

# AI-Powered Skin Cancer Detection with Dual-Phase Deep Learning, Clinical Fusion and Explainable AI

Prof. Sangita Chaudhari<sup>1</sup>, Arjan Pathan<sup>2</sup>, Piyush Bhavsar<sup>3</sup>, Radha Jadhav<sup>4</sup>, Mahesh Patil<sup>5</sup>

Department of Information Technology<sup>1-5</sup>

Sandip Institute of Technology and Research Centre, Nashik, India

An Autonomous Institute Affiliated to Savitribai Phule Pune University, Pune

<sup>1</sup>sangita.chaudhari@sitrc.org, <sup>2</sup>arjanpathan.112023336@sitrc.org, <sup>3</sup>piyushbhavsar.112023273@sitrc.org,

<sup>4</sup>radhajadhav.122024080@sitrc.org, <sup>5</sup>maheshpatil.112022279@sitrc.org

**Abstract:** Skin cancer is among the most common cancers worldwide, and melanoma found late has high mortality, yet dermatologists are scarce in rural and low-resource regions. This project, “AI-Powered Skin Cancer Detection with Dual-Phase Deep Learning, Clinical Fusion and Explainable AI”, builds a decision-support system that works from smartphone photographs and a short clinical questionnaire instead of dermoscopy equipment. Phase 1 is an EfficientNet-B3 classifier trained on ISIC 2019 (8 lesion classes, lesion-level split). Phase 2 is a two-head fusion model that joins the image embedding with 14 clinical features and is fine-tuned on the smartphone dataset PAD-UFES-20 (6 classes, patient-level split). A rule-based red-flag layer covers six rare cancers that have no public training images. Grad-CAM, ABCD dermoscopy scoring and SHAP explain each prediction. Phase 1 reached  $0.647 \pm 0.022$  balanced accuracy and  $0.579 \pm 0.020$  macro F1 in 5-fold cross-validation. Phase 2 reached 0.689 macro F1 and 0.704 balanced accuracy with clinical fusion, against 0.694 and 0.736 for the image-only head, so fusion did not improve accuracy. A fairness audit shows a 0.247 sensitivity gap between light and medium skin tones on DDI, and no malignant case detected in the DermaCon-IN neoplasm subset (52 images, Fitzpatrick III-V only, no light-skin baseline). The system is intended to support triage and referral, not to diagnose.

**Keywords:** Skin Cancer, Melanoma, Deep Learning, EfficientNet, Clinical Fusion, Grad-CAM, SHAP, Explainable AI, Fairness, Smartphone Imaging

## I. INTRODUCTION

Skin cancer is among the most commonly diagnosed cancers worldwide, and melanoma is its deadliest form. Outcomes depend strongly on the stage at diagnosis, so early detection is the most effective way to save lives. In practice, access to a dermatologist is the bottleneck, especially in rural and low-resource regions.

Deep learning has reached dermatologist-level accuracy on curated dermoscopic benchmarks [1], [5], [8]. Most of that work relies on dermoscopy images that need specialist equipment. It also ignores the clinical history a doctor would ask about, explains its output poorly, and is rarely checked across skin tones [2]. Smartphone photographs paired with a short questionnaire are a more realistic input for first-contact screening [3], [6], [7].

This project, “AI-Powered Skin Cancer Detection with Dual-Phase Deep Learning, Clinical Fusion and Explainable AI”, addresses these gaps with five contributions:

- **Dual-phase learning.** An EfficientNet-B3 classifier is trained on ISIC 2019 (8 classes), then transferred into a two-head fusion model that combines the image with 14 clinical features and is fine-tuned on PAD-UFES-20 smartphone photographs.



- **Safety layer.** Six rule-based red-flag checks for rare cancers with no public training images force an urgent referral regardless of the network output. Predictions below 55% confidence are reported as Uncertain.
- **Explainability.** Grad-CAM heatmaps, ABCD dermoscopy scoring and SHAP clinical-feature attributions accompany every prediction.
- **Fairness audit.** Sensitivity is measured by Fitzpatrick skin type on DDI and DermaCon-IN, and the gap is reported rather than hidden.
- **Working system.** A FastAPI and Supabase backend serves patient, doctor and admin roles with JWT authentication.

