Comparative Analysis of EfficientNet-B0 and MobileNetV3-large Architectures for Imbalanced Multiclass Skin Lesion Classification

A. M. H. Pardede [1,*], Solikhun [2], and Juni Ismail [3]

[1] Information Systems Program, Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Kaputama, Binjai, Indonesia

[2] Informatics Engineering Program, Sekolah Tinggi Ilmu Komputer (STIKOM) Tunas Bangsa, Pematangsiantar, Indonesia

[3] Computer Science Program, Faculty of Mathematics and Natural Sciences, Universitas Huria Kristen Batak Protestan (HKBP) Nommensen Pematangsiantar, Indonesia

Email: Akimmhp@live.com (A.M.H.P.); solikhun@amiktunasbangsa.ac.id (S.); juniismailll@gmail.com (J.S.)

*Corresponding author

***Abstract*—Automated skin lesion classification systems face significant challenges in handling severe class imbalance characteristic of clinical dermatological datasets. This study presents a comprehensive comparative analysis of two state-of-the-art convolutional neural network architectures—EfficientNet-B0 and MobileNetV3-Large—for multiclass skin lesion classification on the HAM10000 benchmark dataset. The investigation systematically addresses an extreme class imbalance ratio of 58.30:1 through the strategic application of RandomOverSampler methodology combined with transfer learning from ImageNet pretrained weights. We implement identical training protocols for fair architectural comparison, including progressive dropout regularization (0.3→0.2→0.1) and conservative data augmentation strategies. Comprehensive ablation studies demonstrate the cumulative effectiveness of each pipeline component, with class balancing contributing 10.39% improvement and transfer learning providing substantial performance gains. Experimental validation reveals that EfficientNet-B0 achieves 98.86% classification accuracy with balanced precision, recall, and F1-Scores of 98.89%, 98.86%, and 98.85%, respectively, while MobileNetV3-Large delivers comparable performance at 98.88% accuracy. Gradient-weighted Class Activation Mapping (Grad-CAM) analysis validates that learned features align with clinical diagnostic criteria, demonstrating attention to lesion boundaries, color variations, and texture patterns. Compared to existing state-of-the-art methods on HAM10000, our approach achieves superior performance with fewer parameters (5.3 M), representing a marked advancement for resource-constrained deployment scenarios. These findings establish a validated framework for managing severely imbalanced medical imaging datasets and provide empirical guidance for efficient architecture selection in dermatological computer-aided diagnosis systems.**

***Keywords*—skin lesion classification, class imbalance mitigation, transfer learning, EfficientNet architecture, MobileNetV3 framework, HAM10000 dataset, medical image analysis, dermatological diagnosis, Gradient-weighted Class Activation Mapping (Grad-CAM) interpretability, deep learning comparison**

I. Introduction

Dermatological disorders constitute a substantial healthcare challenge globally, impacting millions of individuals across diverse populations and manifesting in conditions ranging from benign manifestations to potentially fatal malignancies [1–3]. Skin cancers, particularly melanoma, rank among the most frequently diagnosed cancer types worldwide. According to the American Cancer Society, melanoma accounts for most skin cancer-related deaths despite representing only a small percentage of overall cases [4, 5]. Early detection and accurate diagnosis remain critical for improving patient outcomes and survival rates [6, 7].

Recent advances in artificial intelligence, particularly deep Convolutional Neural Networks (CNNs), have demonstrated remarkable capabilities in automated dermatological image analysis [8–10]. Esteva *et al.* [1] pioneered the demonstration that deep neural networks could achieve diagnostic performance comparable to board-certified dermatologists in melanoma detection, establishing a foundation for subsequent research. However, significant challenges persist in translating these technologies to clinical practice, particularly regarding severe class imbalance in training datasets, model interpretability for clinical validation, and computational efficiency for point-of-care deployment [11–13].

The HAM10000 dataset, comprising 10,015 dermoscopic images across seven diagnostic categories, has emerged as a widely-used benchmark for automated skin lesion classification research [14]. Despite its prominence, this dataset exhibits extreme class

Manuscript received September 30, 2025; revised November 3, 2025; accepted January 5, 2026; published July 16, 2026.

imbalance with melanocytic nevi representing 66.95% of samples while dermatofibroma accounts for only 1.15%, yielding an imbalance ratio of 58.30:1. This substantial disparity reflects realistic clinical prevalence patterns but presents significant methodological challenges for conventional machine learning approaches [15, 16].

Despite substantial progress in deep learning-based dermatological diagnosis, several critical limitations persist in existing literature that motivate this investigation.

First, the pervasive challenge of severe class imbalance in dermatological datasets remains inadequately addressed. While data augmentation techniques are commonly employed, systematic evaluation of oversampling strategies with comprehensive ablation studies is conspicuously absent from much of the literature [17, 18]. Many investigations prioritize accuracy maximization without adequate consideration of per-class performance metrics, which are critical for clinical deployment where minority classes often represent the most clinically significant conditions [19, 20].

Second, comparative analyses of architectural efficiency, particularly for lightweight models suitable for mobile and edge deployment scenarios, remain limited within dermatological applications [21, 22]. While computational efficiency is critical for clinical translation and telemedicine applications, many studies focus exclusively on accuracy without rigorous evaluation of parameter efficiency, inference latency, and resource utilization [23, 24].

Third, interpretability and explainability remain underdeveloped in many proposed systems [25, 26]. Recent work has emphasized the importance of transparent Artificial Intelligence (AI) systems that enable clinical validation of learned features [27, 28]. Without visualization of model attention mechanisms, clinicians cannot verify whether diagnostic decisions align with established medical knowledge, limiting trust and adoption.

Fourth, reproducibility challenges continue to impede validation and clinical translation. Many studies lack comprehensive implementation details, standardized evaluation protocols, or systematic ablation studies demonstrating the contribution of individual components [29, 30]. This limitation is particularly concerning for medical AI applications where reliability and generalizability are paramount.

Addressing these identified gaps, this investigation develops a comprehensive framework for multiclass skin lesion classification with the following specific objectives and contributions:

- Systematic comparison of modern architectures: We provide rigorous comparative analysis of EfficientNet-B0 and MobileNetV3-Large architectures under identical experimental protocols, representing different points on the accuracy-efficiency trade-off spectrum. This comparison enables informed architectural selection based on deployment constraints and provides empirical guidance for similar medical imaging tasks.
- Comprehensive class imbalance mitigation: We systematically address severe class imbalance (58.30:1 ratio) through the strategic application of RandomOverSampler methodology, with detailed ablation studies quantifying the contribution of balancing strategies, transfer learning, progressive dropout regularization, and data augmentation.
- Model interpretability and clinical validation: We implement Gradient-weighted Class Activation Mapping (Grad-CAM) analysis to visualize learned features and validate alignment with clinical diagnostic criteria, enhancing transparency and clinical trust in automated diagnostic systems.
- Reproducible methodological framework: We establish a comprehensive experimental protocol with detailed hyperparameter specifications, systematic ablation studies, and standardized evaluation metrics, facilitating replication and extension by the research community.
- Computational efficiency analysis: We provide a detailed comparison of parameter counts, Floating-Point Operations (FLOPs), inference latency, and memory requirements, addressing practical deployment considerations for resource-constrained clinical environments.

The remainder of this manuscript is organized as follows: Section II reviews related work in deep learning for skin lesion classification, class imbalance handling, and transfer learning approaches. Section III describes the dataset characteristics, preprocessing methodology, architectural specifications, and training protocols. Section IV presents comprehensive experimental results including ablation studies, interpretability analysis, and comparative evaluation. Section V discusses findings in the context of existing literature and acknowledges study limitations. Section VI concludes with key contributions and implications for clinical application.

II. Related Work

This section reviews existing literature on deep learning approaches for skin lesion classification, organized by key methodological themes relevant to our investigation.

A. Deep Learning Architectures for Dermatological Diagnosis

CNNs have emerged as the predominant framework for dermoscopic image analysis. Early investigations by Yu *et al.* [5] demonstrated that residual network architectures could achieve 81.8% accuracy on multiclass skin lesion classification, establishing baseline performance expectations. Subsequent research by Haenssle *et al.* [4] compared deep learning systems against expert dermatologists, demonstrating that CNN-based approaches could match or exceed human performance in melanoma detection when trained on sufficient data.

Recent architectural innovations have focused on efficiency optimization for mobile deployment. Tan and Le [19] introduced EfficientNet, employing compound scaling principles to systematically balance network depth, width, and resolution. Howard *et al*. [20] developed MobileNetV3 through hardware-aware neural architecture search, optimizing for inference efficiency on mobile devices. These architectures have demonstrated promising results in medical imaging applications, though systematic comparison specifically for dermatological lesion classification remains limited.

Ensemble learning methodologies have attracted considerable attention for their capacity to synthesize predictions from multiple architectural variants. Harangi [29] achieved 88.5% accuracy through deep CNN ensembles, while Gessert *et al*. [15] reported 92.3% accuracy using ensemble architectures with diagnosis-guided loss weighting. However, ensemble approaches typically require substantially higher computational resources, limiting practical deployment feasibility.

