DermFM: Evaluating Fairness and Generalizability in Skin Lesion Classification

Adnan Ahmed[*]
York University
Algoverse AI Research

Jeremy Lin
Algoverse AI Research

Pardhu Meduri
Algoverse AI Research

Ayush Prasanna
Algoverse AI Research

Sunishchal Dev
Algoverse AI Research

Kevin Zhu
University of California, Berkeley
Algoverse AI Research

Kiran Nijjer[†]
Stanford University
Algoverse AI Research

Abstract

Foundation models are reshaping medical AI by enabling efficient transfer learning from large, pretrained representations. In this work, we evaluate Google Health's Derm Foundation Model for skin lesion classification and fairness in dermatologic imaging. Using pre-encoded embeddings from PAD-UFES-20 and DERM12345, we trained lightweight classifiers for five major conditions: Actinic Keratosis, Basal Cell Carcinoma, Malignant Melanoma, Squamous Cell Carcinoma, and Seborrheic Keratosis. The model achieved high AUCs and consistent performance across sex, age, and lesion characteristics, demonstrating the strength of foundation-model representations for dermatology. However, fairness analysis revealed noticeable lower sensitivity for darker Fitzpatrick skin tones (4-6), indicating bias embedded within the pretrained feature space. Applying importance weighting and group-balanced resampling helped mitigate but did not fully eliminate these disparities. Our findings highlight the need for more diverse pretraining datasets and fairness-aware adaptation strategies to ensure equitable deployment of foundation models in clinical AI applications.

1 Introduction

Foundation models are large-scale pretrained AI models capable of performing diverse tasks, representing a major breakthrough in artificial intelligence [1]. Rather than training specialized models from the ground up, developers can leverage these versatile models as powerful feature extractors. Using pretrained embeddings from a foundation model enables new medical AI applications to be built with far less labeled data and compute. This convenience has fueled increased interest in using foundation models for healthcare tasks, where acquiring extensive labeled data remains difficult.[2].

Dermatology could benefit substantially from recent advances in artificial intelligence. Skin diseases are widespread globally, yet access to dermatologic care remains limited as over 3 billion people lack even basic services, particularly in underserved regions [3]. This shortage leads to delayed diagnoses and worse outcomes, underscoring the need for scalable solutions. Artificial intelligence has emerged

[*]First author.
[†]Senior author / Principal Investigator.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFm.

as a promising tool to bridge this gap. Deep learning models can analyze skin lesion images and assist in diagnosis, in some cases achieving accuracy on par with expert dermatologists [4].

However, there are pressing concerns about fairness and bias in AI in dermatology. If the training data of a model is not demographically representative, its performance can disproportionately favor certain groups of patients. Notably, many skin image classifiers have shown reduced accuracy on darker skin tones [5]. Similar biases could arise for other attributes such as demographics of the patient or presentation of the lesion [6, 7, 8]. Ensuring foundation models perform fairly across all patient groups is vital to prevent widening healthcare disparities. Evaluating their accuracy on diverse skin tones and demographics is a key step toward safe, equitable clinical use.

In this work, we leverage Google's recently released Derm Foundation model to investigate both its effectiveness and fairness in skin lesion classification [9]. We use Derm Foundation model to extract image embeddings and then evaluate the performance of our best model in subgroups stratified by Fitzpatrick skin type, demographic factors, and clinical features of the injury. This analysis allows us to evaluate the general precision of the foundation model in a challenging multiclass dermatological task, while also examining potential biases in its predictions in different patient populations. The findings show how well the model may work in real-world dermatology and where fairness improvements are needed.

2 Related Works

In healthcare, foundation models have been proposed to integrate multiple data types (imaging, clinical notes) and support diverse applications such as diagnosis and treatment planning [10]. Collectively, these works highlight the promise of foundation models for medical imaging, while also noting persistent challenges around bias and the need for domain-specific validation. For example, research has envisioned "generalist medical AI" built through self-supervised training on clinical data [11].

Within dermatology, several recent works have developed specialty foundation models for skin image analysis. Yan et al. [4] Notably, PanDerm, is a multimodal dermatology foundation model pretrained on over 2 million real-world images. PanDerm achieved state-of-the-art performance on 28 dermatology tasks, often outperforming task-specific baselines using only 10% of labeled data. Similarly, Xu et al. [12] introduced DermNIO, a model trained using hybrid semi- and self-supervised pretraining on appoximately 433,000 dermatology images. DermNIO consistently outperformed prior models across diverse tasks, including malignancy classification and segmentation, and showed robustness across diverse skin types and sexes. In parallel, Google's Derm Foundation model offers pretrained skin image embeddings, allowing researchers to develop accurate dermatology classifiers using limited data and compute [9].