The system is designed to support triage and referral. It does not replace a clinical diagnosis.

## II. LITERATURE SURVEY

Research on automated skin lesion analysis has moved from binary melanoma detection on dermoscopy images toward multi-class classification, clinical-data fusion, explainability and fairness evaluation. Eight recent studies (2021 to 2025) were reviewed in detail, and Table I summarises their datasets, methods and reported results.

TABLE I: LITERATURE SURVEY OF DIFFERENT PAPERS

| Sr. No. | Method / Algorithm           | Key Idea                                                                                                                                                                                                        | Advantages                                                                               | Disadvantages                                                                                                       | Author / Source            |
|---------|------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|----------------------------|
| 1       | Transfer-learning comparison | Fine-tune ResNet-50, MobileNet, DenseNet-169, InceptionV3 and InceptionResNetV2 on 3,297 dermoscopic ISIC images for benign versus malignant classification, with augmentation and hyperparameter tuning.       | ResNet-50 was best: 93.5% accuracy, 0.94 precision, 0.861 ROC-AUC.                       | Small dataset, binary labels only, no explainability or external validation.                                        | Islam and Panta, 2024 [1]  |
| 2       | DDI fairness study           | Curate the Diverse Dermatology Images (DDI) set of 656 clinical images spanning Fitzpatrick I-VI and test existing dermatology AI models on it.                                                                 | Exposed a 27%-36% ROC-AUC drop and worse performance on dark skin and uncommon diseases. | Evaluation set only; the gap closed only after fine-tuning on DDI itself, that is, with more diverse training data. | Daneshjou et al., 2022 [2] |
| 3       | Smartphone skin cancer app   | ResNet-50 image encoder fused with clinical metadata (age, location, bleeding, itching) on 2,057 PAD-UFES-20 smartphone images (70% image, 30% clinical features); React Native app with Flask/PyTorch backend. | 85.5% balanced accuracy and 96.4% recall, favouring fewer missed cancers.                | Small dataset; no hospital-scale deployment.                                                                        | Krohling et al., 2021 [3]  |



| Sr. No. | Method / Algorithm                | Key Idea                                                                                                                                                                        | Advantages                                                                                                              | Disadvantages                                                                                         | Author / Source             |
|---------|-----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|-----------------------------|
| 4       | Multi-dataset multiclass ensemble | Six public datasets merged into 32,748 images and 9 classes (including an unknown class); EfficientNet-B4/B5, SE-ResNeXt-101, Inception-ResNet-v2 and NASNet-A-Large ensembled. | Realistic multiclass, outlier-aware setting.                                                                            | Only 63.4% balanced accuracy; no explainability.                                                      | Steppan and Hanke, 2021 [4] |
| 5       | FCDS-CNN lightweight CNN          | Custom CNN with class weighting and augmentation, trained from scratch on HAM10000 (10,015 images, 7 classes) and compared with four pretrained models.                         | 96.66% accuracy, above the pretrained baselines (86.7%-92.0%).                                                          | Dermoscopy only; no explainability or external validation.                                            | Nawaz et al., 2025 [5]      |
| 6       | Hybrid ViT + XGBoost              | Black-hat transform and adaptive thresholding, ViT-B/16 image features, XGBoost on clinical metadata, stacked, on PAD-UFES-20 (2,298 images, 6 classes).                        | 97.61% accuracy, 96.88% precision, 97.63% recall on phone photographs; Grad-CAM and SHAP explanations.                  | Small dataset; only 52 melanoma images.                                                               | El Mertahi et al., 2025 [6] |
| 7       | Multimodal DenseNet-121           | DenseNet-121 image encoder concatenated with a clinical network (age, symptoms, Fitzpatrick type, family history) on PAD-UFES-20, tuned with BOHB.                              | AUC 0.91, above image-only and clinical-only models; reported by Fitzpatrick type; Grad-CAM and permutation importance. | Binary task only; no external validation; few darker-skin samples.                                    | Deng et al., 2025 [7]       |
| 8       | Skin-DeepNet pipeline             | Hair removal, Mask R-CNN with GrabCut segmentation, HRNet-attention and Deep Belief Network features, XGBoost stacking fusion.                                                  | 99.65% (ISIC 2019) and 100% (HAM10000) accuracy.                                                                        | Computationally heavy; dermoscopy only; near-perfect accuracy needs cautious reading for overfitting. | Al-Waisy et al., 2025 [8]   |

