A Comparative Analysis of CNN Models and Vision Transformers for Skin Lesion Classification

Yasmin C. Aguirre, Ashiley Bianca S. de Oliveira,
Rodrigo C. Barros, Lucas S. Kupssinskü

[1]MALTA - Machine Learning Theory and Applications Lab - PUCRS, Brazil

{yasmin.cardozo,ashiley.bianca}@edu.pucrs.br
{rodrigo.barros,lucas.kupssinsku}@pucrs.br

Abstract. *Skin cancer is one of the most common malignancies in Brazil, making accurate diagnosis essential. This study evaluates five convolutional neural networks (CNNs) and two Vision Transformers (ViTs) on the binary classification of dermoscopic images (nevus vs. melanoma) from four public datasets characterized by class imbalance. Using a unified fine-tuning protocol and holdout splits, we observed that ViTs generally outperformed CNNs in F1-score and ROC-AUC. These results suggest that attention-based models may provide more balanced performance under class imbalance.*

1. Introduction

Skin cancer is the most prevalent cancer type in Brazil. According to the Brazilian National Cancer Institute [Santos et al. 2023], an estimated 220,490 new cases of non-melanoma and 8,980 melanoma cases are expected, underscoring its public health relevance and the need for effective early detection of melanoma.

Computer Vision and Deep Learning are being explored in the dermoscopic image classification task as a cheap, noninvasive clinical resource to aid in early skin cancer diagnosis. Comparative evaluations of architectures such as ResNet-50, VGG16, MobileNet, InceptionV3, and DenseNet have reported accuracies above 93% on the ISIC dataset [Islam and Panta 2024]. Segmentation has shown improvement in binary classification performance but may reduce accuracy in multi-class tasks when relevant contextual information is lost [Araújo et al. 2024]. Other studies have explored hybrid strategies, such as combining transfer learning with zero-shot learning, enabling the classification of unseen lesion categories with similar accuracy levels [Chowdhury et al. 2025].

This study presents a multi-dataset comparative analysis of seven deep learning architectures, *ResNet* [He et al. 2015], *EfficientNet* [Tan and Le 2019], *AlexNet* [Krizhevsky et al. 2017], *Inception* [Szegedy et al. 2015], VGG [Simonyan and Zisserman 2015], ViT [Dosovitskiy et al. 2021], and ViT-MAE [He et al. 2022], in the binary classification of malignant and benign skin lesions. Our evaluation considers four dermoscopic datasets, with varying degrees of class imbalance. In our experiments, attention-based models consistently outperformed their convolutional counterparts, particularly in recall-oriented metrics, indicating greater robustness in imbalanced settings.

2. Materials and Methods

To conduct our study we framed dermoscopic image classification as a binary classification task between nevus (benign) and melanoma (malignant) skin lesions. We selected the datasets *HAM10000*, *ISIC2019*, *ISIC2020*, and *PH2* and removed all images from each dataset that were not of nevus or melanoma classes. The resulting binary classification problem was imbalanced, with prevalence of the nevus class in all of the studied datasets. A summary of the datasets used is found in Table 1.

We compared the Accuracy, Precision, Recall, F1-Score, and AUC of five distinct convolution-based neural networks and two Vision Transformers, the models are listed in Table 2. All models were fine-tuned under the same training protocol implemented in PyTorch/PyTorch Lightning, using the Adam optimizer, Binary Cross-Entropy loss, early stopping (patience = 3), and a learning rate scheduler (initial LR = 0.0001, gamma = 0.1). Data were split into 80% training, 10% validation, and 10% test sets, with validation guiding early stopping. Performance was assessed on the hold-out test set using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The source code is available on Github[1] to facilitate reproducibility.

Table 1. Samples per class across datasets

Dataset	Nevus	Melanoma	Ratio
HAM10000	6705	1113	≈1:6
ISIC2019	12875	4522	≈1:3
ISIC2020	5193	584	≈1:9
PH2	160	40	1:4

Table 2. Models architecture, number of parameters, and dataset

Model	Architecture	$\#\theta$	Pretraining dataset
AlexNet	Convolution	≈60M	ImageNet
EfficientNet	Convolution	≈5.3M	ImageNet
Inception	Convolution	≈24M	ImageNet
ResNet	Convolution	≈11.7M	ImageNet
VGG	Convolution	≈138M	ImageNet
ViT	Transformer	≈86M	ImageNet-21k
ViT-MAE	Transformer	≈86M	ImageNet

3. Results and Discussion

The results obtained from evaluating the skin lesion classification models after fine-tuning on the training set are presented in Table 3. Our results show that attention-based vision models consistently outperformed convolution-based models in all of the tested datasets.

Among the attention-based models, *ViT* consistently achieved the highest F1-Score and ROC-AUC across all datasets. One possible factor contributing to this strong performance is that *ViT* was pretrained on ImageNet-21k, which may provide a richer set of transferable features. Nevertheless, *ViT-MAE*, which was pretrained on ImageNet like all CNN baselines, also ranked as the second-best model across datasets. This result indicates that the observed advantage of attention-based architectures is not solely attributable to pretraining scale, but rather reflects that attention-based models may inherently encode features in a way that is less sensitive to imbalanced distributions.

In contrast, CNN-based models displayed a tendency for high accuracy but low *recall*, particularly in datasets with pronounced class imbalances such as *ISIC2020* (≈1:9 ratio) and *HAM10000* (≈1:6 ratio). These models were biased towards the majority class, leading to a higher number of false negatives and poor performance on the minority class. Their best relative performance was on *ISIC2019*, which has a more balanced class distribution of ≈1:3, highlighting their difficulty in generalizing from imbalanced data.

