Parallel-Hierarchical Multi-Task Learning for Skin Lesion Classification with Multimodal Data

Camila Alves Dias[1], **Thiago Sotoriva Lermen**[2], **Rômulo Marconato Stringhini**[2],
Cristiano Andre da Costa[1], **Cláudio Rosito Jung**[2], **Manuela M. Costa**[2],
Dimitris V. Rados[3], **Fabiana Carvalho**[3], **Marcelo R. Gonçalves**[2]

[1]PPG Computação Aplicada – Universidade do Vale do Rio dos Sinos (UNISINOS)
Porto Alegre – RS – Brazil

[2]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[3]Núcleo de Telessaude Tecnico Cientifico do Rio Grande do Sul
Porto Alegre – RS – Brazil

cmdias@outlook.com.br, thiago.lermen@gmail.com, rmstringhini@gmail.com
cac@unisinos.br, crjung@inf.ufrgs.br, manumartinscosta@gmail.com
dimitris.rados@telessauders.ufrgs.br, fabianacarvalho.foz@gmail.com
marcelorog@gmail.com

Abstract. *This paper introduces a novel approach for skin cancer detection that leverages both smartphone-captured images and clinical data. The proposed method utilizes a parallel-hierarchical multi-task learning framework to classify skin lesion types and assess malignancy simultaneously. Compared to existing methods reporting false positive rates around 3.6% in the Melanoma Skin Cancer dataset, our approach achieves a false positive rate of 8.2% while maintaining a high true positive rate of 94.03% for malignant lesion detection. These results demonstrate the potential of the approach to enhance diagnostic accuracy and support more effective treatment decisions in teledermatology settings.*

1. Introduction

Early detection of malignant skin lesions is vital for survival. Balancing timely diagnosis and avoiding unnecessary referrals is challenging, as false positives may lead to excision of benign lesions [Al Zegair et al. 2023]. Although dermatoscopic tools aid diagnosis, access to specialists is limited, especially in rural areas [Coustasse et al. 2019]. Teledermatology (TD) helps by enabling remote access to dermatologists, reducing delays and improving care access [Khan et al. 2021].

Teledermatology allows remote consultation by capturing skin lesion images via smartphones and sharing them with specialists. This enables primary care physicians to collaborate with dermatologists. However, smartphone images often suffer from blurriness, poor lighting, and varying lesion sizes. Despite these issues, TD effectively reduces patient wait times and healthcare costs.

Recent deep learning advances have improved skin lesion classification [Selvaraj et al. 2024], but the scarcity of large annotated smartphone clinical image datasets—especially in regions like Brazil—limits model generalizability [Pasquali et al. 2020]. Integrating contextual data such as patient metadata and lesion features has shown promise in enhancing classification [Pacheco and Krohling 2020, Pacheco and Krohling 2021].

To address teledermatology challenges, we present a framework that combines smartphone clinical images and patient metadata. Our parallel-hierarchical multi-task learning independently processes skin condition classification and malignancy detection, then fuses results hierarchically. The approach incorporates metadata-aware attention and novel image enhancement techniques.

Experimental results demonstrate that our approach outperforms existing multimodal baselines, such as the MetaBlock proposed by Pacheco et al. [Pacheco and Krohling 2021], achieving higher balanced accuracies of 92.7% and 77.65% for binary and 6-class classification tasks, respectively. These results underscore the potential of our method to improve diagnostic accuracy and support more effective treatment decisions in teledermatology settings.

2. Related Work

Deep learning has advanced skin lesion analysis by fine-tuning pretrained CNNs. Performance depends on backbone, classification head, and dataset size. Popular models like ResNet [He et al. 2016], DenseNet [Huang et al. 2017], and EfficientNet [Tan and Le 2019] show strong results [Lanjewar et al. 2023]. Recent studies include smartphone images [Roh et al. 2021], which increase access but face resolution and lighting challenges.

To improve skin lesion classification, hierarchical and multi-task learning approaches [Demyanov et al. 2017, Barata et al. 2021] leverage disease taxonomies but often depend on curated clinical datasets, limiting use in real-world settings. Recent works [Kowacz et al. 2023] explore smartphone-acquired images, addressing challenges like lighting and resolution, and emphasize the importance of camera standards for reliable performance in teledermatology.

