Evaluating Deep Neural Skin Cancer Classifiers With Multiple Images Inputs

Afonso S. Magalhães[1], **Luis A. Souza Jr.**[1], **André G.C. Pacheco**[1]

[1]Department of Informatics and Graduate Program of Informatics
Federal University of Espírito Santo(UFES), Vitória, Espírito Santo, Brasil

afonso.magalhaes@edu.ufes.br, apacheco@inf.ufes.br, la.souza@inf.ufes.br

Abstract. *Skin cancer represents one-third of all globally diagnosed cancers. Despite its generally low mortality rate, late diagnosis remains a significant cause of complications. To mitigate such risks, Computer-Aided Diagnosis (CAD) systems have been developed to provide more accessible and timely diagnostic methods. While the CAD field has demonstrated considerable promise, most existing systems rely on a single image of the lesion, and the impact of using multiple images has not been extensively studied. This work aims to investigate how incorporating multiple images affects the efficiency and accuracy of CAD systems. Specifically, we evaluate the performance of three different deep learning models integrated into a stacking-like strategy that processes multiple image inputs. Notably, we achieved a 6% increase in balanced accuracy, without adding significant training or testing burdens to the existing models.*

Resumo. *O câncer de pele representa um terço de todos os cânceres diagnosticados globalmente. Embora tenha uma taxa de mortalidade geralmente baixa, o diagnóstico tardio continua sendo o principal fator para complicações. Para mitigar esses riscos, sistemas de Diagnóstico Assistido por Computador (CAD) vem sendo desenvolvidos para fornecer métodos de diagnóstico mais acessíveis e oportunos. Embora os CADs vem desmostrando resultados consistentes, a maioria dos sistemas existentes se baseia em uma única imagem da lesão, e o impacto do uso de múltiplas imagens de uma mesma lesão não vem sendo estudado. Este trabalho visa investigar como a incorporação de múltiplas imagens afeta a eficiência e a precisão dos sistemas de CAD. Especificamente, foi avaliado o desempenho de três diferentes modelos de aprendizado profundo integrados em uma estratégia de stacking que processa múltiplas entradas de imagem de uma mesma lesão. De maneira geral, foi observado aumento de até 6% na acurácia balanceada, sem adicionar processamento significativos de treinamento ou de teste aos modelos existentes.*

1. Introduction

Skin cancer is a worldwide health problem, being one of the most prevalent types of cancer globally [WHO 2025]. In Brazil, it accounts for approximately 31.2% of all cancer cases [INCA 2023]. Early diagnosis significantly improves prognosis; however, accurate identification of malignant lesions remains challenging, particularly in remote/rural areas where access to experts is limited [Feng et al. 2018]. Also, in emerging countries such as Brazil, the limited availability of medical instruments makes accurate diagnosis even more

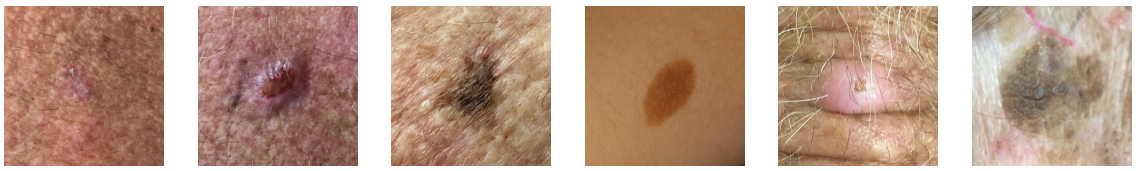
difficult. In this sense, the development of Computer Aided Diagnosis (CAD) systems is well desired to assist in skin cancer detection.

Over the past few years, CADs based on deep learning have been achieving promising results on skin cancer classification [Tuncer et al. 2024, Kumar et al. 2024, Cui et al. 2023, Pacheco and Krohling 2021, Celebi et al. 2019, Brinker et al. 2018]. Most of the proposed methods use only an image of the lesion to provide the diagnosis. There are models that rely on dermoscopic images [Sinz et al. 2017], while others use only clinical images [Maqsood and Damaševičius 2023, Pacheco and Krohling 2020a]. Some models also incorporate medical anamneses along with the images to improve diagnosis [Pacheco and Krohling 2021, Pacheco and Krohling 2020b, Li et al. 2020]. Regardless of the image type, the majority of the proposed models provide a diagnosis based on a single image sample per lesion.

