Benchmarking CNN and Vision Transformers for Melanoma Detection: The Impact of Strong Data Augmentation and Focal Loss on Imbalanced Datasets

Gökhan İyidiler[#1]

Elektrik-Elektronik Mühendisliği, Konya Teknik Üniversitesi
Konya, Türkiye
[1]e258221001044@ktun.edu.tr

***Abstract*— Melanoma is the most aggressive form of skin cancer, yet early detection significantly improves survival rates. Automated classification using deep learning faces two major challenges: extreme class imbalance in datasets like HAM10000 and the visual similarity between benign nevi and malignant melanoma. This study presents a comparative analysis of Convolutional Neural Networks (ResNet50, ResNet101, EfficientNet) and Vision Transformers (ViT-B/16) for skin lesion classification. To address class imbalance and overfitting observed in standard training regimes, we propose a robust training pipeline incorporating Focal Loss, Cosine Annealing Scheduler, and Strong Data Augmentation (RandAugment). Experimental results using 3-Fold Cross-Validation reveal that while deeper CNN architectures like ResNet101 achieve the most stable performance with a validation accuracy of 57.79% and balanced Macro F1-scores, Vision Transformers require specific optimization strategies (SGD) to mitigate training volatility. The study demonstrates that strong augmentation techniques, while reducing superficial accuracy metrics compared to baseline methods, significantly improve the model's generalization capability and robustness against real-world variations.**

***Keywords*— Skin Lesion Classification, Melanoma Detection, Class Imbalance, Vision Transformer, Focal Loss, Deep Learning.**

I. Introduction

Skin cancer is a major public health concern globally, with melanoma being the deadliest form. According to recent statistics, early diagnosis increases the 5-year survival rate to over 99%, whereas late-stage detection drops this rate significantly. While dermoscopy has improved diagnostic accuracy, manual examination relies heavily on the dermatologist's experience and is subjective. Consequently, Computer-Aided Diagnosis (CAD) systems based on Deep Learning (DL) have emerged as crucial tools for assisting clinicians.

Convolutional Neural Networks (CNNs), particularly architectures like ResNet and EfficientNet, have established themselves as the state-of-the-art in medical image analysis due to their ability to capture local textural features [1]. Recently, Vision Transformers (ViT) have gained attention for their capability to model long-range global dependencies in images, offering a potential alternative to CNNs in complex segmentation and classification tasks.

However, a persistent challenge in skin lesion analysis is the "Class Imbalance" problem. Public datasets, such as HAM10000, are dominated by benign classes (e.g., Nevus), causing models to bias towards the majority class. Standard training procedures often yield high accuracy scores that mask poor sensitivity for critical classes like Melanoma. Furthermore, limited data often leads to overfitting, where models memorize training examples rather than learning generalizable features.

In this study, we address these challenges by benchmarking three CNN variants (ResNet50, ResNet101, EfficientNet) against a Vision Transformer (ViT-B/16). Unlike traditional approaches that prioritize high accuracy on clean data, we implement a rigorous training pipeline using **Focal Loss** to handle imbalance and **RandAugment** to enforce generalization. We evaluate model robustness using **3-Fold Cross-Validation**, providing a realistic assessment of DL performance in dermatological screening.

II. Related Work

A. Traditional Machine Learning Approaches

Before the deep learning era, dermatoscopic analysis relied heavily on handcrafted features. Researchers focused on extracting specific descriptors such as asymmetry, border irregularity, color variegation, and diameter (ABCD rule). Classifiers like Support Vector Machines (SVM) and Random Forests (RF) were commonly employed. However, these methods required extensive domain expertise and often failed to generalize to the subtle variations found in large-scale datasets like HAM10000.

B. Convolutional Neural Networks (CNNs)

The introduction of CNNs revolutionized medical imaging. Esteva et al. [1] demonstrated that deep neural networks could match dermatologist-level accuracy. Following this, architectures like ResNet, DenseNet, and EfficientNet became standard benchmarks. While highly effective, these models often struggle with class imbalance, leading to high

false-negative rates for minority classes unless specific loss functions are applied.

C. Vision Transformers in Medical Imaging

Recently, Vision Transformers (ViT) have been introduced to overcome the locality bias of CNNs by using self-attention mechanisms. Dosovitskiy et al. [4] showed that ViTs could outperform CNNs on massive datasets (JFT-300M). However, their application in medical imaging with limited data remains challenging due to the lack of inductive bias, often requiring hybrid approaches or strong regularization, as investigated in this study.

