Enabling safe AI deployment: an automated Fitzpatrick skin type guardrail for out-of-distribution dermatology

Pedro H. G. Bouzon[1], **Eduarda P. Magesk**[1], **Luis A. de Souza Jr.**[1]
André G. C. Pacheco[1]

[1] Graduate Program in Informatics – Federal University of Espírito Santo (UFES) – 29075-910 – Vitória – ES – Brazil

{pedro.bouzon, eduarda.magesk}@edu.ufes.br

{la.souza, apacheco}@inf.ufes.br

Abstract. *Dermatological AI models frequently fail on darker skin tones, posing significant safety and regulatory risks. To prevent unsafe out-of-distribution inference, we propose an interpretable Fitzpatrick Skin Type (FST) classifier that acts as an active algorithmic guardrail. Our Prototype Matching framework classifies clinical images into FST I-IV or V-VI. To ensure robustness, we employ SAM3 pre-segmentation to isolate healthy skin prior to feature extraction. This strictly prevents shortcut learning caused by clinical artifacts and lesion morphology. Evaluated on the Diverse Dermatology Images dataset, our method achieved a Cohen's Kappa of 0.78, Balanced Accuracy of 0.88, and Macro F1-Score of 0.89, substantially outperforming established baselines. By retrieving the nearest prototypes for visual verification, our framework delivers the interpretability required for safe, compliant medical AI deployment.*

Resumo. *Modelos dermatológicos de IA falham frequentemente em tons de pele mais escuros, representando significativos riscos regulatórios e de segurança. Para evitar inferências inseguras fora da distribuição de treinamento (out-of-distribution), propomos um classificador interpretável da Escala de Fitzpatrick (FST) que atua como um mecanismo de proteção algorítmico ativo. Nossa abordagem de Correspondência de Protótipos (Prototype Matching) classifica imagens clínicas em FST I-IV ou V-VI. Para garantir a robustez, empregamos a pré-segmentação com o modelo SAM3 para isolar a pele saudável antes da extração de características. Isso evita rigorosamente o aprendizado por atalhos (shortcut learning) causado por artefatos clínicos e pela morfologia da lesão. Avaliado no conjunto de dados Diverse Dermatology Images (DDI), nosso método alcançou um Kappa de Cohen de 0,78, Acurácia Balanceada de 0,88 e F1-Score Macro de 0,89, superando substancialmente os métodos de referência estabelecidos. Ao recuperar os protótipos mais próximos para verificação visual, nosso framework fornece a interpretabilidade necessária para a implantação segura e em conformidade de IA na área médica.*

1. Introduction

Skin cancer is the most common type of cancer in the world with 1.67 million cases estimated in 2025[IARC 2025]. In tropical countries, such as Brazil, 263,280 new skin cancer cases are expected from 2026 to 2028 [INCA 2026]. To alleviate this burden,

many Computed Aided Diagnostic (CAD) systems have been proposed [Ray et al. 2024]. However, numerous commercial algorithms exhibit significant performance disparities across diverse skin tones. Daneshjou *et al.* [Daneshjou et al. 2022] showed a sensitivity drop from 69% on Fitzpatrick Skin Type (FST) I-II to 23% on FST V-VI. In a real world validation, Kamulegeya *et al.* [Kamulegeya et al. 2023] have shown top-1 accuracy of 17% on dark skin populations in Uganda. Similarly, Barros *et al.* [Barros et al. 2023] evaluated both supervised and self-supervised models fine-tuned on predominantly white skin datasets, demonstrating severe performance degradation when classifying melanoma on out-of-distribution datasets featuring darker skin tones and acral lesions.

To address the risks of out-of-distribution (OOD) inference, regulatory bodies emphasize strict scope enforcement and proactive risk mitigation. In a recent draft, the U.S. Food and Drug Administration (FDA) outlined expectations for AI-enabled devices to be evaluated across specific demographic subgroups, noting that poor performance in unrepresented groups restricts the safe operational scope of the device [U.S. Food and Drug Administration 2025]. In Brazil, ANVISA's RDC No. 657 (2022), for Software as a Medical Device (SaMD), mandates the monitoring of algorithm behavior and authorizes the suspension if performance irregularities are found [Brasil. Agência Nacional de Vigilância Sanitária (ANVISA) 2022]. However, current manufacturers rely on user manuals to indicate the algorithm‘s limitations, which is prone to human error, especially in decentralized point-of-care settings where operators may lack specialized technical support. Therefore, we propose an automated Fitzpatrick Skin Type classifier, based on Prototype Matching, to serve as a guardrail for out-of-distribution inference. The main contributions of this work are:

- An interpretable, prototype-matching-based Fitzpatrick Skin Type classifier that serves as an active algorithmic guardrail, automatically flagging out-of-distribution skin tones to prevent unsafe CAD usage and support regulatory compliance.
- A robust feature extraction pipeline that utilizes SAM3 pre-segmentation to isolate genuine skin tone, successfully mitigating the shortcut learning caused by spurious clinical artifacts.
- An open-source implementation of the proposed framework, available at: `https://github.com/life-ufes/skin-type-guardrail`.