Despite these performance gains, bias and fairness remain critical challenges in dermatology AI. Notably, Daneshjou et al. [13] curated the Diverse Dermatology Images (DDI) dataset and found that state-of-the-art skin lesion classifiers exhibited substantial performance drops (27–36% lower ROC-AUC) on images of dark skin tones and uncommon diseases compared to standard test results. Their analysis revealed that all evaluated models underperformed on darker skin tones. Likewise, Benčević et al. [5] demonstrated that lesion segmentation networks systematically under-segment lesions on darker skin, indicating a pronounced association between skin tone and model accuracy. Conventional bias-mitigation strategies achieved only marginal improvements, highlighting the persistent disparities among higher Fitzpatrick skin types resulting from imbalanced training data.

More generally, evaluations of large pretrained medical imaging models have exposed subgroup performance disparities. For instance, Khan et al. [14] conducted a systematic fairness audit of six medical imaging foundation models and observed that models pretrained on medical images, as opposed to general images, achieved higher overall accuracy but worse subgroup fairness, disproportionately favoring majority racial groups (White, Asian) and underperforming on female patients. This highlghts that scaling with large datasets alone cannot ensure equity. According to Queiroz et al. [15], achieving fairness in foundation models necessaitates systematic interventions throguhout the development pipeline, encompassing data collection, training and deployment.

3 Methods

3.1 Derm Foundation Model

We leverage the Derm Foundation model developed by Google Health to extract domain-specific image embeddings for our dermatologic analysis. The model is built upon a BiT ResNet101x3 architecture and trained using a two-stage process that combines large-scale contrastive pretraining on paired image–text data with supervised fine-tuning on clinical teledermatology datasets [16]. This approach enables the model to encode high-level visual representations of dermatologic features such as lesion morphology. The resulting embeddings serve as feature vectors that facilitate data-efficient training of downstream classifiers for disease categorization.

3.2 Datasets

We utilized two publicly available dermatology datasets, PAD-UFES-20 and DERM12345, for our classification and bias experiments. The PAD-UFES-20 dataset consists of 2,298 clinical images of skin lesions collected from 1,373 patients and includes six main diagnostic classes [17]. Each image is accompanied by detailed clinical metadata such as Fitzpatrick skin tone, patient demographics, and presentation characteristics including bleeding and lesion size. The DERM12345 dataset contains 12,345 dermatological images sourced from multiple clinical centers and annotated within a structured hierarchy of five superclasses, fifteen main classes, and forty subclasses of skin lesions [18]. The pre-encoded embeddings for both datasets are publicly available. For this study, we filter both datasets to include only overlapping diagnostic categories.

3.3 Classification Experiments

We conducted multi-class classification experiments for Actinic Keratosis (ACK), Basal Cell Carcinoma (BCC), Malignant Melanoma (MEL), Squamous Cell Carcinoma (SCC), and Seborrheic Keratosis (SEK). Multiple machine learning classifiers were evaluated on the pre-encoded Derm Foundation embeddings, including both linear and non-linear models. To prevent data leakage, we performed a patient-level split, allocating 70% of the data for training, 15% for validation, and 15% for testing. Model performance was assessed using the macro-average Area Under the ROC Curve (AUC) as well as per-disease AUC values, each reported with 95% confidence intervals estimated through bootstrapping. We report Macro AUC of our models and per-disease AUC of our top performing model. We trained and tested models on PAD-UFES-20, DERM12345 individually, and their combined dataset to evaluate domain-specific performance and cross-domain generalization.

3.4 Bias Experiments

To assess model bias, we evaluated fairness for our top performing model across several predefined demographic and clinical subgroups provided within PAD-UFES-20's metadata. The same 70/15/15 patient-level split was applied; however, the test set was drawn exclusively from the PAD-UFES-20 dataset, as it is the only one that includes detailed metadata. Model performance was stratified by Fitzpatrick skin tone rating, demographic variables and clinical presentation. Based on the distribution of available samples, we defined four groups: Type 1 Type 2, Type 3, and an aggregated Type 4-6 group, since darker skin tones were underrepresented in the dataset. Demographic stratification included patient age ($<$ 55, 55-64, 65-74, and 75+) and sex (male and female).

Finally, we analyzed clinical presentation variables using PAD-UFES-20 metadata fields describing lesion characteristics. The attributes were hurt, bleed, elevation, and lesion size. Lesion area was estimated from the recorded horizontal and vertical diameters, and each sample was assigned to one of three groups: small ($<$40 mm^2), medium (40–110 mm^2), or large ($>$ 110 mm^2). These variables were selected because they can directly affect how lesions appear in images and therefore may influence model behavior. For instance, bleeding can obscure lesion boundaries, elevated lesions can affect lighting and shadowing, and larger lesions may exhibit greater internal variability that can challenge model consistency.