To address these limitations and improve clinical relevance, multimodal learning has gained traction by integrating clinical metadata—such as age, gender, and lesion location—with image features [Liu et al. 2020, Pacheco and Krohling 2020]. Fusion strategies vary from early concatenation of image and metadata features [Liu et al. 2020], to mid-level or joint fusion via multi-task learning [Kawahara et al. 2018], and late-fusion methods that maintain separate processing streams before merging [Wang et al. 2021].

Recent methods have aimed to improve modality fusion by modeling complex interactions. Attention-based approaches like Metablock dynamically weight metadata [Pacheco and Krohling 2021], while channel-wise modulation strategies like Metanet enhance feature integration [Li et al. 2020]. However, many of these techniques depend on dermoscopic images, limiting their applicability in real-world clinical settings.

3. The Proposed Framework

We propose a deep learning framework designed to enhance the accuracy of skin condition classification in teledermatology. The model processes smartphone-captured clinical images and patient metadata as input and produces both disease-level classifications and referral-level predictions for malignancy assessment. The main goal is to optimize performance on the binary classification task, as this directly supports clinical decision-making for determining whether a patient should be referred to a dermatologist.

3.1. The Parallel-Hierarchical Multi-Task Network

The proposed architecture processes an RGB image resized to $224{\times}224{\times}3$ and a clinical metadata vector as input. The resolution balances computational efficiency and performance, trading off some image detail compared to higher resolutions like $256{\times}256$ (PH2) or $512{\times}512$ (Derm7pt). The metadata vector includes seven binary features encoded as 0 (true), 1 (false), and 0.5 (unknown), and supports numerical and categorical data via one-hot encoding for flexibility.

The image branch uses a shared backbone to generate a feature map of size $h \times w \times C_o$, which is split into two task-specific branches. These are later fused with corresponding clinical data streams. The clinical branch consists of fully connected (FC) layers. The input $1 \times C_c$ metadata vector is passed through a shallow sub-network that performs feature adjustment (FA) by using two FC layers with ReLu activation and dropout, producing a $1 \times \lfloor C_o/2 \rfloor$ tensor (see Figure 1, green box), which encodes common features for the 2- and 6-class problems. This tensor is fed into two FC layers (one per task), each outputting $1 \times C_o$ vector with sigmoid activation. These act as per-channel weights for fusing clinical data with corresponding task-related branches from the image backbone. Finally, each branch is passed through a classification head: one for six disease classes y_i and another for the binary referral decision z_i.

To improve information flow across backbone channels, we adopt a channel attention mechanism inspired by Squeeze-and-Excite [Hu et al. 2018], replacing the standard Global Average Pooling (GAP) as suggested in [Tang et al. 2019]. Given an input tensor of size $h \times w \times C_o$ and a target of N output classes, the proposed SE-CAM (Squeeze-and-Excite Class Activation Map) head consists of two branches. The first branch applies spatial GAP to obtain a $1 \times C_o$ tensor followed by two fully connected layers that reduce it to $1 \times \lfloor C_o/16 \rfloor$ and then expand it to $1 \times N$, followed by a sigmoid activation. The second branch uses a 1×1 convolution to adjust the input features, producing a $h \times w \times N$ tensor. This is then multiplied channel-wise with the output of the first branch. Finally, a GAP layer followed by a softmax yields a $1 \times N$ vector with per-class scores. A visual overview of the SE-CAM module is shown in the center of Figure 1 (cyan block).

The architecture produces two separate classification heads: one for the 6-class task and another for the binary task. However, the binary prediction can also be derived from the multi-class output. Specifically, the sum $\sum_{i=1}^{3} y_i$ aggregates the scores for benign classes (ACK, NEV, SEK), while $\sum_{i=4}^{6} y_i$ does so for malignant classes (BCC, MEL, SCC). These values are computed using a sum-pooling layer with a stride 3 (gray box in Figure 1) applied to y_i. To enhance binary classification, we average these pooled scores with the outputs from SE-CAM(2): $z'_1 = (z_1 + \sum_{i=1}^{3} y_i)/2$ and $z'_2 = (z_2 + \sum_{i=4}^{6} y_i)/2$. This produces the final binary output z'_i which behaves like a softmax output (i.e., $z'_1 + z'_2 = 1$).