2. Related Works

2.1. Skin Cancer Classification

Computer-Aided Diagnostic (CAD) systems for skin cancer have evolved from traditional machine learning [Celebi et al. 2007] to Convolutional Neural Networks (CNNs) achieving dermatologist-level accuracy [Esteva et al. 2017]. Recently, the field has shifted toward multimodal frameworks that integrate clinical metadata with images. Attention-driven fusion strategies have demonstrated particular effectiveness in this domain, enabling robust performance even with missing patient information and successfully outperforming human experts [Pacheco and Krohling 2021, Xu et al. 2025, Bouzon et al. 2025].

2.2. Fitzpatrick Skin Type Classification

Over the past years, a few works have been proposed to automatically identify the Fitzpatrick Skin Type (FST). Groh *et al.* [Groh et al. 2021] employed Individual Typology Angles (ITA) to estimate the FST on the Fitzpatrick17K dataset. Their method achieved an off-by-one accuracy of 70.38%. Similarly, Kinyanjui *et al.* [Kinyanjui et al. 2019] utilized segmentation models to isolate non-diseased skin regions and applied the ITA metric derived from the CIELab color space to categorize skin tones into discrete bins. To directly estimate FST, Benčević *et al.* [Benčević et al. 2024] trained a VGG-16 neural network to classify clinical skin images into three aggregated Fitzpatrick categories (I-II, III-IV, and V-VI). Hamrani *et al.* [Hamrani et al. 2024] proposed an unsupervised method, named AIDA, to identify FST in images, achieving an accuracy of 97%. However, their method was evaluated on a small dataset with only 48 subjects, which limits the generalizability of their results. Srimaharaj *et al.* [Srimaharaj and Chaising 2024] proposed a method that combines distance measurements with the Fuzzy Analytic Hierarchy Process (AHP). Evaluated on 1,022 clinical images, their method achieved an accuracy of 93%, precision of 80%, and specificity of 96%. Nevertheless, none of these methods provide interpretable results, which are essential for the acceptance of clinical tools. Therefore, we aim to fill this gap by classifying the skin type with prototype matching.

3. Methods

In this section, we provide an explanation of the dataset, evaluation methodology, and the proposed method, which consists of healthy skin segmentation, embeddings extraction and prototype matching classification.

3.1. Dataset

To evaluate the proposed method, we employed the Diverse Dermatology Images (DDI) dataset [Daneshjou et al. 2022]. This dataset is composed of 656 clinical images labeled into three Fitzpatrick Skin Type (FST) groups: FST I-II (31.7%), FST III-IV (36.7%), and FST V–VI (31.6%). All images were retrospectively collected from Stanford Clinics between 2010 and 2020. To provide a fair comparison between light (FST I-II) and dark (FST V-VI) skin types, the researchers selected images from matching patients across each group, based on diagnostic, age, sex, and date of photography. Therefore, this dataset represents a suitable benchmark for unbiased estimation of Fitzpatrick Skin Type in skin lesion images. For our binary guardrail task, we aggregated the FST I-II and III-IV categories into a single FST I-IV class to be evaluated against FST V-VI.

3.2. Healthy Skin Segmentation

We hypothesize that the skin lesion and background might contribute negatively to the skin type estimation. Therefore, we devised a method to extract healthy-skin patches from raw images. Initially, the Segment-Anything-Model 3 (SAM3) [Carion et al. 2025] model was employed to segment skin and lesions sequentially, using the prompts "skin" and "mole", respectively. Then a logical operation was performed to extract only the healthy skin mask by subtracting the two masks, as seen in Fig. 1A.

Subsequently, as illustrated in Fig. 1B, we employed a binary search to determine the largest square patch that fits entirely within the healthy skin mask, evaluating sizes

between 32×32 and 256×256 pixels via morphological erosion. Once the maximum valid size was established, we extracted the patch whose coordinates minimized the squared Euclidean distance to the center of the bottom-left quadrant. This specific region was targeted to maximize the distance from the primary lesion, which is typically centered in dermatological photographs. If no valid 32×32 patch could be found, we utilized a fallback patch that maximized the intersection with the healthy mask. All resulting patches were resized to 224×224 pixels prior to embedding extraction.

