

Identifying Malignant Skin Diseases Through Deep Learning

Identificação de Doenças de Pele Malignas por Meio de Aprendizado Profundo

Gabriel Sávio de Lima Mota¹, Matheus Vieira da Silva², João Fernando Mari¹,
André Ricardo Backes^{3*}

Abstract: Early and accurate identification of malignant skin diseases is critical to improving treatment outcomes and patient prognosis. Current diagnostic methods heavily rely on visual assessments by medical professionals, followed by biopsies for confirmation. These approaches can be subjective and require significant expertise, while access to specialized care remains limited, particularly in developing regions. Recent advancements in artificial intelligence, specifically deep learning, present an opportunity to improve dermatological diagnostics. This paper explores the application of two deep learning models, ResNet and Vision Transformer (ViT), for the automated detection of malignant skin conditions. Both models were trained on the PAD-UFES-20 dataset, which comprises clinical skin lesion images captured via smartphones. Our findings suggest that these models can offer real-time diagnostic support at reduced costs, with accuracy levels comparable to or exceeding human performance, with the best experiment achieving the F1 score of 79.89% and an accuracy of 80.63%.

Keywords: Deep Learning — Skin Diseases — Medicine — Classification — Data Augmentation

Resumo: A identificação precoce e precisa de doenças malignas de pele é fundamental para melhorar o tratamento e o prognóstico dos pacientes. Os métodos de diagnóstico atuais dependem fortemente de avaliações visuais feitas por profissionais de saúde, seguidas de biópsias para confirmação. Essas abordagens podem ser subjetivas e requerem um nível significativo de especialização, enquanto o acesso a cuidados especializados permanece limitado, especialmente em regiões em desenvolvimento. Avanços recentes em inteligência artificial, especificamente em aprendizado profundo, permitem aprimorar o diagnóstico dermatológico. Este trabalho explora a aplicação de dois modelos de aprendizado profundo, ResNet e Vision Transformer (ViT), para a detecção automatizada de condições malignas de pele. Ambos os modelos foram treinados no conjunto de dados PAD-UFES-20, que inclui imagens clínicas de lesões cutâneas capturadas via smartphones. Nossos achados sugerem que esses modelos podem oferecer suporte diagnóstico em tempo real a custos reduzidos, com níveis de precisão comparáveis ou superiores ao desempenho humano.

Palavras-Chave: Aprendizado Profundo — Doenças de Pele — Medicina — Classificação — Aumento de Dados

¹ Institute of Exact and Technological Sciences, Federal University of Viçosa (UFV), Brazil

² School of Computer Science, Federal University of Uberlândia (UFU), Brazil

³ Department of Computing, Federal University of São Carlos (UFSCar), Brazil

*Corresponding author: arbackes@yahoo.com.br

DOI: <http://dx.doi.org/10.22456/2175-2745.143463> • Received: 23/10/2024 • Accepted: 03/01/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Identifying certain malignant skin diseases accurately in the early stages is crucial for increasing the chances of cure and potentially optimizing the patient's treatment. The current diagnosis of skin diseases relies on a visual medical assessment and biopsy for confirmation. These procedures can be subjective and may require a high level of expertise from the medical professional to perform successfully. Moreover, ac-

cess to qualified specialists can be limited due to financial and geographical constraints, especially in third-world countries [1].

Over the past few years, the research in artificial intelligence has made significant progress, particularly in deep learning, and these advancements have impacted different sectors, such as medicine, agriculture, and industry. In medicine, dermatology stands out, where new deep learning models show potential to streamline and automate specific processes

[2]. In their analyses, these models demonstrate performance equal or superior to humans, often matching or surpassing human capabilities in such tasks. Deep learning models could potentially automate the screening process, enabling faster and more precise interventions [3]. The new architectures, such as CNN [4] and Transformer-based [5], can be employed to analyze medical imaging and give a diagnostic in real-time with a fraction of the cost.

Despite advancements, evaluating and comparing different architectures for skin disease identification with diverse, real-world datasets remains necessary. This study aims to: (1) Assess the effectiveness of Deep Residual Learning for Image Recognition (ResNet) [6] and Vision Transformer (ViT) [7] models; (2) Investigate the impact of class imbalance and evaluate techniques like Focal Loss and data augmentation; (3) Examine how dataset splits affect model generalization. By addressing these aspects, we seek to unlock the full potential of these architectures, enhancing their accuracy and reliability for the early detection of malignant skin diseases. We trained these models on the PAD-UFES-20 dataset, a comprehensive collection of clinical skin lesion images captured using smartphones[8].