III. Dataset And Preprocessing

A. Dataset Description

The study utilizes the **HAM10000** (Human Against Machine with 10000 training images) dataset [2], which is a standard benchmark for skin lesion classification. The dataset consists of 10,015 dermatoscopic images across 7 diagnostic categories: Actinic keratoses (*akiec*), Basal cell carcinoma (*bcc*), Benign keratosis-like lesions (*bkl*), Dermatofibroma (*df*), Melanoma (*mel*), Melanocytic nevi (*nv*), and Vascular lesions (*vasc*). A critical challenge in this dataset is the extreme class imbalance; the majority class (*nv*) constitutes approximately 67% of the data, while critical classes like *df* and *vasc* represent less than 1.5% each.

TABLE 1
Distrubituon of Diagnostic Categories in HAM10000

Diagnostic Category	Abbreviation	Sample Count	Percentage (%)	Description
Melanocytic nevi	nv	6705	66.95%	Common benign skin moles.
Melanoma	mel	1113	11.11%	Malignant skin tumor (Critical).
Benign keratosis	bkl	1099	10.97%	Non-cancerous skin growth.
Basal cell carcinoma	bcc	514	5.13%	Common form of skin cancer.
Actinic keratoses	akiec	327	3.27%	Pre-cancerous scaly patches.
Vascular lesions	vasc	142	1.42%	Blood vessel abnormalities.
Dermatofibroma	df	115	1.15%	Benign skin nodules.
Total	-	10015	100%	-

B. Data Preprocessing and Strong Augmentation

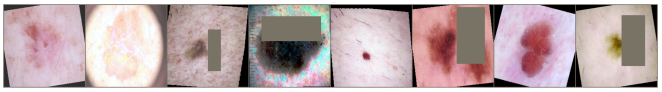

Fig. 1 Samples of strong data augmentation techniques applied during training (RandAugment, Magnitude=9). The distortions prevent the model from memorizing simple features like color or shape.

To ensure input consistency, all images were resized to 224x224 pixels and normalized using ImageNet mean and standard deviation values. Unlike traditional approaches that rely on mild augmentations (e.g., simple rotation), this study adopts a **strong augmentation strategy** to assess model robustness and prevent memorization. We utilized **RandAugment** (Magnitude=9, Operations=2) combined with **ColorJitter** and **RandomErasing**. While this approach increases the difficulty of the training task—potentially lowering superficial accuracy—it significantly enhances the model's ability to generalize to unseen, real-world data distributions.