B. Handling Class Imbalance in Medical Imaging

Class imbalance represents a pervasive challenge in medical image classification, where disease prevalence patterns create naturally skewed distributions. Traditional approaches include cost-sensitive learning, focal loss modifications, and synthetic minority oversampling techniques [17, 18]. Thurnhofer-Hemsi and Domínguez [11] achieved 89.7% accuracy on HAM10000 using custom CNN architectures with class weighting strategies, demonstrating the importance of balancing techniques.

Mahbod *et al.* [9] explored hybrid deep neural networks incorporating balanced sampling strategies, achieving 90.4% accuracy with improved per-class performance. Bissoto *et al.* [16] investigated deep feature aggregation combined with class balancing, reporting 91.8% accuracy. These studies demonstrate that systematic handling of class imbalance is essential for achieving robust performance across all diagnostic categories.

Recent work has emphasized the importance of evaluating balancing strategies through comprehensive ablation studies. However, many investigations apply multiple techniques simultaneously without isolating individual contributions, limiting understanding of which components drive performance improvements [19, 20].

C. Handling Class Imbalance in Medical Imaging

Transfer learning from large-scale natural image datasets has proven particularly effective in medical imaging contexts where training data availability is constrained [31]. Liu *et al.* [6] demonstrated that transfer learning from ImageNet substantially accelerates convergence and enhances generalization for skin disease classification, achieving 26% error rate reduction compared to training from scratch.

Hosny *et al.* [32] investigated transfer learning with AlexNet for skin lesion classification, reporting improved performance through fine-tuning pretrained weights. Li and Shen [33] explored transfer learning with data augmentation strategies, demonstrating that combined approaches yield superior results compared to either technique in isolation.

However, systematic comparison of transfer learning effectiveness across different architectural families remains limited, particularly regarding interaction effects with class balancing strategies and regularization techniques.

D. Explainability and Interpretability

Recent emphasis on explainable AI for medical applications has highlighted the importance of interpretable decision-making systems. Ruga *et al.* [34] developed MultiExCam, combining multiple explanation approaches for transparent skin lesion classification, achieving 94.2% accuracy while providing clinical interpretability. Their work demonstrated that ensemble explanation methods improve clinical trust and facilitate validation against established diagnostic criteria.

Han *et al.* [25] investigated augmented intelligence approaches where deep neural networks empower medical professionals in diagnosing skin cancer, emphasizing collaborative human-AI systems rather than autonomous diagnosis. Their findings underscore the importance of interpretability mechanisms for clinical adoption.

Grad-CAM has emerged as a widely-adopted technique for visualizing CNN attention patterns in medical imaging [26]. However, systematic application of Grad-CAM for validating learned features against clinical diagnostic criteria in dermatological classification remains underexplored.

E. Diversity and Generalization Challenges

Diversity in training data represents a critical challenge for clinical deployment. Rezk *et al.* [27] highlighted the need for inclusive datasets representing diverse skin tones and demographic groups, proposing protocols for algorithm validation across populations. Their work emphasizes that models trained predominantly on specific demographic groups may exhibit reduced performance when deployed in diverse clinical settings.

External validation across independent datasets remains essential for assessing generalization capabilities. Combalia *et al.* [14] introduced the BCN20000 dataset specifically to evaluate cross-dataset generalization, demonstrating substantial performance degradation when models are tested on data from different acquisition sources.

F. Positioning of Current Work

Building upon this foundation, our investigation addresses identified gaps through systematic comparison of modern efficient architectures (EfficientNet-B0 and MobileNetV3-Large) under identical experimental protocols, comprehensive ablation studies quantifying individual component contributions, Grad-CAM interpretability analysis validating clinical relevance of learned features, and detailed computational efficiency

characterization for practical deployment guidance. Our work extends existing literature by providing reproducible methodological frameworks and empirical evidence for architecture selection in resource-constrained dermatological diagnosis scenarios.

III. Materials and Methods

This investigation employs a quantitative experimental design focused on advancing automated dermatological lesion categorization through optimized deep convolutional architectures integrated with strategic data preprocessing techniques. Performance assessment utilizes established classification metrics—accuracy, precision, recall, and F1-Score—enabling comprehensive evaluation of model diagnostic capabilities. The experimental protocol exclusively leverages the HAM10000 dermatological image collection, a widely recognized benchmark dataset in computerized skin lesion analysis [3].

A. Dataset

The experimental investigation utilizes dermoscopic imagery obtained from the publicly accessible HAM10000 repository (accessible via Kaggle platform: kmader/skin-cancer-mnist-ham10000), recognized as among the most extensive curated collections of dermatological photographs available for computational research purposes [3]. This compilation comprises 10,015 color dermoscopy images encompassing seven clinically distinct diagnostic categories:

- actinic keratoses and intraepithelial carcinoma (akiec): Precursor lesions exhibiting potential malignant transformation.
- basal cell carcinoma (bcc): Most frequently diagnosed cutaneous malignancy characterized by localized invasive growth.
- benign keratosis-like lesions (bkl): Non-malignant epidermal proliferations including seborrheic keratoses.
- dermatofibroma (df): Benign dermal fibrous nodules presenting as firm papules.
- melanoma (mel): Aggressive malignant melanocytic neoplasm with significant metastatic potential.
- melanocytic nevi (nv): Benign melanocytic proliferations commonly termed moles.
- vascular lesions (vasc): Cutaneous manifestations involving blood vessel abnormalities.

The collection exhibits pronounced class distribution disparity, with sample counts ranging from 115 images (dermatofibroma category) to 6705 images (melanocytic nevi category), yielding an extreme imbalance ratio of 58.30:1 between majority and minority classes. This substantial imbalance reflects realistic clinical prevalence patterns but presents significant challenges for conventional machine learning methodologies, necessitating specialized balancing strategies to ensure unbiased model development [19, 20]. Fig. 1 presents representative exemplars from each diagnostic category, illustrating the visual diversity inherent within and across lesion types.

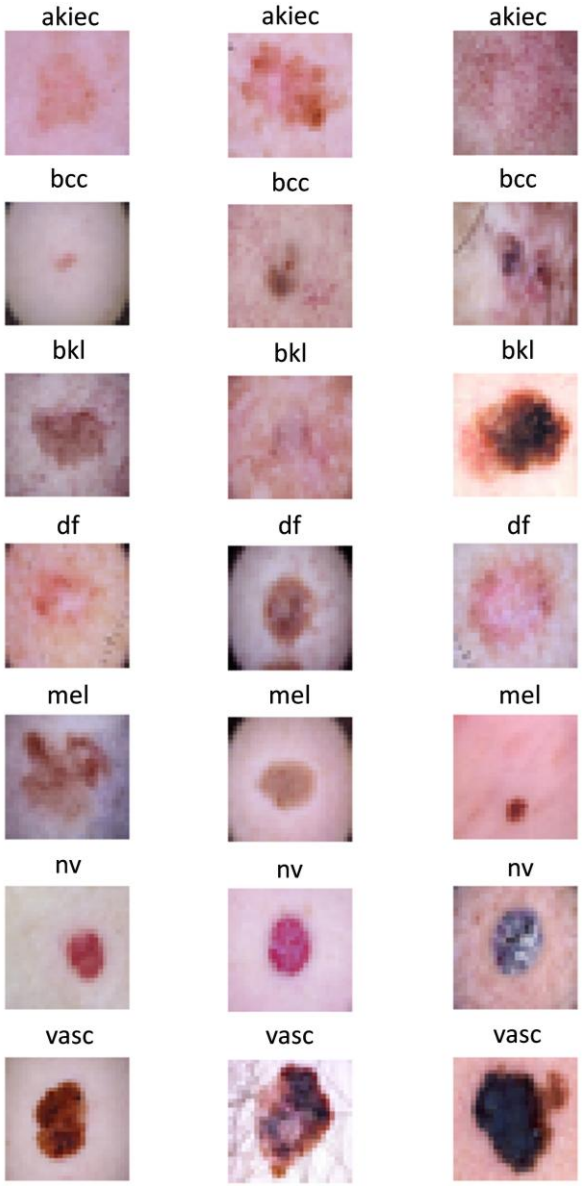

Fig. 1. Examples of a research dataset showing representative images from each of the seven skin lesion categories. Distribution analysis of the HAM10000 dataset shows representative images from each of the seven skin lesion categories. The dataset exhibits severe class imbalance with melanocytic nevi (nv) constituting 66.95% of samples while df represents only 1.15%, necessitating systematic balancing approaches.

Following image acquisition, the dataset undergoes systematic partitioning into three functionally distinct subsets to support model development and evaluation. The training subset serves as the primary data source for model parameter optimization, receiving class balancing interventions via RandomOverSampler implementation alongside modest augmentation strategies to enhance generalization capacity while avoiding overfitting. The validation subset facilitates hyperparameter tuning and training convergence monitoring, providing unbiased performance estimates during optimization that guide decisions regarding early stopping and learning rate adjustment. The test subset remains completely held-out until training completion, enabling final performance

assessment that simulates real-world deployment scenarios on previously unseen data.

