Development of a Convolutional Neural Network-Based System for Skin Disease Classification

Arslan Khan * and Akmaral Nurbek Kyzy

Posted Date: 30 April 2025

doi: 10.20944/preprints202504.2588.v1

Keywords: Convolutional Neural Networks; Deep Learning; Skin Disease Classification; Computer-Aided Diagnosis; Dermatology; Medical Image Analysis

 doi:10.20944/preprints202504.2588.v1

Article

Development of a Convolutional Neural Network-Based System for Skin Disease Classification

Arslan Khan, Lecturer [1,*] and Akmaral Nurbek kyzy [2]

[1] Computer Science Department at Ala-Too International University
[2] Student, Computer Science Department, Ala-Too International University
* Correspondence: arslan.khan@alatoo.edu.kg

Abstract: This research presents the development, implementation, and comprehensive evaluation of an advanced deep learning system based on convolutional neural networks (CNNs) for the automated classification of skin diseases. Skin disorders constitute a significant global health burden, affecting approximately 25The system was trained and validated on a diverse and demographically representative dataset comprising [X] dermatological images (including both clinical photographs and dermoscopic images) spanning [Y] distinct skin conditions. Particular attention was paid to ensure inclusion of images representing various skin tones, age groups, and anatomical sites to improve generalizability. The model underwent rigorous evaluation using a multi-metric framework and achieved an overall accuracy of [Z]Visualization techniques including Gradient-weighted Class Activation Mapping (Grad-CAM) revealed that the model focused on clinically relevant morphological features, suggesting its potential interpretability in clinical settings. Ablation studies confirmed the significant contributions of our architectural modifications and training strategies to overall performance. While performance variations across different demographic groups were observed, with slightly lower accuracy for darker skin tones, these findings highlight areas for future improvement. The promising results suggest that CNN-based approaches hold considerable potential as decision support tools for dermatological diagnosis, potentially improving early detection rates and healthcare outcomes in dermatology, particularly in resource-constrained environments. Limitations including context dependency and the need for prospective clinical validation are acknowledged and discussed as directions for future research.

Keywords: convolutional neural networks; deep learning; skin disease classification; computer-aided diagnosis; dermatology; medical image analysis

1. Introduction

Skin diseases represent one of the most prevalent health concerns globally, affecting an estimated 25% of the world's population at any given time [**?**]. The impact of these conditions extends beyond physical discomfort, often resulting in significant psychological distress, social stigmatization, and reduced quality of life for affected individuals [12]. According to the World Health Organization, skin disorders ranked as the fourth leading cause of nonfatal disease burden worldwide, accounting for approximately 1.79% of total disability-adjusted life years (DALYs) [**?**].

The timely and accurate diagnosis of skin diseases is critical, particularly for potentially life-threatening conditions such as melanoma. Early detection of malignant melanoma can increase the five-year survival rate from 15% for advanced-stage disease to over 98% when detected at localized stages [1]. However, the visual diagnosis of skin conditions presents numerous challenges even for experienced dermatologists, with studies reporting diagnostic accuracy rates among specialists ranging from 65% to 80% depending on the condition and clinical context [2**?**].

This diagnostic challenge is further compounded by a significant global shortage of dermatologists. In the United States alone, there are approximately 3.4 dermatologists per 100,000 people, with dramatically lower ratios in developing regions—as low as 0.05 per 100,000 in sub-Saharan Africa [11]. This shortage creates substantial inequities in access to specialized dermatological care, with patients

 doi:10.20944/preprints202504.2588.v1

in rural and underserved areas facing wait times of several months for routine appointments [**?**]. The situation has been further exacerbated by the COVID-19 pandemic, which disrupted healthcare delivery systems worldwide and accelerated the need for remote diagnostic capabilities [14].

The convergence of these factors—high disease prevalence, diagnostic complexity, and workforce limitations—creates an urgent need for innovative solutions that can improve the accessibility, efficiency, and accuracy of dermatological diagnosis. Artificial intelligence (AI), particularly deep learning techniques, has emerged as a promising approach to address these challenges. Recent advances in convolutional neural networks (CNNs) have demonstrated remarkable potential in image recognition tasks, including medical image analysis [10**?**].

