

Comparison of Leading Deep Learning Architectures on Skin Lesion Classification Using the ISIC 2019 Dataset

Anil Tan Aktan, Alaa Almouradi, Erchan Aptoula and Berrin Yanikoglu
Sabanci University, Istanbul, 34956 Türkiye
{tan.aktan,alaaalmouradi,erchan.aptoula,berrin}@sabanciuniv.edu

Abstract—This paper addresses the problem of automatic classification of dermoscopic skin lesion images into one of eight diagnoses that include multiple cancer types. The goal of this research was to evaluate methods to address problems of class imbalances and image variation in skin lesion classification, as well as to propose an ensemble model that performs well on multi-class metrics. The widely used dataset developed for the International Skin Imaging Collaboration (ISIC) challenge, ISIC 2019, was chosen to test the methods proposed in this paper. Methods evaluated in this paper include modern architectures such as Vision Transformer, Convolutional Neural Networks, and Vision Mamba, as well as training strategies such as data augmentation, test-time augmentation, and specialized loss functions addressing class imbalance. The best performance was obtained using an ensemble approach, achieving an overall accuracy of 74.80% and a macro F1 score of 74.30%. These results compare favorably to the state-of-the-art, where drops in performance with multi-class metrics have typically been observed.

Index Terms—Skin cancer, class imbalance, test-time augmentation, heterogeneous ensemble.

I. INTRODUCTION

Skin cancer is one of the most widespread variants, posing significant threats to public health and burdening healthcare systems. Melanoma, in particular, presents high metastatic potential and rapid progression, making early and accurate detection crucial for successful treatment [1]. Dermoscopic examination performed by dermatologists has been the clinical standard for lesion diagnosis. However, accurate diagnosis is a challenge even for clinicians, as they typically achieve around 60% accuracy in lesion classification [2]. This diagnostic uncertainty, combined with limited availability and unequal geographic distribution of dermatologists, has motivated research towards robust and accurate automated diagnostic methods.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), transformers, and hybrid models, have substantially improved performance in medical image analysis [3]. However, despite these advances, critical challenges persist. First, medical datasets frequently exhibit significant class imbalance. This imbalance can bias model training and undermine performance on minority classes [4]. Second, due to the introduction of many hyperparameters in ensembling techniques and complex models, standard training

methodologies have often overlooked comprehensive hyperparameter optimization, limiting model generalization [5]. Additionally, techniques performing well in other problems are underexplored in the skin lesion problem. Traditional approaches have rarely leveraged test-time augmentation (TTA) fully, resulting in missed opportunities to improve robustness, a critical requirement for this task given the variations commonly encountered in clinical practice [6]. Furthermore, while individual model architectures have been extensively studied, ensembles combining diverse architectures have remained underexplored [7].

The official leaderboard for the International Skin Imaging Collaboration (ISIC) 2019 challenge reveals the uneven performance of top teams across different metrics (Table II). While accuracy measures overall correctness, it can be misleading in imbalanced datasets. Balanced accuracy accounts for class imbalance by averaging recall across classes and was the official evaluation metric of the challenge. F1-score, on the other hand, emphasizes the harmonic mean of precision and recall, offering a more sensitive measure of performance in the presence of skewed class distributions. These evaluation inconsistencies and the underlying performance gaps motivate this research to study methods that deliver balanced and generalizable results across all key metrics.

The contributions of the paper can be summarized as follows:

- Evaluating different methods that address class imbalance and model robustness in the area of skin lesion classification.
- Proposing an ensemble model for skin lesion classification with high performance on multi-class metrics.

II. RELATED WORKS

Skin lesion classification methodologies generally fall into two architecture categories: deep convolutional neural networks and transformer-based architectures. Hybrid approaches and additional architecture components have also been proposed to improve generalization and robustness.

Deep Learning Architectures: CNNs have become widespread in automated skin lesion analysis [3]. Architectures such as EfficientNet [8] have demonstrated superior performance, effectively capturing the local spatial features critical for lesion recognition tasks. However, CNNs have primarily

focused on localized patterns, often lacking sufficient global context, limiting their performance on visually ambiguous lesion categories [9].

To address this, transformer-based models like Vision Transformer (ViT) [9] have introduced self-attention mechanisms that explicitly model long-range dependencies. This global modeling capability complements the local pattern extraction of CNNs and is particularly beneficial in dermatology, in which lesion differentiation may rely on subtle global contextual clues.

More recently, hybrid vision architectures such as MambaVision [10] aim to balance contextual awareness with computational efficiency by integrating ViT-style attention and Mamba's linear-time sequence modeling. While such hybrid architectures have shown promising results in broader medical imaging domains, their application in dermatology remains limited and underexplored.