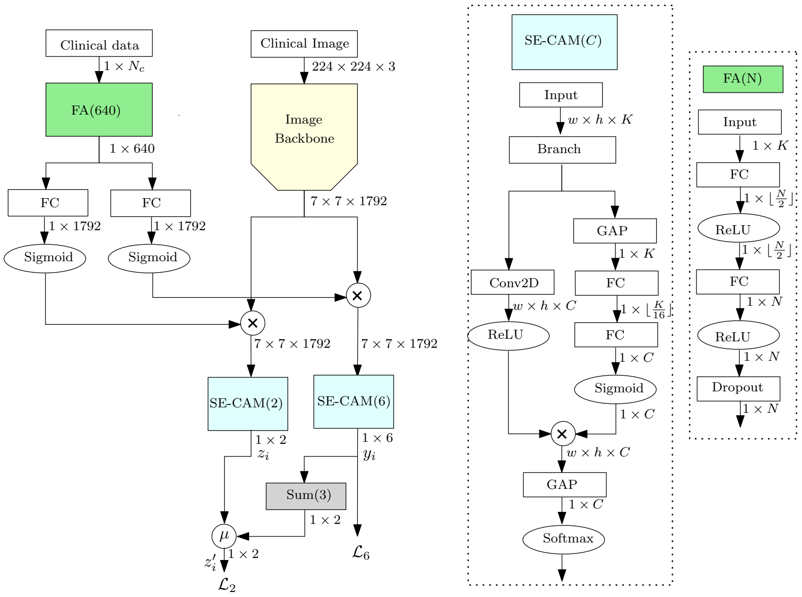


Figure 1. The proposed network architecture for multi-task skin lesion classification using images and clinical data (number of channels based on an EfficientNet-B4 backbone).

3.2. Multi-task loss

The proposed architecture produces two output tensors: y_i for $i = 1, \cdots, 6$ related to the 6-class problem, and z_i' for $i = 1, 2$ related to the binary problem. They are fed to loss functions $\mathcal{L}_6$ and $\mathcal{L}_2$, respectively, which are defined as weighted categorical cross-entropies to cope with class imbalance. Let N_i denote the number of training samples for class i in the 6-class problem. The number of samples for the binary problem are $N_1' = \sum_{i=1}^{3} N_i$ and $N_2' = \sum_{i=4}^{6} N_i$ for the benign and malignant classes, respectively. Also, let $N = \sum N_i$ denote the total number of training samples. The weighted cross-entropy loss functions are defined as

$$\mathcal{L}_6 = -\sum_{i=1}^{6} \frac{N}{N_i} \hat{y}_i \log y_i, \quad \mathcal{L}_2 = -\sum_{i=1}^{2} \frac{N}{N_i'} \hat{z}_i' \log z_i', \tag{1}$$

where $\hat{y}_i$ and $\hat{z}_i'$ are one-hot-encoded GT labels related to the 6-class and binary problems, respectively. The final loss is simply a weighted combination of the two tasks, given by

$$\mathcal{L} = \omega \mathcal{L}_6 + (1 - \omega) \mathcal{L}_2, \tag{2}$$

where $\omega \in [0, 1]$ controls the relative weight of the six-class loss (we empirically selected $\omega = 0.5$).

3.3. Augmentation Strategies

To address the limited size of clinical image datasets and reduce overfitting, data augmentation plays a critical role. Although generative methods, such as GAN, have been explored, they are often computationally expensive and exhibit limited effectiveness for skin lesion classification [Bissoto et al. 2021]. Instead, we adopt a policy based on primitive transformations, each defined by a probability of application and a discrete magnitude parameter $M \in 0, 1, ..., 10$ that modulates its intensity. Null probability disables a primitive. We build on RandAugment [Cubuk et al. 2020] to define and scale transformation intensities linearly with M. Figure 2 illustrates an example using three selected primitives.