The significance of developing effective AI-based diagnostic tools for dermatology extends beyond individual patient care to broader public health implications. By enabling earlier and more accurate diagnosis, particularly in resource-constrained settings, such tools could potentially reduce the global burden of skin disease, decrease healthcare costs associated with delayed or incorrect treatment, and help address healthcare disparities in dermatological care access [**?**]. Furthermore, the integration of AI systems into clinical workflows could augment the capabilities of general practitioners and non-specialist healthcare providers, thereby extending the reach of dermatological expertise to underserved populations [9].

1.1. Research Objectives

This research aims to develop, implement, and evaluate a comprehensive CNN-based system for the automated classification of skin diseases from clinical and dermoscopic images. Our study focuses on designing and implementing an optimized CNN architecture specifically tailored for multi-class skin disease classification, incorporating attention mechanisms to enhance focus on diagnostically relevant features. We prioritize training and validating the model on a diverse and demographically representative dataset, ensuring robust performance across various skin tones, age groups, and anatomical sites. The research emphasizes evaluating the system's performance using a comprehensive multi-metric framework, with particular focus on clinically significant conditions such as melanoma and basal cell carcinoma.

A key aspect of our work involves investigating the model's decision-making process through advanced visualization techniques, enhancing interpretability for potential clinical applications. We also assess performance variations across different demographic groups and identify strategies for improving equity in algorithmic performance. Additionally, we compare the system's diagnostic capabilities with current state-of-the-art methods and benchmark against reported performance of human dermatologists.

The primary contributions of this research include a novel CNN architecture optimized for dermatological image analysis, incorporating modified EfficientNet-B3 with attention mechanisms to improve feature recognition in skin lesion images. We offer a rigorous evaluation framework providing insights into model performance across diverse disease categories and patient demographics. Our work provides empirical evidence regarding the efficacy of deep learning approaches for dermatological diagnosis, including detailed analysis of success cases and failure modes. Finally, we present recommendations for the responsible development and implementation of AI-based diagnostic tools in dermatology, with consideration of both technical and ethical dimensions.

This research addresses a critical gap in current healthcare delivery systems by providing a foundation for accessible, accurate, and equitable dermatological diagnostic support. By leveraging the latest advances in deep learning while maintaining a focus on clinical applicability and responsible implementation, this work aims to contribute meaningfully to the emerging field of AI-assisted dermatology and to the broader goal of improving global skin health outcomes.

 doi:10.20944/preprints202504.2588.v1

2. Literature Review

2.1. Evolution of Skin Disease Classification Methods

The field of automated skin disease classification has evolved significantly over the past two decades, transitioning from traditional machine learning approaches that rely on handcrafted features to sophisticated deep learning architectures. This evolution reflects both technological advancements in computational capabilities and the increasing availability of large dermatological image datasets.

Early approaches to automated skin disease classification emerged in the late 1990s and early 2000s, focusing primarily on the detection and classification of melanoma. These systems typically employed conventional image processing techniques to extract predefined features such as asymmetry, border irregularity, color variegation, and diameter (the ABCD criteria) from dermoscopic images [**?**]. Rubegni et al. [**?**] demonstrated one of the first successful applications of artificial neural networks in this domain, achieving 92% accuracy in distinguishing melanomas from nevi using 13 geometric and colorimetric features. Similarly, Ganster et al. [13] developed a system that combined feature extraction with k-nearest neighbor classification, reporting 87% sensitivity for melanoma detection.

The evolution continued with more sophisticated feature extraction methods coupled with ensemble classifiers. Celebi et al. [4] employed feature selection techniques to identify the most discriminative attributes from a pool of color, texture, and border features, improving classification performance while reducing computational complexity. Maglogiannis and Doukas [**?**] provided a comprehensive review of these conventional machine learning approaches, highlighting their reliance on domain expertise for feature engineering—a limitation that would later be addressed by deep learning methods.

2.2. Deep Learning Approaches in Dermatological Image Analysis

The landscape of skin disease classification was fundamentally transformed with the introduction of deep learning techniques, particularly convolutional neural networks (CNNs). Unlike traditional methods, CNNs can automatically learn hierarchical feature representations directly from raw image data, eliminating the need for manual feature engineering.