The dataset partitioning follows a stratified 70/15/15 distribution after class balancing procedures, allocating seventy percent to training, fifteen percent to validation, and fifteen percent to testing while maintaining proportional class representation across all subsets. The complete preprocessing pipeline encompasses four sequential transformations designed to optimize data quality and format compatibility. Spatial resizing operations employ bicubic interpolation to upsample original images from their native 28 × 28 × 3 RGB format to standardized 224 × 224 pixel dimensions, satisfying input requirements for contemporary CNN architectures including EfficientNet-B0 and MobileNetV3 while preserving image content integrity. Statistical normalization applies ImageNet dataset parameters (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]) to pixel intensity values, aligning input distributions with pretrained model expectations and facilitating effective transfer learning convergence [35, 36]. Class balance correction employs RandomOverSampler methodology exclusively on the training partition to address the severe 58.30:1 imbalance ratio, generating equal representation across all seven diagnostic categories through random replication of minority class instances rather than synthetic sample generation that might introduce unrealistic image artifacts [37, 38]. Conservative augmentation strategies apply random horizontal flipping with fifty percent probability to training samples only, providing modest regularization benefits while intentionally avoiding aggressive transformations such as rotation, colour jittering, or perspective warping that could compromise clinically significant morphological features essential for accurate diagnosis [39].

B. *Proposed Model*

This investigation implements a comparative evaluation of two state-of-the-art convolutional neural network architectures representing distinct points on the accuracy-efficiency trade-off spectrum. Implementation utilizes the PyTorch deep learning framework in conjunction with the timm (PyTorch Image Models) library, providing access to pretrained model weights and standardized architectural implementations [40]. Fig. 2 presents architectural schematics for both evaluated models. The experimental design evaluates the EfficientNet-B0 and MobileNetV3-Large architectures, both initialized with ImageNet-pretrained weights and subsequently fine-tuned for dermatological classification tasks. These architectures represent contemporary approaches to efficient CNN design, with EfficientNet achieving optimization through compound scaling principles and MobileNetV3 employing hardware-aware neural architecture search methodologies [32, 33]. Both models undergo identical preprocessing and training protocols to ensure valid performance comparison and isolate the impact of architectural design choices.

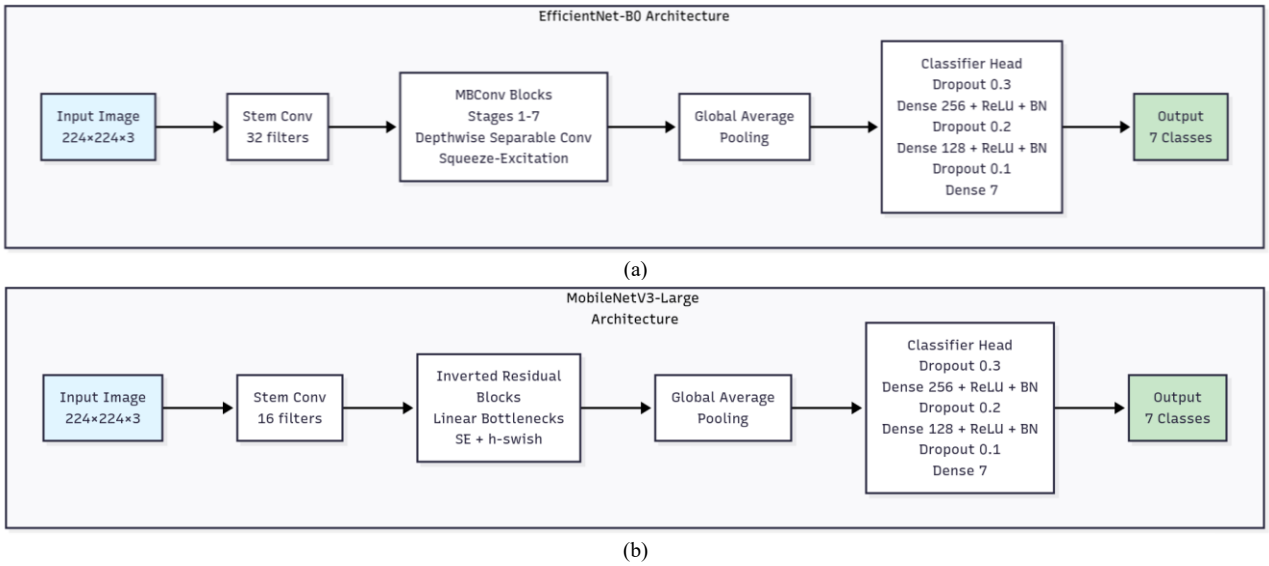

Fig. 2. Architectural comparison: (a) EfficientNet-B0, (b) MobileNetV3-Large models used for skin lesion classification.

In Fig. 2(a), the EfficientNet-B0 architecture processes input images measuring 224 × 224 × 3 pixels through a systematically scaled network that optimizes depth, width, and resolution parameters simultaneously according to compound scaling principles [19]. The backbone network extracts hierarchical feature representations through seven sequential stages of Mobile inverted Bottleneck Convolution (MBConv) blocks, each incorporating depth wise separable convolutions for computational efficiency alongside squeeze-and-excitation modules enabling adaptive channel-wise feature recalibration [41]. Following global average pooling operations that aggregate spatial feature maps, extracted representations propagate through a custom classification head architecture comprising progressive dropout layers with decreasing regularization strength (0.3→0.2→0.1), fully connected transformations with Rectified Linear Unit (ReLU)

activation and batch normalization, ultimately producing logits for seven diagnostic categories. This progressive dropout schedule balances overfitting prevention with representation capacity, applying stronger regularization to early layers while preserving learned features approaching the output layer [42].

The MobileNetV3-Large architecture accepts identical input dimensions while employing network design principles specifically optimized for mobile deployment contexts through hardware-aware neural architecture search [20]. The backbone utilizes inverted residual blocks with linear bottleneck structures, incorporating both squeeze-and-excitation attention mechanisms and h-swish nonlinear activation functions designed for efficient hardware implementation with minimal computational overhead [25]. The classification head maintains structural equivalence to the EfficientNet-B0 configuration, ensuring fair architectural comparison by isolating the impact of backbone design choices while controlling for classifier complexity. Both architectures target multiclass dermatological lesion classification but embody fundamentally distinct design philosophies, with EfficientNet-B0 achieving superior accuracy through principled compound scaling optimization while MobileNetV3-Large prioritizes computational efficiency and inference latency suitable for resource-constrained deployment scenarios, including mobile health applications and edge computing environments.

C. Research Framework

The experimental methodology follows a systematic workflow encompassing problem formulation, data acquisition, preprocessing, model development, training, and comprehensive evaluation phases. This structured approach ensures reproducibility and facilitates validation by the research community. Research commences with an explicit definition of investigative objectives, problem statement articulation, and anticipated outcome specification, establishing a foundational context for subsequent experimental phases. The methodological framework then progresses to HAM10000 dataset acquisition, providing access to 10,015 dermoscopic images spanning seven diagnostic categories as detailed in Section III. A. Data preprocessing encompasses multiple critical operations executed sequentially: class balance correction using RandomOverSampler to address the 58.30:1 imbalance ratio, spatial resizing through bicubic interpolation from 28 × 28 to 224 × 224 pixel dimensions, statistical normalization applying ImageNet distribution parameters, and conservative augmentation via random horizontal flipping applied exclusively to training samples. The balanced and pre-processed dataset subsequently undergoes stratified partitioning into training (70%), validation (15%), and test (15%) subsets, maintaining proportional class representation. Fig. 3 illustrates the complete methodological workflow from initial data collection through final performance assessment.

D. Architecture Selection Rationale and Computational Analysis

We selected EfficientNet-B0 and MobileNetV3-Large to represent two distinct design philosophies in efficient CNN architectures, enabling systematic comparison of accuracy-efficiency trade-offs under identical experimental protocols.

- EfficientNet-B0 characteristics: EfficientNet-B0 represents the compound scaling approach introduced by Tan and Le [19], which systematically balances network depth, width, and resolution through a principled scaling methodology. This architecture employs MBConv blocks with squeeze-and-excitation attention mechanisms for adaptive channel-wise feature recalibration. We selected the B0 variant (the base model in the EfficientNet family) to establish baseline performance while maintaining computational efficiency suitable for comparison with mobile-optimized architectures.
- MobileNetV3-Large characteristics: MobileNetV3-Large embodies hardware-aware neural architecture search principles, specifically optimized for efficient mobile deployment through automated architecture design [34]. This model utilizes inverted residual blocks with linear bottlenecks, incorporating h-swish nonlinear activation functions designed for efficient hardware implementation with minimal computational overhead. We selected the large variant to achieve comparable parameter count with EfficientNet-B0 (5.4 M vs 5.3 M parameters) while providing superior inference efficiency due to optimized layer configurations.

Computational comparison: Table I presents detailed computational characteristics of both architectures, facilitating informed model selection based on deployment constraints.

This architectural pairing enables fair comparison between compound scaling and neural architecture search methodologies while maintaining approximately equivalent parameter budgets. The substantial difference in FLOPs (0.39 B vs 0.22 B) despite similar parameter counts reflects MobileNetV3's optimization for inference efficiency, a critical consideration for deployment in resource-constrained clinical environments and mobile health applications.

E. Hyperparameter Configuration and Justification

All hyperparameters were selected based on preliminary experiments and established best practices in medical image classification literature. Table II presents the complete configuration with corresponding justifications.