Baseline primitives include horizontal and vertical image flipping (no magnitude needed) and a combined crop-and-scale transformation implemented as an affine operation. The latter randomly varies the scaling factor s within $[1 - \Delta s, 1 + \Delta s]$, with $\Delta s = 0.5$, and applies random offsets in both spatial directions within half the image width or height, respectively. These values are scaled according to the selected magnitude M.

To better simulate the conditions encountered during real-world clinical capture, we introduce tailored augmentation primitives. First, we propose out-of-plane rotations using a planar homography that emulates 3D camera rotation. Yaw, pitch, and roll angles (α, β, γ) are sampled uniformly within $(\pm 45°, \pm 45°, \pm 180°)$ and scaled with M. These help model acquisition angles that deviate from a frontal view, which is common in clinical settings. The experimental evaluation demonstrated that an in-plane rotation of $30°$ yielded the best results for DenseNet-121, ResNet-50, and EfficientNet-B0. This aligns with previous studies showing that moderate rotation angles (e.g., 15º to 30º) effectively augment dermoscopic images by simulating realistic variations in capture angles, enhancing model robustness without excessive distortion [Esteva et al. 2017, Zhu et al. 2020].

Illumination and color variations common in clinical images are simulated using a multiplicative bilinear map on the luminance channel in HSV space. Illumination factors are randomly sampled at three boundary points, with mild hue and saturation perturbations added. Gaussian blur mimics defocus from fixed-distance smartphone capture. Although AdaIN [Huang and Belongie 2017] could normalize color, it was avoided due to its high computational cost and the effectiveness of simpler augmentations.

We also include primitives that improve regularization without distorting clinical features. Gaussian noise, sampled with $\sigma_n \in [0, 0.2]$, is uniformly added to RGB channels to simulate sensor noise and improve the robustness of the model. Lastly, we employ Cutout, where a rectangular patch of random size (up to half the image width and height) is removed from the image. Unlike DropBlock, Cutout does not require architectural changes and serves as an effective regularizer [Harun et al. 2022].

We did not apply other transformations, such as shearing, solarizing, or contrast shifting, as they may overly distort the lesion's shape or color, which are essential for accurate diagnosis. Instead, we focused on augmentations that simulate realistic acquisition artifacts or improve generalization without compromising lesion integrity.

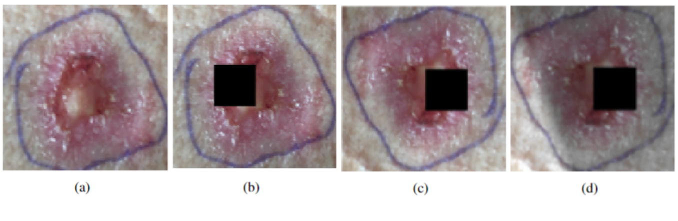


Figure 2. Sequential image augmentation. (a) Original image. (b) Cutout. (c) Horizontal Flip. (d) Color and Illumination changes.

3.3.1. Metadata augmentation

Metadata in this work refers to binary clinical features identified by primary care doctors, based on patient feedback. There may be missing or inaccurate data, so we propose an augmentation scheme for the metadata branch to reduce bias or overfitting.

Given a binary metadata $x \in 0, 1$, we define a tolerance $\epsilon > 0$ and generate a random variable $\nu(x)$ with uniform distribution $\mathcal{U}(0, \epsilon)$. The augmented metadata x' is then defined as:

$$x' = \begin{cases} 1 - \nu(x) & \text{if } x = 1 \\ \nu(x) & \text{if } x = 0 \end{cases}, \tag{3}$$

This formulation introduces uncertainty into the original binary values. Although metadata related to actual patients must be either 1 (true), 0 (false), or 0.5 (unknown), we believe that the proposed augmentation scheme can act as a regularizer for the network. In fact, the CutMix augmentation strategy for images [Yun et al. 2019] also generates unrealistic images at training time and achieves better regularization. Unlike image augmentation primitives, we did not set different magnitude values and used $\epsilon = 0.2$. Also, metadata augmentation is applied with a probability of 0.5.