The remainder of the paper is organized as follows: Section 2 provides an overview of related works. Section 3 describes the dataset and deep learning architectures used in this study. The experimental design is outlined in Section 4. In Section 5, we present and discuss the results. Finally, Section 6 offers our conclusions and future works.

2. Related Work

Meanwhile, Shehzad et al. [9] introduced a deep learning ensemble model architecture, utilizing EfficientNetV2S and Swin-Transformer architectures. They initially detected the focal area of the cancer, demonstrating efficiency in identifying dark places, achieving an accuracy score of around 99.10%.

In this study, Flosdorf et al. [10] utilized Vision Transformer (ViT) architectures to improve skin cancer detection, comparing them with CNNs and other simpler models. The ViT-L32 model achieved an accuracy of 91.57% and a recall of 58.54% for melanoma detection, while the ViT-L16 model obtained a slightly higher accuracy of 92.79% but a lower recall of 56.10%. These results suggest that while the ViT models show promise in accuracy, there is room for improvement in their ability to identify melanoma.

Sakl et al. [11] propose a multi-stage learning approach using a Lightweight CNN (LWCNN) architecture. The model classifies skin lesions into two groups, malignant and benign, with further subclassifications into five benign and three malignant classes. They evaluated the model on the HAM10000 dataset [12], achieving an Area Under the Curve (AUC) of 0.9799 for the first method and 0.9981 for the second method.

In their study, Carmo et al. [13] conducted a comparison between different convolutional neural networks (CNNs) for the classification of pigmented skin lesions. Four CNN archi-

tectures were selected for comparison: ResNet-50, VGG-16, Inception-v3, and DenseNet-121, applying different training strategies, including data augmentation techniques. Among the architectures, DenseNet-121 performed the best, achieving an accuracy of 90% on the HAM10000 dataset [12].

Auxilia et al. [14] proposed the weighted extreme learning optimal diagonal-kernels convolution (WEL-ODKC) model for automatic skin cancer detection, addressing the issue of imbalanced data and eliminating operator variability, showing promising results on the HAM10000 [12] and PAD-UFES-20 [8] datasets. Nijhawan et al. [15] integrated patient history and different modalities with deep learning models for accurate diagnosis of skin lesions, outperforming other deep learning models in terms of performance. Lima and Krohling [16] explored advancements in transformer-based networks for skin lesion diagnosis, achieving state-of-the-art results on the PAD-UFES-20 dataset.

3. Material and Methods

3.1 Dataset

The PAD-UFES-20 dataset [8] consists of information from 1,373 patients, with 2,298 images collected through smartphones and clinical data from these patients with around 21 attributes. The dataset covers six different types of skin diseases and cancers. The classes are Basal Cell Carcinoma (BCC) with 845 samples, Squamous Cell Carcinoma (SCC) with 192 samples, Actinic Keratosis (ACK) with 730 samples, and Seborrheic Keratosis (SEK) with 230 samples. The other class Melanoma (MEL) has 52 samples, while Nevus (NEV) contains 244 samples.

They collected the images in the dataset under different conditions, reflecting real-world analysis conditions, which increases the robustness and applicability of the solutions developed with it. The dataset is accompanied by metadata on patients' demographic information, such as the history and characteristics of the lesions.

3.2 Architectures

CNNs are deep learning models well-suited for processing grid-like data, such as images. The CNN architecture consists of convolutional layers that learn spatial hierarchies of features by applying learnable filters (kernels) across the input data. The convolution operation to generate a feature map can be expressed as:

$$Y(i, j) = \sum_{m=-1}^a \sum_{n=-a}^a K(m, n)X(i+m, j+n), \quad (1)$$

where X is the input, K is the convolutional kernel with size $2a+1 \times 2a+1$, and Y is the resulting feature map. Pooling layers typically follow convolutional layers to reduce the spatial dimensions of the feature maps, preserving important features while decreasing computational complexity.

LeNet, an early example of CNN developed by LeCun et al. [4] for handwritten digit recognition, consists of alternating

convolutional and pooling layers, followed by fully connected layers and a softmax output layer. The softmax function, which converts the final layer's outputs into class probabilities, is defined as:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}}, \quad (2)$$

where z_i is the input score for class i , and C is the number of classes. LeNet's architecture demonstrated the effectiveness of CNNs for image classification and laid the groundwork for more advanced and deeper networks.

In this work, we implemented two primary architectures, ResNet and Vision Transformers (ViT), to detect skin diseases and address problems with computer vision.

