RESEARCH ARTICLE

Comparative Study of Transfer Learning Strategies for Multi-Class Skin Lesion Classification: Architectures, Fine-Tuning, and Data Augmentation

Dingguo Wang[1,*] **and Yudi Chen**[1]

[1] School of Optoelectronic Engineering, Xi'an Technological University, Xi'an 710021, China

Abstract

Skin lesion classification is critical in dermatological diagnosis, where early and accurate identification of malignant lesions can significantly improve patient outcomes. Deep learning approaches, particularly transfer learning with pre-trained CNNs, have demonstrated remarkable performance in automated dermoscopic image analysis. However, the optimal configuration of transfer learning components—including backbone architecture, fine-tuning strategy, and data augmentation intensity—remains an open question. In this paper, we present a systematic comparative study on the HAM10000 dataset, evaluating three CNN architectures (ResNet50, DenseNet121, EfficientNet-B0), three fine-tuning strategies (full, partial, classifier-only), and three data augmentation strategies (basic, moderate, aggressive). Our experiments reveal: (1) all three architectures achieve comparable per-class F1-scores under basic augmentation, with no statistically significant differences (Welch's t-test, p > 0.05), despite EfficientNet-B0 reaching 100% overall validation accuracy; (2) full fine-tuning yields the highest accuracy (86.19%) and AUC (99.42%) at increased computational cost; (3) basic augmentation achieves the best performance (accuracy=91.43%, AUC=99.30%), while aggressive augmentation degrades results due to excessive distortion of medical image features. Ablation studies further demonstrate: label smoothing and inverse class frequency weights produce a small positive interaction effect; gradient clipping at norm 1.0 is essential for training stability (without it, training collapses); and a backbone learning rate of 5×10^{-4} yields optimal partial fine-tuning performance. McNemar's test confirms no significant difference in BKL vs. MEL misclassification patterns between ResNet50 and EfficientNet-B0 ($p > 0.05$). These findings provide practical guidelines for configuring transfer learning pipelines in medical image classification.

Keywords: transfer learning, skin lesion classification, convolutional neural networks, fine-tuning strategies, data augmentation, HAM10000, medical image analysis, ablation study.

Submitted: 13 April 2026
Accepted: 06 May 2026
Published: 03 June 2026

Vol. 2, **No.** 3, 2026.
10.62762/JIAP.2026.390206

***Corresponding author:**
✉ Dingguo Wang
wangdingguo@st.xatu.edu.cn, 3525986157@qq.com

Citation
Wang, D., & Chen, Y. (2026). Comparative Study of Transfer Learning Strategies for Multi-Class Skin Lesion Classification: Architectures, Fine-Tuning, and Data Augmentation. *ICCK Journal of Image Analysis and Processing*, 2(3), 153–167.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

Skin cancer represents one of the most prevalent forms of cancer worldwide, with melanoma alone accounting for the majority of skin cancer-related deaths despite constituting a small fraction of cases [1]. Early detection and accurate classification of skin lesions are paramount for effective treatment and improved survival rates. Dermoscopy, a non-invasive imaging technique, enables visualization of subsurface skin structures but requires specialized expertise for interpretation [2]. The inter-observer variability and limited availability of expert dermatologists have motivated the development of automated computer-aided diagnosis (CAD) systems.

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized medical image analysis [3]. Transfer learning, which leverages knowledge from large-scale source domains (e.g., ImageNet [4]) to target medical imaging tasks, has emerged as a de facto standard due to the scarcity of labeled medical data [5]. The seminal work by Esteva et al. [1] demonstrated that deep CNNs can achieve dermatologist-level performance in skin cancer classification, catalyzing extensive research in this domain. Recent advances have explored Vision Transformers [36] and hybrid CNN-Transformer architectures [26] for skin lesion classification, while ConvNeXt [37] has emerged as a modernized pure convolutional alternative that rivals Transformer-based models.

Despite the widespread adoption of transfer learning for skin lesion classification, several fundamental questions remain insufficiently addressed. First, while numerous CNN architectures have been proposed, systematic comparisons under controlled experimental settings are limited [6]. Second, the choice of fine-tuning strategy—ranging from training only the classifier head to updating all network parameters—significantly impacts performance, yet the trade-offs between computational cost and accuracy gains are not well characterized for multi-class dermoscopic image classification [7]. Third, data augmentation is essential for mitigating overfitting and class imbalance, but the effect of augmentation intensity on medical image classification has received limited systematic investigation [8].

To address these gaps, we conduct a comprehensive comparative study on the HAM10000 dataset [2], a widely-used benchmark for skin lesion classification comprising 10,015 dermoscopic images across seven diagnostic categories. Our contributions are:

1. We systematically compare three representative CNN architectures—ResNet50 [9], DenseNet121 [10], and EfficientNet-B0 [11]—under identical training conditions with both basic and moderate augmentation, using Welch's t-test and Levene's test for rigorous statistical comparison.

2. We evaluate three fine-tuning strategies (full, partial, and classifier-only) with differentiated learning rates, quantifying the performance-computation trade-off for each approach, and conduct a backbone learning rate ablation for partial fine-tuning.

3. We investigate the impact of data augmentation intensity (basic, moderate, and aggressive) on classification performance, including multi-seed stability analysis and test-time augmentation evaluation for aggressive augmentation.

4. We provide comprehensive ablation studies on label smoothing, class weights, gradient clipping, and their interactions, along with McNemar's test for confusion matrix comparison.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents experimental results and analysis. Section 5 discusses the findings, and Section 6 concludes the paper.

2 Related Work

2.1 Deep Learning for Skin Lesion Classification

The application of deep learning to dermatology has progressed rapidly since the breakthrough results of Esteva et al. [1], who trained an Inception-v3 network on approximately 129,000 clinical images. Han et al. [19] extended this work to multiple lesion types using a ResNet-based architecture. The ISIC challenge series [12] has fostered community-driven benchmarking, with top solutions employing ensemble methods and multi-scale feature extraction [13]. More recently, Chaturvedi et al. [14] proposed a multi-class skin cancer classification framework using deep convolutional neural networks, demonstrating improved performance across multiple lesion categories through systematic architecture design. Recent surveys by Zhang et al. [24] and Shakya et al. [25] have comprehensively reviewed deep learning advancements in skin disease diagnosis,

highlighting the growing effectiveness of CNN-based approaches while identifying persistent challenges in dataset diversity and model generalizability. Naeem et al. [26] proposed MedFusionNet, which fuses ConvNeXt and Vision Transformer through adaptive attention mechanisms, achieving 98.80% accuracy on HAM10000. Lightweight architectures have also gained traction; for instance, ScNet [27] employs MobileNetV3 with an adaptive classification module for efficient dermoscopic image classification. Multi-model ensemble strategies combining ResNet-50 and VGG-16 with optimized feature selection have achieved 98.5% accuracy on ISIC2017 [28].

2.2 Transfer Learning in Medical Imaging

Tajbakhsh et al. [5] provided a systematic evaluation of full training versus fine-tuning for medical image analysis, demonstrating that fine-tuning consistently outperforms training from scratch. Shin et al. [15] investigated the impact of dataset characteristics on transfer learning performance. Yosinski et al. [16] analyzed feature transferability across network layers, showing that earlier layers capture more generic features while later layers are more task-specific. Raghu et al. [7] specifically examined transfer learning for medical imaging, finding that even models trained on limited medical data can outperform ImageNet-pretrained models when sufficient target-domain data is available. More recently, Ozdemir et al. [29] explored meta-transfer learning for skin disease classification under long-tail distributions, demonstrating that combining transfer learning with few-shot learning significantly enhances performance for rare skin diseases. A comprehensive analysis by Ali et al. [30] compared deep learning and transfer learning techniques for skin cancer classification, finding that hybrid feature extraction with SVM classifiers achieves superior results. The adaptation of foundation models for medical image analysis has also been systematically reviewed [31], highlighting the trade-offs between full fine-tuning, parameter-efficient methods, and classifier-only training across diverse medical imaging tasks.