When dermatologists assess skin lesions, they typically examine the lesion from multiple angles and light conditions to obtain a more comprehensive evaluation. This observation suggests that CAD systems could benefit from using multiple images of the same lesion, rather than relying on a single image, to improve diagnostic accuracy. In fact, some studies have reported using multiple images of skin lesions for classification purposes [Tanaka et al. 2021, Liu et al. 2020, Chen et al. 2016]. However, these works lack detailed explanations regarding model design and do not provide comparative metrics between using multiple images and single-image approaches. For example, in [Liu et al. 2020], the authors disclose that their model may accept up to six images as input. However, they do not provide details on how the model processes these images, whether it can operate with a single image, nor present experiments evaluating the performance based on varying the number of input images. The absence of such analysis may hinder the understanding of the impact of multiple image inputs on the performance of deep learning models. Although the improvement of results when using more images per analysis is expected, the magnitude of such improvement is still underreported in current literature.

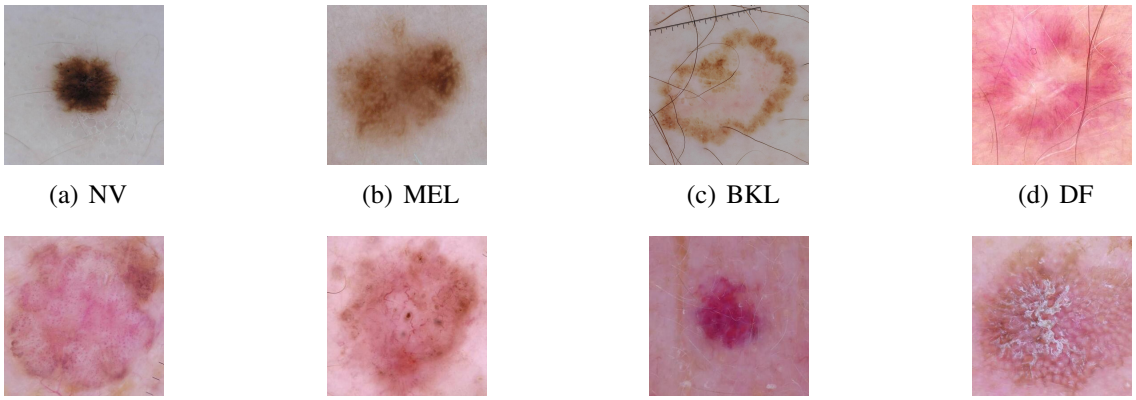
This work aims to perform a quantitative evaluation to assess the impact of using multiple images of the same skin lesion in a deep learning model. Specifically, we investigate if incorporating additional views of the lesion improves classification performance at a considerable rate. To achieve this, we employ Convolutional Neural Networks (CNNs), common in the field of skin lesion CAD systems, to analise the multiple views of the lesion, leveraging their capacity to extract intricate patterns from medical images. By systematically evaluating the model with varying numbers of input images, we aim to provide insights into the benefits of multi-image approaches for skin cancer diagnosis.

The remainder of this paper is organized as follows: Section 2 details the data used in this research, including its quantity, source, distribution, and the partitioning methods applied. Section 3 introduces the system developed for model comparison and analysis, describing the training process and performance evaluation with and without the multi-input approach. In Section 4, we present the results obtained from these evaluations, along with a statistical analysis of the data. Section 5 presents the discussion surrounding the results presented in the previous section. Finally, in Section 6, we conclude the experiment's findings and discuss potential directions for future research.



(a) ACK (b) BCC (c) MEL (d) NEV (e) SCC (f) SEK

Figure 1. An example for each type of skin disorder present in PAD-UFES-20+ dataset



(a) NV (b) MEL (c) BKL (d) DF

(e) SCC (f) BCC (g) SEK (h) SEK

Figure 2. An example for each type of skin disorder present in ISIC-19 dataset

2. Data

In this work, we employ two different datasets of skin lesions: PAD-UFES-20+ and ISIC-19. An expansion of PAD-UFES-20 [Pacheco et al. 2020], the PAD-UFES-20+ is a database composed of clinical skin lesions, along with the respective clinical information. The dataset contains 15,112 clinical images, of a total 5,589 patients, collected from smartphone devices, comprising 21 patient clinical features such as age, gender, anatomical region, cancer history, skin prototype, among others, and six different skin disorders, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Melanoma (MEL), and Nevus (NEV).