3.2.1 ResNet

ResNet (Residual Network) is a deep neural network architecture introduced by He et al. in 2015 to address the vanishing gradient problem that often occurs in deeper networks [17]. The critical innovation of ResNet is the use of residual blocks, which introduce shortcut connections to bypass one or more layers. The residual learning function can be expressed as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}, \quad (3)$$

where $\mathbf{x}$ is the input to the residual block, $\mathcal{F}(\mathbf{x}, \{W_i\})$ represents the residual function to be learned (typically composed of a series of convolutional layers), and $\mathbf{y}$ is the output of the block. The shortcut connection adds the input directly to the output of the second convolutional layer, enabling the gradient to flow more quickly through the network. This formulation helps avoid the degradation problem when training deeper networks, facilitating the training of deeper networks with hundreds of layers.

The ResNet architecture includes several variations, such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, which differ primarily in the number of layers and the complexity of their residual blocks. ResNet-18 and ResNet-34 use basic residual blocks with two 3×3 convolutional layers, while the deeper versions like ResNet-50, ResNet-101, and ResNet-152 use more advanced bottleneck blocks. These bottleneck blocks contain three layers (1×1 , 3×3 , 1×1) that help reduce dimensionality and computational cost while preserving deep learning capacity.

Each variation of ResNet offers a trade-off between depth, complexity, and computational efficiency. Shallower networks like ResNet-18 and ResNet-34 are well-suited for more straightforward tasks or environments with limited computational resources. At the same time, deeper models like ResNet-50, ResNet-101, and ResNet-152 are better for handling complex tasks such as object detection, image segmentation, and large-scale classification, where capturing intricate patterns and features is crucial.

3.2.2 Vision Transformer (ViT)

The Vision Transformer (ViT) architecture, proposed by Dosovitskiy et al. [18], represents a shift from the traditional CNNs that have been predominantly used for image recognition tasks. Unlike CNNs, the Vision Transformer applies attention mechanisms to sequences of image patches, enabling the model to capture relationships between different parts of the image. This characteristic is particularly useful for analyzing medical images, where spatial context is crucial for accurately identifying patterns.

The ViT model divides an input image into a grid of non-overlapping patches, each treated as a token similar to words in natural language processing tasks. These patches are then linearly embedded into a sequence of vectors and combined with positional encodings to retain information about their original location in the image. The model applies the Transformer architecture, specifically the self-attention mechanism, to these embedded patches. The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q , K , and V represent the query, key, and value matrices derived from the input patch embeddings, and d_k is the dimensionality of the key vectors. This mechanism allows the ViT to dynamically weigh the importance of each patch concerning every other patch, effectively modeling long-range dependencies across the entire image.

By leveraging the self-attention mechanism, Vision Transformers can achieve competitive performance on image classification tasks without relying on convolutional layers. The model's ability to focus on different parts of the image, regardless of their spatial distance, makes it a powerful tool for various applications, especially those involving complex and high-dimensional data.

3.3 Model evaluation

To evaluate the models, we implemented Accuracy and the F1 Score. Accuracy measures the proportion of correctly classified instances out of the total cases and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the counts of each type of prediction. However, accuracy can be misleading in imbalanced datasets, as it may not reflect the model's ability to predict minority classes accurately. The F1 Score provides a more balanced evaluation by considering Precision and Recall. It is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

The F1 Score is particularly useful when false positives and negatives have different implications.

3.4 Computational resources

We conducted the experiments on a PC with an i5 processor running at 3.0 GHz, 32 GB of RAM, and an NVIDIA 1080 Ti GPU with 11 GB of memory. The development environment is based on the Python 3.12 programming language and PyTorch 2.4.0 with CUDA 12.2. We also used the libraries torchvision 0.19.0, scikit-learn 1.5.1, pandas 2.2.2, and Matplotlib 3.9.1.

4. Experiment design

We applied network architectures such as AlexNet, ViT, MedViT, MedMamba, and Resnet to evaluate early datasets in some initial experiments. The preliminary results indicated that architectures like ResNet-101 and ViT were more promising, motivating us to focus on these in subsequent experiments.

