONCLUSION

This review highlights a sustained engineering focus on efficiency in edge-based dermatological imaging systems. Lightweight CNN architectures, characterized by depthwise and grouped convolutions and compact attention modules, form the backbone of most on-device solutions due to their favorable balance of accuracy and computational cost. Hybrid CNN–transformer models extend this foundation by incorporating global context modeling, offering enhanced representational capacity when carefully scaled and compressed.

Model compression techniques—particularly quantization, pruning, and knowledge distillation—are central to bridging the gap between high-performing architectures and the strict memory, latency, and power constraints of edge platforms. The literature increasingly reports deployment-relevant metrics, underscoring a shift toward hardware-aware evaluation alongside traditional predictive performance measures.

Taken together, the surveyed work demonstrates that practical, real-time dermatological image analysis on mobile and embedded systems is achievable through coordinated architectural design and systematic efficiency optimization. The prevailing trend is not the pursuit of ever-larger models, but the refinement of compact, deployment-ready networks that align with the operational realities of edge computing environments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Behara, K., Bhero, E., & Agee, J. T. (2024). AI in dermatology: A comprehensive review into skin cancer detection. *PeerJ Computer Science, 10*, e2530. https://doi.org/10.7717/peerj-cs.2530

[2] Valova, I., Dinh, P., & Gueorguieva, N. (2023). Mobile application for skin cancer classification using deep learning. *Journal of Machine Intelligence and Data Science, 4*, 15–26. https://doi.org/10.11159/jmids.2023.003

[3] Pirahandeh, M. (2025). Dermatological health: A high-performance, embedded, and distributed system for real time facial skin problem detection. *Electronics, 14*(7), Article 1319. https://doi.org/10.3390/electronics14071319

[4] Pan, J., Li, R., Wang, Y., & Liu, Z. (2022). EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers. In *Proceedings of the European Conference on Computer Vision*. https://doi.org/10.1007/978-3-031-20083-0_18

[5] Yu, P., Zhang, H., & Li, W. (2025). MedLiteNet: Lightweight hybrid medical image segmentation model. *arXiv*. https://doi.org/10.48550/arXiv.2509.03041

[6] Baig, A. R., Khan, S., & Ahmed, M. (2023). Light-Dermo: A lightweight pretrained convolution neural network for the diagnosis of multiclass skin lesions. *Diagnostics, 13*(3), Article 385. https://doi.org/10.3390/diagnostics13030385

[7] Cheng, H., Lian, J., & Jiao, W. (2024). Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism. *Scientific Reports, 14*, Article 28850. https://doi.org/10.1038/s41598-024-80087-w

[8] Nawaz, K., Khan, S. A., & Ahmad, S. (2025). Skin cancer detection using dermoscopic images with convolutional neural network. *Scientific Reports, 15*, Article 7252. https://doi.org/10.1038/s41598-025-91446-6

[9] Xu, Y., Liu, J., & Wang, H. (2025). Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. *Artificial Intelligence Review, 58*, Article 93. https://doi.org/10.1007/s10462-024-11033-5

[10] Jiao, W., Cheng, H., & Lian, J. (2025). HCViT-Net: Hybrid CNN and multi-scale query transformer network for dermatological image segmentation. *Journal of Applied Clinical Medical Physics, 26*(12), e70385. https://doi.org/10.1002/acm2.70385

[11] Fiaz, M., Ali, S., & Khan, F. (2025). An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. *Frontiers in Medicine, 12*, Article 1681542. https://doi.org/10.3389/fmed.2025.1681542

[12] Winata, A., Santoso, B., & Wijaya, D. (2025). MTAKD: Multi-teacher agreement knowledge distillation for edge AI skin disease diagnosis. *Scientific Reports, 15*, Article 44314. https://doi.org/10.1038/s41598-025-27038-1

[13] Tjong, V., Darian, G., & Surantha, N. (2025). Performance evaluation of convolutional neural network (CNN) for skin cancer detection on edge computing devices. *Applied Sciences, 15*(6), Article 3077. https://doi.org/10.3390/app15063077

[14] Cabrejos-Yalán, V. M., & Rodriguez, C. (2024). Convolutional neural network model for skin cancer diagnosis in a dermatological center. *Mathematical Modelling and Engineering Problems, 11*(11), 2997–3005. https://doi.org/10.18280/mmep.111112

[15] Kobayashi, K., Tanaka, Y., & Suzuki, H. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *SN Computer Science, 5*, Article 1234. https://doi.org/10.1007/s10916-024-02105-8

[16] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6848–6856). https://doi.org/10.1109/CVPR.2018.00716

[17] Ferdous, M., Islam, R., & Hossain, S. (2025). MedNet: A lightweight attention-augmented CNN for medical image classification. *Scientific Reports, 15*, Article 41936. https://doi.org/10.1038/s41598-025-25857-w

[18] Nirupama, N., & Virupakshappa, V. (2024). MobileNet-V2: An enhanced skin disease classification by attention and multi-scale features. *Journal of Imaging Informatics in Medicine, 38*(3), 1734–1754. https://doi.org/10.1007/s10278-024-01271-y

[19] Tang, K., Li, S., & Wang, Y. (2024). SkinSwinViT: A lightweight transformer-based method for multiclass skin lesion classification with enhanced generalization capabilities. *Applied Sciences, 14*(10), Article 4005. https://doi.org/10.3390/app14104005

[20] He, F., Liu, Z., & Chen, Y. (2025). Skin lesion classification network based on improved MobileViT. *Engineering Applications of Artificial Intelligence, 139*, Article 111726. https://doi.org/10.1016/j.engappai.2025.111726

[21] Sinha, A., Kumar, P., & Singh, R. (2024). DermSynth3D: Synthesis of in-the-wild annotated dermatology images. *Medical Image Analysis, 97*, Article 103145. https://doi.org/10.1016/j.media.2024.103145

[22] Córdova-Cárdenas, R., Lopez, M., & Garcia, J. (2025). Edge AI in practice: A survey and deployment framework for neural networks on embedded systems. *Electronics, 14*(24), Article 4877. https://doi.org/10.3390/electronics14244877

[23] Nie, Y., Liu, S., & Zhang, Y. (2023). A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss. *Diagnostics, 13*(1), Article 72. https://doi.org/10.3390/diagnostics13010072

[24] Al Mamun, A., Hossain, M., & Rahman, M. (2025). Optimizing deep learning for skin cancer classification: A computationally efficient CNN with minimal accuracy trade-off. *arXiv*. https://doi.org/10.48550/arXiv.2505.21597

[25] Islam, N., Hasan, K., & Chowdhury, A. (2024). Leveraging knowledge distillation for lightweight skin cancer classification. *arXiv*. https://doi.org/10.48550/arXiv.2406.17051.