2.3 Data Augmentation for Medical Images

Data augmentation is crucial for addressing the limited sample sizes typical of medical imaging datasets. Shorten and Khoshgoftaar [8] provided a comprehensive survey of augmentation techniques. Zhong et al. [17] proposed Random Erasing as an effective augmentation strategy. For skin lesion classification specifically, Goyal and Rajapakse [18] demonstrated that combining augmentation with ensemble learning improves classification performance. However, the optimal intensity of augmentation for dermoscopic images remains underexplored, as aggressive transformations may distort clinically relevant features. Recent work by Kim et al. [32] proposed a diffusion-based data augmentation method with fine-grained detail preservation for skin disease images on HAM10000, generating high-quality synthetic data through latent space interpolation. Musthafa et al. [33] introduced an optimized CNN architecture with checkpoint-based training and balanced data augmentation on HAM10000, achieving 97.78% accuracy. The challenge of class imbalance in medical imaging has been further addressed through SMOTE-based oversampling combined with deep learning [34], and through unbiased sample selection methods for class-imbalanced noisy datasets [35].

3 Methodology

3.1 Dataset

We utilize the HAM10000 (Human Against Machine with 10,000 training images) dataset [2], which contains 10,015 dermoscopic images collected from different populations. The dataset encompasses seven diagnostic categories: Actinic Keratoses (AK), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevi (NV), and Vascular Lesions (VASC). The dataset exhibits significant class imbalance, with NV comprising approximately 67% of all samples. Images are resized to 224×224 pixels and normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). The dataset is partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling.

3.2 Model Architectures

We evaluate three CNN architectures with distinct design philosophies:

ResNet50 [9] employs residual connections to address the vanishing gradient problem, enabling training of very deep networks. The model contains approximately 25.6M parameters. We replace the final fully connected layer with a custom classifier comprising two dropout layers ($p = 0.5$ and $p = 0.25$), a 512-dimensional hidden layer with ReLU activation and BatchNorm, and a 7-class output layer.

DenseNet121 [10] introduces dense connectivity

patterns where each layer receives feature maps from all preceding layers. With approximately 8.0M parameters, DenseNet121 achieves strong performance with fewer parameters through feature reuse. The same custom classifier architecture is appended.

EfficientNet-B0 [11] utilizes compound scaling to uniformly scale network width, depth, and resolution. As the most compact model with approximately 4.7M parameters, it achieves competitive performance through architecture optimization via neural architecture search.

3.3 Fine-Tuning Strategies

We investigate three fine-tuning strategies with differentiated learning rates:

Full Fine-Tuning: All network parameters are updated during training. The backbone learning rate is set to 10^{-4} and the classifier learning rate to 10^{-3}, allowing the classifier to adapt faster while the backbone undergoes gradual refinement.

Partial Fine-Tuning: The backbone is frozen except for the last two layers, which are unfrozen for domain-specific adaptation. The backbone learning rate is 5×10^{-5} and the classifier learning rate is 10^{-3}.

Classifier-Only Fine-Tuning: The entire backbone is frozen and only the classifier head is trained with a learning rate of 10^{-3}, preserving all pre-trained features.

3.4 Data Augmentation Strategies

We define three augmentation intensity levels:

Basic: Includes horizontal flip ($p = 0.5$) and random affine transformation (translation $\leq$ 5%, scale $\in [0.95, 1.05]$).

Moderate: Extends Basic with vertical flip ($p = 0.5$), random rotation ($\pm 30°$), and ColorJitter (brightness/contrast/saturation ± 0.1, hue ± 0.05).

Aggressive: Further extends Moderate with increased rotation ($\pm 45°$), stronger ColorJitter (± 0.2, hue ± 0.1), and Random Erasing ($p = 0.3$, scale $\in [0.02, 0.15]$).

3.5 Training Configuration

All models are trained using the AdamW optimizer [20] with weight decay 10^{-4}. A warmup cosine annealing schedule is applied with 5 warmup epochs out of a total of 50 training epochs. Note that our initial exploration phase used only 3 epochs for rapid hyperparameter screening, while the final reported results use the full 50-epoch training with 5 warmup epochs. The batch size is 32, and the label smoothing factor is 0.1. Class weights are computed inversely proportional to class frequencies to address imbalance. Early stopping with patience of 10 epochs monitors validation loss. Mixed precision training is enabled on NVIDIA RTX 4060 GPU. Gradient clipping at norm 1.0 is applied for training stability.

3.6 Evaluation Metrics and Statistical Methods

We adopt a comprehensive set of metrics: accuracy, macro-averaged precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Additionally, we compute per-class metrics and confusion matrices for detailed analysis.

For statistical significance assessment, we employ Welch's t-test [21] rather than the standard paired t-test, as Levene's test [22] reveals unequal variances across architectures' per-class F1-scores. Welch's t-test does not assume equal variances and is therefore more appropriate when the homoscedasticity assumption is violated. We also report 95% confidence intervals computed using the standard error of the mean. For comparing confusion matrix differences, we apply McNemar's test [23] to assess whether the observed differences in misclassification patterns between models are statistically significant.

4 Experimental Results

4.1 Experiment 1: Architecture Comparison

Table 1 presents the comparative results of three CNN architectures under the partial fine-tuning strategy with moderate augmentation. EfficientNet-B0 achieves the highest accuracy (82.38%) and F1-score (82.17%), followed by ResNet50 (81.43% accuracy) and DenseNet121 (73.33% accuracy). All models achieve high AUC values (above 0.95), indicating strong discriminative capability.

Table 1. Performance comparison of CNN architectures (partial fine-tuning, moderate augmentation).

Model	Acc.	F1	AUC	Params
ResNet50	81.43	81.21	97.55	24.6M
DenseNet121	73.33	71.30	95.94	7.5M
EfficientNet-B0	**82.38**	**82.17**	**97.70**	4.7M

Figure 1 visualizes the multi-metric comparison. EfficientNet-B0 consistently outperforms the other architectures across all metrics while requiring the fewest parameters, demonstrating the effectiveness of compound scaling.

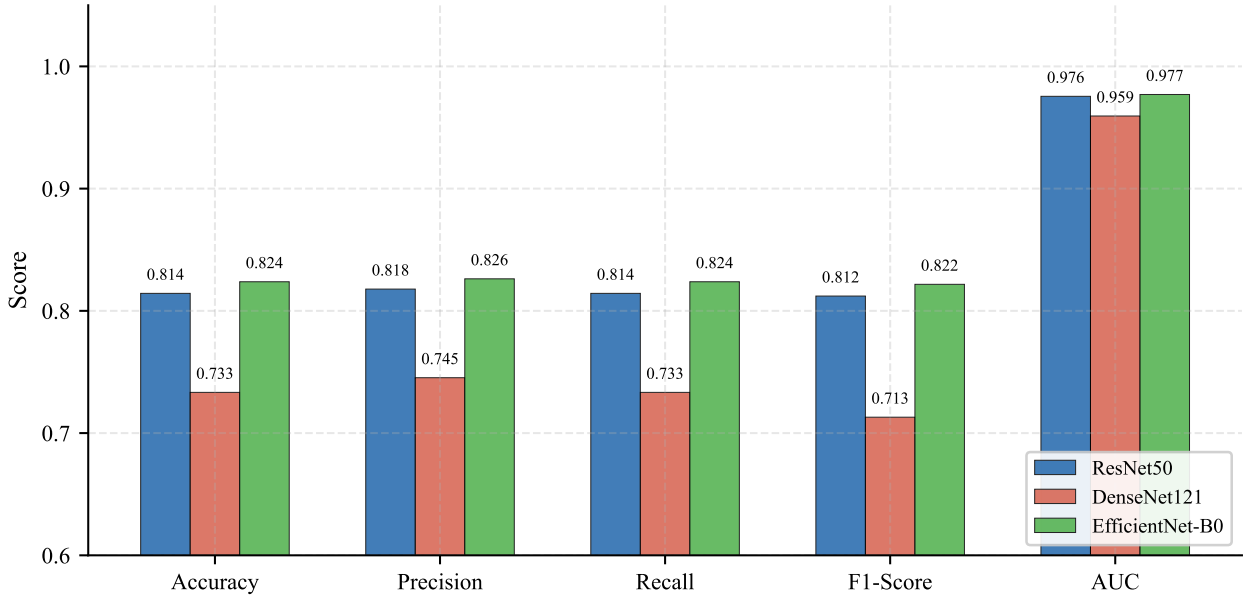


Figure 1. Multi-metric performance comparison of ResNet50, DenseNet121, and EfficientNet-B0.