We began with a detailed analysis and segmentation of the experiments in the dataset, applying several initial optimizations to the ResNet-101 and ViT networks to tackle the class imbalance problem. These optimizations involved penalizing classes using Focal Loss and enhancing the dataset using data augmentation techniques. The Focal loss is defined in the following equation:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (7)$$

where p_t is the probability predicted by the model for the correct class t , defined as:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases} \quad (8)$$

where p represents the probability assigned to the positive class. α_t is the balancing factor to be adjusted to handle class imbalance. γ is the focusing parameter, which controls the focus of the loss function. For $\gamma = 0$, Focal Loss simplifies to traditional cross-entropy. As the value of γ increases, the modulating term $(1 - p_t)^\gamma$ assigns less weight to correctly classified examples (those with high p_t) and more weight to hard-to-classify examples (those with low p_t).

The choice of the hyperparameters was carried out based on well-established practices and the characteristics of the dataset. We used a batch size (bs): 32, which was selected to achieve a balance in GPU memory usage. A learning rate (lr): 0.0001 and a momentum (mm): 0.9 were chosen to maintain stable convergence, given the amount of data and the existing risk of overfitting. We chose a maximum of 200 epochs (ep) for training, and included an Early Stopping mechanism to prevent overfitting if no improvement was observed for 20 consecutive training epochs. Regarding the hyperparameters of the Focal Loss technique, we chose an α : 0.25 and a γ : 2.0; these are the standard values defined based on the relevant literature for this strategy.

To further address the challenges caused by class imbalance, we implemented K-Fold cross-validation with $K = 5$. This approach enabled us to identify specific patterns within

certain data splits, which, in turn, guided the development of additional strategies to be evaluated during training.

We created the following splits from the main dataset: PAD-UFES, PAD-UFES-FULL, PAD-UFES-IS, and PAD-UFES-AUG. For the PAD-UFES, we randomly split the main dataset, with 80% of the samples from the main dataset allocated for training and 20% for testing. We created the PAD-UFES-FULL set by randomly removing 15% of the images from the main dataset. We split the remaining 85% into 70% for training and 15% for testing. For the PAD-UFES-IS set, we applied the same strategy as PAD-UFES-FULL but used a different random seed. All the splits were performed proportionally for each class.

To create the PAD-UFES-AUG set, we randomly split 80% of the main dataset for training and 20% for testing. However, we oversampled the training set using data augmentation techniques. These augmentations were applied probabilistically to each image, ensuring a diverse range of synthetic variations.

Specifically, we implemented brightness adjustments with a range of $\pm 50\%$ and contrast adjustments of $\pm 50\%$ to simulate varying lighting conditions. We also applied geometric transformations, including horizontal flipping (with a 50% probability), random rotations up to ± 30 degrees, scaling between 80% and 120%, translations up to 20 pixels in both x and y directions, and shearing with angles up to ± 0.3 radians to introduce variability in the spatial orientation of the images. We also applied color jittering, modifying the hue by ± 10 units and altering saturation and value by 80% to 120%. Gaussian blur was applied using size 3, 5, or 7 kernels to replicate image imperfections.

Subsequently, for each of these new sets, we trained the models using 5-fold cross-validation. The models generated in each iteration of the K-Fold process were then tested on their respective test sets, which had been previously set aside for each split. We evaluated the performance of the models using metrics such as accuracy, sensitivity, precision, and F1-score, calculated individually for each fold. The results of these metrics were then aggregated to provide a comprehensive evaluation of the overall model performance. Figure 1 illustrates the training process for each split.

5. Results and discussion

In this section, we discuss the results of our experiments conducted on the ResNet-101 and ViT architectures on the PAD-UFES-20 dataset. We seek to understand how techniques such as Focal Loss and data augmentation impacted the performance of the models in each of the different dataset splits. The results are evaluated based on metrics such as accuracy and F1-score, emphasizing how each model dealt with the class imbalance challenges present in each dataset split.

Table 1 shows the results of ResNet-101 for each of the splits made in PAD-UFES-20 and the techniques applied in each fold. Figure 2 shows the average evaluation results and standard deviation among folds. We notice that the appli-

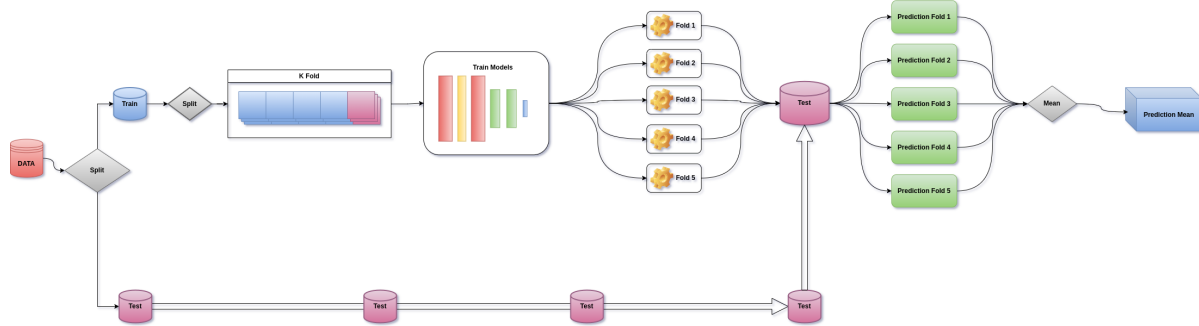


Figure 1. Illustration of the training process.