Strategies for Class Imbalance and Robustness: Addressing class imbalance has been a significant research area, with methods such as Weighted Cross-Entropy and Focal Loss [4], [11] widely applied to balance model attention towards minority classes. Data augmentation methods have also been extensively studied as a means to improve generalization performance [12]. Test-time augmentation, although effective in other domains, has rarely been optimized in the context of dermatology [6]. Thus, systematic exploration and evaluation of augmentation techniques, combined with robust loss functions, have remained limited in dermoscopic image tasks.

Ensemble Learning: Ensemble methods have proven highly effective in various machine learning tasks by leveraging model diversity to enhance predictive performance and robustness [7]. However, heterogeneous ensembles combining fundamentally different architectures (CNNs, transformers, hybrids) have remained under-explored in dermatological image classification, motivating the explicit exploration present in this paper.

In summary, while significant advancements have been made, critical aspects such as class imbalance, robust augmentation strategies, and heterogeneous ensemble modeling have remained insufficiently addressed. This paper explicitly targets these gaps, systematically integrating and evaluating these strategies for robust, accurate, and generalizable skin lesion classification.

III. PROPOSED APPROACH

The primary goal of the proposed methodology is to achieve robust and generalizable multiclass image classification by addressing several critical challenges, including high class imbalance, limited data for underrepresented classes, sensitivity to image variability, and combining the strengths of different deep learning architectures. In this section, each design choice is explicitly justified by detailing its expected contribution towards solving these specific challenges.

Loss Functions for Class Imbalance: To improve model performance on skewed class distributions, the proposed

method explicitly incorporates loss functions designed to address class imbalance, including standard cross-entropy, Weighted cross-entropy, and focal loss [4] functions. Focal Loss was selected due to its focus on difficult and minority-class samples.

Test-Time Augmentation (TTA): Another issue is to obtain robustness with different viewpoints and variations in appearance caused by lighting, orientation, or scale. Test-time augmentation is implemented at inference time to improve the robustness of predictions by averaging model outputs across augmented versions of each input image. The augmentation strategies selected include horizontal flips, rotations at increments of 90°, and additional small-degree rotations, color jittering, and resizing.

The TTA approach aims to reduce prediction variance caused by alterations in images obtained in a clinical scenario. Indeed, producing predictions from multiple augmented views improves the output stability and model accuracy, as demonstrated in our experiments. The enhanced reliability and generalizability are important in real-world clinical deployment scenarios.

Hyperparameter Optimization: Model performance and generalization capability critically depend on careful hyperparameter tuning [5]. To systematically address this challenge, hyperparameters were optimized by comparing validation macro-F1 scores using Optuna, a state-of-the-art optimization framework that efficiently explores parameter spaces via Bayesian optimization methods [5]. Key hyperparameters that were tuned include: learning rate, weight decay, and focal loss gamma.

Learning rate impacts convergence behavior and final accuracy; weight decay controls regularization and prevents overfitting; and focal loss gamma determines the trade-off between focusing on challenging examples and ignoring easily classified ones.

Heterogeneous Ensemble Learning: Different deep learning architectures inherently possess distinct internal biases, capturing complementary visual features [7]. Therefore, a heterogeneous ensemble comprising three fundamentally different architectures (EfficientNet-B5 (CNN-based) [8], Vision Transformer (ViT-base-patch16-224, transformer-based) [9], and MambaVision (ViT-Mamba hybrid) [10] is proposed in this work. Aggregating predictions from these complementary architectures was implemented through simple mean ensembling, i.e., averaging their class probability outputs, which enhances overall accuracy, robustness, and generalizability by leveraging their diverse strengths.

EfficientNet-B5 was chosen due to its parameter-efficient scaling, enabling effective processing of high-resolution images critical for capturing detailed lesion features [8]. ViT was selected for its capability to model global context and dependencies between image regions, aiding recognition of overall lesion structures and patterns [9]. MambaVision, as a hybrid architecture, was selected because it combines the global context modeling of Vision Transformers with the efficient sequential state-space modeling of Mamba layers,

effectively capturing both global dependencies and structured representations [10].

IV. EXPERIMENTS AND RESULTS

Experiments are conducted to evaluate different methods that address challenges such as class imbalance, sensitivity to image variability, and other previously mentioned challenges in the skin lesion classification problem. These methods include: different architectures, data stratification, data augmentation, test-time augmentation, and different losses. All experiments are conducted using the ISIC 2019 dataset.