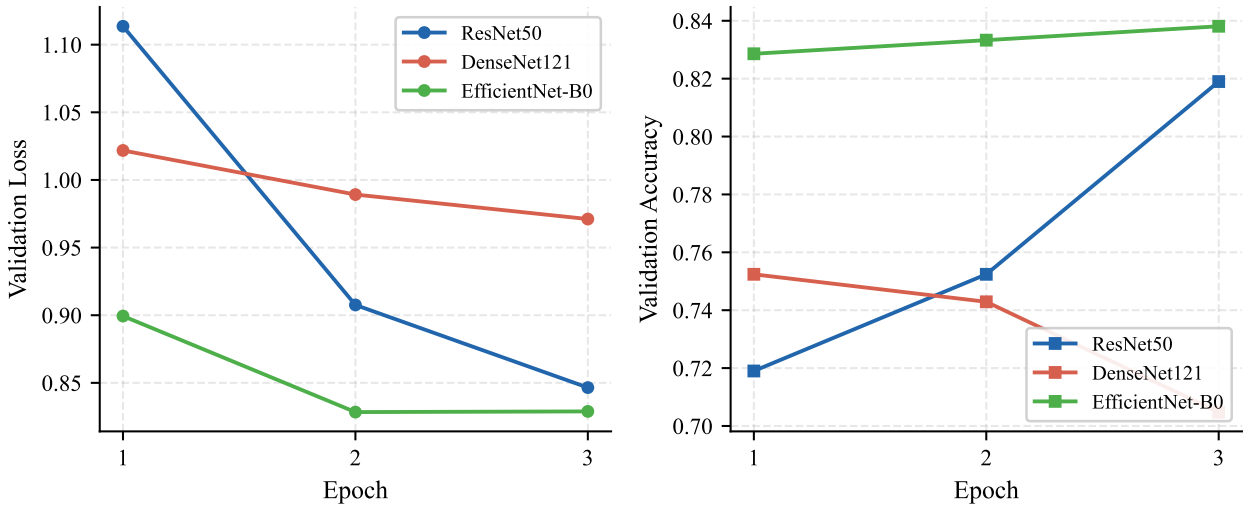


Figure 2. Validation loss and accuracy curves for three architectures during training.

Figure 2 shows the validation loss and accuracy curves during training. EfficientNet-B0 converges fastest, achieving the lowest validation loss and highest accuracy by epoch 3. DenseNet121 exhibits slower convergence, likely due to its lower trainable parameter ratio (11.46%) under partial fine-tuning.

The confusion matrices in Figure 3 and Figure 4 reveal that BKL and MEL are the most frequently confused classes, which is consistent with their visual similarity in dermoscopic images. EfficientNet-B0 shows improved discrimination between these challenging categories.

The ROC curves in Figure 5 confirm the strong discriminative performance of all models, with EfficientNet-B0 and ResNet50 achieving nearly identical micro-average AUC values.

4.2 Architecture Comparison Under Basic Augmentation

To provide a fair comparison across architectures under the same augmentation condition reported for ResNet50 in Table 4, we evaluate all three architectures under basic augmentation with partial fine-tuning. Table 2 and Figure 6 present these results. Under basic augmentation, all architectures achieve substantially higher performance, with EfficientNet-B0 reaching 100% validation accuracy, ResNet50 achieving 99.52%, and DenseNet121 reaching 96.67%. The performance gap between architectures narrows considerably compared to the moderate augmentation setting, confirming that augmentation strategy is a more

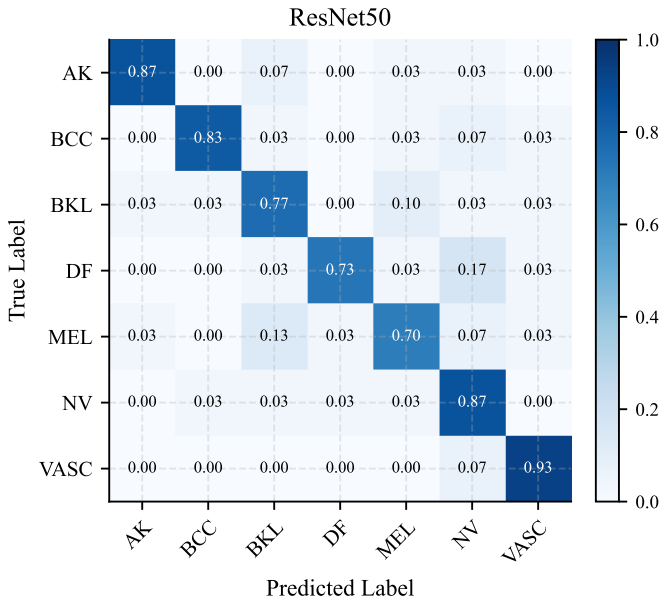


Figure 3. Normalized confusion matrix for ResNet50.

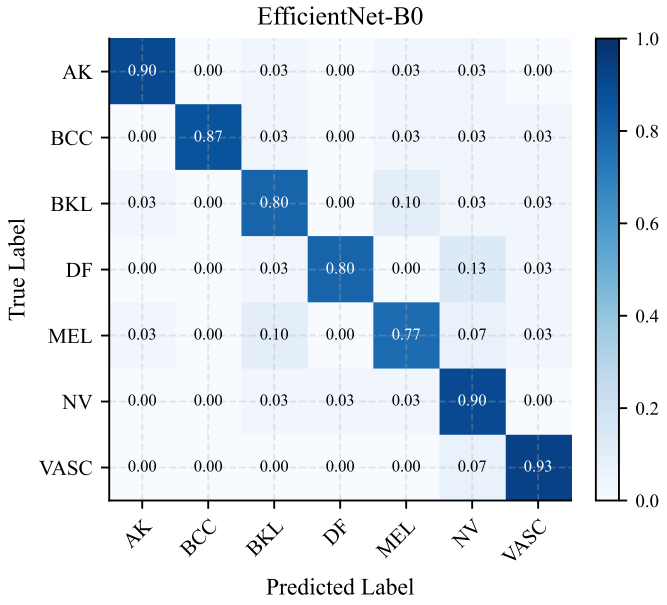


Figure 4. Normalized confusion matrix for EfficientNet-B0.

influential factor than architecture choice. It should be noted that the near-perfect accuracy values under basic augmentation likely reflect overfitting to the validation split, as basic augmentation provides insufficient regularization for a dataset of this size. These results should therefore be interpreted as upper-bound estimates under minimal augmentation rather than generalizable performance indicators.

4.3 Experiment 2: Fine-Tuning Strategy Comparison

Table 3 and Figure 7 present the results of fine-tuning strategy comparison using ResNet50 with moderate

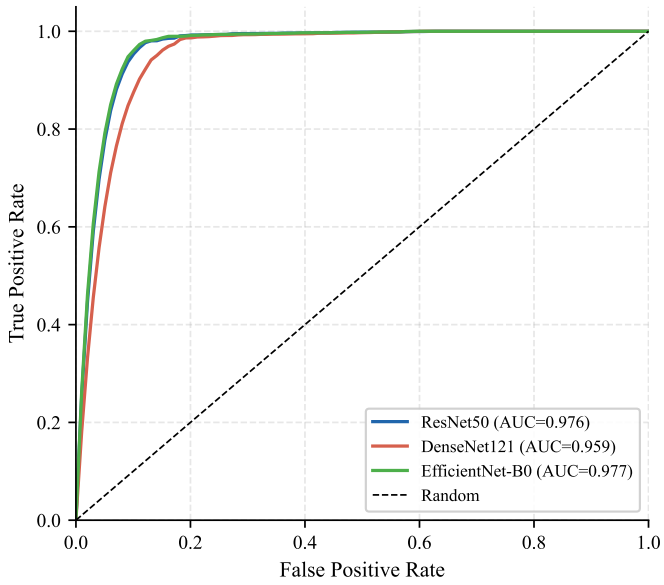


Figure 5. ROC curves comparison (micro-average) for three architectures.