cation of Focal Loss improved performance in most splits. This is mostly notable for the PAD-UFES-FULL split, where the model achieved an average accuracy of 73.68% with the technique, compared to 72.34% without Focal Loss. This highlights a small but significant improvement in overall performance. The average F1 score also improved from 70.39% (without Focal Loss) to 71.87% (with Focal Loss), indicating a good balance between precision and recall. Precision scores are relatively stable across the dataset variants. Recall scores are also fairly consistent but slightly lower on average than Precision. The F1-Score shows some variability, with higher scores generally seen in the PAD-UFES-FULL and PAD-UFES-AUG datasets. When Focal Loss is applied, the model achieves its highest F1 on PAD-UFES-AUG (up to 0.7452), implying that the balance between Precision and Recall is best maintained here.

In the PAD-UFES-AUG split, the use of data augmentation resulted in an increase in average accuracy to 73.10% with Focal Loss and a drop to 72.28% without Focal Loss. These results indicate that data augmentation may have slightly reduced class imbalance issues, diminishing the benefit of Focal Loss in this scenario.

The PAD-UFES-IS split showed the lowest performance, with an average accuracy of 68.12% using Focal Loss and 69.45% without it. This indicates an accentuated class imbalance, and while Focal Loss helped, it did not completely solve the problem.

Overall, the application of Focal Loss has a stronger impact on the PAD-UFES and PAD-UFES-IS splits, where class imbalance was more severe. The low standard deviation across folds indicates that performance was balanced across different cross-validation splits.

Table 2 presents the results of ViT for each of the folds, while Figure 2 shows the averages and standard deviation of these results. They follow a slightly similar pattern to ResNet-101 but with some differences. The ViT model performed better, specifically in the PAD-UFES-IS split. With the application of Focal Loss, ViT achieved an average accuracy of 71.59%, surpassing ResNet-101 in this split. This result indicates that ViT is better prepared to handle certain characteristics from the PAD-UFES-IS split.

Interestingly, ViT achieved the highest accuracy of 80.63%

in this PAD-UFES-IS split without the application of Focal Loss. The result suggests that the model generalized well without additional mechanisms such as class balancing, possibly related to its architectural advantages over ResNet-101. However, in the PAD-UFES-FULL split, the model's performance with Focal Loss was close to that of ResNet-101, with an average accuracy of 73.92%.

ViT model generally maintains a high F1-Score across most dataset variants. It tends to have a marginally better F1-score and Recall on the PAD-UFES-IS dataset than ResNet-101, particularly without Focal Loss, where it achieves a higher consistency across folds (F1 up to 0.8270). In contrast, ResNet-101 struggles more on this dataset variant, with lower F1 and Recall values, especially when Focal Loss is applied. Focal Loss improved the overall F1 score for most splits, demonstrating its effectiveness in balancing precision and recall, especially in scenarios where the dataset was more imbalanced. Performance in the PAD-UFES-AUG split was stable, with average accuracies of 73.68% (with Focal Loss) and 72.98% (without Focal Loss).

For both models, ResNet-101 and ViT, the application of Focal Loss resulted in consistent improvements in the F1 score, indicating an improved ability to handle situations with class imbalance. However, the degree of improvement varied depending on the dataset split adopted. Although ResNet-101 and ViT showed similar performances in some specific splits, ViT showed a slight advantage, becoming more evident in the PAD-UFES-IS split, where its accuracy without Focal Loss surpassed that of ResNet-101.

Both models showed that data augmentation techniques, demonstrated in the PAD-UFES-AUG split, increase accuracy, reinforcing the importance of data diversity in model training.