The ISIC-19 Challenge dataset [Tschandl et al. 2018, Codella et al. 2017, Hernández-Pérez et al. 2024] is widely used in the current literature for skin lesion diagnosis. The diagnosis present in ISIC-19 that exceeds PAD-UFES-20+ are Vascular Lesions (VASC) and Dermatofibroma (DF), both labels combined total less than 2% of the images avaliable in the dataset.

Table 2 shows the diagnostic distribution of all datasets, and Figures 1 and 2 present an example of each type of skin disorder present on each dataset. While the datasets share some of its diagnostic labels, the type of image collected in both are different, PAD-UFES-20+ presents clinical images, and ISIC-19 presents dermatoscopic images.

All employed datasets provide skin lesion evaluations with multiple images per lesion for some samples. To analyze the impact of multiple images on diagnosis, lesions

Label	PAD-UFES-20+	ISIC-19
ACK	8400	867
BCC	3042	3323
MEL	346	4522
NEV	253	12875
SCC	2282	628
SEK	689	2624
VASC	-	253
DF	-	239
Total	**15112**	**25331**

Table 1. Diagnostic distribution per images in each dataset

with more than one image were isolated from the main dataset, enabling model evaluation after training, as shown in Table 2. As a result, PAD-UFES-20+ contains 12,521 lesions with a single image and 660 lesions with at least two images, while ISIC-19 includes 8,872 single-image lesions and 3,393 with two or more images. The existence of lesions with a single image and others with multiple reflects the reality in witch the colletion process finds itself, this variability shows the value of a system capable of processing both use cases.

No. of Images	No. of Lesions	
	PAD-UFES-20+	ISIC-19
1	12,521	8,872
2 plus	660	3393

Table 2. The number of lesions containing a single image or at least two image samples.

3. Methodology

As previously mentioned, the main goal of this work is to quantitatively evaluate the impact of using multiple images of the same lesion in the classification of skin lesions using deep neural networks. To do so, we design the following experiment protocol:

- We employed three well-known Convolutional Neural Network (CNN) architectures: ResNet-50, ResNet-101 [He et al. 2016], and MobileNetV2 [Sandler et al. 2018] to classify skin lesions. These architectures were selected because they are widely used in the literature and have shown good performance in the classification of skin lesions [Souza et al. 2024, Pacheco and Krohling 2021]. All three models were pre-trained on the ImageNet dataset [Deng et al. 2009] and then fine-tuned on each employed dataset using samples with a single image of each lesion.
- An ensemble stacking-like strategy [Pavlyshenko 2018] was employed for each model, during the inference phase, to classify skin lesions using multiple images of the same lesion. Next, we compare the results using only a single image of each lesion with the results using multiple images of the same lesion.

Each CNN architecture was fine-tuned for each dataset using the samples with a single image of each lesion, described in Table 2, the lesions with multiple images used in the next stage of this study were not used in this tunning phase. We employed a 5-fold cross-validation strategy, stratified by lesion frequency and patient. The models were trained separately, the ISIC-19 tested model was trained[1] using the SGD optimizer with a learning rate of 0.0001, momentum of 0.9, and weight decay of 0.01. For ISIC-19, the learning rate was reduced by a factor of 0.1 if no improvement was observed for 10 consecutive epochs, with early stopping triggered if this pattern persisted for 15 epochs. For PAD-UFES-20+, these thresholds were set to 5 and 7 epochs, respectively. To address class imbalance in the dataset, we used a weighted cross-entropy loss function, with weights assigned based on label frequencies. All images were resized to 224×224, and data augmentation was performed using common image transformations, including horizontal and vertical flips, brightness, contrast, and saturation adjustments, image scaling, and random noise addition [Pacheco and Krohling 2020a, Gessert et al. 2020].