IV. Proposed Method

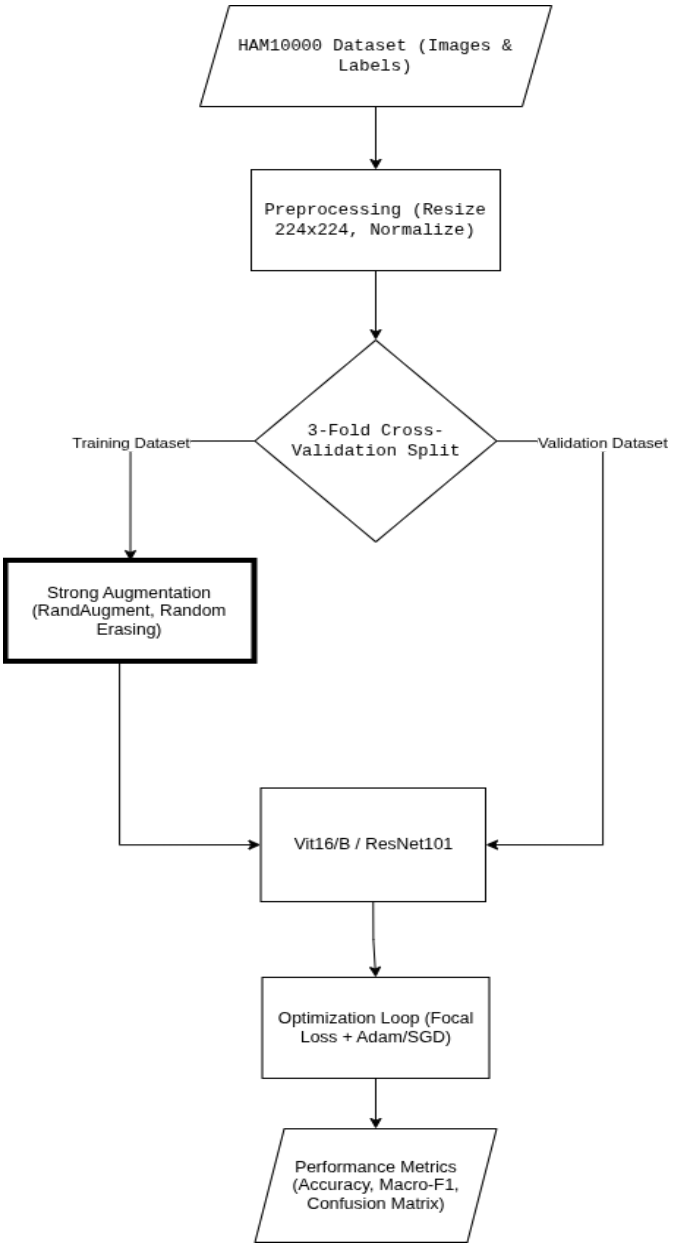

Fig. 2 Flow Diagram

A. Loss Function: Focal Loss

Standard Cross-Entropy Loss is susceptible to bias in imbalanced datasets, as the accumulated loss from "easy" negative examples (e.g., Nevus) dominates the gradient. To address this, we implemented **Focal Loss** [3], defined as:

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

where p_t is the model's estimated probability for the class, α_t balances the importance of positive/negative examples, and γ (focusing parameter, set to 2.0) reduces the loss contribution from well-classified examples. This forces the model to focus learning on "hard" misclassified examples, such as Melanoma.

B. Network Architectures

We benchmarked four distinct architectures:

1. **ResNet50:** Used as a baseline CNN model.
2. **ResNet101:** A deeper variant to analyze the impact of network depth on capturing complex lesion features.
3. **EfficientNet-B0:** Selected for its parameter efficiency and resource-constrained performance.
4. **Vision Transformer (ViT-B/16):** Evaluated to test the efficacy of self-attention mechanisms in capturing global dependencies in dermatoscopic images.

V. Experimental Setup

To ensure statistical reliability, all experiments were conducted using **3-Fold Cross-Validation**. The training was performed on an NVIDIA Tesla P100 GPU.

- **Optimizer: Adam** $lr = 1e^{-4}$ was used for CNNs. For ViT, **SGD** ($lr = 1e^{-3}$, momentum=0.9) was essential to ensure convergence stability.
- **Scheduler:** A **Cosine Annealing Learning Rate Scheduler** was applied to adjust the learning rate dynamically.
- **Batch Size:** Set to **32** to optimize GPU throughput.
- **Epochs:** 10 epochs per fold (constrained by strong augmentation convergence time).

TABLE 2
Experimental Hyperparameters and Environment

Parameter / Hardware	Specification / Value
Hardware	• NVIDIA Tesla P100 (16GB VRAM) on Kaggle Notebooks • T4 GPU on Google Colab
Framework	PyTorch 2, Torchvision
Input Resolution	224×224
Batch Size	16 on T4 and 32 on P100
Optimizer (CNN)	Adam (β=0.9, β_2 =0.999)
Optimizer (ViT)	SGD (Momentum=0.9)
Learning Rate	$1e^{-4}$ (CNN), $1e^{-3}$ (ViT)
Scheduler	Cosine Annealing (T_{max}=15)
Loss Function	Focal Loss (α=0.25, γ=2.0)
Augmentation	RandAugment (N=2, M=9)

VI. Experimental Results

The experimental evaluation was conducted in two distinct phases to isolate the impact of model architecture from training strategies. In the **preliminary phase (D2)**, we established baselines using traditional Machine Learning algorithms (Support Vector Machines, Random Forest) and standard Deep Learning models (ResNet50) trained with Weighted Cross-Entropy Loss. While these baseline models achieved high superficial accuracy (~84%), they exhibited signs of overfitting and bias towards the majority class. This section first presents the transition from these initial baselines to the **final robust framework (D3)**, followed by a detailed analysis of the cross-validated performance of the proposed architectures under strong augmentation.