Across the eight studies three patterns emerge. First, the highest accuracies (93.5% to 100%) come from studies trained and tested only on dermoscopic images [1], [5], [8], while studies on smartphone images report more modest and more realistic results [3], [6], [7]. Second, every study that fused image and clinical data [3], [6], [7] outperformed its image-only comparison, which supports a multimodal design. Third, fairness across skin tone is rarely evaluated, and the one study that tested it directly [2] found 27% to 36% lower ROC-AUC and worse results on darker skin; explainability appears in only [6] and [7]. This project combines smartphone input, clinical fusion, explanation and a fairness audit in one pipeline.



### III. DISCUSSION

The system classifies lesions reasonably well on dermoscopy, gains no accuracy from clinical data on smartphone photographs, and explains its output. It also shows a clear accuracy gap on darker skin, which is the main limitation reported here.

#### A. Interpretation of Results

TABLE II: CLASSIFICATION PERFORMANCE OF THE TWO PHASES (PHASE 2 TRAINED ON 1,830 IMAGES, EVALUATED ON 231)

| Model                    | Evaluation set                      | Classes | Balanced accuracy | Macro F1          |
|--------------------------|-------------------------------------|---------|-------------------|-------------------|
| Phase 1, EfficientNet-B3 | ISIC 2019, 5-fold cross-validation  | 8       | $0.647 \pm 0.022$ | $0.579 \pm 0.020$ |
| Phase 2, image-only head | PAD-UFES-20 validation (231 images) | 6       | 0.736             | 0.694             |
| Phase 2, fused head      | PAD-UFES-20 validation (231 images) | 6       | 0.704             | 0.689             |

Phase 1 reached a macro AUC-ROC of 0.780 on a single held-out test split of the 8-class ISIC 2019 task (the other Phase 1 figures above are from 5-fold cross-validation), with a best validation macro F1 of 0.6108 at epoch 50 out of 60 total epochs. Cross-validation spread is small ( $\pm 0.022$  balanced accuracy), so the result is stable across folds. The weakest class is squamous cell carcinoma, with 33.8% recall.

Phase 2 transferred all 576 of 576 backbone tensors from Phase 1, so the smartphone model starts from dermoscopy features. Adding the 14 clinical features did not improve accuracy. Macro F1 moved by  $-0.005$ , a statistical tie on 231 images, but balanced accuracy also fell from 0.736 to 0.704 with fusion, consistent with the clinical encoder adding training noise on this small dataset. Fusion is therefore justified by the clinical explanation it enables, not by an accuracy gain.

The fairness audit is the most important finding for deployment (Table III).

TABLE III: SENSITIVITY BY SKIN TONE AT A FIXED DECISION THRESHOLD

| Evaluation set       | Skin type                                               | Malignant sensitivity |
|----------------------|---------------------------------------------------------|-----------------------|
| DDI (n = 656)        | Fitzpatrick I-II                                        | 0.612                 |
| DDI (n = 656)        | Fitzpatrick III-IV                                      | 0.365                 |
| DDI (n = 656)        | Fitzpatrick V-VI                                        | 0.375                 |
| DermaCon-IN (n = 52) | Neoplasm subset, FST III-V only, no light-skin baseline | 0.000                 |

Sensitivity falls by 0.247 between Fitzpatrick I-II and III-IV, and the model misses every malignant case in the small DermaCon-IN set at the same threshold. ISIC 2019 is dominated by light-skinned European and Australian patients, which is the likely cause and matches earlier findings [2]. DermaCon-IN is the neoplasm subset only, with 52 images, all Fitzpatrick III-V and no light-skin cases, so it has no internal light-skin baseline and that zero is a warning rather than a precise estimate.