Figure 3 illustrates the stacking-like strategy employed in this work. This approach involves the use of multiple images of one skin lesion. Formally, let us consider a sample with multiple images as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, where k represent the number of images for a single lesion. Each image is processed by a CNN model, producing a set of predictions $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \ldots \hat{\mathbf{y}}_k\}$:

$$\hat{\mathbf{Y}} = \mathbf{CNN}(\mathbf{X}) \Rightarrow \hat{\mathbf{y}}_k = \mathbf{CNN}(\mathbf{x}_k) \tag{1}$$

with $\hat{\mathbf{Y}} \in \mathbb{R}^{k \times n}$, where n is the number of classification labels present in the dataset. In other words, each label n of each image k is associated with a prediction probability $(\hat{y}_n^k)$. Next, we aggregate all prediction probabilities associated to the same image k as follows:

$$\hat{y}_n = \sum_{i=0}^{k} \hat{\mathbf{y}}_n^k \tag{2}$$

As such, we have the aggregated predictions considering all k image samples ($\hat{\mathbf{y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$). The final prediction ($\hat{y}_{\text{final}}$) is computed according to the following equation:

$$\hat{y}_{\text{final}} = \underset{n}{\operatorname{argmax}}\, \sigma(\hat{\mathbf{y}}) \tag{3}$$

where $\sigma()$ is the softmax function computed as $\sigma(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{k} e^{y_j}}$. This straightforward yet effective strategy enables the model to take into account any number of images of the same lesion for classification.

4. Experimental Evaluation

In this section, we present the results of the different CNN architectures using the stacking-like strategy. We compare the results obtained using only a single image of

[1]All models were implemented using PyTorch, and the code will be made publicly available after the review phase.

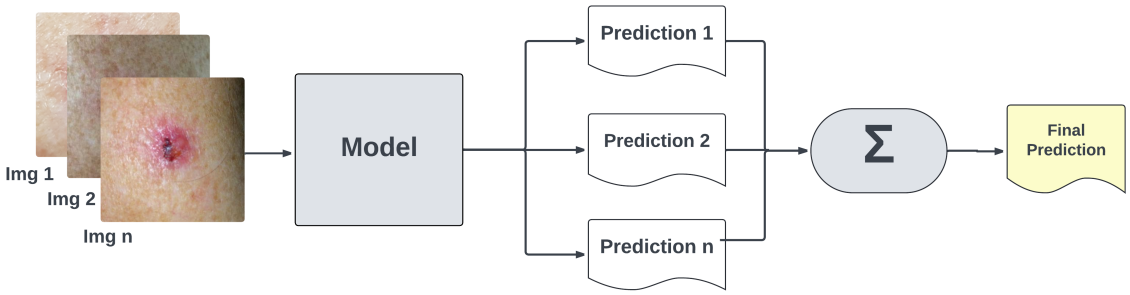


Figure 3. The schematic of the stacking strategy pipeline used to evaluate multiple images per lesion in the classification of skin lesions using CNNs

Model	Input	ACC	BACC	F1	AUC
MobileNet	Single	0.684 ± 0.015	0.670 ± 0.013	0.696 ± 0.014	0.803 ± 0.008
	Multi	0.724 ± 0.011	0.687 ± 0.011	0.738 ± 0.010	0.815 ± 0.006
Resnet50	Single	0.655 ± 0.017	0.635 ± 0.014	0.670 ± 0.017	0.782 ± 0.008
	Multi	0.714 ± 0.008	0.695 ± 0.013	0.726 ± 0.008	0.818 ± 0.007
Resnet101	Single	0.686 ± 0.012	0.674 ± 0.023	0.698 ± 0.010	0.805 ± 0.013
	Multi	0.728 ± 0.017	0.689 ± 0.008	0.738 ± 0.013	0.816 ± 0.006

Table 3. Performance of the CNN models for the PAD-UFES-20+ dataset using a single and multiple images of the same skin lesion

each lesion with the results obtained using multiple images of the same lesion, as seen in Table 3 for the PAD-UFES-20+, and Table 4 for the ISIC-19. The data used as the single image portion of tests is part of the multiple images set, in order to make a precise comparision between the methods. To measure the performance, we computed the average and standard deviation of the following metrics: accuracy (ACC), balanced accuracy (BACC), F1-score (F1) and the aggregated area under the curve (AUC). As the dataset is imbalanced, we consider BACC as the main metric [Gessert et al. 2020]. We also performed a statistical analysis using the Wilcoxon signed-rank test with a significance level of 0.05 [Derrac et al. 2011] to compare the results obtained using a single image of each lesion with the results obtained using multiple images of the same lesion. The result for the statistical test is presented in Table 5.