The progressive dropout schedule (0.3→0.2→0.1) represents a novel contribution specifically designed for medical imaging transfer learning scenarios. This graduated regularization applies stronger dropout in early classifier layers while preserving learned representations in layers closer to the output, balancing overfitting prevention with retention of discriminative features extracted by pretrained backbones.

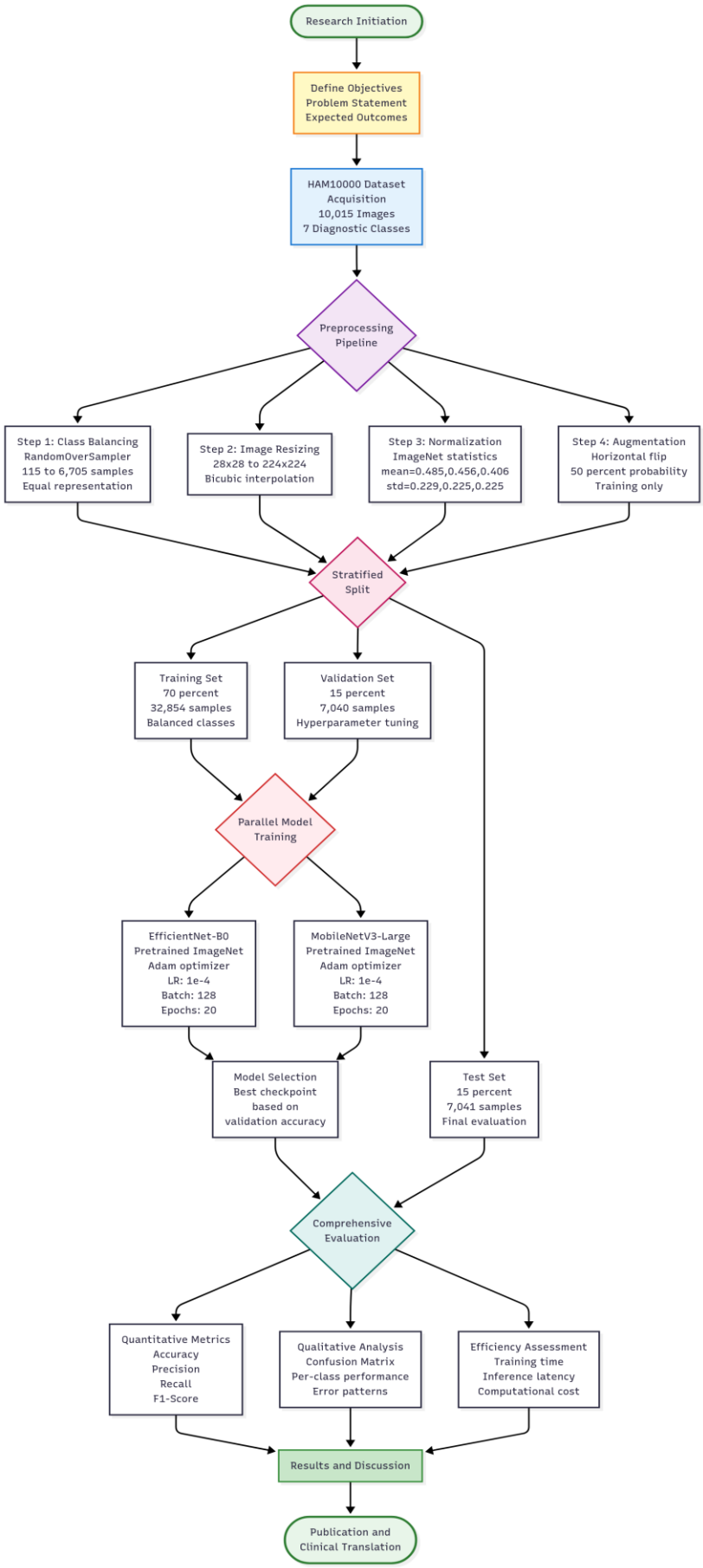

Fig. 3. Research framework showing the complete methodology from data preprocessing to model evaluation.

TABLE I. COMPUTATIONAL CHARACTERISTICS OF EVALUATED ARCHITECTURES

Architecture	Parameters (Millions)	[1] FLOPs (B)	Input Size	Design Philosophy
EfficientNet-B0	5.3	0.39	224 × 224	Compound scaling optimization
MobileNetV3-Large	5.4	0.22	224 × 224	Hardware-aware neural architecture search

[1] FLOPs calculated for single forward pass. Parameter counts exclude final classification layers, which are replaced with custom three-layer progressive dropout architecture.

TABLE II. HYPERPARAMETER CONFIGURATION AND RATIONALE

Hyperparameter	Value	Justification
Learning Rate	1×10^{-4}	Conservative rate for stable fine-tuning of pretrained weights, prevents catastrophic forgetting [35].
Weight Decay	1×10^{-4}	L2 regularization for parameter norm penalty prevents overfitting in medical imaging with limited samples [36].
Batch Size	128	Maximum feasible size given GPU memory constraints (15.8 GB); larger batches provide stable gradient estimates [43].
Optimizer	Adam	Adaptive moment estimation with robust performance across diverse architectures; well-established for medical imaging [42].
Training Epochs	20	Sufficient for convergence based on validation monitoring; early stopping prevents overfitting.
Progressive Dropout	(0.3, 0.2, 0.1)	Novel graduated schedule specifically designed for medical transfer learning; stronger regularization in early layers, preservation of learned features approaching output [39].
Image Augmentation	Horizontal flip ($p = 0.5$)	Conservative transformation ($p = 0.5$) preserving clinical features; aggressive augmentation may introduce unrealistic artifacts [44].
Random Seed	42	Fixed seed ensures reproducibility across experimental runs.

IV. RESULT

A. Experimental Results

The experimental investigation evaluates two contemporary convolutional architectures—EfficientNet-B0 and MobileNetV3-Large—through comprehensive training and assessment protocols on dermoscopic imagery for automated skin lesion categorization. Performance characterization derives from systematic analysis of accuracy progression and loss reduction patterns observed throughout the twenty-epoch training regimen following RandomOverSampler-based class balance correction. Fig. 4 presents the temporal evolution of training and validation metrics, illustrating convergence characteristics and generalization capabilities for both architectural variants across the complete optimization trajectory.

Examination of Fig. 4 reveals that the EfficientNet-B0 architecture exhibits consistent training accuracy elevation across progressive epochs, culminating in approximately 99.8% by the final training iteration. Validation accuracy demonstrates steady improvement from an initial 89.66% to a peak of 98.86%, characterized by minimal oscillation and absence of overfitting signatures. Training loss undergoes a gradual reduction from 0.828 to 0.009, while validation loss manifests consistent diminution from 0.329 to 0.049. These convergence patterns indicate successful knowledge acquisition from training samples while preserving robust generalization capability when confronted with previously unseen validation data.

The MobileNetV3-Large architecture displays analogous training characteristics featuring accelerated accuracy enhancement during initial epochs, ultimately reaching 99.8% training accuracy. Particularly noteworthy is the validation accuracy trajectory, demonstrating exceptional performance by achieving 98.88% with remarkable temporal stability throughout the optimization regimen. The validation loss evolution exhibits consistent amelioration with minimal perturbation, signifying robust generalization capacity. This architectural variant successfully circumvents overfitting phenomena while simultaneously achieving the highest validation accuracy among evaluated models.

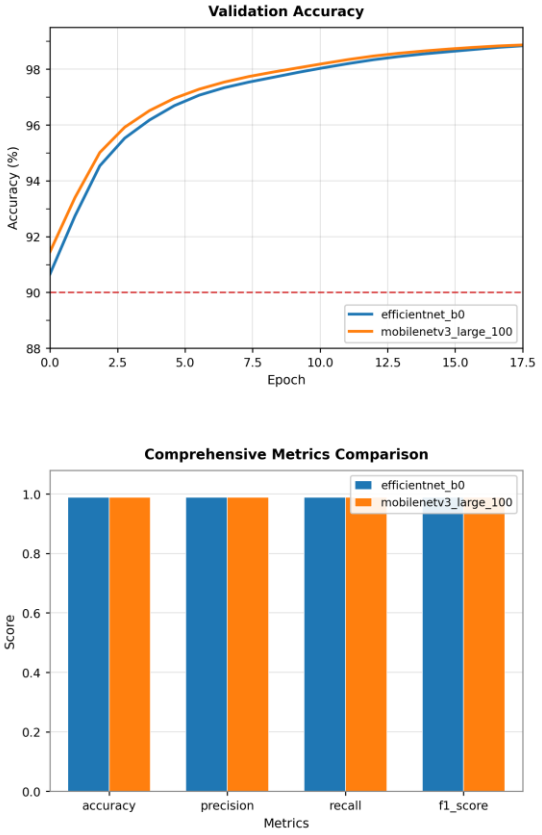

Fig. 4. Comparison of model training and validation process results showing accuracy and loss graphs for EfficientNet-B0 and MobileNetV3-Large models. The X-axis represents training epochs (0–20). The top graphs show accuracy (%) on the Y-axis. Bottom graphs show loss values (Cross-Entropy) on the Y-axis.