6. Conclusions

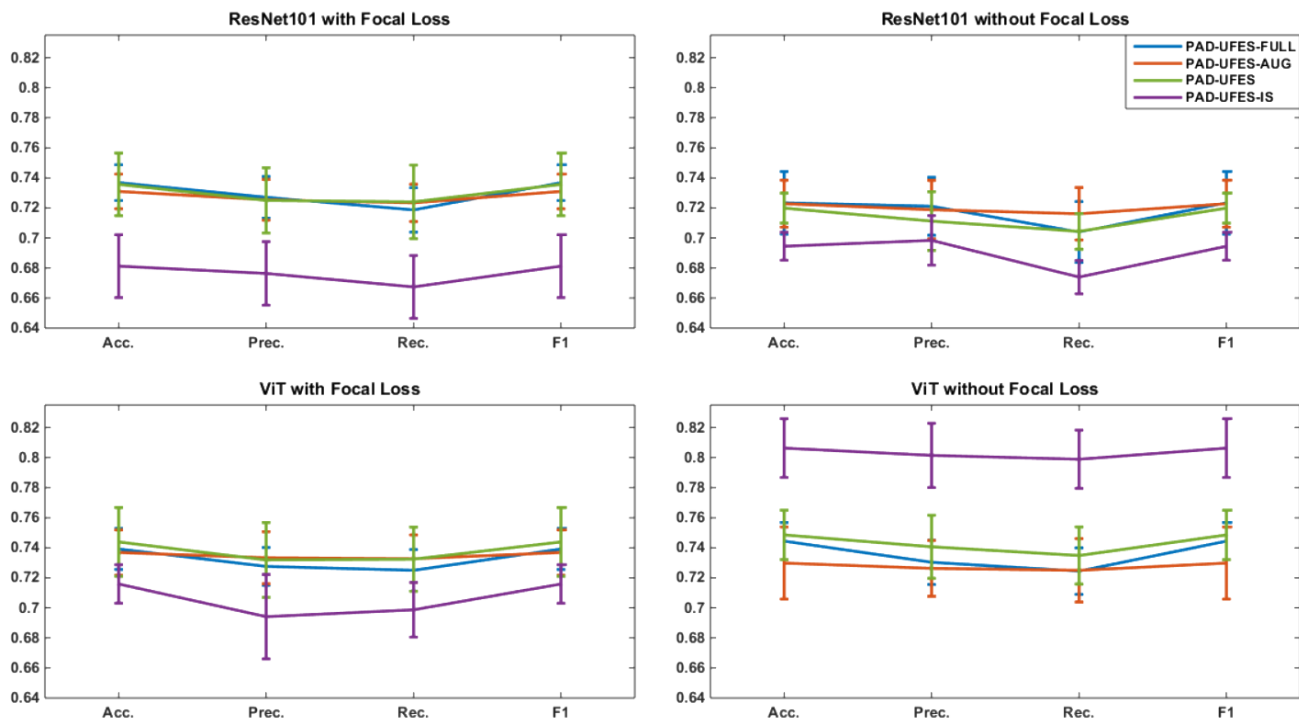
In this article, we addressed the automated detection of malignant skin conditions using deep learning. We conducted a comparative analysis between the ResNet-101 and ViT models applied to the PAD-UFES-20 dataset. To handle class imbalance, we explored techniques such as Focal Loss and data augmentation. Additionally, we examined different dataset splits to evaluate the impact of the presence and absence of these approaches on the models' performance. The results

Table 1. Test set evaluation for ResNet-101 with KFold ($k = 5$) and Focal Loss.

Focal Loss	Fold	PAD-UFES-FULL				PAD-UFES-AUG				PAD-UFES				PAD-UFES-IS			
		Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.
TRUE	1	0.7368	0.7208	0.7101	0.7368	0.7515	0.7485	0.7452	0.7515	0.7047	0.6943	0.6868	0.7047	0.6599	0.6485	0.6439	0.6599
	2	0.7398	0.7326	0.7183	0.7398	0.7251	0.7195	0.7187	0.7251	0.7632	0.7556	0.7556	0.7632	0.6657	0.6767	0.6528	0.6657
	3	0.7222	0.7193	0.7118	0.7222	0.7281	0.7264	0.7216	0.7281	0.7368	0.7232	0.7262	0.7368	0.6916	0.6875	0.6780	0.6916
	4	0.7544	0.7488	0.7443	0.7544	0.7251	0.7163	0.7157	0.7251	0.7339	0.7256	0.7278	0.7339	0.6772	0.6652	0.6654	0.6772
	5	0.7310	0.7139	0.7089	0.7310	0.7251	0.7163	0.7157	0.7251	0.7398	0.7263	0.7238	0.7398	0.7118	0.7040	0.6969	0.7118
FALSE	1	0.7310	0.7086	0.7056	0.7310	0.7456	0.7495	0.7452	0.7456	0.7251	0.7252	0.7069	0.7251	0.6945	0.7095	0.6782	0.6945
	2	0.7310	0.7292	0.7096	0.7310	0.7193	0.7062	0.7080	0.7193	0.7076	0.6803	0.6869	0.7076	0.6916	0.6890	0.6702	0.6916
	3	0.6901	0.6947	0.6757	0.6901	0.7076	0.7027	0.7032	0.7076	0.7339	0.7304	0.7201	0.7339	0.6830	0.7191	0.6591	0.6830
	4	0.7193	0.7297	0.6973	0.7193	0.7310	0.7269	0.7195	0.7310	0.7164	0.7094	0.7023	0.7164	0.6945	0.6773	0.6728	0.6945
	5	0.7456	0.7433	0.7316	0.7456	0.7105	0.7088	0.7046	0.7105	0.7164	0.7108	0.7056	0.7164	0.7089	0.6971	0.6895	0.7089