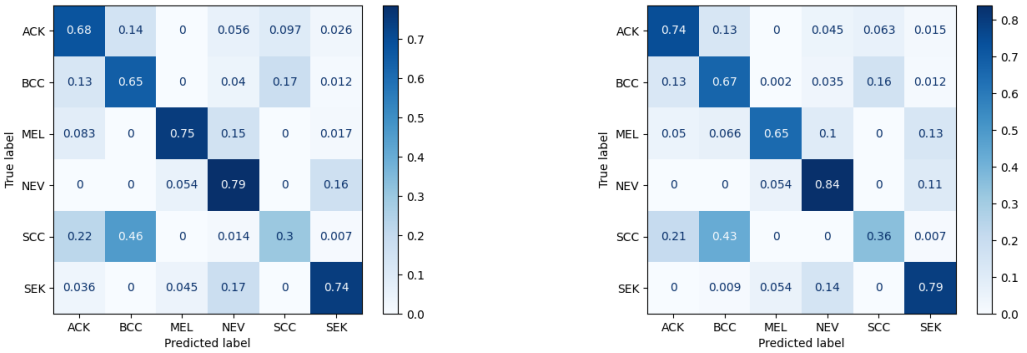
As we may observe, the results for the three CNN architectures show that em-

Model	Input	ACC	BACC	F1	AUC
MobileNet	Single	0.653 ± 0.011	0.588 ± 0.022	0.660 ± 0.011	0.767 ± 0.012
	Multi	0.704 ± 0.007	0.631 ± 0.025	0.708 ± 0.007	0.792 ± 0.013
Resnet50	Single	0.675 ± 0.015	0.575 ± 0.024	0.678 ± 0.015	0.762 ± 0.013
	Multi	0.716 ± 0.016	0.605 ± 0.040	0.716 ± 0.014	0.780 ± 0.021
Resnet101	Single	0.674 ± 0.022	0.605 ± 0.010	0.676 ± 0.021	0.776 ± 0.007
	Multi	0.723 ± 0.016	0.639 ± 0.024	0.722 ± 0.019	$0,797 \pm 0.013$

Table 4. Performance of the CNN models for the ISIC-19 dataset using a single and multiple images of the same skin lesion

Model	p_{value}	
	PAD-UFES-20+	ISIC-19
MobileNetV2	$\sim 10^{-09}$	$\sim 10^{-08}$
Resnet50	$\sim 10^{-31}$	$\sim 10^{-04}$
Resnet101	$\sim 10^{-16}$	$\sim 10^{-04}$

Table 5. The result of the Wilcoxon pairwise test for all models. Each comparison pair is given by the model using a single and multiple images of the same skin lesion



(a) Single input

(b) Multiple inputs

Figure 4. The confusion matrix for each type of input, considering the MobileNetV2 model

ploying multiple images of the same lesion slightly improves all classification metrics for all models, varying marginally between datasets. The balanced accuracy increases approximately 3% for the models regarding the ISIC-19 dataset and, in the best case of PAD-UFES-20+, we observe an improvement of 6%. This analysis is on pair with the results of the statistical tests, which show that the differences between the results obtained using a single image of each lesion and multiple images of the same lesion are statistically significant in all cases.

Figure 4 presents the confusion matrix for each input type using the MobileNetV2 architecture. The matrices show that employing multiple images of the same lesion improves diagnostic accuracy across all classes, except for melanoma. This is a concerning result, since melanoma is the deadliest type of skin cancer. This outcome may be attributed to the limited number of melanoma samples in the dataset; hence, any misclassification of a melanoma sample significantly impacts the overall performance. Nonetheless, the results suggest that, in some cases, incorporating additional images of the same lesion can lead to mistakes in the classification process.