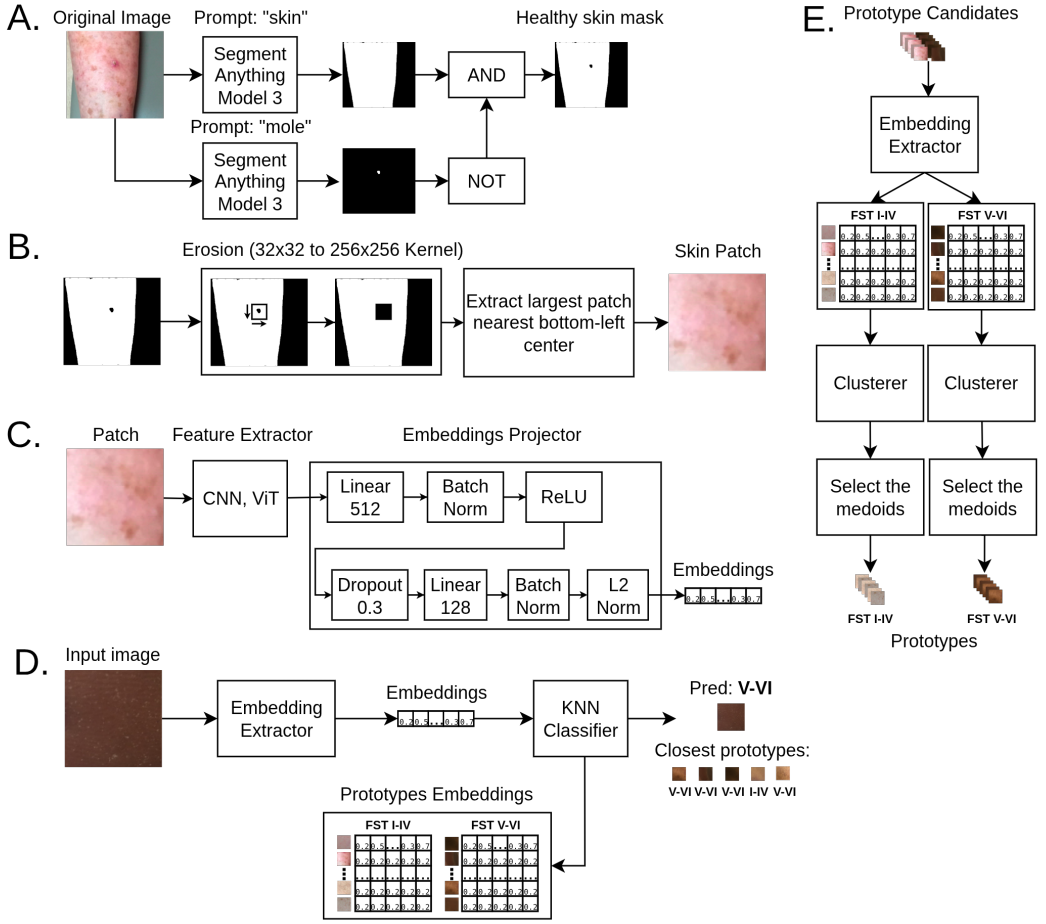

Figure 1. Overview of the proposed Fitzpatrick Skin Type classification framework. (A) SAM3 segmentation using "skin" and "mole" prompts to isolate healthy skin masks. (B) Extraction of the largest valid healthy skin patch nearest the bottom-left quadrant to avoid the central lesion. (C) Feature extraction architecture optimized via ArcFace loss to produce a 128-dimensional skin-tone embedding space. (D) Real-time FST classification via K-Nearest Neighbors matching against learned prototypes. (E) Offline prototype generation by clustering training set embeddings and selecting the medoids of each class.

3.3. Embeddings Extraction

As a first step towards classification, we employed a deep learning model to extract embeddings from the skin patches. This model, depicted in Fig. 1C, is defined by two parts: an image feature extractor and an embeddings projector. To extract features from images, we evaluated four competitive models: MobileNetV3 [Howard et al. 2019],

ResNet-50 [He et al. 2015], EfficientNetV2-S [Tan and Le 2021], and DinoV3 (ViT-S) [Siméoni et al. 2025]. The embedding projector consists of a combination of Linear layers, Batch Normalization [Ioffe and Szegedy 2015], ReLU [Agarap 2019], Dropout [Srivastava et al. 2014], and L2 normalization, mapping the features to a unit hypersphere as required by margin-based softmax losses [Deng et al. 2019].

To optimize the network, we employed the Additive Angular Margin Loss (ArcFace) [Deng et al. 2019]. Unlike traditional metric learning approaches that rely on sampling pairs or triplets, ArcFace directly enhances the discriminative power of the features by simultaneously maximizing intra-class compactness and inter-class discrepancy. It achieves this by adding an angular margin penalty m to the angle between the deep features and their corresponding ground-truth class weights. Equation 1 describes the ArcFace loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{C}e^{s\cos\theta_j}} \tag{1}$$

where N is the batch size, C is the total number of classes, θ_{y_i} is the angle between the i-th L2-normalized feature embedding and the normalized weight of the ground-truth class y_i, s is the feature scale parameter, and m is the additive angular margin. Following the original implementation [Deng et al. 2019], we fixed the scale parameter at $s = 64$. Furthermore, to rigorously enforce decision boundaries between different Fitzpatrick Skin Types in the embedding space, we experimented with angular margin m values of 28.6° (equivalent to 0.5 radians, the standard default), 45°, and 90°.