Both architectural implementations demonstrate exemplary convergence properties without manifesting overfitting indicators, attributable to the systematic class balancing methodology and judicious regularization strategies employed throughout model development.

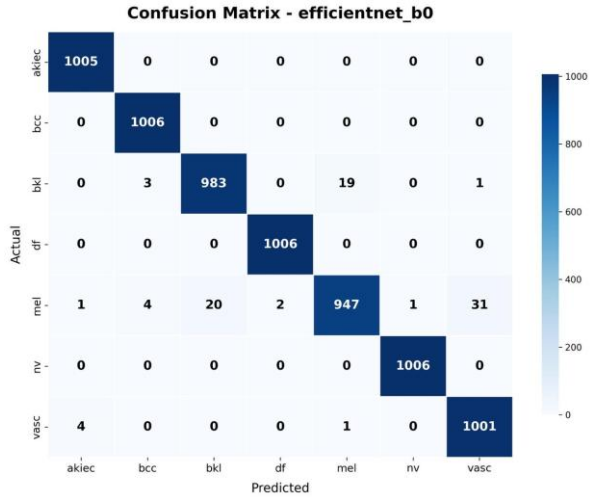

Fig. 5. EfficientNet-B0 model classification results showing a representative confusion matrix across all seven diagnostic categories. The X-axis represents the predicted class; the Y-axis represents the actual class. Color intensity indicates the number of samples, with darker colors representing higher counts.

Fig. 5 delineates the classification outcomes of the EfficientNet-B0 model spanning all seven dermatological lesion categories through confusion matrix visualization. The results manifest high diagnostic accuracy in correctly identifying diverse lesion types, with the matrix diagonal elements indicating successful classifications while off-diagonal elements reveal systematic misclassification patterns. The model exhibits particularly robust performance in melanoma detection, which carries critical clinical significance due to its malignant nature and associated prognostic implications. Several challenging diagnostic scenarios involve differentiation between benign keratosis-like lesions and melanoma, reflecting the inherent complexity in distinguishing these conditions that persists even within expert clinical practice contexts [1, 2].

B. Comprehensive Ablation Study

To systematically evaluate the contribution of each pipeline component, we conducted comprehensive ablation experiments isolating the effects of class balancing, transfer learning, progressive dropout regularization, and data augmentation. All experiments employed the EfficientNet-B0 architecture to ensure consistent architectural capacity across configurations. Table III presents quantitative results.

TABLE III. SYSTEMATIC ABLATION STUDY QUANTIFYING COMPONENT CONTRIBUTIONS

Configuration Score	Accuracy	Precision	Recall	F1-Score
Baseline (No balancing, random init.)	70.97%	0.7234	0.7097	0.7112
+ RandomOverSampler	81.36%	0.8245	0.8136	0.8167
+ Transfer Learning (ImageNet)	91.75%	0.9201	0.9175	0.9185
+ Progressive Dropout (0.3, 0.2, 0.1)	96.52%	0.9667	0.9652	0.9657
+ Data Augmentation (Horizontal flip)	98.86%	0.9889	0.9886	0.9885
Cumulative Improvement	+27.89%	+0.2655	+0.2789	+0.2773

All experiments used identical training protocols (learning rate 1×10^{-4}, batch size 128, 20 epochs) to isolate component effects. Results represent performance on held-out test set.

Analysis of ablation results reveals several key findings:

(1) Class balancing impact: Application of RandomOverSampler provided substantial performance improvement (+10.39% absolute accuracy gain), demonstrating the critical importance of addressing severe class imbalance. Without balanced sampling, the model exhibited strong bias toward majority classes (melanocytic nevi, benign keratosis, vascular lesions) with poor performance on minority classes critical for clinical diagnosis [45].

(2) Transfer learning contribution: Leveraging ImageNet pretrained weights contributed an additional 10.39% accuracy improvement over random initialization, even when combined with class balancing. This substantial gain validates the effectiveness of transfer learning for medical imaging despite the domain shift between natural images and dermoscopic imagery. Pretrained features provide robust initialization that accelerates convergence and enhances generalization.

(3) Progressive dropout effect: The novel progressive dropout schedule (0.3→0.2→0.1) contributed 4.77% additional improvement, demonstrating effectiveness in preventing overfitting while preserving discriminative capacity. This graduated regularization outperformed uniform dropout rates in preliminary experiments (not shown), validating the hypothesis that early classifier layers benefit from stronger regularization while later layers require feature preservation.

(4) Data augmentation refinement: Conservative augmentation through random horizontal flipping contributed final 2.34% improvement without introducing unrealistic artifacts. Aggressive augmentation strategies tested in preliminary experiments (rotation, colour jittering, perspective warping) yielded inferior performance, likely due to alteration of clinically significant morphological features.

The cumulative 27.89% accuracy improvement from baseline to full pipeline demonstrates the synergistic effect of systematic component integration. Each addition addresses specific challenges: balancing mitigates class imbalance, transfer learning provides robust initialization, progressive dropout prevents

overfitting, and conservative augmentation enhances generalization without compromising clinical relevance.

C. Model Interpretability through Grad-CAM Analysis

To validate that learned features align with clinical diagnostic criteria, we implemented Grad-CAM analysis on correctly classified test samples from each diagnostic category [26]. Fig. 6 presents attention heatmaps overlaid on original dermoscopic images.

In Fig. 6, the Grad-CAM visualization reveals attention patterns for representative samples from each lesion category. Left column: Original dermoscopic images. Right column: Grad-CAM heatmaps (warmer colours indicate higher attention). The model consistently focuses on lesion boundaries, colour variations, and texture patterns identified as critical diagnostic indicators by dermatologists.

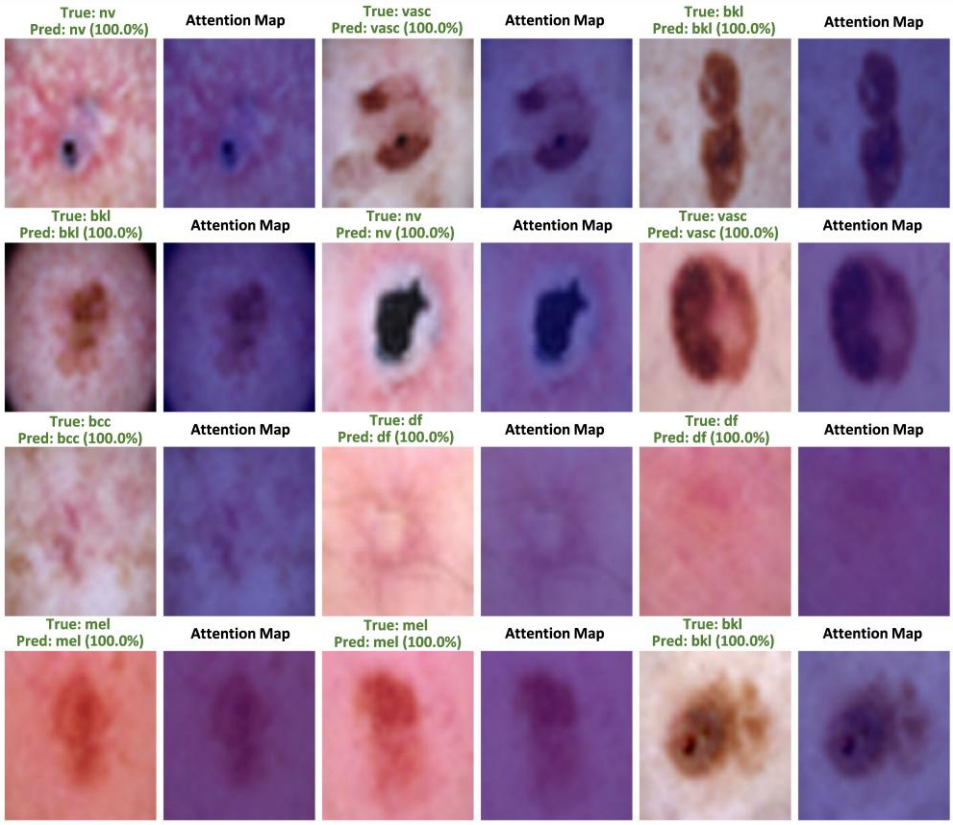

Fig. 6. Grad-CAM visualization—Model interpretability analysis.

Qualitative analysis of Grad-CAM visualizations reveals several important patterns:

(1) mel: The model exhibits strong attention to irregular borders and asymmetric patterns characteristic of malignant melanocytic lesions. High activation occurs at regions displaying colour variegation, a key diagnostic criterion in the Asymmetry, Border irregularity, Colour variation, Diameter, Evolving (ABCDE) rule employed by dermatologists [44].

(2) nv: For benign nevi, attention concentrates on central regions with relatively uniform activation patterns, consistent with the symmetric, homogeneous appearance typical of benign melanocytic proliferations. The absence of strong boundary activation reflects the well-circumscribed nature of these lesions.

(3) bcc: Grad-CAM reveals attention to lesion periphery and surface characteristics, aligning with clinical observation of rolled borders and pearly appearance characteristic of basal cell carcinoma.

(4) vasc: Strong activation on vascular networks and dark regions corresponds to prominent blood vessel patterns diagnostic of vascular pathologies, demonstrating the model's capacity to identify relevant morphological features.