The explanation tools behave as expected in most cases. Grad-CAM focuses on the lesion for melanoma, nevus, basal cell carcinoma, dermatofibroma and vascular lesions, is only partly focused on diffuse actinic keratosis, and fails on squamous cell carcinoma and on benign keratosis, where a ruler in the frame attracts attention. Globally, two missing-indicator columns (Fitzpatrick missing, rank 2; diameter missing, rank 3) rank above the symptoms, but they are excluded as non-clinical signals. Among real clinical features, SHAP ranks age first, elevation second, bleeding third and sex fourth. A direct logit-flip test confirmed that all six symptom inputs change the output by 0.09 to 0.19.



## *B. Explanation of System Architecture*

### *1) Data Collection*

Two public datasets are used. ISIC 2019 [9]-[11] provides 25,331 training and 8,238 test dermoscopy images in 8 classes: melanoma, nevus, basal cell carcinoma, actinic keratosis, benign keratosis, squamous cell carcinoma, dermatofibroma and vascular lesion. PAD-UFES-20 [12] provides smartphone photographs with patient data in 6 classes, with no dermatofibroma or vascular class. A patient-level split gives 1,830 training, 231 validation and 237 held-out test images (2,298 in total). At inference the patient answers a short questionnaire (age, sex, lesion site, elevation, and whether the lesion has grown, changed, bled, itched or hurt).

### *2) Splitting and Preprocessing*

Images are resized to 300 x 300 pixels. ISIC 2019 is split by lesion and PAD-UFES-20 by patient, so no lesion or patient appears in both training and test data. This prevents the leakage that inflates results when several photographs of one lesion are scattered across splits.

### *3) Phase 1: EfficientNet-B3 Classifier*

Phase 1 is an EfficientNet-B3 network of about 12 million parameters, trained for 60 epochs on an RTX 3050 6 GB GPU. The loss is focal loss ( $\gamma = 2$ ) with label smoothing to handle class imbalance and noisy labels. Output is a probability over the 8 classes.

### *4) Phase 2: Two-Head Fusion Model*

The Phase 1 backbone is reused to produce a 512-dimensional image embedding. A small network maps the 14 clinical features to a 32-dimensional embedding. The two are concatenated (544 dimensions), passed through a 256-unit dense layer and classified into 6 classes. An image-only head is trained alongside it, which makes the ablation in Table II possible.

### *5) Red-Flag Safety Layer*

Six rare cancers have no public training images, so they are handled by rules taken from published criteria: Merkel cell carcinoma (AEIOU, at least 3 of 5 features), dermatofibrosarcoma protuberans, Kaposi sarcoma, cutaneous T-cell lymphoma, sebaceous carcinoma and angiosarcoma. If any rule fires, the case is forced to Urgent Referral regardless of the network output.

### *6) Risk Categorisation*

Predicted classes map to risk levels. Melanoma, basal cell carcinoma and squamous cell carcinoma give Urgent Referral. Actinic keratosis gives High Risk. Nevus, benign keratosis, dermatofibroma and vascular lesions give Low Risk. Any prediction below 55% confidence is reported as Uncertain, with advice to see a dermatologist.

### *7) Explainability*

Three tools explain each result. Grad-CAM, taken at the last block of the EfficientNet-B3 backbone, shows where the model looked. The ABCD module scores asymmetry (flip overlap, threshold 0.85), border irregularity (compactness index, 0 to 8), colour diversity (k-means in LAB space,  $k = 3$  to 6, chosen by silhouette) and diameter (bounding box in pixels), then combines them into a Total Dermoscopy Score. Diameter is measured in pixels because no physical calibration target was present, and segmentation uses GrabCut with a centre-seeded rectangle instead of a trained segmentation model. ABCD scores therefore demonstrate interpretability and are indicative only, not clinical measurements. SHAP, computed with a kernel explainer over a PCA-decorrelated background, ranks the clinical features that moved the prediction.