3.4. Prototype Matching Classification

To classify the Fitzpatrick Skin Type, we used prototype matching. This task consists of comparing an image to a set of references (prototypes), which are representatives of each class. The FST was then estimated as the most common class in the n_p-nearest prototypes. Figure 1D illustrates the classification process. In this work we used 20 prototypes for each class and $n_p = 5$. To obtain the prototypes, we used a clustering algorithm, within each class, to create the 20 groups, as depicted in Fig. 1E. Finally, the prototypes were defined as the images closer to the center of each cluster (medoids). The following clustering algorithms were evaluated: Agglomerative Clustering [Jr. 1963] and K-Means [MacQueen 1967].

3.5. Experiments Methodology

In order to evaluate the model, the DDI dataset was randomly split into train and test sets with a 80/20 ratio. We then performed 5-Fold stratified cross-validation on the training set to compare different hyperparameters and clustering techniques. All models were trained using Pytorch, with weights sourced from the TIMM API, pre-trained on the ImageNet-V1, except for the DinoV3 (ViT-S) which was pre-trained on 1,689 million images of the private LVD-1689M dataset. The models were fine-tuned for 50 epochs, with a learning rate of 0.0001, batch size of 24, and using the AdamW optimizer with 0.01 weight decay. To avoid over-fitting, early stopping was triggered if no decrement on validation loss was observed for 10 consecutive epochs, and the learning rate was reduced by a factor of 0.1

if no improvement was observed for 5 epochs. On the test set evaluation, we employed a soft-voting ensemble, composed of the five models trained during cross-validation.

We utilized four metrics for model assessment: Cohen's Kappa (Kappa), Balanced Accuracy (BACC), Mean Absolute Error (MAE), and Macro F1-Score (F1). Cohen's Kappa measures the agreement between raters and is widely used in the medical literature [McHugh 2012]. The Balanced Accuracy is defined as the macro-averaged sensitivity. The Mean Absolute Error quantify the average numerical distance of the predictions and the ground-truth skin type. Finally, the F1-Score is the harmonic average of the sensitivity and precision. To statistically evaluate the experimental results, we first employed the non-parametric Friedman test to detect overall performance differences among the evaluated configurations. Upon establishing statistical significance, the best-performing configuration was designated as the control. Post-hoc comparisons were then conducted exclusively between this control and the remaining methods using the Wilcoxon signed-rank test [Derrac et al. 2011]. To strictly control the family-wise error rate during these specific control-versus-others comparisons, the Bonferroni correction was applied, adjusting the significance threshold to maintain an overall $\alpha = 0.05$. The F1-Score was utilized as the base metric for all statistical evaluations. Furthermore, when comparing various hyperparameters and clustering algorithms, metrics were aggregated across all vision backbones to ensure the statistical findings remained independent of any specific architecture.

4. Results and Discussion

4.1. Cross-Validation and Hyperparameter Analysis

To determine the optimal architecture and embedding space configuration, we evaluated four vision backbones, two clustering algorithms (Agglomerative and K-Means), and three ArcFace angular margins ($m \in \{28.6°, 45°, 90°\}$) using 5-fold cross-validation on the training set.

As shown in Table 1, EfficientNetV2-S consistently outperformed other backbones across all metrics, achieving a peak Macro F1 Score of 0.85 ± 0.04 and a Balanced Accuracy of 0.85 ± 0.04 at $m = 28.6°$. MobileNetV3 and DinoV3 (ViT-S) demonstrated competitive but slightly lower performance, while ResNet-50 struggled, particularly at higher margin penalties.

The angular margin (m) of the ArcFace loss played a decisive role in shaping the embedding space. To rigorously evaluate this, we aggregated the metrics across all vision backbones ($n = 20$ samples per configuration) and designated the $m = 28.6°$ margin with K-Means clustering as our statistical control. Increasing the angular margin to $90°$ significantly degraded model performance across the Macro F1 Score ($p = 0.015$), Balanced Accuracy ($p = 0.005$), and Cohen's Kappa ($p = 0.021$). Furthermore, the $90°$ margin with K-Means resulted in a significantly worse Mean Absolute Error compared to the $28.6°$ control ($p = 0.042$). While the moderate $45°$ margin showed no significant difference in F1 Score or Kappa, it yielded a statistically significant drop in Balanced Accuracy when paired with Agglomerative clustering ($p = 0.045$). This indicates that while a moderate angular penalty helps enforce decision boundaries, excessively large margins disrupt the feature learning process.