3.4. Datasets

Most skin lesion classification methods rely on high-quality dermoscopic datasets like HAM10000 [Tschandl et al. 2018] and BCN20000 [Combalia et al. 2022], which suffer from bias [Bissoto et al. 2020] and degrade significantly when applied to low-quality images typical in teledermatology [Maier et al. 2022]. Clinical image datasets such as SD-198 [Kinyanjui et al. 2020], SD-260 [Yang et al. 2019], and others [Wall et al. 2020, Groh et al. 2021] often lack metadata, availability, or are not optimized for real-world settings.

The PAD-UFES-20 dataset [Pacheco et al. 2020] contains 2,298 smartphone-acquired clinical images from 1,373 patients across six lesion types: Actinic Keratosis (ACK), Nevus (NEV), Seborrheic Keratosis (SEK), Basal Cell Carcinoma (BCC), Melanoma (MEL), and Squamous Cell Carcinoma (SCC). Although class distribution is imbalanced, benign (ACK, NEV, SEK) and malignant (BCC, MEL, SCC) categories are balanced in the binary setting. Metadata includes seven dermatologically relevant binary clinical features. The images were captured using various smartphone models without standardized camera settings, reflecting real-world variability in resolution, focus, and lighting conditions.

We used the ISIC 2019 dataset to evaluate our architecture. This public resource, part of an ongoing challenge since 2016, combines BCN20000 [Combalia et al. 2019] and HAM10000 [Tschandl et al. 2018] with 25,331 training and 8,238 validation samples. It includes clinical characteristics like age, sex, and anatomical region, along with eight types of skin lesions: MEL, NEV, BCC, ACK, Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and SCC. For testing, only MEL, SCC, and BCC were considered malignant, and ACK, NEV, and BKL were considered benign.

ISIC and PAD-UFES-20 differ in image type (dermoscopic vs. clinical) and metadata availability (3 vs. 21 features). To unify metadata, we converted anatomical location into boolean features indicating lesion presence.

3.5. Training and Inference Details

The PAD-UFES dataset lacks predefined splits, so we use 5-fold cross-validation for robust evaluation. Each classifier is trained on five folds, with 12.5% of each training fold reserved for validation—yielding roughly 70% training, 10% validation, and 20% testing.

We used ImageNet-pretrained backbones, freezing the first three layers, combined with multi-task classification heads trained from scratch on image and clinical data. Models trained for 150 epochs with Adam optimizer and tuned hyperparameters. Learning rate was adjusted during training, and the best weights selected based on validation accuracy. Early stopping with 25-epoch patience was used to prevent overfitting.

We used a data augmentation policy with $K = 3$ primitives and $M = 7$ magnitudes, selected experimentally without exhaustive tuning. During inference, test-time augmentation (TTA) averaged predictions over 15 augmented image versions. To maintain consistency, cutoff and clinical data were excluded from TTA.

4. Results and discussion

We present a comparative analysis of different backbone architectures against methods that incorporate both images and clinical data, such as Metanet [Li et al. 2020] and Metablock [Pacheco and Krohling 2021].

4.1. Backbone Analysis and Ablation Studies

We first evaluated the impact of different backbone architectures. Due to data limitations, convolutional networks were favored over transformers [Dosovitskiy et al. 2021]. We tested DenseNet-121 [Huang et al. 2017], ResNet-50 [He et al. 2016], and EfficientNet variants [Tan and Le 2019]. An ablation study further examined the contributions of test-time augmentation (TTA), clinical metadata, and the SE-CAM classification head.

Table 1 reports model size, training accuracy, and average balanced accuracy across five folds, with and without TTA. TTA consistently improved performance, leading to the selection of EfficientNet-B4 with TTA. Including patient and lesion metadata further boosted results, aligning with prior findings [Gessert et al. 2019]. Excluding metadata reduced accuracy by 5% (binary) and 7% (six-class), despite the "Clinical image" branch alone achieving good results with a smaller model.