[1] Github: https://github.com/4gu1rr3/Skin-Lesion-Classification

Table 3. Metrics computed on test sets for Experiment 1. Best AUC and F1-Score for each model within each dataset are highlighted in bold.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUC
HAM10000	AlexNet	0.8838	0.6585	0.2596	0.3724	0.6195
	EfficientNet	0.5326	0.0957	0.2981	0.1449	0.4333
	Inception	0.8672	0.0000	0.0000	0.0000	0.5000
	ResNet	0.8672	0.0000	0.0000	0.0000	0.5000
	VGG	0.8646	0.4615	0.1154	0.1846	0.5474
	ViT	0.9374	0.8571	0.6346	**0.7293**	**0.8092**
	ViT-MAE	0.9119	0.7333	0.5288	0.6145	0.7497
ISIC2019	AlexNet	0.8156	0.6977	0.4255	0.5286	0.6832
	EfficientNet	0.8231	0.7170	0.4492	0.5523	0.6961
	Inception	0.8156	0.7965	0.3239	0.4605	0.6487
	ResNet	0.8168	0.7261	0.3948	0.5115	0.6735
	VGG	0.8145	0.6984	0.4161	0.5215	0.6792
	ViT	0.9150	0.8590	0.7778	**0.8164**	**0.8684**
	ViT-MAE	0.8851	0.8055	0.6950	0.7462	0.8206
ISIC2020	AlexNet	0.9067	0.4000	0.0377	0.0690	0.5160
	EfficientNet	0.3489	0.0622	0.4340	0.1087	0.3871
	Inception	0.9085	0.0000	0.0000	0.0000	0.5000
	ResNet	0.9085	0.0000	0.0000	0.0000	0.5000
	VGG	0.9033	0.2000	0.0189	0.0345	0.5056
	ViT	0.9568	0.8500	0.6415	**0.7312**	**0.8151**
	ViT-MAE	0.9413	0.6863	0.6604	0.6731	0.8150
PH^2	AlexNet	0.8500	0.0000	0.0000	0.0000	0.5000
	EfficientNet	0.3000	0.1765	1.0000	0.3000	0.5882
	Inception	0.8500	0.0000	0.0000	0.0000	0.5000
	ResNet	0.8000	0.0000	0.0000	0.0000	0.4706
	VGG	0.8500	0.0000	0.0000	0.0000	0.5000
	ViT	0.9500	1.0000	0.6667	**0.8000**	**0.8333**
	ViT-MAE	0.9500	1.0000	0.6667	**0.8000**	**0.8333**

The *PH2* dataset, however, complicates this pattern. Despite its milder imbalance (1:4), it yielded the poorest results overall. With only 200 images, including 40 melanomas, *PH2* offered too few examples for stable training and evaluation. Under these conditions, most CNNs collapsed completely into majority-class prediction, while *EfficientNet* degenerated in the opposite direction by predicting almost exclusively the minority class. These outcomes suggest that data scarcity can undermine performance even more severely than imbalance.

While dataset size shaped outcomes, model size did not show the same effect. *ResNet*, with ≈11.7M parameters, performed almost identically to *VGG*, a model with ≈138M parameters. This is best observed on the *ISIC2019* dataset, where *ResNet* and *VGG* achieved *F1-Scores* of 0.5115 and 0.5215, respectively. This suggests that model architecture plays a more critical role than size alone.

4. Conclusion

This study presented a comparative analysis of CNNs and ViTs for binary skin lesion classification on four dermoscopic datasets with varying malignant-to-benign ratios. Overall, attention-based models outperformed CNN-based counterparts in terms of F1-score and ROC-AUC, reinforcing the advantage of Transformer architectures in this context.

Model size appears less important than model architecture for handling class imbalance. For instance, ResNet and VGG produced similar results despite differences in parameter count, while EfficientNet—the smallest CNN tested—was among the few CNNs that avoided collapsing to majority-class predictions.

Our study has two main limitations: 1 - Dermoscopic image as a binary classification problem does not assess model performance on less predominant classes like vascular lesions; 2 - Our evaluation relied on a single stratified hold-out split rather than cross-validation, which prevents a more robust estimate of variance in model performance. Regardless of those limitations, in our experiments, Vision-Transformers proved to be superior to convolutional-based architectures for nevus and melanoma classification.

References

Araújo, R. L., de S. Luz, D., de Lima, B. V., Marques, J. V. M., de M. S. Veras, R., de C. Filho, A. O., Araújo, F. H. D., and e Silva, R. R. V. (2024). Quantifying the effects of segmentation in image classification for melanoma recognition. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2024)*, pages 400–411. SBC.

Chowdhury, T. A., Wagner, E., Motzki, P., and Lehser, M. (2025). Enhanced transfer learning algorithm with zero-shot components for dermatological diagnosis using the ham10000 dataset. In *Proceedings of SPIE Medical Imaging*, volume 13292, page 132920G. SPIE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Islam, M. S. and Panta, S. (2024). Skin cancer images classification using transfer learning techniques. *CoRR*, abs/2406.12954.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Santos, M. d. O., Lima, F. C. d. S. d., Martins, L. F. L., Oliveira, J. F. P., Almeida, L. M. d., and Cancela, M. d. C. (2023). Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, 69:e–213700.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.