### *8) Backend and Output Layer*

A FastAPI service exposes the pipeline through a /predict endpoint. Supabase provides authentication, a PostgreSQL database (profiles, lesion cases, predictions, doctor reviews, doctor-patient links) and image storage for lesion photographs and heatmaps. Three roles are supported: patients submit images and view their history, doctors review linked patients' cases and record a verdict, and admins manage users. 13 of 24 designed endpoints are live; an end-to-end test passed 27 of 29 checks, and a prediction returns in under 3 seconds on the RTX 3050. The overall data flow is shown in Fig. 1.



Every image passes safety rules, a classifier and three explanation tools

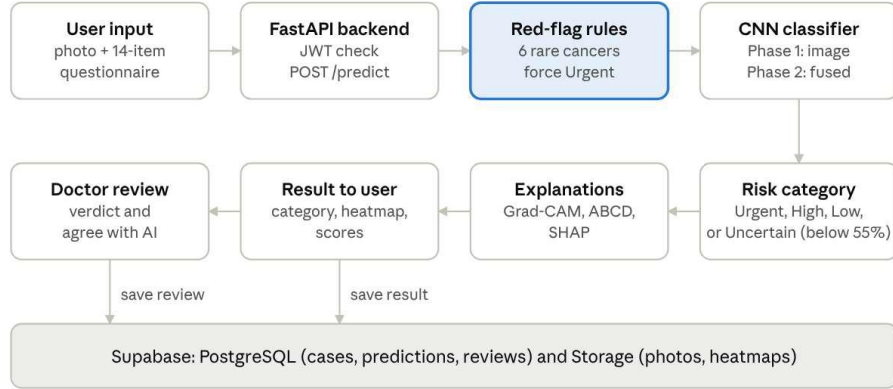


Fig. 1 System Architecture

#### IV. EQUATIONS

The task is modelled as multi-class lesion classification with a clinical side input and a rule-based safety override.

**Image and clinical embeddings.** An image  $x$  is mapped by the EfficientNet-B3 backbone to a 512-dimensional embedding. The 14 clinical features  $c$  are mapped by a small network to a 32-dimensional embedding.

$$z_{img} = f_{\theta}(x) \in \mathbb{R}^{512}, z_{clin} = g_{\phi}(c) \in \mathbb{R}^{32}, c \in \mathbb{R}^{14}$$

**Fusion and classification.** The embeddings are concatenated (544 dimensions), passed through a 256-unit layer and classified into  $K = 6$  classes by a softmax.

$$z = [z_{img}; z_{clin}] \in \mathbb{R}^{544}, h = \text{ReLU}(W_1 z + b_1) \in \mathbb{R}^{256}, P(y = k | x, c) = \text{softmax}(W_2 h + b_2)_k$$

**Loss function.** Training uses focal loss with label smoothing, where  $p_k$  is the predicted probability of class  $k$ ,  $\gamma = 2$  and  $\epsilon$  is the smoothing factor.

$$\tilde{y}_k = (1 - \epsilon) y_k + \frac{\epsilon}{K}, \mathcal{L} = - \sum_{k=1}^K \tilde{y}_k (1 - p_k)^{\gamma} \log p_k$$

**Decision rule.** Let  $R(c) = 1$  when any of the six red-flag rules fires. The final category combines the network output, its confidence and the rule override.

$$\text{category} = \begin{cases} \text{Urgent Referral} & \text{if } R(c) = 1 \\ \text{Uncertain} & \text{if } \max_k p_k < 0.55 \\ \text{risk}(\arg\max_k p_k) & \text{otherwise} \end{cases}$$

**Total Dermoscopy Score.** A, B, C and D are the asymmetry, border, colour and diameter scores.

$$TDS = 1.3A + 0.1B + 0.5C + 0.5D$$

**Grad-CAM.** For class  $c$  and feature map  $A^k$ , the channel weight is the spatially averaged gradient, and the heatmap is the rectified weighted sum of maps.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

**SHAP.** The attribution of clinical feature  $i$  is its average marginal contribution over all subsets  $S$  of the feature set  $F$ .