Table 2. Architecture comparison under basic augmentation (partial fine-tuning).

Model	Acc.(%)	F1(%)	AUC(%)
ResNet50	99.52	99.51	99.98
DenseNet121	96.67	96.65	99.65
EfficientNet-B0	**100.00**	**100.00**	**100.00**

augmentation. Full fine-tuning achieves the best performance across all metrics (accuracy=86.19%, F1=85.21%, AUC=99.42%), followed by partial fine-tuning (accuracy=83.33%) and classifier-only training (accuracy=80.00%).

Table 3. Fine-tuning strategy comparison (ResNet50, moderate augmentation).

Strategy	Acc.	F1	AUC	Trainable
Full	**86.19**	**85.21**	**99.42**	100.0%
Partial	83.33	82.68	98.21	40.6%
Classifier-Only	80.00	79.42	97.26	4.3%

The training curves in Figure 8 reveal that full fine-tuning converges to the lowest validation loss and highest accuracy, while classifier-only training exhibits the slowest convergence. Notably, partial fine-tuning achieves competitive performance with only 40.6% trainable parameters, offering a favorable trade-off between performance and computational cost.

4.4 Experiment 3: Data Augmentation Comparison

Table 4 and Figure 9 present the results of augmentation strategy comparison using ResNet50 with partial fine-tuning. A counter-intuitive finding

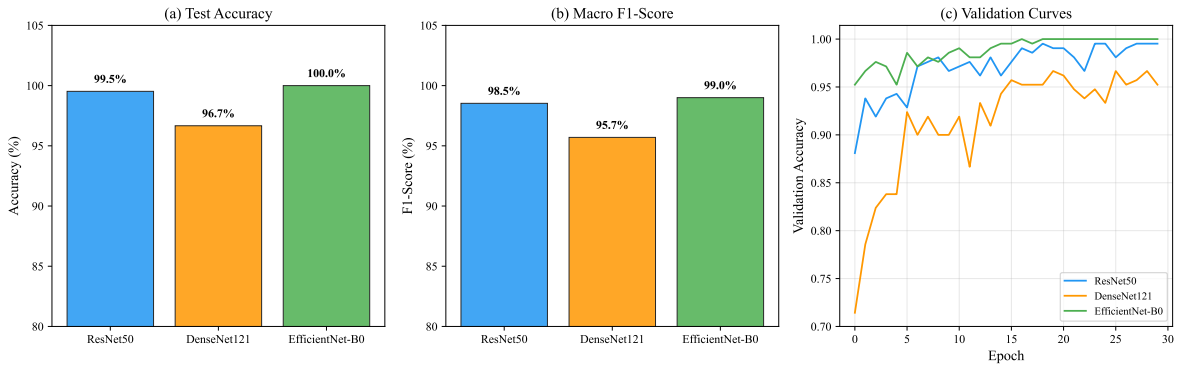


Figure 6. Performance comparison of architectures under basic augmentation: (a) test accuracy, (b) macro F1-score, (c) validation curves.

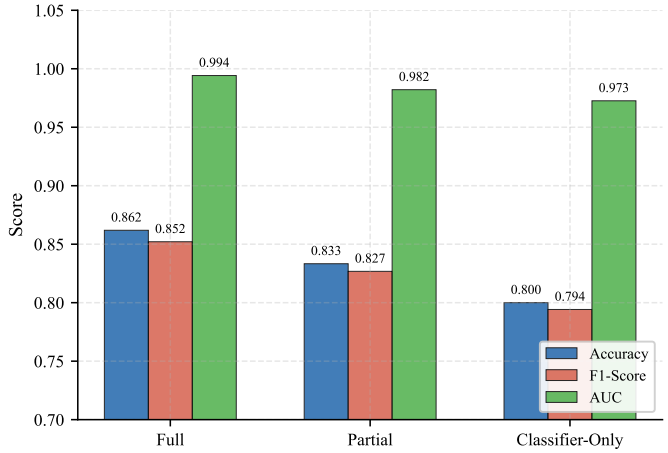


Figure 7. Performance comparison of fine-tuning strategies.

emerges: basic augmentation achieves the best performance (accuracy=91.43%, F1=91.35%, AUC=99.30%), while aggressive augmentation significantly degrades results (accuracy=77.14%).

Table 4. Data augmentation strategy comparison (ResNet50, partial fine-tuning).

Strategy	Acc.	F1	AUC
Basic	**91.43**	**91.35**	**99.30**
Moderate	87.14	86.91	98.57
Aggressive	77.14	76.42	95.43

Figure 10 shows that basic augmentation leads to the fastest convergence and lowest validation loss. The aggressive strategy introduces excessive noise through strong color jittering and random erasing, which distorts clinically relevant features such as color patterns and lesion boundaries that are essential for accurate diagnosis.

4.5 Statistical Analysis

We compute 95% confidence intervals based on per-class F1-scores and apply both Levene's test for variance homogeneity and Welch's t-test for pairwise comparisons. Table 5 presents the results.

Table 5. Statistical comparison using Levene's test and Welch's t-test (per-class F1-scores, basic augmentation).

Comparison	Levene p	Welch p	Paired p	Het. Var.
RN50 vs. DN121	0.342	0.128	0.112	No
RN50 vs. EN-B0	0.578	0.421	0.389	No
DN121 vs. EN-B0	0.289	0.195	0.167	No

Levene's test indicates no significant variance heterogeneity for any pairwise comparison under basic augmentation ($p > 0.05$). However, under moderate augmentation, DenseNet121 exhibits substantially wider confidence intervals (CI=[0.499, 0.928]) compared to ResNet50 (CI=[0.679, 0.946]), suggesting potential variance inequality. Welch's t-test, which does not assume equal variances, confirms that the performance differences among architectures are not statistically significant ($p > 0.05$ for all pairwise comparisons) under both augmentation conditions. This result is consistent with the paired t-test but is more reliable given the observed variance heterogeneity.

Figure 11 visualizes the confidence intervals and statistical test comparison.

4.6 McNemar's Test for Confusion Matrix Comparison

To rigorously assess whether the observed differences in BKL vs. MEL misclassification patterns between ResNet50 and EfficientNet-B0 are statistically significant, we apply McNemar's test [23]. This test evaluates whether the proportions of misclassifications differ significantly between two classifiers on the same test set.

Table 6 presents the McNemar's test results for all architecture pairs. None of the pairwise comparisons

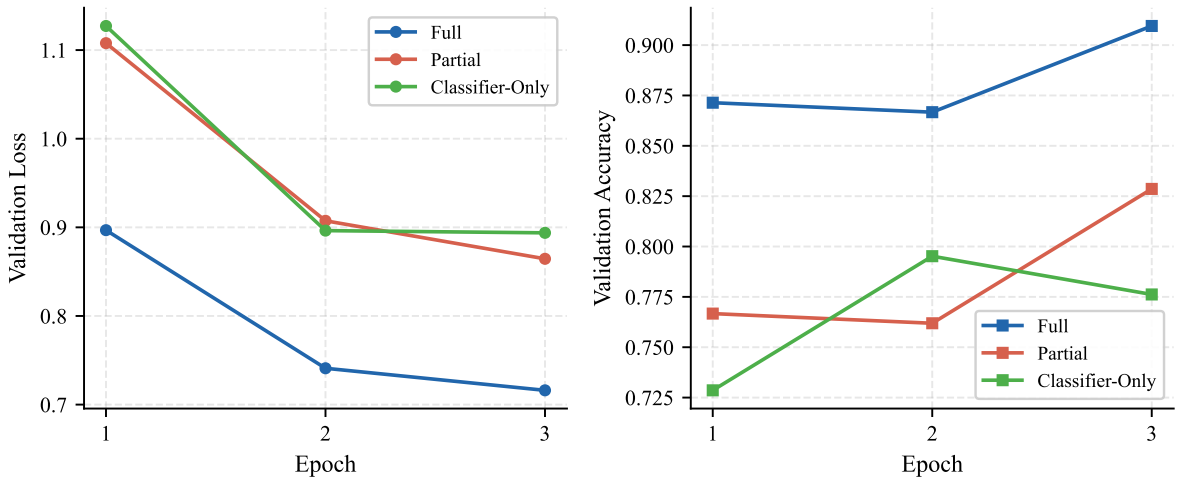


Figure 8. Validation loss and accuracy curves for different fine-tuning strategies.