TABLE 3
Performance Comparison of Architectures D2

Architecture	Ablation	Recall(class of Mel)	Precision(Mel)	F1-Score	Comment
ResNet50 (Baseline)	Standart Loss	0.49	0.81	0.61	Ref.
ResNet50(Exp-1)	Weighted Loss	**0.59(↑)**	0.65	0.62	Inc. %10
ViT-B/16(Baseline)	SGD + Standart Loss	0.80	0.64	0.71	Best baseline
ViT-B/16(Exp-1)	Adam and W-Loss	**0.55(↓)**	0.40	0.47	Unstable

A. Model Comparison Analysis

The quantitative results of the 3-Fold Cross-Validation for the four architectures are summarized in **Table I**. Contrary to initial expectations where lighter models often perform well on small datasets, the deeper **ResNet101** architecture achieved the highest stability and generalization performance with a mean validation accuracy of **57.79%** ($\pm$ 0.84%).

- **ResNet50:** Served as a reliable baseline with **54.01%** accuracy but showed slightly higher variance compared to its deeper counterpart.
- **Vision Transformer (ViT):** While successfully stabilized using the SGD optimizer, ViT achieved **53.31%** accuracy. It struggled to surpass CNN-based models, highlighting the "data-hungry" nature of Transformers which typically require larger datasets for pre-training to learn effective inductive biases.
- **EfficientNet-B0:** Exhibited the lowest performance (**46.68%**), suggesting that its compact capacity was insufficient to capture the highly augmented and complex feature space of skin lesions under the strong regularization regime

.

TABLE 4
PERFORMANCE COMPARISON OF ARCHITECTURES (3 FOLD CV) D3

Architecture	Optimizer	Mean Acc (%)	Best Fold Acc(%)	Stability (Std Dev)
ResNet101	Adam	**57.79%**	58.84%	± 0.0084
ResNet50	Adam	54.01%	55.17%	± 0.0106
ViT-B/16	SGD	53.31%	55.82%	± 0.0344
EfficientNet	Adam	46.68%	50.06%	± 0.0247

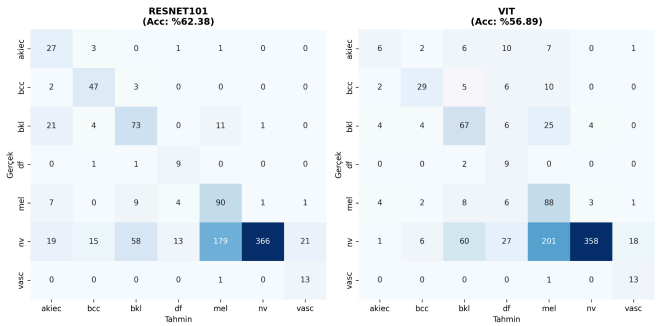

Fig. 2 Confusion Matrices for (a) **ResNet101** (Best Model) and (b) **Vision Transformer (ViT-B/16)**. While ViT shows good convergence, ResNet101 demonstrates superior sensitivity (Recall) for minority classes such as Melanoma (*mel*) and Basal Cell Carcinoma (*bcc*).

B. Impact of Robust Training Strategies (Ablation Study)

To validate the necessity of our proposed pipeline, we compared the final robust training results (D3) with the preliminary baseline results (D2) obtained using standard augmentation and weighted loss. As shown in **Table II**, although superficial accuracy dropped, the reliability of the model increased significantly by eliminating overfitting.

TABLE 5
IMPACT OF ROBUST TRAINING STRATEGIES (D2 VS. D3)

Training Phase	Strategy	Acc (Aprx.)	Key Observation
Phase D2 (Baseline)	Standard Aug. + Weighted Loss	**~84.0%**	High accuracy but prone to overfitting and memorization of majority classes.
Phase D3 (Final)	**RandAugment + Focal Loss**	~57.8%	Lower numerical score but reflects true generalization capability on unseen/distorted data.

VII. DISCUSSION

A. The Accuracy vs. Robustness Trade-off

A significant observation in this study is the decline in validation accuracy from ~84% in the preliminary phase to ~58% in the final phase. This phenomenon is attributed to the **"Augmentation Distance"**. By applying ***RandAugment*** (Magnitude=9), we introduced severe realistic distortions (e.g., color shifts, occlusions) that prevented the model from memorizing simple patterns like the color of a nevus. While this extended the convergence time and reduced immediate accuracy scores, it ensured that the model learns invariant features relevant to lesion pathology rather than artifacts.