Table 2. Test set evaluation for ViT with KFold ($k = 5$) and Focal Loss.

Focal Loss	Fold	PAD-UFES-FULL				PAD-UFES-AUG				PAD-UFES				PAD-UFES-IS			
		Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.	Acc.	Prec.	F1	Rec.
TRUE	1	0.7427	0.7247	0.7254	0.7427	0.7398	0.7346	0.7362	0.7398	0.7251	0.6988	0.7087	0.7251	0.7176	0.7059	0.7063	0.7176
	2	0.7485	0.7432	0.7399	0.7485	0.7251	0.7193	0.7191	0.7251	0.7135	0.7116	0.7097	0.7135	0.6945	0.6450	0.6669	0.6945
	3	0.7281	0.7176	0.7130	0.7281	0.7193	0.7138	0.7152	0.7193	0.7632	0.7462	0.7463	0.7632	0.7291	0.7163	0.7133	0.7291
	4	0.7544	0.7380	0.7373	0.7544	0.7427	0.7425	0.7387	0.7427	0.7602	0.7543	0.7525	0.7602	0.7205	0.7016	0.7031	0.7205
	5	0.7222	0.7144	0.7092	0.7222	0.7573	0.7564	0.7542	0.7573	0.7573	0.7484	0.7447	0.7573	0.7176	0.7017	0.7038	0.7176
FALSE	1	0.7485	0.7413	0.7321	0.7485	0.7602	0.7489	0.7511	0.7602	0.7515	0.7487	0.7389	0.7515	0.7810	0.7717	0.7736	0.7810
	2	0.7281	0.7131	0.7050	0.7281	0.7251	0.7199	0.7198	0.7251	0.7661	0.7627	0.7516	0.7661	0.8012	0.7962	0.7931	0.8012
	3	0.7485	0.7265	0.7221	0.7485	0.7047	0.7061	0.7023	0.7047	0.7398	0.7342	0.7299	0.7398	0.8069	0.8053	0.8045	0.8069
	4	0.7368	0.7214	0.7170	0.7368	0.7105	0.7138	0.7091	0.7105	0.7251	0.7077	0.7045	0.7251	0.8069	0.8028	0.7963	0.8069
	5	0.7602	0.7493	0.7459	0.7602	0.7485	0.7426	0.7424	0.7485	0.7602	0.7498	0.7492	0.7602	0.8357	0.8314	0.8270	0.8357

**Figure 2.** Metrics comparison between ViT and ResNet-101 with and without Focal Loss.

show that both ResNet-101 and ViT are effective for classification tasks on the PAD-UFES-20 dataset, particularly when combined with techniques to manage imbalanced data, with ViT slightly outperforming ResNet-101. The ViT achieved an average accuracy of 80.63% on the PAD-UFES-IS split without the application of Focal Loss, indicating its superior generalization capability for certain data characteristics. Meanwhile, ResNet-101 reached an increase in average ac-

curacy to 73.68% with Focal Loss on the PAD-UFES-FULL split, compared to 72.34% without using it.

The selection of architecture and the implementation of class balancing techniques is done by carefully considering the dataset characteristics and the degree of class imbalance. In cases of significant class imbalance, Focal Loss, especially when used with data augmentation strategies, has proven beneficial. These findings are important for the field, indicating

that advanced deep learning architectures like ViT, when combined with methods to address class imbalance, significantly enhance accuracy in skin lesion classification. This improvement is vital for the early detection of malignant conditions, leading to better patient outcomes through precise interventions. By fully utilizing the capabilities of advanced architectures, this approach maximizes diagnostic accuracy and efficiency, unlocking the full potential of these technologies in clinical settings.