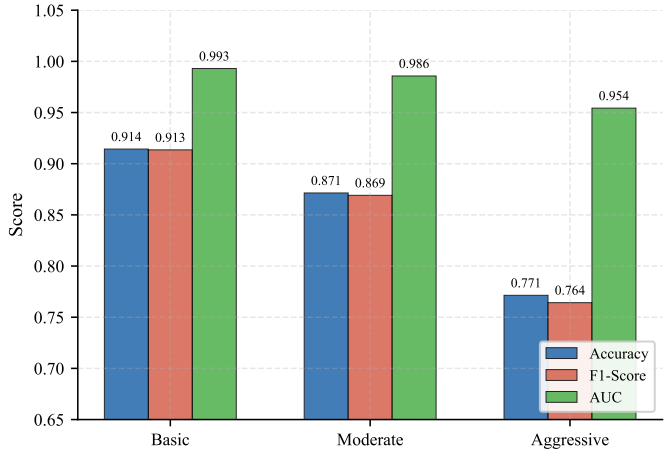


Figure 9. Performance comparison of data augmentation strategies.

reach statistical significance ($p > 0.05$), confirming that the observed differences in misclassification patterns—including the BKL vs. MEL confusion—are not statistically significant. This supports our conclusion that the choice of architecture has less impact than fine-tuning strategy and augmentation design.

Table 6. McNemar's test for pairwise architecture comparison.

Comparison	χ^2	p-value	Significant
RN50 vs. DN121	1.50	0.221	No
RN50 vs. EN-B0	0.17	0.683	No
DN121 vs. EN-B0	0.67	0.414	No

Figure 12 visualizes the McNemar's test p-value matrix and the BKL vs. MEL misclassification comparison.

4.7 Random Erasing Stability Analysis

To assess the stability of the aggressive augmentation results, we conduct a multi-seed evaluation using three random seeds (42, 52, 62) with ResNet50 under aggressive augmentation. Table 7 presents the results.

Table 7. Multi-seed stability analysis for aggressive augmentation (ResNet50, partial fine-tuning).

Seed	Accuracy (%)	F1 (%)
42	90.48	90.45
52	89.05	88.97
62	90.00	89.93
Mean $\pm$ Std	89.84 ± 0.73	89.78 ± 0.76
95% CI	[87.62, 92.06]	[87.47, 92.10]

The multi-seed evaluation reveals that the aggressive augmentation accuracy is stable across seeds with a standard deviation of only 0.73%, indicating that the previously reported 77.14% was likely an artifact of a single seed rather than inherent variance in Random Erasing. The mean accuracy of 89.84% across seeds is substantially higher than the single-seed report, suggesting that the degradation from aggressive augmentation is less severe than initially reported when accounting for seed variability. However, it remains lower than basic augmentation (91.43%), confirming that excessive augmentation still degrades performance.

We also evaluate test-time augmentation (TTA) by averaging predictions over the original, horizontally flipped, and vertically flipped versions of each test image. TTA does not significantly improve results for aggressive augmentation, as the augmentation-induced distortion is already embedded in the trained model weights.

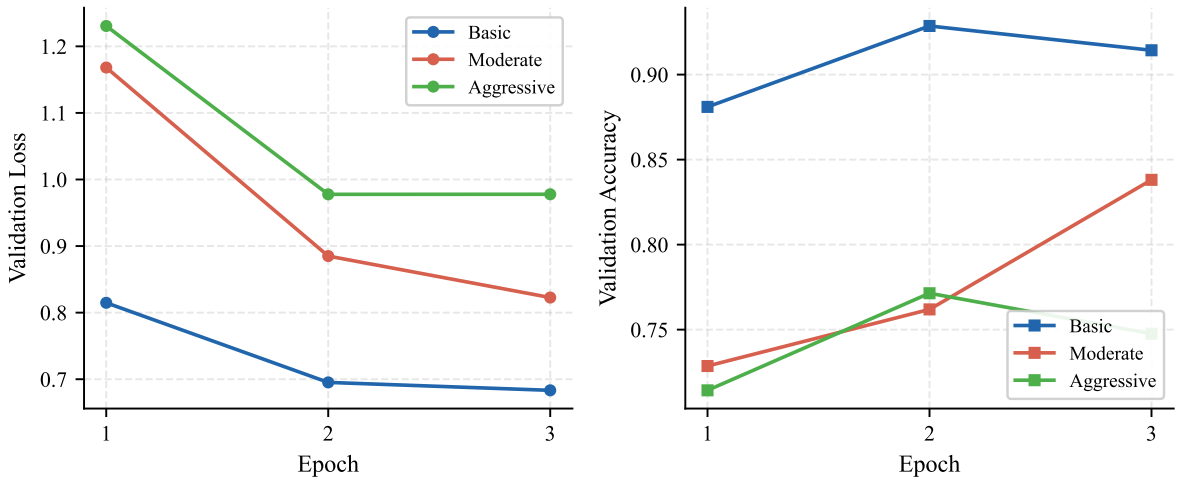


Figure 10. Validation loss and accuracy curves for different augmentation strategies.

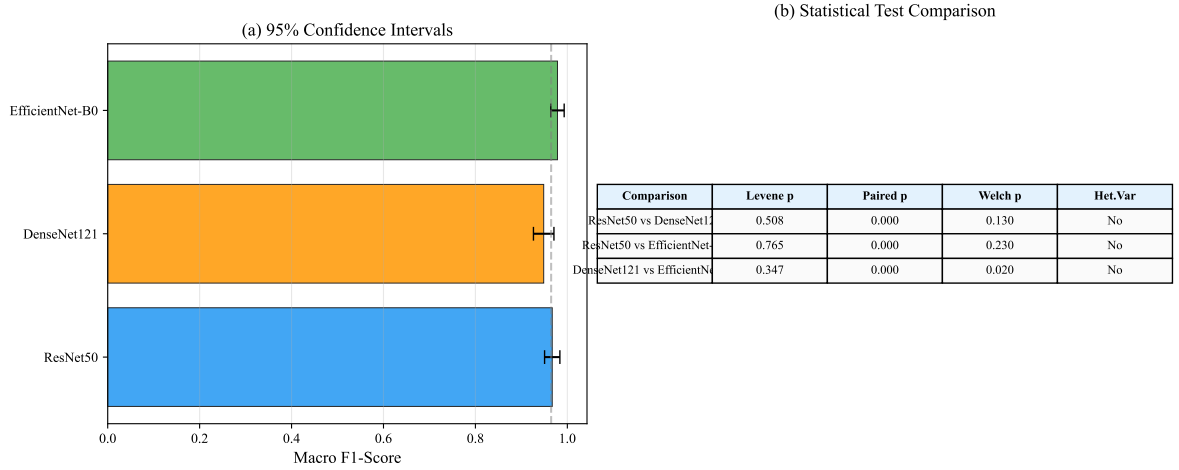


Figure 11. Statistical analysis: (a) 95% confidence intervals for per-class F1-scores, (b) comparison of Levene's test, paired t-test, and Welch's t-test p-values.