A seminal contribution to this field came from Esteva et al. [10], who demonstrated for the first time that deep neural networks could achieve dermatologist-level accuracy in classifying skin cancer. Training an Inception v3 CNN architecture on a dataset of 129,450 clinical images spanning 2,032 different diseases, their system achieved performance comparable to 21 board-certified dermatologists for the tasks of keratinocyte carcinoma classification and melanoma recognition. This groundbreaking work established the potential of deep learning in dermatology and catalyzed subsequent research in this domain.

Building upon this foundation, Han et al. [**?**] developed a classifier based on a ResNet-152 architecture that could differentiate between 12 skin diseases, including both malignant and benign conditions. Their system achieved a mean area under the curve (AUC) of 0.95, performing comparably to 16 dermatologists on a validation set of 1,300 images. Similarly, Haenssle et al. [17] reported that a CNN-based system outperformed 58 international dermatologists in melanoma detection from dermoscopic images, achieving a higher sensitivity (95% vs. 86.6%) while maintaining comparable specificity.

Transfer learning has emerged as a particularly effective approach in medical image analysis, where large annotated datasets may be limited. Kawahara et al. [**?**] demonstrated the efficacy of fine-tuning pre-trained CNNs for skin lesion classification, achieving 81.8% accuracy across 10 categories using the Dermofit Image Library. Menegola et al. [**?**] explored various transfer learning strategies using models pre-trained on ImageNet, showing significant performance improvements compared to training from scratch.

Recent years have witnessed further architectural innovations tailored specifically for dermatological applications. Li and Shen [**?**] proposed a dense-attention network that combines features from different CNN layers with attention mechanisms, achieving state-of-the-art performance on the ISIC 2017 skin lesion classification challenge. Gessert et al. [15] introduced a multi-resolution approach

 doi:10.20944/preprints202504.2588.v1

that processes dermoscopic images at different scales, capturing both fine-grained details and broader contextual information to improve classification accuracy.

2.3. Multi-Class Classification of Diverse Skin Conditions

While early deep learning research in dermatology focused primarily on binary classification tasks (particularly melanoma detection), more recent efforts have addressed the more challenging problem of multi-class classification across diverse skin conditions.

Yang et al. [**?**] developed one of the first comprehensive systems for classifying clinical images across multiple common skin diseases, training an ensemble of deep learning models on a dataset of over 6,000 images spanning 198 skin disease classes. Their approach achieved a top-1 accuracy of 67.1% and top-5 accuracy of 85.3%, demonstrating the feasibility of multi-class classification even with relatively limited data per class.

The SD-198 dataset introduced by Sun et al. [**?**] has become an important benchmark for evaluating multi-class skin disease classification systems. Utilizing this dataset, Groh et al. [16] explored various CNN architectures and training strategies, achieving 77.8% accuracy across all 198 classes. Their work highlighted the challenges of distinguishing between visually similar conditions and the impact of class imbalance on model performance.

More recently, Liu et al. [**?**] proposed a dual-branch CNN architecture that processes both clinical and dermoscopic images when available, showing improved performance compared to single-image modality approaches. Similarly, Zhang et al. [**?**] developed an attention-guided network that selectively focuses on discriminative regions within skin lesion images, achieving 89.0% accuracy on a dataset comprising seven common skin diseases.

2.4. Addressing Diversity and Fairness in Skin Disease Classification

A critical limitation of many existing skin disease classification systems is their development and validation on datasets that lack demographic diversity. Several studies have demonstrated that this can lead to algorithmic biases and performance disparities across different population groups.

Daneshjou et al. [8] evaluated the performance of deep learning models across different skin tones, finding significant accuracy disparities between lighter and darker skin types. On independent test sets, they observed a 10-15% reduction in accuracy for images of darker skin tones compared to lighter ones. Similarly, Kinyanjui et al. [**?**] analyzed the distribution of skin tones in publicly available dermatology datasets, finding severe underrepresentation of darker skin types.