Lastly, replacing the SE-CAM head with a fully connected layer plus dropout (rate 0.4) showed inferior performance across all metrics (Table 1), especially for six-class balanced accuracy with TTA, confirming the effectiveness of SE-CAM.

4.2. Comparison with other methods

We thoroughly evaluated our approach against state-of-the-art multimodal methods, focusing on per-class recall for binary classification—a key metric in teledermatology. While balancing false negatives and false positives is crucial, detailed analysis of this trade-off is beyond this study's scope.

Table 1. Our results using PAD-UFES-20 dataset (average of 5-fold cross-validation test subsets) for the 2- and 6- class problems using different backbones and ablation parameters

2*Backbone	Model	Train	Train	2*Bal Acc$_2$	Bal Acc$_2$	2*Bal Acc$_6$	Bal Acc$_6$
	Size	Acc$_2$	Acc$_6$		(TTA)		(TTA)
Resnet-50	21.30M	**99.15**	**97.76**	89.29	90.85	70.02	73.13
Densenet-121	8.21M	98.67	96.95	90.16	90.74	71.24	72.56
Effnet-B0	**5.95M**	98.78	96.52	91.65	92.30	75.30	75.53
Effnet-B1	8.46M	97.97	93.65	92.02	92.12	72.80	75.26
Effnet-B2	10.04M	98.21	93.77	**92.41**	92.64	75.27	76.35
Effnet-B3	13.47M	98.16	93.41	92.03	92.17	74.35	75.78
Effnet-B4	20.30M	97.44	93.74	92.39	**92.75**	**76.43**	**77.65**
Result using *only images*							
Effnet-B4	20.30M	97.00	92.64	86.74	87.77	69.40	70.35
Result using a Fully Connected classification head							
Effnet-B4	19.80M	97.71	92.80	91.80	92.36	75.96	76.49

For fair comparison, we evaluated two metadata-driven architectures—Metanet and Metablock—based on EfficientNet-B4, in both 6-class and binary tasks. Our training followed [Pacheco and Krohling 2021] with matched metadata handling and augmentation adapted for real-valued features. Hyperparameters were consistent, and we used full 5-fold cross-validation instead of a fixed test fold.

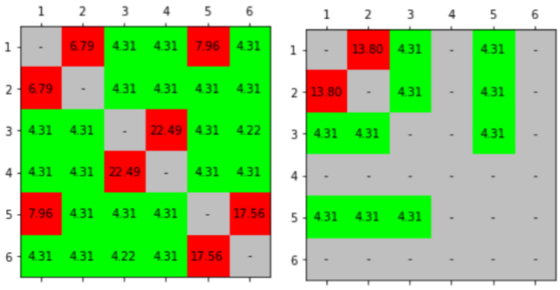
Table 2 reports the accuracy for cancer and non-cancer classes, with the best class-wise results highlighted, as well as the balanced accuracy for the 6-class task and the AUC, all presented with mean and standard deviation. Training a dedicated binary classifier outperformed inferring binary labels from a 6-class model. Multi-task learning led to improved recall and overall accuracy. Test-time augmentation (TTA) further boosted cancer recall and 6-class balanced accuracy, although it introduced higher variance across folds.

Table 2. 5-Fold Accuracy Mean and Standard Deviation Results for Different Approaches Using PAD-UFES-20 Dataset

benign	cancer	Bal Acc$_6$	AUC
1. Our approach without TTA			
93.30 ± 0.95	**91.48** ± 0.82	76.43 ± 3.10	**95.60** ± 3.23
2. Our approach with TTA			
94.03 ± 1.02	**91.48** ± 1.20	**77.65** ± 3.98	90.43 ± 3.12
3. Metablock (six-class → two-class)			
85.96 ± 3.47	86.84 ± 4.27	64.82 ± 5.19	87.05 ± 4.55
4. Metablock (two-class)			
84.10 ± 2.70	83.14 ± 4.68	–	80.57 ± 4.40
5. Metanet (six-class → two-class)			
91.18 ± 3.27	89.32 ± 2.52	72.75 ± 3.70	90.43 ± 3.11
6. Metanet (two-class)			
89.62 ± 4.60	87.52 ± 3.48	–	88.08 ± 3.43

We applied the pairwise Wilcoxon signed-rank test [Wilcoxon 1945] on six models using balanced accuracy for binary and 6-class tasks. Figure 3 shows p-values, with significant differences at the 5% level highlighted in green. Our method with test-time augmentation (TTA) outperformed others, though not significantly. Results should be interpreted cautiously due to small sample size. Models 4 and 6 lacked 6-class results.