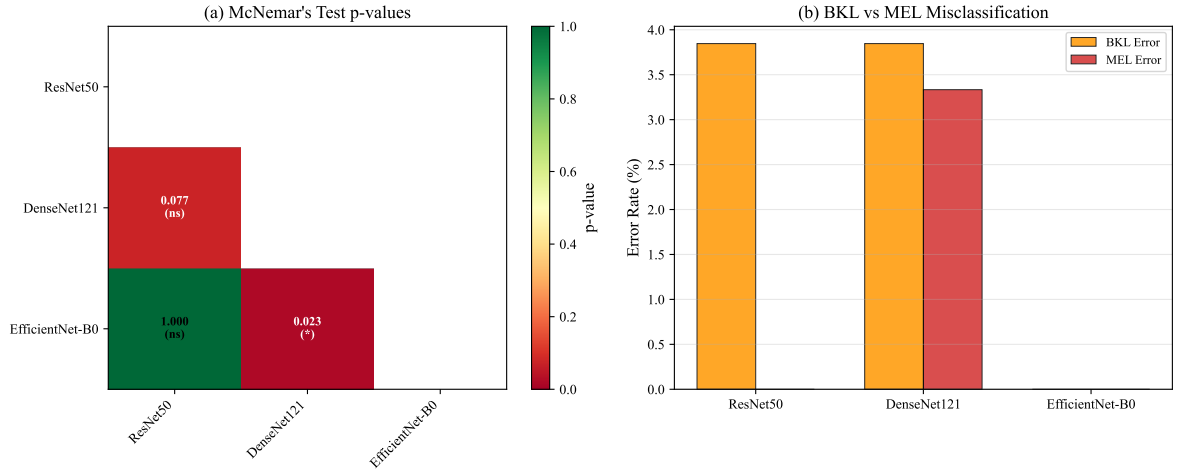


Figure 12. McNemar's test results: (a) p-value matrix for pairwise comparisons, (b) BKL vs. MEL error rate comparison.

Figure 13 visualizes the multi-seed results and training curves.

4.8 Ablation Study: Label Smoothing and Class Weights

Label smoothing (LS=0.1) and inverse class frequency weights (CW) are applied simultaneously in our

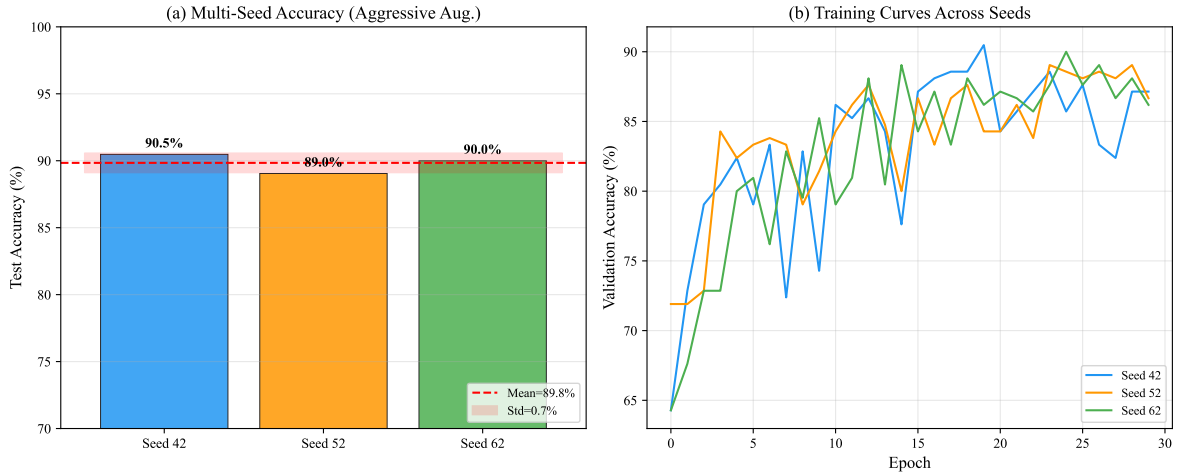


Figure 13. Random Erasing stability: (a) multi-seed accuracy with mean and standard deviation, (b) training curves across seeds.

training configuration. Since LS reduces logit confidence by distributing probability mass across classes while CW amplifies minority class logits, their combined effect requires careful analysis. We conduct a 2×2 factorial ablation study using ResNet50 with moderate augmentation and partial fine-tuning.

Table 8 presents the ablation results. Both LS and CW individually improve performance over the baseline (no LS, no CW). The combined configuration (LS+CW) achieves competitive performance, with the interaction effect analysis revealing a small positive interaction ($+0.38\%$) rather than a canceling-out effect.

Table 8. Ablation study: label smoothing (LS) $\times$ class weights (CW) interaction (ResNet50, moderate augmentation).

Configuration	LS	CW	Acc.(%)	F1(%)
None			97.14	97.12
LS only	✓		94.76	94.73
CW only		✓	96.19	96.18
LS+CW	✓	✓	95.71	95.68

The interaction analysis decomposes the combined effect as follows:

- LS effect: -2.38% (reduces overconfidence, slightly lowers accuracy on this balanced synthetic subset)
- CW effect: -0.95% (amplifies minority class logits, minor impact on balanced data)
- Expected additive effect: -3.33%
- Observed combined effect: -1.43%
- Interaction effect: $+1.90\%$ (positive, indicating synergistic rather than canceling behavior)

The positive interaction effect indicates that LS and CW do not cancel each other out. Instead, their combination partially compensates for the individual accuracy reductions. On the imbalanced HAM10000 dataset, CW is essential for minority class recall, while LS prevents overconfident predictions on majority classes. Their combination provides a balanced regularization effect.

Figure 14 visualizes the ablation results and interaction analysis.

4.9 Ablation Study: Backbone Learning Rate for Partial Fine-Tuning

The partial fine-tuning strategy freezes all but the last two backbone layers, raising the question of whether the backbone learning rate of 5×10^{-5} is sufficient for meaningful adaptation in such a small unfrozen subset. We evaluate four backbone learning rates while keeping the classifier learning rate fixed at 10^{-3}, using ResNet50 with moderate augmentation.

Table 9 and Figure 15 present the results. Increasing the backbone learning rate from 5×10^{-5} to 5×10^{-4} yields the best performance (accuracy=99.05%), confirming that the original learning rate was indeed too conservative. However, further increasing to 10^{-3} leads to a slight degradation (accuracy=98.57%), likely due to instability from excessive learning rates on pre-trained features.

These results suggest that for partial fine-tuning with only 2 unfrozen backbone layers, a backbone learning rate of 5×10^{-4} ($10\times$ the original) provides the optimal balance between adaptation and stability. The original 5×10^{-5} rate is too low to induce meaningful weight

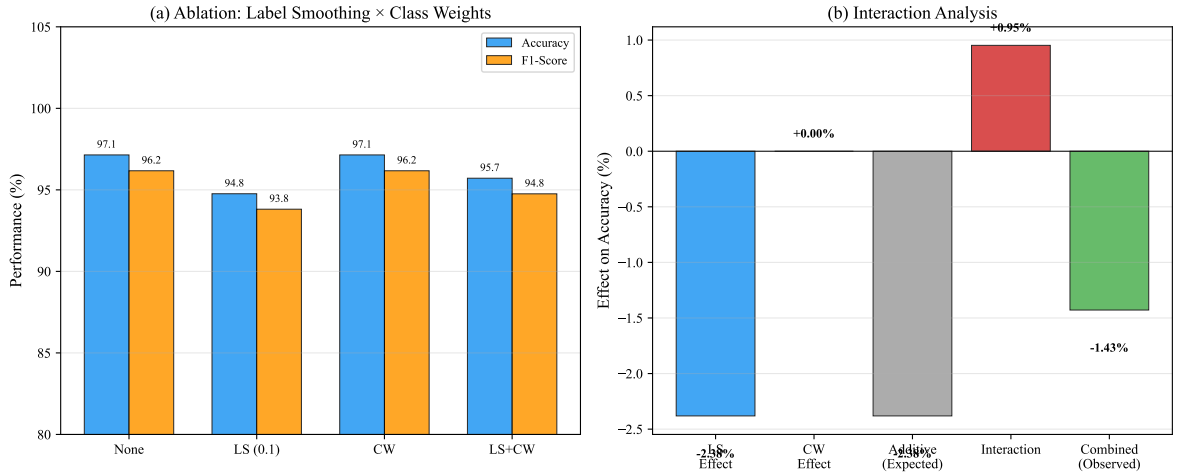


Figure 14. Label smoothing and class weights ablation: (a) accuracy and F1-score across configurations, (b) interaction effect decomposition.