Recent efforts have begun to address these disparities through more inclusive dataset curation and targeted algorithmic approaches. Groh et al. [16] employed techniques such as balanced sampling and loss reweighting to mitigate performance gaps across demographic groups. Oktay et al. [**?**] proposed a fairness-aware learning algorithm that explicitly penalizes disparate error rates between predefined sensitive groups during model training.

The Fitzpatrick 17k dataset, introduced by Groh et al. [16], represents an important step toward more equitable dermatological AI, containing over 16,577 images with balanced representation across different Fitzpatrick skin types. Models trained on this dataset have demonstrated more consistent performance across demographic groups, though disparities have not been entirely eliminated.

2.5. Interpretability and Explainability in Dermatological AI

As deep learning models become increasingly integrated into clinical dermatology, the need for interpretable and explainable systems has gained prominence. Several approaches have been developed to provide insights into the decision-making processes of dermatological classification models.

Visualization techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM), introduced by Selvaraju et al. [**?**], have been widely adopted to highlight regions of interest that influence model predictions. Young et al. [**?**] applied these techniques to melanoma classifica-

 doi:10.20944/preprints202504.2588.v1

tion, demonstrating that CNNs often focus on clinically relevant features similar to those used by dermatologists, though sometimes relying on unexpected image regions.

Tschandl et al. [**?**] conducted a human-computer collaboration study in which dermatologists were provided with model explanations during the diagnostic process. They found that interpretable AI assistance improved diagnostic performance by an average of 13% compared to either humans or AI alone, highlighting the potential of explainable systems as clinical decision support tools.

More recently, Barata et al. [3] proposed a concept-based explanation framework that identifies high-level dermatological attributes (such as pigment network, streaks, and globules) associated with model predictions. This approach provides explanations in clinically familiar terms, potentially increasing trustworthiness and adoption among medical practitioners.

2.6. Comparative Analysis of Current Approaches

Table 1 presents a comparative analysis of recent deep learning approaches for skin disease classification, highlighting their architectural choices, dataset characteristics, performance metrics, and limitations.

Table 1. Comparative analysis of recent deep learning approaches for skin disease classification

Study	Architecture	Dataset	Performance	Limitations
Esteva et al. [10] (2017)	Inception v3	129,450 images, 2,032 diseases	AUC 0.96 for keratinocyte carcinoma, 0.94 for melanoma	Limited demographic diversity; binary classification focus
Han et al. [**?**] (2018)	ResNet-152	15,408 images, 12 diseases	Mean AUC 0.95 across all classes	Limited to common conditions; single-center dataset
Tschandl et al. [**?**] (2019)	Ensemble of ResNet-50, SE-ResNeXt-50	HAM10000 dataset, 7 skin conditions	Accuracy 87.3%, mean AUC 0.93	Limited to dermoscopic images; seven disease classes only
Liu et al. [**?**] (2020)	Dual-branch CNN with SENet	8,545 images (clinical and dermoscopic), 9 diseases	Accuracy 89.7%, F1-score 0.87	Requires both clinical and dermoscopic images for optimal performance
Gessert et al. [15] (2020)	Multi-resolution ResNet-50	ISIC 2019 dataset, 8 disease classes	Balanced accuracy 63.9%, AUC 0.93 for melanoma	Limited to dermoscopic images; moderate performance on rare classes
Daneshjou et al. [8] (2021)	DenseNet-121	Multi-source dataset with Fitzpatrick skin type annotations	10-15% lower accuracy on darker skin tones	Highlights disparities but does not fully resolve them
Wu et al. [**?**] (2022)	EfficientNet-B4 with attention	45,000 images, 26 disease classes	Top-1 accuracy 82.6%, top-3 accuracy 95.7%	Limited evaluation across demographic subgroups

 doi:10.20944/preprints202504.2588.v1

As evidenced by this comparative analysis, significant progress has been made in developing deep learning systems for skin disease classification, with performance metrics approaching or exceeding those of expert dermatologists for specific tasks. However, several limitations and challenges persist across current approaches.

First, most systems have been evaluated on relatively homogeneous datasets that do not adequately represent the full spectrum of skin tones, age groups, and anatomical sites encountered in clinical practice. This limits generalizability and raises concerns about performance disparities across different patient populations.