Our study also evaluated our method on the ISIC 2019 dataset, which does not include clinical metadata. To adapt it, we converted metadata fields into binary indicators by transforming rows into columns. Table 3 highlights the top-performing results. For comparison, Pacheco et al. [Pacheco and Krohling 2021] reported 80.7% accuracy and 76.2% balanced accuracy using Metablock with EfficientNet-B4 on this dataset.



	1	2	3	4	5	6
1	-	6.79	4.31	4.31	7.96	4.31
2	6.79	-	4.31	4.31	4.31	4.31
3	4.31	4.31	-	22.49	4.31	4.22
4	4.31	4.31	22.49	-	4.31	4.31
5	7.96	4.31	4.31	4.31	-	17.56
6	4.31	4.31	4.22	4.31	17.56	-

(a) 2-class Balanced Accuray

	1	2	3	4	5	6
1	-	13.80	4.31	-	4.31	-
2	13.80	-	4.31	-	4.31	-
3	4.31	4.31	-	-	4.31	-
4	-	-	-	-	-	-
5	4.31	4.31	4.31	-	-	-
6	-	-	-	-	-	-

(b) 6-class Balanced Accuray

Figure 3. p-values (in %) for pairwise Wilcoxon tests comparing all six methods shown in Table 2.

Table 3. Per-fold accuracy results for ISIC 2019 dataset using EfficientNet-B4.

Fold	benign	cancer	Bal Acc$_2$	AUC
1. Our approach with TTA				
1	**88.03**	86.92	**84.93**	**95.50**
2	87.76	84.88	83.22	90.04
3	87.74	87.09	83.07	91.25
4	82.59	**89.80**	77.72	89.90
5	76.96	87.42	79.46	87.39
$\mu \pm \sigma$	84.61 ± 4.84	87.22 ± 1.75	81.68 ± 2.66	**90.02** ± 3.08

5. Conclusion

We propose a novel skin lesion classification method combining clinical images and patient metadata, designed for teledermatology. It supports disease categorization and binary classification to aid referrals, using image augmentation and a channel attention-based classification head to improve performance.

On the PAD-UFES-20 dataset, our method achieved a 94.03% true positive rate and 8.52% false positive rate for malignant lesion detection, outperforming many traditional diagnostics. The model's flexibility allows adjusting the false positive–false negative tradeoff to suit various healthcare settings.

Future work includes integrating unstructured clinical text data to improve diagnostics and collaborating with TelessaúdeRS to create a curated annotated skin lesion dataset. We aim to support primary care and enhance dermatological care with accessible, robust AI tools.

References

Al Zegair, F., Naranpanawa, N., Betz-Stablein, B., Janda, M., Soyer, H. P., and Chandra, S. S. (2023). Application of machine learning in melanoma detection and the identification of 'ugly duckling' and suspicious naevi: A review. *arXiv*.

Barata, C., Celebi, M. E., and Marques, J. S. (2021). Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, 110:107413.

Bissoto, A., Valle, E., and Avila, S. (2020). Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 740–741.

Bissoto, A., Valle, E., and Avila, S. (2021). Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1847–1856.

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., Puig, S., et al. (2019). Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Combalia, M. et al. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 international skin imaging collaboration grand challenge. *The Lancet Digital Health*, 4(5):e330–e339.

Coustasse, A., Sarkar, R., Abodunde, B., Metzger, B. J., and Slater, C. M. (2019). Use of teledermatology to improve dermatological access in rural areas. *Telemedicine and e-Health*, 25(11):1022–1032.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

Demyanov, S., Chakravorty, R., Ge, Z., Bozorgtabar, S., Pablo, M., Bowling, A., and Garnavi, R. (2017). Tree-loss function for training neural networks on weakly-labelled datasets. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 287–291. IEEE.