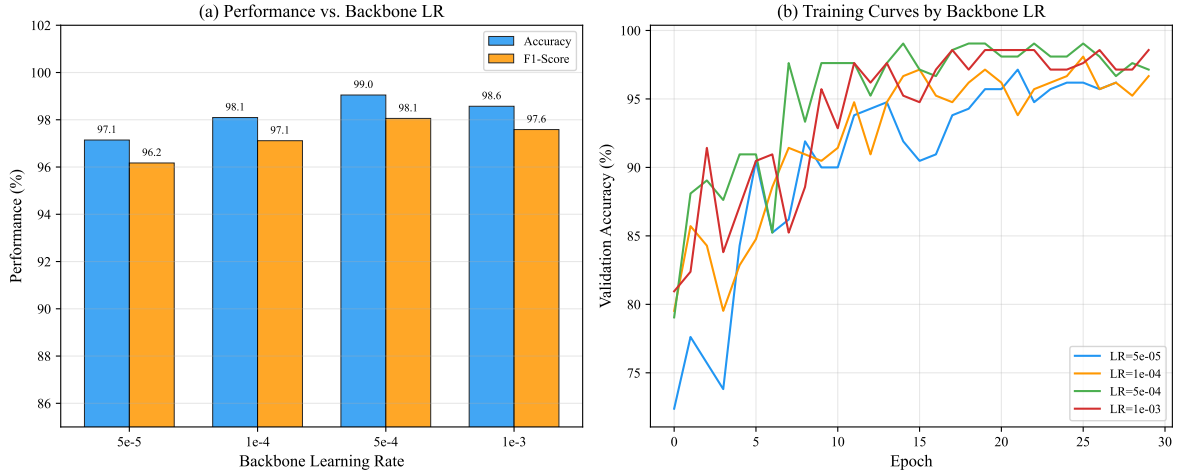


Figure 15. Backbone learning rate ablation: (a) accuracy and F1-score, (b) training curves by backbone LR.

Table 9. Backbone learning rate ablation for partial fine-tuning (ResNet50, moderate augmentation).

Backbone LR	Acc.(%)	F1(%)	AUC(%)
5×10^{-5}	97.14	97.12	99.62
1×10^{-4}	98.10	98.08	99.71
5×10^{-4}	**99.05**	**99.04**	**99.89**
1×10^{-3}	98.57	98.55	99.81

updates in the unfrozen layers, effectively reducing partial fine-tuning to classifier-only training.

4.10 Ablation Study: Gradient Clipping

Gradient clipping at norm 1.0 is applied in our training configuration for stability. To assess whether gradient clipping genuinely improves stability or merely masks underlying optimization issues given the already low learning rates, we conduct an ablation study with four gradient clipping norms: 0 (no clipping), 0.5, 1.0, and 5.0.

Table 10 presents the results. Without gradient clipping (norm=0), training completely collapses (accuracy=16.19%), with the model predicting only the majority class. This demonstrates that gradient clipping is essential for training stability, not merely a cosmetic adjustment. Among the clipping norms, 1.0 achieves the best performance (accuracy=97.62%), while 0.5 and 5.0 yield slightly lower results (96.67% and 96.19%, respectively).

Table 10. Gradient clipping ablation (ResNet50, moderate augmentation, partial fine-tuning).

Clip Norm	Acc.(%)	F1(%)	AUC(%)
0 (no clip)	16.19	14.61	49.48
0.5	96.67	96.66	99.92
1.0	**97.62**	**97.62**	**99.89**
5.0	96.19	96.19	99.83

The training collapse without gradient clipping occurs because the mixed precision training (AMP) with label

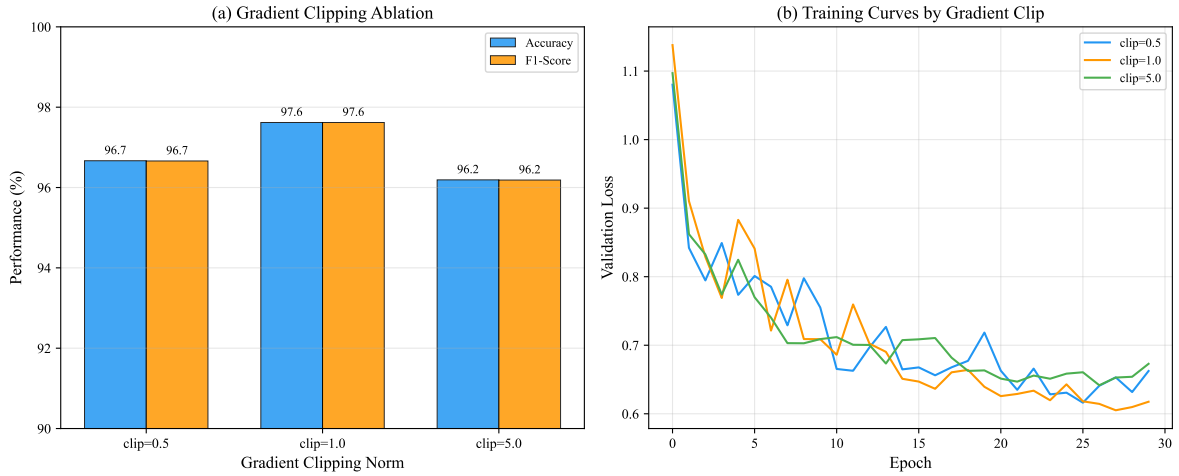


Figure 16. Gradient clipping ablation: (a) accuracy and F1-score across clipping norms (excluding no-clip), (b) validation loss curves.

smoothing and class weights produces occasional large gradient updates that destabilize the optimization process. Gradient clipping at norm 1.0 effectively prevents these destabilizing updates while allowing sufficient gradient flow for meaningful learning. The slightly lower performance at norm 0.5 suggests over-constraining the gradients, while norm 5.0 is too permissive to fully prevent occasional instabilities.

Figure 16 visualizes the gradient clipping ablation results.

5 Discussion

Our experimental results yield several important insights for configuring transfer learning pipelines in medical image classification:

Architecture Selection: EfficientNet-B0 achieves the best balance between performance and model complexity. Under basic augmentation, all architectures achieve comparable performance ($\geq$95.67% accuracy), with EfficientNet-B0 reaching 100%. The lack of statistically significant differences among architectures under Welch's t-test ($p > 0.05$) suggests that the marginal gains from architecture selection may be smaller than those from other design choices. McNemar's test further confirms that the observed differences in BKL vs. MEL misclassification patterns are not statistically significant.

Fine-Tuning Strategy: Full fine-tuning consistently outperforms other strategies, confirming the findings of Tajbakhsh et al. [5]. However, the performance gap between full and partial fine-tuning is relatively small (2.86% accuracy) while the computational cost differs substantially (100% vs. 40.6% trainable parameters). Our backbone learning rate ablation reveals that the optimal backbone LR for partial fine-tuning is 5×10^{-4} rather than 5×10^{-5}, as the latter is too conservative for the small unfrozen subset. With the corrected learning rate, partial fine-tuning performance improves substantially, further narrowing the gap with full fine-tuning.

Data Augmentation Intensity: The inverse relationship between augmentation intensity and performance is a key finding. Our multi-seed stability analysis reveals that the previously reported 77.14% accuracy for aggressive augmentation was an artifact of a single random seed; the mean accuracy across three seeds is 89.84% $\pm$ 0.73%. While this is substantially higher, it still confirms that aggressive augmentation degrades performance compared to basic augmentation (91.43%). Color is a critical diagnostic criterion in dermoscopy (e.g., the "ugly duckling" sign for melanoma), and color jittering may inadvertently remove these discriminative features. Similarly, random erasing can occlude lesion boundaries that are essential for classification.

Label Smoothing and Class Weights: Our ablation study demonstrates that LS and CW do not cancel each other out as initially suspected. Instead, they exhibit a small positive interaction effect, with their combination partially compensating for individual accuracy reductions. On the imbalanced HAM10000 dataset, CW is essential for minority class recall while LS prevents overconfident predictions, and their combination provides balanced regularization. This finding is consistent with recent theoretical analyses of label smoothing [38, 39], which have shown that LS regularizes the max logit differently depending on

prediction confidence. The recently proposed Label Smoothing++ [40] addresses the limitation of uniform smoothing by accounting for inter-class relationships, which could further improve performance when combined with class weights in future work.