Loss margin (m) Clusterer	28.6° **Agglomerative**	28.6° **K-Means**	45° **Agglomerative**	45° **K-Means**	90° **Agglomerative**	90° **K-Means**
Cohen's Kappa						
MobileNetV3	0.65 ± 0.04	0.66 ± 0.07	0.60 ± 0.10	0.64 ± 0.04	0.62 ± 0.07	0.66 ± 0.04
ResNet-50	0.60 ± 0.10	0.55 ± 0.14	0.58 ± 0.09	0.57 ± 0.07	0.48 ± 0.13	0.44 ± 0.17
EfficientNetV2-S	0.70 ± 0.08	0.71 ± 0.07	0.66 ± 0.07	0.66 ± 0.12	0.61 ± 0.03	0.60 ± 0.11
DinoV3 (ViT-S)	0.61 ± 0.08	0.66 ± 0.08	0.60 ± 0.05	0.68 ± 0.15	0.58 ± 0.11	0.56 ± 0.06
Balanced Accuracy						
MobileNetV3	0.83 ± 0.02	0.84 ± 0.03	0.81 ± 0.05	0.84 ± 0.02	0.81 ± 0.05	0.83 ± 0.02
ResNet-50	0.83 ± 0.05	0.80 ± 0.09	0.82 ± 0.03	0.82 ± 0.03	0.76 ± 0.07	0.74 ± 0.09
EfficientNetV2-S	0.85 ± 0.04	0.85 ± 0.04	0.83 ± 0.04	0.83 ± 0.07	0.79 ± 0.02	0.78 ± 0.05
DinoV3 (ViT-S)	0.82 ± 0.04	0.84 ± 0.04	0.81 ± 0.02	0.85 ± 0.07	0.78 ± 0.07	0.78 ± 0.03
Mean Absolute Error						
MobileNetV3	0.15 ± 0.02	0.15 ± 0.03	0.17 ± 0.04	0.16 ± 0.03	0.16 ± 0.02	0.15 ± 0.02
ResNet-50	0.19 ± 0.05	0.20 ± 0.05	0.20 ± 0.05	0.21 ± 0.03	0.24 ± 0.07	0.25 ± 0.07
EfficientNetV2-S	0.13 ± 0.03	0.12 ± 0.02	0.15 ± 0.03	0.14 ± 0.04	0.16 ± 0.01	0.16 ± 0.04
DinoV3 (ViT-S)	0.18 ± 0.04	0.15 ± 0.04	0.18 ± 0.02	0.14 ± 0.06	0.17 ± 0.04	0.19 ± 0.03
Macro F1 Score						
MobileNetV3	0.82 ± 0.02	0.83 ± 0.03	0.80 ± 0.05	0.82 ± 0.02	0.81 ± 0.04	0.83 ± 0.02
ResNet-50	0.80 ± 0.05	0.77 ± 0.07	0.79 ± 0.05	0.78 ± 0.03	0.73 ± 0.07	0.71 ± 0.09
EfficientNetV2-S	0.85 ± 0.04	0.85 ± 0.03	0.83 ± 0.04	0.83 ± 0.06	0.80 ± 0.01	0.80 ± 0.05
DinoV3 (ViT-S)	0.80 ± 0.04	0.83 ± 0.04	0.80 ± 0.02	0.84 ± 0.07	0.79 ± 0.06	0.78 ± 0.03

Table 1. Comparison of the 5-fold cross-validation Cohen's Kappa, Balanced Accuracy, Mean Absolute Error, and Macro F1 Score for different clustering algorithms, ArcFace loss margins, and feature extractor backbones.

Regarding the prototype selection strategy, the tests revealed no statistically significant differences between Agglomerative and K-Means clustering at the optimal 28.6° margin across any metric (e.g., F1 Score $p = 0.996$, Balanced Accuracy $p = 0.628$, Cohen's Kappa $p = 0.996$). Given this statistical parity, we theoretically favored Agglomerative clustering because, unlike K-Means, it does not assume that clusters form equally sized, spherical blobs. This algorithmic flexibility is better suited for capturing the natural, unevenly distributed variance of skin tones in the embedding space. Coupled with its peak nominal performance on our top vision backbone, we selected EfficientNetV2-S with Agglomerative clustering and $m = 28.6^\circ$ as the optimal configuration for the final test set evaluation.

4.2. Test Set Evaluation and Comparison with the State-of-the-art

Table 2 compares our optimal Prototype Matching configuration against two established baselines: a deep learning approach utilizing VGG-16 [Benčević et al. 2024] and a heuristic approach based on Individual Typology Angle (ITA) [Kinyanjui et al. 2019]. Our method substantially outperformed both baselines across all metrics, yielding a Cohen's Kappa (CK) of 0.78, compared to 0.65 for VGG-16 and 0.23 for ITA.