Second, while binary classification tasks (particularly melanoma detection) have received extensive attention, multi-class classification across diverse skin conditions remains challenging, with performance typically decreasing as the number of classes increases. This reflects both the inherent difficulty of distinguishing between visually similar conditions and the limited availability of training data for rare diseases.

Third, despite advances in visualization and explanation techniques, the interpretability of deep learning models in dermatology remains imperfect. Current approaches can highlight relevant image regions but often fall short of providing clinically meaningful explanations that align with dermatological reasoning processes.

Fourth, most existing systems rely solely on visual information from single images, without incorporating contextual factors such as patient demographics, medical history, or lesion evolution that dermatologists typically consider in their diagnostic assessments.

Our proposed approach aims to address these limitations through a modified EfficientNet architecture with integrated attention mechanisms, trained on a demographically diverse dataset spanning multiple disease categories. By incorporating visualization techniques and evaluating performance across different patient subgroups, we seek to develop a system that not only achieves high classification accuracy but also addresses considerations of fairness, interpretability, and clinical applicability.

3. Research Methodology

This section details our methodological approach to developing a convolutional neural network-based system for skin disease classification. We present a comprehensive overview of the dataset acquisition and preparation process, the proposed CNN architecture, training and optimization strategies, and the evaluation framework.

3.1. Dataset Acquisition and Preparation

3.1.1. Data Sources

To ensure robust performance across diverse skin conditions and patient demographics, we compiled a comprehensive dataset from multiple sources:

The primary components of our dataset include:

The HAM10000 dataset [**?**], which contains 10,015 dermatoscopic images across seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). This dataset provides detailed metadata including patient age, sex, and anatomical site.

The ISIC 2019 Challenge dataset [7], comprising 25,331 dermoscopic images of skin lesions across eight diagnostic categories. This dataset represents an extension of previous ISIC challenge datasets with additional images and refined ground truth annotations.

The SD-198 dataset [**?**], containing 6,584 clinical images across 198 skin disease classes. This dataset offers exceptional disease diversity, including both common and rare conditions, captured under varying imaging conditions.

 doi:10.20944/preprints202504.2588.v1

The Fitzpatrick 17k dataset [16], which includes 16,577 images with balanced representation across different Fitzpatrick skin types (I-VI). This dataset was specifically designed to address algorithmic fairness concerns and enable evaluation across diverse skin tones.

To supplement these public resources and address specific gaps in demographic or disease representation, we collaborated with the Dermatology Department at [University/Hospital Name] to collect and annotate an additional 3,500 clinical images. These images represent patients of diverse ethnicities, age groups, and skin conditions, with special attention to conditions that were underrepresented in the public datasets. All images were collected with informed consent and institutional review board approval, following appropriate anonymization protocols.

3.1.2. Data Preprocessing and Augmentation

All images underwent a standardized preprocessing pipeline to ensure consistency and optimize model training:

Initial quality assessment: Images were reviewed for quality, with those exhibiting severe artifacts, excessive blur, or improper framing excluded from the dataset.

Standardization: Images were resized to a uniform dimension of 224 × 224 pixels, consistent with the input requirements of the EfficientNet-B3 architecture. Color normalization was applied to address variations in lighting conditions and camera specifications.

Class balancing: To address the inherent class imbalance in dermatological datasets, where common conditions typically have substantially more examples than rare ones, we implemented a combination of oversampling for minority classes and undersampling for majority classes. For classes with fewer than 100 examples, we applied more aggressive augmentation to generate additional synthetic examples.

Data augmentation: To improve model generalization and robustness, we applied the following augmentation techniques to the training set:

- Random rotations (±30°)
- Horizontal and vertical flips
- Random brightness and contrast adjustments (±15%)
- Random zoom (0.8-1.2x)
- Slight color jitter (hue and saturation shifts of ±10%)
- Random cropping with minimum 85% area coverage
- Cutout regularization with random 32×32 pixel patches

These augmentation techniques were applied with varying probabilities during training to create a diverse set of training examples while preserving the essential diagnostic characteristics of the skin lesions. For validation and testing, only resize and normalization operations were performed to maintain the integrity of the evaluation process.