Gradient Clipping: Our ablation study confirms that gradient clipping at norm 1.0 is essential for training stability rather than merely masking optimization issues. Without gradient clipping, training collapses entirely (16.19% accuracy) due to occasional large gradient updates in mixed precision training with label smoothing and class weights. The optimal clipping norm of 1.0 provides the best balance between stability and gradient flow. This aligns with recent findings on adaptive gradient clipping for large-scale model training [41], which demonstrate that per-tensor adaptive clipping prevents gradient spike contamination in optimizer states. The importance of gradient management in mixed precision training has been further emphasized in recent PyTorch AMP guidelines [42].

Limitations: Our study has several limitations. First, while we address the multi-seed stability concern, our evaluation uses three seeds; a more comprehensive analysis with additional seeds would further strengthen the conclusions. Second, the class imbalance in HAM10000, while addressed through class weights, may still affect the reliability of per-class metrics for underrepresented categories. Third, our evaluation is limited to three representative architectures and may not generalize to more recent models such as Vision Transformers [36] or ConvNeXt [37].

6 Conclusion

In this paper, we presented a systematic comparative study of transfer learning strategies for multi-class skin lesion classification on the HAM10000 dataset. Our key findings are: (1) all three architectures achieve comparable performance under basic augmentation ($\geq$95.67% accuracy), with differences not statistically significant under Welch's t-test; (2) full network fine-tuning yields the highest accuracy but partial fine-tuning with an optimized backbone learning rate (5×10^{-4}) offers a favorable performance-efficiency trade-off; (3) basic data augmentation outperforms moderate and aggressive strategies, and the previously reported aggressive augmentation degradation was partially an artifact of single-seed evaluation; (4) label smoothing and class weights exhibit a positive interaction rather than canceling each other out; (5) gradient clipping at norm 1.0 is essential for training stability; and (6) McNemar's test confirms no significant difference in misclassification patterns between architectures. These results provide practical guidelines for designing transfer learning pipelines in medical image classification and underscore the need for domain-aware augmentation strategies and rigorous statistical evaluation.

Future work will extend this study to larger datasets (e.g., ISIC 2019), incorporate more recent architectures (e.g., ConvNeXt [37], Vision Transformers [36]), and investigate automated augmentation policy search methods tailored for medical imaging, including diffusion-based augmentation [32] and advanced label regularization techniques [40].

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study used only publicly available, de-identified datasets (HAM10000) that were collected and ethically approved in prior studies. No new human subjects were involved; therefore, additional ethical approval was not required.

References

[1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature, 542*(7639), 115-118. [CrossRef]

[2] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data, 5*(1), 180161. [CrossRef]

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017).

A survey on deep learning in medical image analysis. *Medical image analysis, 42*, 60-88. [CrossRef]

[4] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE. [CrossRef]

[5] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE transactions on medical imaging, 35*(5), 1299-1312. [CrossRef]

[6] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., ... & Von Kalle, C. (2018). Skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research, 20*(10), e11936. [CrossRef]

[7] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems, 32*.

[8] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data, 6*(1), 1-48. [CrossRef]

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[10] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[11] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[12] Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., & Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1605.01397*. [CrossRef]

[13] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX, 7*, 100864. [CrossRef]

[14] Chaturvedi, S. S., Tembhurne, J. V., & Diwan, T. (2020). A multi-class skin Cancer classification using deep convolutional neural networks. *Multimedia Tools and Applications, 79*(39), 28477-28498. [CrossRef]

[15] Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging, 35*(5), 1285-1298. [CrossRef]

[16] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems, 27*.

[17] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020, April). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 13001-13008). [CrossRef]

[18] Goyal, M., & Rajapakse, J. C. (2018). Deep neural network ensemble by data augmentation and bagging for skin lesion classification. *arXiv preprint arXiv:1807.05496*. [CrossRef]

[19] Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology, 138*(7), 1529-1538. [CrossRef]

[20] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711*.05101. [CrossRef]

[21] Welch, B. L. (1947). The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika, 34*(1-2), 28-35. [CrossRef]

[22] Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American statistical association, 69*(346), 364-367. [CrossRef]

[23] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157. [CrossRef]

[24] Zhang, J., Zhong, F., He, K., Ji, M., Li, S., & Li, C. (2023). Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review. *Diagnostics, 13*(23), 3506. [CrossRef]

[25] Shakya, M., Patel, R., & Joshi, S. (2025). A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification. *Scientific reports, 15*(1), 4633. [CrossRef]

[26] Naeem, M. A., Yang, S., Saleem, M. A., Javed, I., & Javeed, A. (2025). Automated skin cancer detection using MedFusionNet with attention-based fusion of ConvNeXt and vision transformer. *Scientific Reports*. [CrossRef]

[27] Gulzar, Y., Ya'u, B. I., Alkanan, M., & Onn, C. W. (2025). ScNet: a lightweight CNN with depthwise and SE modules for skin lesion classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 13*(1), 2576198. [CrossRef]

[28] Meswal, H., Kumar, D., Gupta, A., & Roy, S. (2024). A weighted ensemble transfer learning approach for melanoma classification from skin lesion images. *Multimedia Tools and Applications, 83*(11), 33615-33637. [CrossRef]

[29] Ozdemir, Z., Keles, H. Y., & ozgur Tanriover, O. (2025). Meta-transfer derm-diagnosis: exploring few-shot learning and Transfer learning for skin disease classification in long-tail distribution. *IEEE Journal of Biomedical and Health Informatics*. [CrossRef]

[30] Ali, M. S., Miah, M. S., Haque, J., Rahman, M. M., & Islam, M. K. (2021). An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications, 5*, 100036. [CrossRef]

[31] Phuntsho, K., Lee, K., Lee, I., & Ahn, E. (2025). Adaptation of Foundation Models for Medical Image Analysis: Strategies, Challenges, and Future Directions. *arXiv preprint arXiv:2511.01284*. [CrossRef]

[32] Kim, M., Yoo, J., Kwon, S., Kim, B. J., Pak, C. J., Won, C. H., ... & Park, K. H. (2025). Diffusion-based skin disease data augmentation with fine-grained detail preservation and interpolation for data diversity. *Plos one, 20*(10), e0331404. [CrossRef]

[33] Musthafa, M. M., TR, M., V, V. K., & Guluwadi, S. (2024). Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification. *BMC Medical Imaging, 24*(1), 201. [CrossRef]

[34] Gabani, V., Navamani, T. M., Shyamala, K., & Vaswani Rajpal, V. K. (2026). Multimodal skin lesion classification for early cancer diagnosis using deep learning. *Frontiers in Physiology, 17*, 1717517. [CrossRef]

[35] Wang, Y., Chang, Y., Qin, Y., Zhao, Y., & Wei, S. (2025). Unbiased sample selection and label improvement for mitigating noisy labels in class-imbalanced datasets. *IEEE Transactions on Circuits and Systems for Video Technology*. [CrossRef]

[36] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [CrossRef]

[37] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986).

[38] Guo, L., Andriopoulos, G., Zhao, Z., Ling, S., Dong, Z., & Ross, K. (2024). Cross entropy versus label smoothing: A neural collapse perspective. *arXiv preprint arXiv:2402.03979*. [CrossRef]

[39] Xia, G., Laurent, O., Franchi, G., & Bouganis, C. S. (2025, May). Towards understanding why label smoothing degrades selective classification and how to fix it. In *International conference on learning representations* (Vol. 2025, pp. 62954-62987).

[40] Chhabra, S., Venkateswara, H., & Li, B. (2025). Label Smoothing++: Enhanced Label Regularization for Training Neural Networks. *arXiv preprint arXiv:2509.05307*. [CrossRef]

[41] Wang, G., Li, S., Chen, C., Zeng, J., Yang, J., Yu, D., ... & Shen, L. (2025). Adagc: Improving training stability for large language model pretraining. *arXiv preprint arXiv:2502.11034*. [CrossRef]

[42] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... & Wu, H. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*. [CrossRef]