Importantly, our method effectively diminished the algorithmic bias against darker skin tones observed in previous models. As illustrated by the confusion matrices in Figure 2, the VGG-16 baseline struggled significantly with darker skin, erroneously classifying 28.9% of FST V-VI images as light skin (I-IV). The ITA method failed fundamentally on lighter skin, operating near random chance (52.9% accuracy for FST I-IV). In contrast, our

Model	Kappa ↑	BACC ↑	MAE ↓	Macro F1 Score ↑
EfficientNetV2-S + Agglomerative (ours)	0.78 [0.64-0.88]	0.88 [0.81-0.94]	0.10 [0.05-0.15]	0.89 [0.82-0.94]
VGG-16 [Benčević et al. 2024]	0.65 [0.50-0.78]	0.82 [0.74-0.88]	0.15 [0.09-0.22]	0.83 [0.75-0.89]
ITA [Kinyanjui et al. 2019]	0.23 [0.08-0.37]	0.63 [0.55-0.71]	0.40 [0.32-0.45]	0.59 [0.51-0.67]

Table 2. Quantitative evaluation of FST classification performance on the test set. Our optimal cross-validation configuration employs an EfficientNetV2-S backbone trained on SAM3-segmented images using ArcFace loss ($m = 28.6°$). Performance is compared against established deep learning (VGG-16) and heuristic (Individual Typology Angle) baselines. Test predictions for our method and VGG-16 are generated using a soft-voting ensemble of the five models trained across cross-validation folds. Bracketed values indicate 95% confidence intervals estimated via bootstrapping ($N = 10,000$). Arrows indicate whether higher (↑) or lower (↓) values are optimal.

Prototype Matching framework maintained a robust 94.3% accuracy on FST I-IV while simultaneously improving the classification of FST V-VI to 82.2%. This demonstrates the viability of our framework as a reliable regulatory guardrail capable of identifying untested populations without sacrificing performance on majority groups.

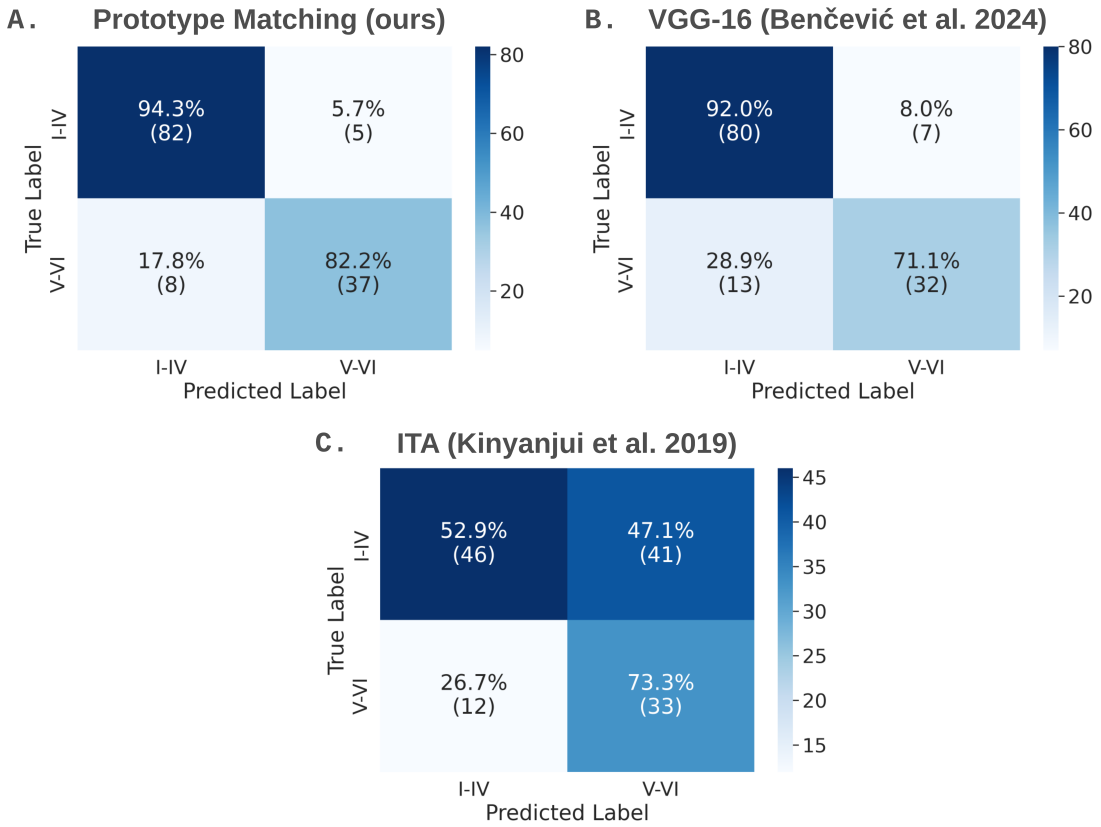

Figure 2. Confusion matrices for binary FST classification. (A) Our proposed Prototype Matching framework achieves the highest accuracy across both categories, successfully reducing the predictive bias toward lighter skin tones observed in (B) the VGG-16 and (C) ITA baselines.

4.3. The Role of Segmentation and Interpretability

Beyond raw accuracy, our approach offers essential interpretability for clinicians. By presenting the K-nearest prototypes alongside the prediction, the system allows physicians to visually verify the semantic reasoning behind the classification. In a clinical setting, this transparency fosters trust in the guardrail's decisions, enabling operators to understand why an image was flagged as out-of-distribution. Figure 3 demonstrates this capability while highlighting the necessity of our SAM3 pre-segmentation step.