3.1.3. Dataset Stratification

The combined dataset was carefully stratified to ensure representative distribution across training, validation, and testing sets:

Training set (70%): Used for model parameter optimization Validation set (15%): Used for hyperparameter tuning and early stopping Test set (15%): Used exclusively for final performance evaluation

Stratification was performed to maintain consistent proportions of each disease class across all three sets. Additionally, we ensured that images from the same patient were assigned to the same set to prevent data leakage, a critical consideration in medical image analysis where multiple images may come from the same individual.

Furthermore, we created demographically balanced test subsets to enable targeted evaluation of model performance across different skin tones, age groups, and anatomical sites. This stratification

 doi:10.20944/preprints202504.2588.v1

strategy allowed us to assess potential performance disparities and ensure that the model performs consistently across diverse patient populations.

3.2. CNN Architecture and Implementation

3.2.1. Base Architecture Selection

After evaluating several state-of-the-art CNN architectures through preliminary experiments, we selected EfficientNet-B3 [**?**] as our base architecture due to its optimal balance between computational efficiency and representational capacity. EfficientNet employs compound scaling that uniformly scales network width, depth, and resolution with a fixed set of scaling coefficients, resulting in models that achieve higher accuracy with fewer parameters compared to conventional CNNs.

The EfficientNet-B3 architecture consists of 7 mobile inverted bottleneck MBConv blocks with varying expansion ratios and kernel sizes, totaling 12 million parameters. Its efficient design makes it suitable for potential deployment in resource-constrained healthcare environments, while its performance on image classification benchmarks demonstrates its capacity to learn complex visual patterns relevant to dermatological diagnosis.

3.2.2. Architectural Modifications

We implemented several key modifications to the base EfficientNet-B3 architecture to optimize it specifically for skin disease classification:

Attention mechanism integration: We incorporated Squeeze-and-Excitation (SE) blocks [**?**] at multiple levels of the network to enhance feature selectivity. These blocks recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels, allowing the network to emphasize informative features and suppress less useful ones. This is particularly valuable in dermatological image analysis, where specific textural and morphological patterns are often diagnostically significant.

The SE blocks were implemented following each MBConv block, with a reduction ratio of 16. The integration process involves:

1. Global average pooling to squeeze spatial information
2. A two-layer fully connected network with bottleneck structure to generate channel-wise attention weights
3. Rescaling of the original feature maps using these weights

Custom classification head: We replaced the original classification layer with a custom head designed to improve discrimination between visually similar skin conditions:

1. Global average pooling layer to aggregate spatial information
2. Dropout layer (rate = 0.5) to reduce overfitting
3. Fully connected layer with 1024 units and ReLU activation
4. Batch normalization layer to stabilize training
5. Dropout layer (rate = 0.3) for additional regularization
6. Final fully connected layer with softmax activation, outputting probabilities for N disease classes

Multi-scale feature fusion: To capture both fine-grained textural details and broader contextual information, we implemented a feature pyramid network (FPN) inspired approach that combines features from different levels of the network. This allows the model to simultaneously analyze skin lesions at multiple spatial resolutions, which is crucial for accurate diagnosis as certain conditions are characterized by patterns at different scales.

3.2.3. Implementation Details

The model was implemented using the TensorFlow 2.6 framework with Keras API. For reproducibility, we set a fixed random seed (42) for all random operations. All experiments were conducted on an NVIDIA Tesla V100 GPU with 32GB memory, using mixed precision training to optimize computational efficiency.

 doi:10.20944/preprints202504.2588.v1

The implementation code, including data preprocessing, model architecture, training procedures, and evaluation metrics, has been made publicly available on GitHub at [repository URL] to facilitate reproducibility and further research in this area.