Despite the advances presented, the study still has certain limitations. The imbalance between the classes in the dataset, especially the Melanoma class, is a significant challenge due to the low number of images. Applying techniques such as Focal Loss and increasing the image data, the problem still exists, causing a negative influence on the generalization of the model. The fact that the dataset was collected using smartphones (under a specific condition) makes it necessary to evaluate the applicability of the models in other scenarios.

Future research should focus on advanced augmentation and oversampling techniques to effectively address class imbalance and enhance model performance, maximizing the potential of architectures for accurate and efficient skin lesion classification.

Acknowledgements

André R. Backes gratefully acknowledges the financial support of CNPq (Grant #307100/2021-9). We gratefully acknowledge the support of NVIDIA Corporation, USA with the donation of the TITAN Xp GPU used for this research. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Author contributions

Conceptualization: G.S.L.M. and M.V.S. Methodology: G.S.L.M., M.V.S., J.F.M., and A.R.B. Code Development: G.S.L.M. and M.V.S. Experiments: G.S.L.M. and M.V.S. Discussion of Results: G.S.L.M., M.V.S., J.F.M., and A.R.B. Writing - Original Draft: G.S.L.M. and M.V.S. Writing - Review & Editing: G.S.L.M., M.V.S., J.F.M., and A.R.B. Supervision: J.F.M. and A.R.B.

References

- [1] DEKONINCK, B.; CHRISTENBERY, T. Skin cancer screening in the medically underserved population: A unique opportunity for apns to make a difference. *Journal of the American Association of Nurse Practitioners*, LWW, v. 27, n. 9, p. 501–506, 2015.
- [2] WILLEMINK, M.; ROTH, H.; SANDFORT, V. Toward foundational deep learning models for medical imaging in the new era of transformer networks. *Radiology: Artificial Intelligence*, Radiological Society of North America, v. 4, n. 5, p. e210284, 2022.
- [3] SAEED, M. et al. The power of generative ai to augment for enhanced skin cancer classification: A deep learning approach. *IEEE Access*, IEEE, 2023.
- [4] LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.
- [5] HAN, K. et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 45, n. 1, p. 87–110, 2022.
- [6] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.
- [7] DOSOVITSKIY, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] PACHECO, A. G. et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, v. 32, p. 106221, 2020. ISSN 2352-3409.
- [9] SHEHZAD, K. et al. A deep-ensemble-learning-based approach for skin cancer diagnosis. *Electronics*, v. 12, n. 6, 2023. ISSN 2079-9292.
- [10] FLOSDORF, C. et al. *Skin Cancer Detection utilizing Deep Learning: Classification of Skin Lesion Images using a Vision Transformer*. 2024. Disponível em: <<https://arxiv.org/abs/2407.18554>>.
- [11] SAKLI, M. et al. Deep learning-based multi-stage analysis for accurate skin cancer diagnosis using a lightweight cnn architecture. In: *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*. [S.l.: s.n.], 2023. p. 1–6.
- [12] TSCHANDL, P. *The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions*. [S.l.]: Harvard Dataverse, 2018.
- [13] CARMO, N.; MARI, J. A comparative study of convolutional neural networks for classification of pigmented skin lesions. In: *Anais do XVII Workshop de Visão Computacional*. Porto Alegre, RS, Brasil: SBC, 2021. p. 171–176. ISSN 0000-0000.
- [14] NANCY, V. A. O.; PRABHAVATHY, P.; ARYA, M. S. Wel-odkc: weighted extreme learning optimal diagonal-kernels convolution model for accurate classification of skin lesions. *The Imaging Science Journal*, Taylor & Francis, v. 71, n. 4, p. 362–377, 2023.
- [15] NIJHAWAN, R.; BHATNAGAR, D.; ROY, S. Diagnosing skin lesion using multi-modal analysis. In: *2022 5th International Conference on Computational Intelligence and Networks (CINE)*. [S.l.: s.n.], 2022. p. 1–5.

- [16] LIMA, L. M. de; KROHLING, R. A. Exploring advances in transformers and cnn for skin lesion diagnosis on small datasets. In: XAVIER-JUNIOR, J. C.; RIOS, R. A. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2022. p. 282–296. ISBN 978-3-031-21689-3.
- [17] HE, K. et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Disponível em: <<http://arxiv.org/abs/1512.03385>>.
- [18] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. Disponível em: <<https://arxiv.org/abs/2010.11929>>.