Lastly, Figure 5 illustrates an example of the variation in prediction probabilities when the model is fed with single and multiple images of the same lesion. In this example, the MobileNetV2 model classifies the skin lesion as BCC with a 45.9% probability when using a single image. When two images are used, the probability increases to 68.1%, and with three images, it reaches 98.6%. This example demonstrates how the model's confidence improves as more images of the same lesion are supplied. Thus, this experiment

(a) 45.9% probability of a BCC

(b) 68.1% probability of BCC using a) and b)

(c) 98.6% probability of BCC using a), b), and c)

Figure 5. An illustration showing the variation in prediction probabilities when the MobileNetV2 model is provided with one (a), two (b), and three (c) images of the same skin lesion

shows that the use of multiple images may also lead to a more confident classification, which may be useful in clinical practice.

5. Discussion

Regarding the benefits of the method, one can observe its improvements, once the behavior could be maintained regardless of its backbone or dataset. The bigger the amount of images presented of the same lesion, the preciser the prediction, as presented in Figure 5. One of our method's benefits is its implementation that does not require additional cost or extension in the CNN architectures we employ. The major difference we propose is the introduction of multiple samples in the input, keeping the trivial architecture of the CNN models to be trained and further evaluated. Hence, in terms of model size, our method does not present any harming effect to any architecture do be used, facilitating its implementation to a wide range of different backbones.

As Tables 3 and 4 depict, the use of multiple images indeed enhances the correct classification of skin lesions not only in absolute values but in statistical terms, as shown by the Wilcoxon statistical evaluation presented in Table 5, the use of multiple images not only statistically improved the correct classification of skin lesion, but also suggests improvements in the classifier confidence since more instances of the same lesions are used during the inference task. Indeed, this process of multiple instances being used refines the capability of the classifier to understand and extract patterns of the current lesion being classified. Figure 5 clearly highlights this process, where different images of the same lesions make the generalization of features more robust to changes in illumination and angulation when extracting and learning patterns of skin lesions.

In principle, the inference of samples based on multiple images should enhance the correct classification and also the confidence related to the classification itself since more are used and the same classification of multiple samples corroborates such decision. In fact, we understand that, according to the quality and aspect of the different samples from the same lesion, different classification outcomes could confuse the final and correct prediction. We address such limitation in our method, understanding its dependence on high-quality samples for the inference input set. Even considering that, when high-

quality samples are used in inference time, the classification confidence is improved in computational and clinical terms.

Finally, we address one more limitation in our method: the necessity of multiple samples for the inference. Not every dataset presents more than one sample for the same subject, and such a lack may represent a bottleneck of our approach. Nonetheless, we still defend our method, leveraging the CNN model generalization capability without adding extra layers or transformations to its architecture, in a design that makes the final skin lesion class decision more robust and trustworthy since it is based on multiple instances of the same evaluated lesion.

6. Conclusions

In this work, we evaluated the impact of employing multiple images of the same lesion on the performance of deep learning models for skin lesion classification. Our experiments, conducted using the PAD-UFES-20+ and ISIC-19 datasets and three well-known CNN architectures (MobileNetV2, ResNet-50, and ResNet-101), demonstrated that incorporating multiple images leads to a consistent, improvement in classification metrics for all models. Also, employing multiple images contributed to a more confident classification, as demonstrated by the increase in prediction probabilities, which may be useful in clinical practice. On the other hand, our analysis also highlighted potential limitation, as seen in the case of melanoma classification, where additional images did not consistently improve diagnostic performance. Overall, our findings suggest that using multiple images may enhance the confidence and accuracy of skin lesion classification models up to 6% in balanced accuracy. This improvement could be beneficial in clinical settings, where confident diagnoses are important. Future research should focus on optimizing models to fully exploit the benefits of multi-image inputs, especially for underrepresented classes such as melanoma, datasets, and explore other aggregation methods. Also, including the use of multiple images in the models' training process, ensuring that the previously considered redundant data can be used at its full potential.