3.3. Training Strategy and Optimization

3.3.1. Transfer Learning Approach

Given the relative scarcity of large-scale dermatological image datasets compared to natural image collections, we employed transfer learning to leverage knowledge from pre-training on the ImageNet dataset. Our training process followed a three-phase approach:

Phase 1 - Feature extraction (10 epochs):

- The EfficientNet-B3 base model weights were frozen
- Only the custom classification head was trained
- Learning rate: 1e-3, Optimizer: Adam

Phase 2 - Partial fine-tuning (20 epochs):

- The final 50 layers of the base model were unfrozen
- Both these layers and the classification head were trained
- Learning rate: 5e-4, Optimizer: Adam with weight decay (1e-5)

Phase 3 - Full fine-tuning (30 epochs):

- The entire network was unfrozen and trained end-to-end
- Learning rate: 1e-4, Optimizer: Adam with weight decay (1e-5)
- Cosine annealing learning rate schedule with warm restarts

This progressive unfreezing approach allowed the model to gradually adapt from general image recognition to the specific characteristics of dermatological images while reducing the risk of catastrophic forgetting of useful pre-trained features.

3.3.2. Loss Function and Class Weighting

To address the inherent class imbalance in our dataset, we employed a weighted categorical cross-entropy loss function. Class weights were computed inversely proportional to class frequencies using the formula:

$$w_c = \frac{N}{C \times n_c} \tag{1}$$

where N is the total number of samples, C is the number of classes, and n_c is the number of samples in class c.

Additionally, we incorporated focal loss components [**?**] to further address class imbalance by dynamically adjusting the contribution of easy and hard examples during training:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{2}$$

where p_t is the model's estimated probability for the ground truth class, α_t is the class weight, and γ is the focusing parameter (set to 2.0 in our experiments).

3.3.3. Regularization and Early Stopping

To prevent overfitting and improve generalization, we implemented multiple regularization techniques:

Dropout at two levels in the classification head (rates 0.5 and 0.3) L2 weight regularization (1e-5) on all convolutional and fully connected layers Batch normalization after convolutional layers and in the classification head Mixup data augmentation [**?**] with alpha=0.2, which creates virtual training examples by linear interpolation of input images and labels

 doi:10.20944/preprints202504.2588.v1

We implemented early stopping based on validation loss with a patience of 7 epochs and model checkpointing to save the best-performing model according to validation F1-score. This strategy ensured that training could proceed long enough to converge while preventing overfitting to the training data.

3.4. Evaluation Framework

3.4.1. Performance Metrics

We employed a comprehensive set of metrics to evaluate our model's performance across different dimensions:

Overall accuracy: Proportion of correctly classified samples across all classes Per-class precision, recall, and F1-score: To assess performance on individual disease categories Macro-averaged F1-score: Unweighted mean of per-class F1-scores, giving equal importance to all classes regardless of their support Weighted F1-score: Average of per-class F1-scores weighted by support, reflecting overall performance while accounting for class imbalance Confusion matrix: To identify specific patterns of misclassification between similar disease categories Area Under the Receiver Operating Characteristic curve (AUROC): For each class in a one-vs-all setting Mean Average Precision (mAP): To evaluate ranking quality across classes

For clinically critical conditions such as melanoma and other malignancies, we additionally reported sensitivity, specificity, and negative predictive value, as these metrics have direct relevance to clinical decision-making.

3.4.2. Comparative Analysis

To contextualize our results, we conducted a comparative analysis across multiple dimensions:

Architecture comparison: We evaluated several state-of-the-art CNN architectures (ResNet50, DenseNet121, InceptionV3, EfficientNet-B0/B3/B5) using identical training procedures and datasets to isolate the impact of architectural choices.

Ablation studies: We systematically assessed the contribution of each key component of our approach through ablation experiments, including:

- Impact of attention mechanisms (with vs. without SE blocks)
- Effect of multi-scale feature fusion
- Contribution of different data augmentation strategies
- Influence of transfer learning and progressive unfreezing

External validation: To evaluate generalizability, we tested our model on external datasets not used during training, including a subset of the Dermofit Image Library and a prospectively collected set of clinical images from a different medical center.

Comparison with dermatologists: We conducted a reader study in which a subset of 300 test images was independently evaluated by 6 board-certified dermatologists with varying levels of experience. This allowed direct comparison between our model and human experts on identical test cases.

3.4.3. Fairness and Bias Assessment

To assess potential algorithmic biases and performance disparities, we evaluated model performance across different demographic subgroups:

Skin tone analysis: Performance stratified by Fitzpatrick skin types I-VI Age group analysis: Performance compared across pediatric, adult, and geriatric populations Anatomical site analysis: Performance evaluated based on lesion location (face, trunk, extremities, etc.)