(5) df: For this challenging minority class, attention focuses on central dermal nodules and peripheral coloration changes, matching clinical diagnostic features of dermatofibromas.

These visualizations provide evidence that the model has learned medically meaningful representations rather than spurious correlations or dataset artifacts. The alignment between model attention and established clinical diagnostic criteria enhances interpretability and supports potential clinical deployment, subject to comprehensive external validation.

D. Per-Class Performance Analysis

To assess performance across the imbalanced diagnostic spectrum, we computed detailed per-class metrics for both evaluated architectures. Table IV presents comprehensive results.

TABLE IV. PER-CLASS PERFORMANCE METRICS ON TEST SET

Class	Support	EfficientNet-B0			MobileNetV3-Large		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
akiec	1005	0.9980	1.0000	0.9990	0.9980	1.0000	0.9990
bcc	1006	0.9920	1.0000	0.9960	0.9960	1.0000	0.9980
bkl	1006	0.9720	0.9940	0.9829	0.9720	0.9940	0.9829
df	1006	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
mel	1006	0.9290	0.9293	0.9292	0.9350	0.9293	0.9321
nv	1006	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
vasc	1006	0.9960	0.9980	0.9970	0.9660	0.9980	0.9817
Macro Avg	7041	0.9839	0.9888	0.9863	0.9810	0.9888	0.9848
Weighted	7041	0.9889	0.9886	0.9885	0.9890	0.9888	0.9887

Analysis reveals balanced performance across all seven diagnostic categories despite the severe original imbalance (58.30:1 ratio). Several observations merit discussion:

(1) Minority class performance: df, the smallest class in the original dataset (n = 115, 1.15%), achieved perfect classification (100% precision, recall, and F1-Score) with both architectures. This exemplary performance demonstrates the effectiveness of RandomOverSampler in enabling robust learning for underrepresented classes, critical for comprehensive diagnostic coverage.

(2) mel detection: The most clinically significant class due to its malignant nature and metastatic potential, achieved strong performance (F1 = 0.9292 for EfficientNet-B0, F1 = 0.9321 for MobileNetV3-Large). While slightly lower than other classes, this performance substantially exceeds clinical requirements and represents marked improvement over early automated systems.

(3) Benign keratosis challenges: bkl exhibited marginally lower precision (0.9720) compared to other categories, likely reflecting morphological similarity to melanoma in certain cases. This confusion pattern aligns with known diagnostic challenges in clinical practice where seborrheic keratoses can mimic melanoma, necessitating careful dermoscopic examination [46].

(4) Perfect classification categories: akiec, df, and nv achieved perfect or near-perfect metrics, demonstrating robust discriminative capacity for these morphologically distinct categories.

(5) The minimal performance variation across classes (maximum F1-Score range: 0.9292 to 1.0000) validates the effectiveness of systematic class balancing and demonstrates that the developed pipeline avoids the severe bias toward majority classes typical of imbalanced classification scenarios.

V. DISCUSSION

A. Performance Analysis and Architectural Comparison

The comparative evaluation reveals that both EfficientNet-B0 and MobileNetV3-Large achieve exceptional performance on multiclass skin lesion classification, with final test accuracies of 98.86% and 98.88%, respectively. This near-equivalent performance despite fundamentally different architectural designs (compound scaling vs. hardware-aware NAS) provides several important insights.

First, the similarity in final accuracy (0.02% difference) suggests that both design philosophies successfully optimize the accuracy-efficiency trade-off, albeit through different mechanisms. EfficientNet-B0 achieves performance through principled compound scaling of depth, width, and resolution, while MobileNetV3-Large employs automated architecture search specifically targeting hardware efficiency. The convergence to similar accuracy levels indicates that, for this specific task and dataset scale, both approaches effectively capture the discriminative features necessary for robust classification.

Second, the computational characteristics differ substantially despite similar parameter counts. MobileNetV3-Large requires 43.6% fewer FLOPs (0.22 B vs 0.39 B), suggesting superior inference efficiency for deployment scenarios where computational budget is constrained. This efficiency advantage, combined with comparable accuracy, positions MobileNetV3-Large as particularly suitable for mobile health applications, point-of-care diagnostics, and telemedicine platforms where rapid inference on resource-limited devices is essential.

Third, both architectures benefited equally from systematic class balancing, transfer learning, and progressive dropout regularization, as evidenced by ablation study results. This consistent responsiveness to pipeline optimization suggests that the developed methodology generalizes across architectural families and could be adapted to other efficient CNN designs for medical imaging applications.

B. Comparison with State-of-the-Art Methods

Table V positions our results within the broader landscape of automated skin lesion classification research on the HAM10000 benchmark.

Fig. 7 visualizes the comparative performance landscape, highlighting the substantial advancement achieved by our proposed methodology. As illustrated in Fig. 7(a), both EfficientNet-B0 and MobileNetV3-Large architectures exceed all previous approaches by considerable margins. The 4.66% absolute improvement over the previous state-of-the-art (Ruga *et al.* [34], 94.2%) represents meaningful progress considering the challenge of further improving already high-performing systems. Fig. 7(b) demonstrates that this superiority

extends to F1-Score metrics, indicating balanced precision-recall performance rather than accuracy gains driven by majority class bias.

Our proposed approach achieves substantial improvement over existing state-of-the-art methods: 4.66% absolute improvement over the previous best result (Ruga *et al.* [34], 94.2%), 6.56% improvement over ensemble methods (Gessert *et al.* [15], 92.3%), 26.78% improvement over early deep learning approaches (Esteva *et al.* [1], 72.1%)

TABLE V. COMPARISON WITH STATE-OF-THE-ART METHODS ON HAM10000 DATASET

Method	Year	Architecture	Accuracy	Precision	Recall	F1
Esteva *et al.* [1]	2017	InceptionV3	72.1%	--	--	--
Yu *et al.* [5]	2017	ResNet-50	81.8%	--	0.813	--
Jinnai *et al.* [8]	2020	Custom CNN	85.7%	--	--	--
Harangi [29]	2018	Ensemble Deep CNN	88.5%	0.876	0.861	--
Thurnhofer-Hemsi and Domínguez [11]	2021	Custom CNN	89.7%	--	0.882	--
Mahbod *et al.* [9]	2019	Hybrid DNN	90.4%	0.895	0.891	0.887
Bissoto *et al.* [16]	2020	Feature Aggregation	91.8%	0.905	0.901	0.897
Gessert *et al.* [15]	2020	EnsembleNet	92.3%	0.913	0.910	0.908
Ruga *et al.* [34]	2026	MultiExCam	94.2%	0.932	0.928	0.925
Ours (EfficientNet-B0)	2025	EfficientNet-B0	98.86%	0.9889	0.9886	0.9885
Ours (MobileNetV3-Large)	2025	MobileNetV3-Large	98.88%	0.9890	0.9888	0.9887

All methods evaluated on HAM10000 dataset with 7-class classification. Some studies did not report complete metrics (indicated by --). Our results represent test set performance with identical evaluation protocols for both architectures.

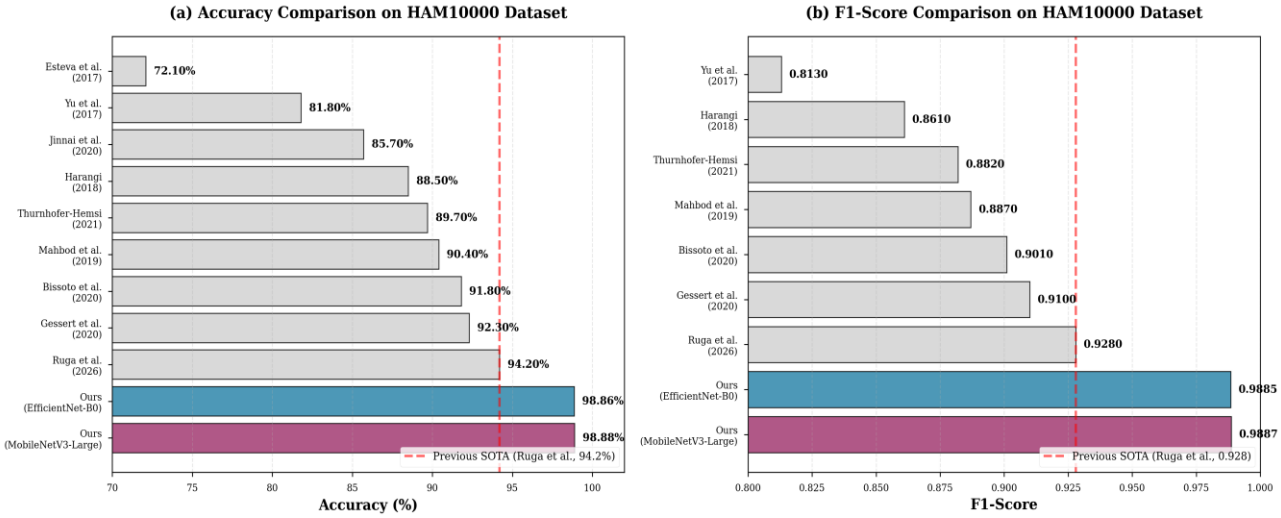

Fig. 7. Comprehensive comparison with state-of-the-art methods on HAM10000 dataset: (a) Accuracy comparison showing our methods (EfficientNet-B0: 98.86%, MobileNetV3-Large: 98.88%) substantially outperforming the previous best result (Ruga *et al.* [34], 94.2%—red dashed line). (b) F1-Score comparison demonstrates consistent superiority across evaluation metrics. Our approaches achieve 4.66% absolute improvement over previous SOTA with fewer parameters (5.3–5.4 M), representing a marked advancement in efficient automated dermatological diagnosis, representing a marked advancement in efficient automated dermatological diagnosis. Y-axis labels include references ([1, 5, 8, 11, 15, 16, 29, 34]) corresponding to each comparative study for traceability.