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$



**Evaluation metrics.** Balanced accuracy and macro F1 average over classes. The fairness gap is the drop in malignant sensitivity between skin-tone groups.

$$\text{BalAcc} = \frac{1}{K} \sum_{k=1}^K \frac{TP_k}{TP_k + FN_k}, \quad \text{Macro-}F_1 = \frac{1}{K} \sum_{k=1}^K F_{1,k}$$

$$\Delta = \text{Sens}_{\text{I-II}} - \text{Sens}_{\text{III-IV}} = 0.247$$

## V. CONCLUSION

This work shows that a smartphone-based skin cancer screening pipeline can combine a deep image classifier, clinical questionnaire data, rule-based safety checks and layered explanations in one deployed system. The EfficientNet-B3 classifier reached  $0.647 \pm 0.022$  balanced accuracy and  $0.579 \pm 0.020$  macro F1 on 8 ISIC 2019 classes. The fusion model did not improve on the image-only model on PAD-UFES-20 (macro F1 0.689 against 0.694, balanced accuracy 0.704 against 0.736), so clinical data adds explanation rather than accuracy at this dataset size. Red-flag rules extend coverage to six rare cancers that have no public image data.

The fairness audit is the main limitation. Malignant sensitivity falls by 0.247 from light to medium skin tones, and no malignant case in the DermaCon-IN neoplasm subset (52 images, Fitzpatrick III-V only) was detected at the same threshold. The system is therefore a decision-support and referral aid, not a diagnostic tool, and it has not been clinically validated.

**Future work.** Planned steps are: (1) retraining or fine-tuning with more images of darker skin and re-testing per skin tone; (2) completing the remaining 11 backend endpoints and the admin panel; (3) building the Flutter mobile app for patients and doctors; (4) prospective evaluation with dermatologists on real smartphone photographs.

## ACKNOWLEDGMENT

The authors thank their project guide and faculty mentors for their guidance, and Sandip Institute of Technology and Research Centre, Nashik for the computing resources and a supportive environment. We also thank the maintainers of the ISIC archive, PAD-UFES-20, the Stanford DDI and DermaCon-IN datasets, and the open-source tools PyTorch, pytorch-grad-cam, SHAP, FastAPI and Supabase that made this work possible.

## REFERENCES

- [1] M. S. Islam and S. Panta, "Skin cancer images classification using transfer learning techniques," *arXiv preprint arXiv:2406.12954*, Jun. 2024.
- [2] R. Daneshjou et al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science Advances*, vol. 8, no. 32, art. no. eabq6147, 2022.
- [3] B. Krohling, P. B. C. Castro, A. G. C. Pacheco, and R. A. Krohling, "A smartphone based application for skin cancer classification using deep learning with clinical images and lesion information," *arXiv preprint arXiv:2104.14353*, Apr. 2021.
- [4] J. Steppan and S. Hanke, "Analysis of skin lesion images with deep learning," *arXiv preprint arXiv:2101.03814*, Jan. 2021.
- [5] K. Nawaz et al., "Skin cancer detection using dermoscopic images with convolutional neural network," *Scientific Reports*, vol. 15, art. no. 7252, 2025.
- [6] A. El Mertahi, H. Ezzine, S. Douzi, and K. Douzi, "AI-driven skin cancer detection from smartphone images: A hybrid model using ViT, adaptive thresholding, black-hat transformation, and XGBoost," *PLOS ONE*, vol. 20, no. 7, art. no. e0328402, Jul. 2025.
- [7] J. Deng, E. Guo, H. J. Zhao, K. Venugopal, and M. Moskalyk, "Development of a transfer learning-based, multimodal neural network for identifying malignant dermatological lesions from smartphone images," *Cancer Informatics*, vol. 23, pp. 1-15, 2025.



- [8] A. S. Al-Waisy, S. Al-Fahdawi, M. I. Khalaf, M. A. Mohammed, B. Al-Attar, and M. N. Al-Andoli, "A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images," *Scientific Reports*, vol. 15, art. no. 31234, 2025.
- [9] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, art. no. 180161, 2018.
- [10] N. C. F. Codella et al., "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 preprint arXiv:1710.05006*, 2017.
- [11] M. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.
- [12] A. G. C. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, art. no. 106221, 2020