For each subgroup analysis, we reported the same comprehensive set of performance metrics to identify any significant disparities. We quantified performance gaps using the Equalized Odds Difference (EOD) metric, which measures the maximum difference in false positive and false negative rates between demographic groups.

 doi:10.20944/preprints202504.2588.v1

3.4.4. Model Interpretability Analysis

To enhance transparency and trust in model predictions, we implemented multiple visualization and explanation techniques:

Gradient-weighted Class Activation Mapping (Grad-CAM) [**?**]: To generate heatmaps highlighting image regions most influential to the model's predictions Integrated Gradients [**?**]: To attribute predictions to specific input features with pixel-level granularity Concept-based explanations: Using concept bottleneck models to identify high-level dermatological attributes (e.g., pigment networks, blue-white structures) associated with specific predictions

To validate the clinical relevance of these explanations, we conducted a qualitative assessment in which three experienced dermatologists reviewed a sample of 100 Grad-CAM visualizations, rating their concordance with clinically significant regions and dermatological reasoning patterns.

This comprehensive evaluation framework enabled us to assess not only the overall accuracy of our approach but also its performance consistency across diverse patient populations, its comparative advantages relative to existing methods, the contributions of individual components, and its potential for integration into clinical workflows through interpretable predictions.

References

1. American Cancer Society, "Cancer facts & figures 2023," American Cancer Society, Atlanta, GA, 2023.
2. G. Argenziano, et al., "Twenty years of dermoscopy," Journal of the American Academy of Dermatology, vol. 81, no. 4, pp. 1088-1086, 2019.
3. C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," Pattern Recognition, vol. 110, pp. 107413, 2021.
4. M. E. Celebi, et al., "A methodological approach to the classification of dermoscopy images," Computerized Medical Imaging and Graphics, vol. 31, no. 6, pp. 362-373, 2007.
5. M. E. Celebi, H. A. Kingravi, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, R. H. Moss, J. M. Malters, J. M. Grichnik, A. A. Marghoob, H. S. Rabinovitz, and S. W. Menzies, "Border detection in dermoscopy images using statistical region merging," Skin Research and Technology, vol. 14, no. 3, pp. 347-353, 2008.
6. N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "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)," IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168-172, 2018.
7. N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.
8. R. Daneshjou, M. Vodrahalli, V. Novoa, M. Marsch, M. Alam, J. Y. Lee, S. Abrouk, H. Rabinovitz, M. Frazier, S. Sadeghpour, A. Young, L. Phillips and J. Zou, "Disparities in dermatology AI: Fewer skin of color images, lower performance, and fairness warnings," Journal of the American Academy of Dermatology, vol. 86, no. 1, pp. 103-114, 2021.
9. R. Daneshjou, R. Zakeri, and J. Zou, "Artificial intelligence and dermatology: opportunities, challenges, and future directions," JAMA Dermatology, vol. 158, no. 3, pp. 318-324, 2022.
10. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.
11. H. Feng, J. Berk-Krauss, P. W. Feng, and J. A. Stein, "Comparison of dermatologist density between urban and rural counties in the United States," JAMA Dermatology, vol. 154, no. 11, pp. 1265-1271, 2018.
12. A. Y. Finlay, R. J. Hay, N. C. Dlova, S. A. Garg, R. Joshipura, and S. Lulla, "Global alliance for patients with serious skin diseases: a way forward," Journal of the American Academy of Dermatology, vol. 76, no. 2, pp. 368-370, 2017.
13. H. Ganster, A. Pinz, R. Röhrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," IEEE Transactions on Medical Imaging, vol. 20, no. 3, pp. 233-239, 2001.
14. A. Garza-Mayers and K. C. McClain, "Telemedicine in deep disparities: a persistent pandemic illuminates the need to address a perennial problem," Journal of the American Academy of Dermatology, vol. 83, no. 6, pp. e401-e402, 2020.

 doi:10.20944/preprints202504.2588.v1

15. N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," IEEE Transactions on Biomedical Engineering, vol. 67, no. 2, pp. 495-503, 2020.
16. M. Groh, C. Harris, A. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1820-1828, 2021.
17. H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, "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.