Acknowledgments

The authors thank the Espírito Santo Research Foundation (FAPES); the Capixaba Institute of Education, Research and Innovation (ICEPi); the National Council for Scientific and Technological Development (CNPq); the Brazilian Ministry of Health (MoH); and Brazilian National Program of Genomics and Precision Health (Genomas Brasil).

References

Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A. H., and von Kalle, C. (2018). Skin cancer classification using convolutional neural networks: systematic review. *Journal of Medical Internet Research*, 20(10):e11936.

Celebi, M. E., Codella, N., and Halpern, A. (2019). Dermoscopy image analysis: overview and future directions. *IEEE Journal of Biomedical and Health Informatics*, 23(2):474–478.

Chen, R. H., Snorrason, M., Enger, S. M., Mostafa, E., Ko, J. M., Aoki, V., and Bowling, J. (2016). Validation of a skin-lesion image-matching algorithm based on computer vision technology. *Telemedicine and e-Health*, 22(1):45–50.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., and Halpern, A. (2017). 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). *arXiv:1710.05006*.

Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. A., and Huo, Y. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Derrac, J., García, S., Molina, D., and Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18.

Feng, H., Berk-Krauss, J., Feng, P. W., and Stein, J. A. (2018). Comparison of dermatologist density between urban and rural counties in the united states. *JAMA Dermatology*, 154:1265—-1271.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., and Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, pages 1–8.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N. C., Rotemberg, V., Halpern, A. C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., Puig, S., Vilaplana, V., and Malvehy, J. (2024). Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*, 11(1):641.

INCA (2023). Incidência do câncer no Brasil. Instituto Nacional do Câncer (INCA). Available on: https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2023.pdf. Last access: 11 Mar 2025.

Kumar, A., Kumar, M., Bhardwaj, V. P., Kumar, S., and Selvarajan, S. (2024). A novel skin cancer detection model using modified finch deep cnn classifier model. *Scientific Reports*, 14(1):11235.

Li, W., Zhuang, J., Wang, R., Zhang, J., and Zheng, W.-S. (2020). Fusing metadata and dermoscopy images for skin disease diagnosis. In *IEEE International Symposium on Biomedical Imaging*, pages 1996–2000.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G. S., Peng, L. H., Webster, D. R., Ai, D., Huang, S. J., Liu, Y., Dunn, R. C., and Coz, D. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908.

Maqsood, S. and Damaševičius, R. (2023). Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural Networks*, 160:238–258.

Pacheco, A. G. and Krohling, R. A. (2020a). The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545.

Pacheco, A. G. and Krohling, R. A. (2020b). Learning dynamic weights for an ensemble of deep models applied to medical imaging classification. In *IEEE International Joint Conference on Neural Networks*, pages 1–8.

Pacheco, A. G., Lima, G. R., Salomão, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., Alves Jr, F. C., Esgario, J. G., Simora, A. C., Castro, P. B., et al. (2020). PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:1–10.

Pacheco, A. G. C. and Krohling, R. (2021). An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE journal of biomedical and health informatics*. In press.

Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 255–258.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Pecognition*, pages 4510–4520.

Sinz, C., Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J.-Y., Kreusch, J., et al. (2017). Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109.

Souza, L., Pacheco, A., Angelo, G., Oliveira-Santos, T., Palm, C., and Papa, J. (2024). Liwterm: A lightweight transformer-based model for dermatological multimodal lesion detection. In *Anais da XXXVII Conference on Graphics, Patterns and Images*, Porto Alegre, RS, Brasil. SBC.

Tanaka, M., Saito, A., Shido, K., Fujisawa, Y., Yamasaki, K., Fujimoto, M., Murao, K., Ninomiya, Y., Satoh, S., and Shimizu, A. (2021). Classification of large-scale image database of various skin diseases using deep learning. *International Journal of Computer Assisted Radiology and Surgery*, 16(11):1875–1887.

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161.

Tuncer, T., Barua, P. D., Tuncer, I., Dogan, S., and Acharya, U. R. (2024). A lightweight deep convolutional neural network model for skin cancer image classification. *Applied Soft Computing*, page 111794.

WHO (2025). Skin Cancer. World Health Organization (WHO). Available on: https://www.iarc.who.int/cancer-type/skin-cancer/. Last access: 11 Mar 2025.