When operating on raw, unsegmented data (Fig. 3 A and B), the embedding extractor fell victim to shortcut learning. Rather than clustering based on generalized skin tone, the model partitioned the embedding space based on spurious clinical artifacts (such as surgical ink and rulers) and localized morphological features (the moles themselves). Consequently, the model routinely matched queries to prototypes with similar anatomical backgrounds or clinical markings, regardless of the actual Fitzpatrick Skin Type. By enforcing healthy skin segmentation prior to embedding extraction (Fig. 3 C and D), we constrained the network to focus primarily on the semantic representation of skin color, thereby enhancing the robustness and potential clinical utility of the proposed pipeline as a regulatory guardrail.

5. Limitations

A constraint of our framework is its reliance on the Fitzpatrick Skin Type (FST) scale. Future work will transition to the Monk Skin Tone (MST) scale [Monk 2023], which better captures the variance of darker skin tones by categorizing visual appearance rather than physiological reactions to sunburn. Additionally, the computational overhead of SAM3 may limit deployment on resource-constrained devices. To mitigate this, future iterations will employ knowledge distillation to train lightweight segmentation networks using SAM3 masks as teacher guidance.

6. Conclusion

In this work, we proposed an automated, prototype-matching-based Fitzpatrick Skin Type guardrail designed to prevent dermatological AI diagnostic tools from performing unsafe inference on out-of-distribution skin tones. Evaluated across several competitive visual backbones and clustering algorithms, our optimal configuration achieved a Cohen's Kappa of 0.78, a Balanced Accuracy of 0.88, and a Macro F1-Score of 0.89 on the test set, substantially outperforming established baselines. By integrating SAM3 pre-segmentation into the pipeline, we effectively isolated genuine skin tone representations and mitigated the severe shortcut learning typically caused by spurious clinical artifacts. Moreover, the proposed method is highly interpretable, allowing physicians to visually examine the specific prototypes responsible for a given prediction. These results demonstrate the framework's viability for real-world usage, serving as an active algorithmic guardrail to enforce regulatory compliance and representing a vital step toward the safe, equitable clinical deployment of medical AI tools. To further refine this system, future works should explore the integration of clinical metadata, such as age and body region, and evaluate the framework using the more inclusive Monk Skin Tone (MST) scale.