Dataset: The ISIC challenge was organized to benchmark automated classification algorithms using the ISIC 2019 dataset [13]–[15] - a publicly available collection of dermoscopic skin lesion images. ISIC 2019 was selected since it is the most established multi-class benchmark, unlike newer ISIC datasets that are larger but framed differently (e.g., binary classification in 2020). The challenge leaderboard uses multi-class balanced accuracy as the metric with which to compare algorithms. The ISIC 2019 challenge has an official training and test split, consisting of 25,331 images for training and 8,238 images for testing. There is no official split for validation.

The class distribution of the ISIC 2019 training set is as follows: Melanocytic nevus (NV) 50.8%, Melanoma (MEL) 17.9%, Basal cell carcinoma (BCC) 13.1%, Benign keratosis (BKL) 10.4%, Actinic keratosis (AK) 3.4%, Squamous cell carcinoma (SCC) 2.5%, Vascular lesion (VASC) 1.0%, and Dermatofibroma (DF) 0.9%. As can be seen, the distribution is highly biased; NV makes up over half of all samples, while DF and VASC contribute about 1% each. This resembles the real-life distribution of these skin lesions, which makes the ISIC 2019 dataset appropriate to evaluate methods that address class imbalance. To maintain class distribution during training and validation, a stratified splitting was used, specifically, 80% for training and 20% for validation. As for the final evaluation, the official test set was used.

All images in the dataset are color dermoscopic photographs in three-channel RGB format. There is no fixed aspect ratio; the majority of images fall within the resolution range of 576×768 pixels to 1024×1024 pixels. This variability reflects the diversity of capture devices and clinical settings encountered in real-world practice.

Data Augmentation and Transformations: Data augmentation was applied to improve model generalization given the class imbalance and variations typically encountered in clinical imaging scenarios. A series of stochastic and geometric transformations was applied during training only. The transformations included:

- Random rotation up to 45°
- Horizontal and vertical flips
- Random affine transformations (translation up to 10%, scaling between 1.0 and 1.05)
- Color jitter with brightness and contrast adjustments (0.1)
- Normalization using ImageNet statistics ($\text{mean} = [0.485, 0.456, 0.406]$, $\text{std} = [0.229, 0.224, 0.225]$)

- Center square crop: Crops the central square region of the image, to preserve aspect ratio.

For validation and testing, only deterministic transformations were applied (center crop, resize, and normalization) to ensure consistent evaluation conditions.

Model Training: Models were initialized with ImageNet pre-trained weights, then trained for 25 epochs on the ISIC 2019 dataset using AdamW. As found by hyperparameter tuning on EfficientNet-B5, ViT, and MambaVision, the learning rate was set to (8.4E-05, 1.00E-05, 4.56E-05) and weight decay to (8.6E-03, 8.6E-03, 4.9E-05). Training and validation accuracies, losses, and macro F1-scores were monitored. The best checkpoints were selected by the highest validation macro F1.

The loss function and TTA configuration used for the final evaluations were selected based on the results of the experiments conducted on different loss functions and TTA configurations, as explained as follows, with the gamma hyperparameter for focal loss being further tuned to 2.41.

Loss Functions: Experiments were conducted on 3 loss functions: standard cross-entropy, weighted cross-entropy, and focal loss, all conducted on the ViT model. The gamma parameter for focal loss was set to the default (2) [4]. Other parameters for training were set as mentioned previously. As seen in Table I, focal loss markedly outperformed both standard and weighted cross-entropy, achieving up to a 1.75% point gain in weighted F1 score on the class-imbalanced dataset.

Test-Time Augmentation (TTA): To evaluate the applicability of TTA, 4 configurations were tested with 8 and 16 views, using stochastic transformations of flips, 90° rotations, additional 10 – 15° rotations, and color jitter of 10%. All experiments were conducted with focal loss, which gave the best results in initial experiments. Test-time augmentation yielded a 0.79% point improvement in F1 compared to no augmentation, highlighting the value of inference-time transformations in enhancing model robustness.

TABLE I
PERFORMANCE OF ViT MODEL WITH DIFFERENT SETTINGS

Loss Function	Test F1 (%)	Acc (%)
Focal Loss (gamma=2)	72.15	72.41
Weighted Cross-Entropy	71.21	71.54
Cross-Entropy	70.40	71.72
Test-time Augmentation	Test F1 (%)	Acc (%)
8x (Flip + Rotate90 + Rot10)	72.15	72.41
8x (Flip + Rotate90 + Rot15)	72.12	72.38
16x (Flip + Rotate90 + Rot15 + ColorJitter)	72.00	72.30
No test-time augmentation	71.36	72.01

Ensemble Results: As can be seen from Table II, a heterogeneous ensemble of three architectures resulted in a 1.67% and 1.02% point increase in F1 score and accuracy, respectively, compared to our best-performing single model. Furthermore, the confusion matrix in Table III confirms

these improvements by demonstrating accurate classification of major classes and consistent performance on minority classes. Moreover, the proposed ensemble represents notable improvements over leading literature and leaderboard methods, with the highest F1-score (0.743) and AUC (0.942). Overall, the ensemble method effectively leverages the complementary strengths of different architectures, demonstrating enhanced robustness and generalization performance in skin lesion classification tasks.