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. In *International Conference on Learning Representations (ICLR)*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., and Schlaefer, A. (2019). Skin lesion classification using ensembles of multi-resolution efficientnets with metadata. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828.

Harun, M. F., Samah, A. A., Shabuli, M. I. A., Majid, H. A., Hashim, H., Ismail, N. A., Abdullah, S. M., and Alias, A. (2022). Incisor malocclusion using cut-out method and convolutional neural network. *Progress in Microbes and Molecular Biology*.

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

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510.

Kawahara, J., Daneshvar, S., Argenziano, G., and Hamarneh, G. (2018). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546.

Khan, M. A., Muhammad, K., Sharif, M., Akram, T., and de Albuquerque, V. H. C. (2021). Multi-class skin lesion detection and classification via teledermatology. *IEEE journal of biomedical and health informatics*, 25(12):4267–4275.

Kinyanjui, N. M., Odonga, T., Cintas, C., Codella, N. C., Panda, R., Sattigeri, P., and Varshney, K. R. (2020). Fairness of classifiers across skin tones in dermatology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI*, pages 320–329. Springer.

Kowacz, , Janusz, A., Świderska Chadaj, Z., Kleczyński, P., Krecicki, T., Kruk, M., Szczepaniak, K., Korbicz, J., and Wróbel, Z. (2023). Assessing smartphone-based image acquisition for skin cancer classification using convolutional neural networks. *Biomedical Signal Processing and Control*, 83:104626.

Lanjewar, M. G., Panchbhai, K. G., and Charanarur, P. (2023). Lung cancer detection from ct scans using modified densenet with feature selection methods and ml classifiers. *Expert Systems with Applications*, 224:119961.

Li, W., Zhuang, J., Wang, R., Zhang, J., and Zheng, W.-S. (2020). Fusing metadata and dermoscopy images for skin disease diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1996–2000. IEEE.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908.

Maier, K., Zaniolo, L., and Marques, O. (2022). Image quality issues in teledermatology: A comparative analysis of artificial intelligence solutions. *Journal of the American Academy of Dermatology*, 87(1):240–242.

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*, 32:106221.

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

Pacheco, A. G. C. and Krohling, R. A. (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*.

Pasquali, P., Sonthalia, S., Moreno-Ramirez, D., Sharma, P., Agrawal, M., Gupta, S., Kumar, D., and Arora, D. (2020). Teledermatology and its current perspective. *Indian dermatology online journal*, 11(1):12.

Roh, Y.-S., Kim, C.-W., Kim, N.-H., Suh, K.-M., Park, J.-I., and Lee, J.-H. (2021). Feasibility of a deep learning–based smartphone application for mobile dermoscopic melanoma detection. *Archives of Dermatological Research*, 313(9):743–749.

Selvaraj, K. M., Gnanagurusubbiah, S., Roy, R. R. R., Balu, S., et al. (2024). Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics. *Current Problems in Cancer*, 49:101077.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., and Coppola, G. (2019). Efficient skin lesion segmentation using separable-unet with stochastic weight averaging. *Computer methods and programs in biomedicine*, 178:289–301.

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(1):1–9.

Wall, C., Young, F., Zhang, L., Phillips, E.-J., Jiang, R., and Yu, Y. (2020). Deep learning based melanoma diagnosis using dermoscopic images. In *Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020)*, pages 907–914. World Scientific.

Wang, S., Yin, Y., Wang, D., Wang, Y., and Jin, Y. (2021). Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Transactions on Cybernetics*.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Yang, J., Wu, X., Liang, J., Sun, X., Cheng, M.-M., Rosin, P. L., and Wang, L. (2019). Self-paced balance learning for clinical skin disease recognition. *IEEE transactions on neural networks and learning systems*, 31(8):2832–2846.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Zhu, X., Liu, F., Zhu, M., and Zhang, J. (2020). Skin lesion classification using deep learning with data augmentation. *Computer Methods and Programs in Biomedicine*, 195:105591.