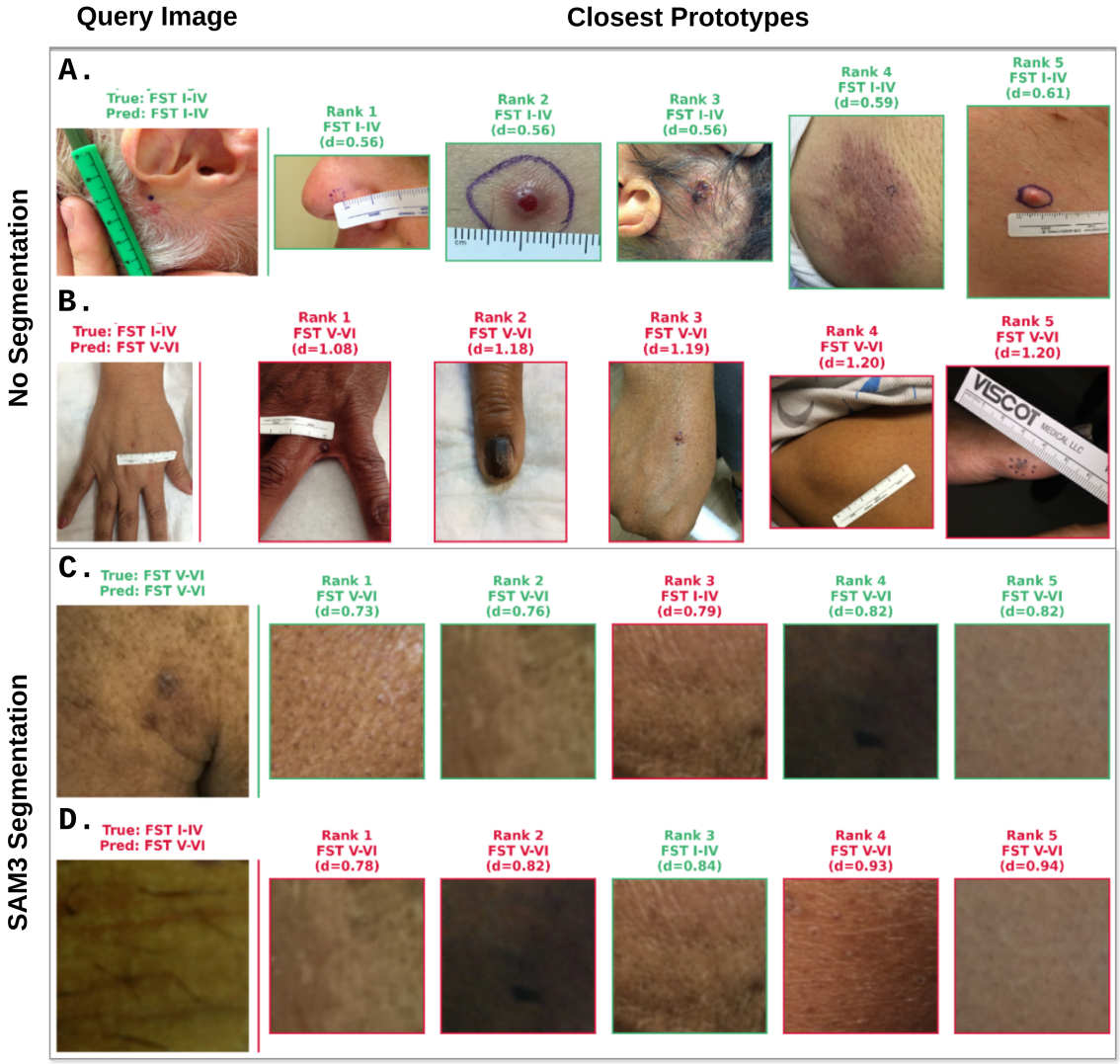

Figure 3. Prototype-based FST classification. Query images (left) and their five nearest prototypes (right). Green and red borders indicate correct and incorrect class matches, respectively. (A-B) Segmented data forces the model to rely on genuine skin tone. (C-D) Unsegmented data triggers shortcut learning, where the model erroneously matches spurious clinical artifacts (e.g., rulers, surgical ink) and localized anatomical features rather than the underlying skin type.

7. Acknowledgments

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

References

Agarap, A. F. (2019). Deep learning using rectified linear units (relu).

Barros, L., Chaves, L., and Avila, S. (2023). Assessing the generalizability of deep neural networks-based models for black skin lesions. In *Iberoamerican Congress on Pattern Recognition*, pages 1–14. Springer.

Benčević, M., Habijan, M., Galić, I., Babin, D., and Pižurica, A. (2024). Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044.

Bouzon, P. H. G., Rocha, W. F. d., Souza, L. A., and Pacheco, A. G. C. (2025). Metablock-se: A method to deal with missing metadata in multimodal skin cancer classification. *IEEE Journal of Biomedical and Health Informatics*, 29(12):8855–8862.

Brasil. Agência Nacional de Vigilância Sanitária (ANVISA) (2022). Resolução da diretoria colegiada - rdc nº 657, de 24 de março de 2022. Diário Oficial da União. Acesso em: 13 fev. 2026.

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., et al. (2025). Sam 3: Segment anything with concepts.

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., et al. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., et al. (2022). Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694.

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.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

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

Hamrani, A., Leizaola, D., Vedere, N. K. R., Kirsner, R. S., et al. (2024). AI dermatochroma analytica (AIDA): Smart technology for robust skin color classification and segmentation. *Cosmetics*, 11(6):218.

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

Howard, A., Sandler, M., Chu, G., Chen, L.-C., et al. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324.

IARC (2025). Skin cancer. International Agency for Research on Cancer. Accessed on: Jul. 30, 2025.

INCA (2026). Incidência do câncer no Brasil. Instituto Nacional do Câncer (INCA). Last accessed 12 Fev 2026.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, pages 448–456. JMLR.org.

Jr., J. H. W. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Kamulegeya, L., Bwanika, J., Okello, M., Rusoke, D., Nassiwa, F., et al. (2023). Using artificial intelligence on dermatology conditions in uganda: a case for diversity in training data sets for machine learning. *African Health Sciences*, 23(2):753–763.

Kinyanjui, N. M., Odonga, T., Cintas, C., Codella, N. C. F., et al. (2019). Estimating skin tone and effects on classification performance in dermatology datasets.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Monk, E. (2023). The monk skin tone scale. *SocArXiv*.

Pacheco, A. G. 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*, 25(9):3554–3563.

Ray, A., Sarkar, S., Schwenker, F., and Sarkar, R. (2024). Decoding skin cancer classification: perspectives, insights, and advances through researchers' lens. *Scientific Reports*, 14(1):30542.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., et al. (2025). Dinov3.

Srimaharaj, W. and Chaising, S. (2024). Distance-based integration method for human skin type identification. *Computers in Biology and Medicine*, 178:108575.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR.

U.S. Food and Drug Administration (2025). Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. Draft guidance for industry and food and drug administration staff, U.S. Department of Health and Human Services.

Xu, J., Huang, K., Zhong, L., Gao, Y., et al. (2025). Remixformer++: A multimodal transformer model for precision skin tumor differential diagnosis with memory-efficient attention. *IEEE Transactions on Medical Imaging*, 44(1):320–337.