TABLE II
COMPARISON OF ISIC 2019 PERFORMANCE: LEADERBOARD, LITERATURE, AND PROPOSED METHODS (BEST IN BOLD, SECOND BEST UNDERLINED)

Method	Weighted F1	AUC	Bal. Accuracy
<i>ISIC2019 Leaderboard:</i>			
DaisyLAB [16]	0.685	<u>0.913</u>	0.636
DysionAI [17]	0.703	0.780	<u>0.607</u>
AImageLab			
-PRHLT [18]	0.709	0.886	0.593
<i>Literature:</i>			
Doctors' Average [2]	0.597	—	—
Skin-Net [19]	<u>0.713</u>	—	—
Kassem <i>et al.</i> [20]	0.575	—	—
<i>This work:</i>			
EfficientNet-B5	0.736	0.925	0.598
ViT	0.722	0.912	0.587
MambaVision	0.723	0.918	0.591
Proposed Ensemble	0.743	0.942	<u>0.607</u>

TABLE III
CONFUSION MATRIX FOR THE PROPOSED ENSEMBLE MODEL.

	MEL	NV	BCC	AK	BKL	DF	VAS	SCC
MEL	918	225	67	40	69	0	7	1
NV	119	2216	77	21	52	1	3	6
BCC	27	29	828	46	26	3	5	11
AK	20	11	84	168	87	0	0	4
BKL	83	94	60	70	333	10	0	10
DF	1	15	16	1	3	53	0	2
VASC	11	8	24	0	0	1	54	6
SCC	9	4	62	15	13	1	0	61

V. CONCLUSION

Our work presents the performance of several deep learning architectures optimized to address class imbalance and improve robustness in multi-class skin lesion classification. The results demonstrate the value of focal loss, systematic hyperparameter optimization, and an optimized test-time augmentation strategy, each contributing measurable gains in generalization.

This study also introduces an ensemble model that achieves strong performance across multi-class metrics. The proposed ensemble model achieved the highest F1-score (0.743) and AUC (0.942) among all evaluated methods, outperforming both the leading literature and the official ISIC 2019 leaderboard submissions. Future work will extend this framework to

larger and more recent datasets and explore the use of metadata and external data sources to further improve the recognition of underrepresented classes.

REFERENCES

- [1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2021," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.
- [2] P. Tschandl, C. Rosendahl, and H. Kittler, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, Jul. 2019.
- [3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE TPAMI*, vol. 42, no. 2, pp. 318–327, 2020.
- [5] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the ACM SIGKDD*, 2019, pp. 2623–2631.
- [6] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttat, "When and why test-time augmentation works," *arXiv:2011.11156*, 2021.
- [7] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.
- [8] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of ICML*, 2019, pp. 6105–6114.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of ICLR*, 2021.
- [10] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *Proceedings of IEEE CVPR*, 2025, pp. 25 261–25 270.
- [11] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of IEEE CVPR*, 2019, pp. 9268–9277.
- [12] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- [13] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018.
- [14] N. C. F. Codella, D. A. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. 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)," *CoRR*, vol. abs/1710.05006, 2017.
- [15] C. Hernández-Pérez, M. Combalia, S. Podlipnik, N. C. Codella, V. Rotemberg, A. C. Halpern, O. Reiter, C. Carrera, A. Barreiro, B. Helba, S. Puig, V. Vilaplana, and J. Malvehy, "Bcn20000: Dermoscopic lesions in the wild," *Scientific Data*, vol. 11, no. 1, p. 641, 2024.
- [16] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schläefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, p. 100864, 2020.
- [17] S. Zhou, Y. Zhuang, and R. Meng, "Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles," *DysionAI*, Tech. Rep., 2019.
- [18] F. Pollastri, J. Maronas, M. Parreno, F. Bolelli, R. Paredes, C. Grana, and A. Albiol, "AImageLab-PRHLT at ISIC challenge 2019," *Università degli Studi di Modena e Reggio Emilia and Universitat Politècnica de València*, Tech. Rep., 2019.
- [19] Y. S. Alsaifi, M. A. Kassem, and K. M. Hosny, "Skin-Net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier," *Journal of Big Data*, vol. 10, p. 105, 2023.
- [20] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020.