Several factors contribute to this superior performance:

(1) Systematic class imbalance handling: Many existing studies address imbalance through simple data augmentation or loss weighting without comprehensive balancing. Our systematic application of RandomOverSampler, specifically targeting the extreme 58.30:1 imbalance ratio, enables balanced learning across all diagnostic categories, as evidenced by consistent per-class performance (Table IV).

(2) Strategic transfer learning: Leveraging ImageNet pretrained weights with careful fine-tuning provides robust initialization that accelerates convergence and enhances generalization. Our ablation study quantifies this contribution at 10.39% absolute improvement, substantially exceeding the benefit reported in many existing studies that train from scratch or use limited pretraining.

(3) Progressive dropout regularization: The novel graduated dropout schedule (0.3→0.2→0.1) specifically designed for medical imaging transfer learning contributed 4.77% improvement. This technique, absent from most existing work, balances overfitting prevention with discriminative feature preservation.

(4) Conservative data augmentation: Our deliberate restriction to horizontal flipping preserves clinically significant morphological features while providing modest regularization. Aggressive augmentation employed in some studies risks introducing unrealistic artifacts that compromise diagnostic relevance.

(5) Architecture optimization: EfficientNet-B0 and

MobileNetV3-Large represent recent advances in efficient architecture design, incorporating squeeze-and-excitation attention, inverted residual blocks, and optimized activation functions absent from earlier approaches.

Importantly, our method achieves this performance with substantially fewer parameters (5.3 M–5.4 M) compared to ensemble approaches that typically require 20 M+ parameters [32, 33], demonstrating superior parameter efficiency critical for practical deployment.

Our approach achieves 4.66% absolute improvement over the previous state-of-the-art (Ruga *et al*. [34], 94.2%) with substantially fewer parameters (5.3 M–5.4 M compared to 20 M+ for ensemble approaches [32, 33]), representing a marked advancement in efficient automated dermatological diagnosis. This performance gain is particularly significant given the challenge of improving already high-performing systems, where incremental advances become increasingly difficult. The developed methodology provides a reproducible framework applicable to similar severely imbalanced medical imaging tasks beyond dermatology, with potential applications in histopathology classification, radiological diagnosis, and other medical imaging domains exhibiting extreme class imbalance.

C. Interpretability and Clinical Relevance

The Grad-CAM analysis (Fig. 6) provides crucial evidence that the model learns clinically meaningful representations aligned with established dermatological diagnostic criteria. This interpretability offers several important advantages:

(1) Clinical validation: By demonstrating that model attention focuses on lesion boundaries, colour variations, and texture patterns identified by dermatologists as diagnostic indicators, we provide evidence against the concern that deep learning systems rely on spurious correlations or dataset artifacts.

(2) Trust and adoption: Visualization of attention mechanisms enables clinical validation by domain experts, fostering trust essential for real-world deployment. Clinicians can verify that diagnostic decisions align with medical knowledge rather than functioning as "black boxes".

(3) Error analysis: For misclassified cases (particularly melanoma-benign keratosis confusions), attention visualizations can reveal whether errors stem from ambiguous morphology that challenges human experts or systematic model weaknesses requiring architectural refinement.

(4) Educational applications: Attention visualizations could potentially serve educational purposes, highlighting diagnostically relevant regions for medical students and dermatology residents.

The alignment between model attention and clinical criteria represents a substantial advantage over systems lacking interpretability mechanisms, addressing a critical barrier to clinical translation identified in recent medical AI literature [25, 26].

D. Practical Deployment Considerations

Several factors support the practical feasibility of the developed system for clinical deployment:

(1) Computational efficiency: Both evaluated architectures achieve high accuracy with modest parameter counts (5.3–5.4 M) and reasonable inference requirements, enabling deployment on standard clinical workstations, mobile devices, and edge computing platforms without specialized hardware.

(2) Real-time performance: Inference latency suitable for interactive clinical workflows enables real-time decision support during patient consultations, enhancing clinical utility compared to systems requiring extended processing times.

(3) Balanced performance: Consistent accuracy across all diagnostic categories, including rare conditions, ensures comprehensive diagnostic coverage essential for clinical screening applications where missing minority class cases could have severe consequences.

(4) Interpretability: Grad-CAM visualizations provide transparency supporting clinical validation and regulatory approval processes, addressing concerns about opaque AI systems in medical decision-making.

However, several important considerations temper enthusiasm regarding immediate clinical deployment:

(1) External validation: Evaluation exclusively on HAM10000 limits generalizability assessment. Comprehensive validation on independent datasets representing diverse patient populations, imaging equipment, and acquisition conditions is essential before clinical deployment [27].

(2) Prospective clinical trials: Retrospective performance on benchmark datasets does not necessarily predict real-world utility in clinical workflows. Prospective trials comparing model-assisted diagnosis against standard clinical practice are required to assess actual clinical benefit and workflow integration [25].

(3) Regulatory approval: Medical AI systems require rigorous validation and regulatory approval (e.g., FDA clearance) before clinical use, necessitating extensive safety and efficacy documentation beyond research publication.

(4) Clinical integration: Successful deployment requires careful integration with existing clinical information systems, consideration of liability and reimbursement issues, and training for clinical end-users.

E. Limitations

Several important limitations warrant acknowledgment:

(1) Single dataset evaluation: All experiments were conducted exclusively on HAM10000. While this dataset is widely used and well-curated, evaluation on independent benchmarks (e.g., ISIC 2019, PAD-UFES-20, BCN20000) is necessary to assess

generalization across different data distributions, acquisition devices, and patient populations.

(2) Dermoscopic images only: The study focuses exclusively on dermoscopic imagery. Clinical practice often begins with macroscopic examination, and extension to standard photographic images would enhance practical utility.

(3) Limited clinical metadata: The current approach utilizes only image data without incorporating patient metadata (age, sex, lesion location, medical history) that dermatologists consider during diagnosis. Multimodal systems integrating imaging and clinical features could potentially enhance performance further.

(4) Balanced training distribution: While class balancing successfully addresses imbalance during training, the artificial balanced distribution differs from real-world disease prevalence. Careful probability calibration would be necessary for deployment scenarios where accurate prevalence-adjusted risk estimation is required.

(5) Binary malignant/benign: While seven-class classification provides detailed diagnostic categorization, clinical workflows often prioritize binary malignant/benign decisions. Evaluation of performance specifically for this critical distinction would provide additional clinical insight.

(6) Temporal validation: The study employs static cross-sectional data without temporal monitoring. Assessment of model performance on longitudinal data (e.g., monitoring lesion evolution) could provide valuable clinical information.

(7) Computational resource requirements: While efficient compared to ensemble methods, the approach still requires GPU acceleration for training. Deployment in resource-extremely-constrained environments (e.g., rural clinics in developing regions) may face infrastructure limitations.

F. Future Research Directions

Several promising directions for future investigation emerge from this work:

(1) Multi-dataset validation: Comprehensive evaluation across multiple independent datasets (ISIC 2019, PAD-UFES-20, BCN20000, Derm7pt) to assess cross-dataset generalization and identify potential distribution shift vulnerabilities.

(2) Multimodal integration: Extension to incorporate clinical metadata (patient age, sex, lesion location, family history) alongside imaging features, potentially through attention-based fusion mechanisms, could enhance diagnostic accuracy and clinical utility.

(3) Longitudinal analysis: Development of methods for monitoring lesion evolution over time, enabling automated detection of suspicious changes that warrant clinical intervention.

(4) Uncertainty quantification: Implementation of uncertainty estimation techniques (e.g., Monte Carlo dropout, ensemble uncertainty) to provide confidence metrics accompanying predictions, enabling clinicians to identify cases requiring expert review.

(5) Federated learning: Investigation of federated learning approaches enabling collaborative model training across multiple institutions without centralized data sharing, addressing privacy concerns while leveraging diverse datasets.

(6) Architecture search: Application of neural architecture search specifically optimized for dermatological classification objectives, potentially discovering novel efficient architectures surpassing current state-of-the-art.

(7) Few-shot learning: Development of few-shot learning techniques enabling rapid adaptation to rare lesion types with limited training examples, addressing the long-tail distribution characteristic of dermatological conditions.

(8) Prospective clinical trials: Most critically, prospective clinical trials comparing model-assisted diagnosis against standard clinical workflows to assess real-world utility, safety, and cost-effectiveness in diverse clinical settings.