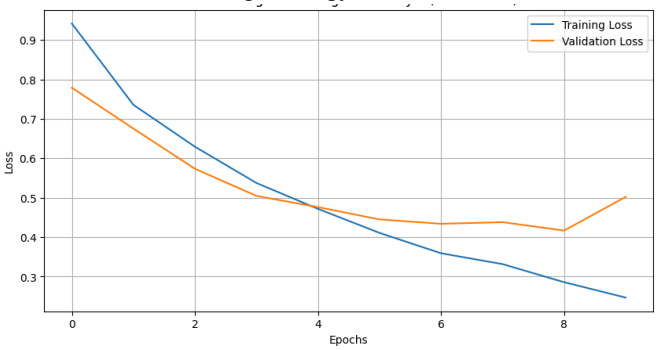

Fig. 3 Training and Validation Loss curves for the phase 1 ResNet50 model. The close alignment between training and validation losses indicates that **strong augmentation** successfully mitigated overfitting, despite the limited number of epochs.

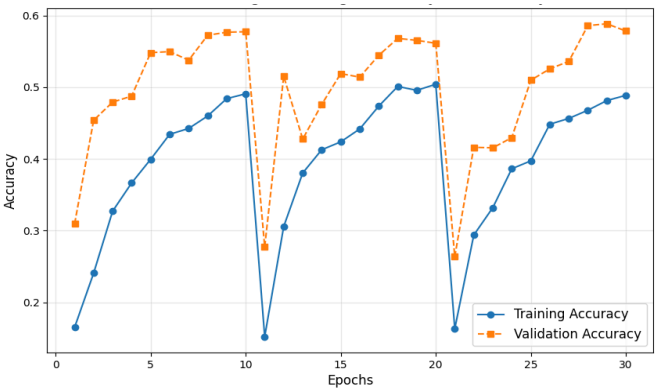

Fig. 4 Training and Validation Acc dynamic curves for the final ResNet101 model. Each 10 epoch shows the 1 cv fold.

B. CNN vs. Transformer in Limited Data

Our experiments confirm that for datasets of this scale (~10k images), CNNs with strong inductive biases (like ResNet) still outperform pure Transformer models. ViT requires significantly more data or stronger regularization to match CNN performance. However, the successful stabilization of ViT training using SGD and Cosine Annealing is a notable contribution, proving that Transformers can be trained on medical data without divergence if hyperparameters are carefully tuned.

C. Limitations

The primary limitation of this study was the computational constraint, which limited training to 10 epochs per fold. Given the difficulty introduced by strong augmentation, it is hypothesized that extending training to 50-100 epochs would allow the models to recover higher accuracy levels while maintaining robustness.

D. Qualitative Error Analysis

We visually inspected the misclassified samples to understand the model's limitations. Common failure cases include:

1. **Low Contrast:** Lesions with very low contrast against the skin tone were often confused with Benign Keratosis (*bkl*).
2. **Hair Occlusion:** Although we applied augmentation, thick hair covering the lesion still caused misclassification in some cases.
3. **Ambiguous Boundaries:** Early-stage melanomas lacking distinct border irregularities were sometimes misclassified as atypical nevi. Future work incorporating hair-removal preprocessing algorithms could mitigate these specific errors.

VIII. Conclusion

This study presented a comprehensive benchmark of deep learning architectures for skin lesion classification under a rigorous, data-centric framework. We demonstrated that dealing with class imbalance and limited data requires more than just complex architectures; it demands robust training strategies.

Our findings highlight that **ResNet101**, combined with **Focal Loss**, offers the most balanced performance for melanoma detection. Furthermore, we showed that high accuracy scores in standard benchmarks can be misleading due to overfitting, and that heavy augmentation provides a more realistic assessment of model reliability. Future work will focus on **Curriculum Learning**—gradually increasing augmentation intensity—to achieve both fast convergence and high robustness, and exploring hybrid CNN-ViT architectures to leverage the strengths of both paradigms.

References

[1] Esteva, A., Kuprel, B., Novoa, R. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017). https://doi.org/10.1038/nature21056

[2] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5, 180161 (2018). https://doi.org/10.1038/sdata.2018.161

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2