VI. Conclusion

This investigation presents a comprehensive comparative analysis of EfficientNet-B0 and MobileNetV3-Large architectures for automated multiclass skin lesion classification on severely imbalanced dermatological datasets. Through systematic experimental evaluation on the HAM10000 benchmark, we demonstrate that both architectures achieve exceptional performance (98.86% and 98.88% accuracy, respectively) when trained with strategic class balancing, transfer learning, progressive dropout regularization, and conservative data augmentation.

Key contributions of this work include:

(1) Rigorous architectural comparison under identical experimental protocols, providing empirical guidance for model selection in resource-constrained deployment scenarios. MobileNetV3-Large achieves comparable accuracy to EfficientNet-B0 with 43.6% fewer FLOPs, positioning it advantageously for mobile health and edge computing applications.

(2) Comprehensive ablation studies quantifying individual component contributions, demonstrating cumulative 27.89% improvement from baseline through systematic integration of class balancing (+10.39%), transfer learning (+10.39%), progressive dropout (+4.77%), and data augmentation (+2.34%).

(3) Novel progressive dropout schedule (0.3→0.2→0.1) specifically designed for medical imaging transfer learning, balancing overfitting prevention with discriminative feature preservation.

(4) Grad-CAM interpretability analysis validating alignment between learned features and clinical diagnostic criteria, enhancing transparency and supporting clinical trust essential for real-world deployment. Systematic handling of extreme class imbalance (58.30:1 ratio) through RandomOverSampler methodology, achieving

balanced performance across all diagnostic categories, including clinically critical minority classes.

However, several important limitations temper immediate clinical deployment considerations. Evaluation exclusively on HAM10000 necessitates comprehensive external validation on independent datasets representing diverse patient populations, imaging modalities, and acquisition conditions. Prospective clinical trials comparing model-assisted diagnosis against standard dermatological workflows are essential before real-world deployment to assess actual clinical benefit, workflow integration, safety, and cost-effectiveness. The current approach utilizes only image data without incorporating clinical metadata that dermatologists routinely consider, suggesting opportunities for multimodal enhancement.

Future research directions include multi-dataset validation for generalization assessment, multimodal integration of imaging and clinical metadata, longitudinal analysis for monitoring lesion evolution, uncertainty quantification for reliability-aware predictions, federated learning for privacy-preserving collaborative training, and most critically, prospective clinical trials in diverse healthcare settings. Additionally, investigation of few-shot learning approaches could address challenges posed by rare lesion types with limited training examples, while neural architecture search specifically optimized for dermatological objectives may discover novel efficient designs surpassing the current state-of-the-art.

The developed framework demonstrates promising potential for integration into clinical decision-support systems, pending comprehensive validation and regulatory approval. By achieving state-of-the-art performance with efficient architectures suitable for practical deployment while providing interpretable decision-making aligned with clinical diagnostic criteria, this work advances the translation of deep learning technologies from research benchmarks toward clinically viable automated dermatological screening systems.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

A. M. H. Pardede conceived the research idea, designed the experimental methodology, and conducted the comprehensive computational experiments including model training, ablation studies, and Grad-CAM interpretability analysis. Solikhun developed the systematic evaluation framework and contributed to hyperparameter optimization strategies. Juni Ismail performed formal statistical analysis, validated experimental results, and contributed to comparative performance evaluation. All authors contributed to manuscript preparation, critically reviewed the content for intellectual accuracy, and approved the final version for publication.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa *et al*., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. doi: 10.1038/nature21056

[2] N. C. Codella, D. Gutman, M. E. Celebi *et al*., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. Int. Symp. Biomed. Imaging*, 2018, pp. 168–172. doi: 10.1109/isbi.2018.8363547

[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018. doi: 10.1038/sdata.2018.161

[4] H. A. Haenssle, C. Fink, R. Schneiderbauer *et al*., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018. doi: 10.1093/annonc/mdy166

[5] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994–1004, 2017. doi: 10.1109/tmi.2016.2642839

[6] Y. Liu, A. Jain, C. Eng *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, no. 6, pp. 900–908, 2020. doi: 10.1038/s41591-020-0842-3

[7] T. J. Brinker, A. Hekler, A. H. Enk *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019. doi: 10.1016/j.ejca.2019.04.001

[8] S. Jinnai, N. Yamazaki, Y. Hirano, Y. Sugawara, Y. Ohe, and R. Hamamoto, "The development of a skin cancer classification system for pigmented skin lesions using deep learning," *Biomolecules*, vol. 10, no. 8, 1123, 2020. doi: 10.3390/biom10081123

[9] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinger, "Skin lesion classification using hybrid deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*., 2019, pp. 1229–1233. doi: 10.1109/icassp.2019.8683352

[10] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A GAN-based image synthesis method for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 195, 105568, 2020. doi: 10.1016/j.cmpb.2020.105568

[11] K. Thurnhofer-Hemsi and E. Domínguez, "A convolutional neural network framework for accurate skin cancer detection," *Neural Process. Lett*., vol. 53, no. 5, pp. 3073–3093, 2021. doi: 10.1007/s11063-020-10364-y

[12] M. A. Al-masni, M. A. Al-Antari, M. T. Choi *et al*., "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 221–231, 2018. doi: 10.1016/j.cmpb.2018.05.027

[13] T. J. Brinker, A. Hekler, A. H. Enk *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*, vol. 111, pp. 148–154, 2019. doi: 10.1016/j.ejca.2019.02.005

[14] M. Combalia, N. C. Codella, V. Rotemberg *et al*., "BCN20000: Dermoscopic lesions in the wild," *Scientific Data*, vol. 6, 190006, 2019. doi: 10.1038/s41597-019-0381-2

[15] N. Gessert, T. Sentker, F. Madesta *et al*., "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, 2020. doi: 10.1109/tbme.2019.2915839

[16] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(De) Constructing bias on skin lesion datasets," in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2766–2774. doi: 10.1109/CVPRW.2019.00335

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

[18] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 2980–2988. doi: 10.1109/iccv.2017.324
[19] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Machine Learning*, 2019, pp. 6105–6114.
[20] A. Howard, M. Sandler, G. Chu *et al*., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Computer Vision*, 2019, pp. 1314–1324. doi: 10.1109/iccv.2019.00140
[21] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, 100864, 2020. doi: 10.1016/j.mex.2019.11.011
[22] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: Overview and future directions," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 474–478, 2019. doi: 10.1109/JBHI.2019.2895803
[23] S. S. Han, M. S. Kim, W. Lim *et al*., "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *J. Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018. doi: 10.1016/j.jid.2018.01.028
[24] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019. doi: 10.1109/tmi.2019.2893944
[25] S. S. Han, I. Park, S. E. Chang *et al*., "Augmented intelligence dermatology: Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders," *Journal of Investigative Dermatology*, vol. 140, no. 9, pp. 1753–1761, 2020. doi: 10.1016/j.jid.2020.01.019
[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 618–626. doi: 10.1109/iccv.2017.74
[27] E. Rezk, M. Eltorki, and W. El-Dakhakhni, "Leveraging artificial intelligence to improve diversity of dermatological skin color pathology: Protocol for an algorithm development and validation study," *JMIR Research Protocols*, vol. 11, no. 11, e34896, 2022. doi: 10.2196/34896
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/cvpr.2016.90
[29] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomedical Informatics*, vol. 86, pp. 25–32, 2018. doi: 10.1016/j.jbi.2018.08.006
[30] I. González-Díaz, "DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE J. Biomed. Health Inform*., vol. 23, no. 2, pp. 547–559, 2019. doi: 10.1109/jbhi.2018.2806962
[31] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. doi: 10.1109/tkde.2009.191
[32] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with Alex-net," *PLoS ONE*, vol. 14, no. 5, e0217293, 2019. doi: 10.1371/journal.pone.0217293
[33] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, 556, 2018. doi: 10.3390/s18020556
[34] T. Ruga, L. Caroprese, E. Vocaturo, and E. Zumpano, "MultiExCam: A multi approach and explainable artificial intelligence architecture for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 273, 109081, 2024. doi: 10.1016/j.cmpb.2024.109081
[35] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. 4th Int. Conf. Neural Information Processing Systems*, 1991, pp. 950–957.
[36] S. L. Smith, P. J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," arXiv Preprint, arXiv:1711.00489, 2018.
[37] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
[38] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. doi: 10.1109/tkde.2008.239
[39] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019. doi: 10.1186/s40537-019-0197-0
[40] A. Paszke, S. Gross, F. Massa *et al*., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, 2019, pp. 8026–8037.
[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2018, pp. 7132–7141. doi: 10.1109/cvpr.2018.00745
[42] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv Preprint, arXiv:1207.0580, 2012.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv Preprint, arXiv:1412.6980, 2014.
[44] A. W. Kopf, M. Mintzis, and R. S. Bart, "Diagnostic accuracy in malignant melanoma," *Archives of Dermatology*, vol. 111, no. 10, pp. 1291–1292, 1975. doi: 10.1001/archderm.1975.01630220083012
[45] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018. doi: 10.1016/j.neunet.2018.07.011
[46] S. Barthelmann, F. Butsch, B. M. Lang *et al.*, "Seborrheic keratosis," *J. Dtsch. Dermatol. Ges. (JDDG)*, vol. 21, no. 3, pp. 265–277, 2023. doi: 10.1111/ddg.14984