3.5 Fairness Metrics

To assess fairness, we used the Fairlearn framework with custom bootstrap resampling for confidence estimation. We report three main fairness metrics. The True Positive Rate (TPR) Disparity quantifies differences in sensitivity across subgroups. The Equalized Odds (EO) gap measures disparity across Fitzpatrick skin tone groups by capturing the largest difference in true and false positive rates between any two groups for each disease, summarizing the overall balance of prediction errors. Finally, the Underdiagnosis Rate measures how often the model fails to predict any condition among patients who truly have a positive diagnosis; to summarize this disparity, we compute the max–min gap, representing the largest observed difference in these rates between any two groups for each disease class. For all three metrics, we calculate 95% confidence intervals using bootstrap resampling.

To ensure consistent evaluation across diseases, we use disease-specific thresholds optimized to maximize the F1 score on the validation set, treating precision and recall equally. This procedure aligns with prior fairness studies in medical AI [19, 20], ensuring balanced decision boundaries that fairly reflect both false negatives and false positives in downstream fairness analysis.

3.6 Bias Mitigation Strategies

The Fitzpatrick skin tone analysis revealed the largest per- formance disparities, whereas demographic and clinical variables showed minimal or inconsistent effects. Accordingly, subsequent bias mitigation efforts focused on skin tone, the primary source of in- equity in model predictions.

The first mitigation strategy importance weighting, assigned each training sample a weight inversely proportional to the joint frequency of its Fitzpatrick group and disease label. This weighting increased the influence of underrepresented group disease combinations, allowing darker skin tones and less common diagnoses to contribute more effectively during optimization. The importance weighting scheme was integrated into our best-performing model by incorporating sample-specific weights during training, enabling balanced learning across the full dataset wile compensating for group imbalance.

The second strategy, group-balanced resampling, aimed to equalize the distribution of training examples across Fitzpatrick groups. Minority group–disease pairs were oversampled until an approximately balanced representation was achieved, while perserving proportionality among disease classes. Unlike importance weighting, which adjusts each sample's contributiong, resampling modifies the training data composition to increase model exposure to underrepredented skin tones.

Fairness outcomes were assessed using the same performance and equity metrics described earlier. To quanify the effect of each intervention, we reported changes in these fairness metrics relative to the unmitigated baseline model.

Collectively these experiments, sought to identify which mitigation strategy most effectively reduced disparities in diagnostic sensitivity and underdiagnosis across skin tone groups, while maintaining overall model performance.

4 Results

4.1 Disease classification performance using Derm Foundation embeddings

As shown in Figure 4, the random forest classifier achieved the highest overall performance, with a macro-averaged AUC of 0.94 (95% CI 0.91–0.96). As reported in Table 1, models trained on the aggregrated dataset outperformed those trained on individual sources across most disease classes, indicating that dataset integration enhanced generalization and mitigated dataset-specific bias. Notably, SEK maintained near perfect classification accuracy under all configurations, whereas classes such as SCC and BCC showed substantial gains from data aggregation. These findings suggest that the Derm Foundation embeddings encode transferable visual representations that generalize well across datasets, supporting robust and consistent skin lesion classification.

Table 1: Classification performance of our best model measured by ROC–AUC (95% CI). Highest per-disease AUC and best Macro AUC are highlighted in light orange.

Disease	PAD-UFES-20	DERM12345	Aggregated
ACK	0.93 (0.89–0.97)	0.93 (0.85–0.99)	0.95 (0.92–0.97)
BCC	0.90 (0.84–0.94)	0.96 (0.91–0.99)	0.91 (0.87–0.95)
SEK	0.99 (0.98–1.00)	0.99 (0.96–1.00)	0.99 (0.98–1.00)
SCC	0.86 (0.77–0.93)	0.86 (0.70–0.97)	0.90 (0.83–0.95)
MEL	0.92 (0.81–1.00)	0.84 (0.68–0.99)	0.94 (0.88–0.99)
Macro AUC	0.92 (0.89–0.95)	0.92 (0.85–0.97)	0.94 (0.91–0.96)

4.2 Fairness Evaluation across Fitzpatrick Skin Tone Groups

Across Fitzpatrick groups, we observed pronounced performance variations for specific lesion types. In particular, darker skin tones (Groups 4–6) exhibited the largest declines in true positive rate (TPR) for ACK, SEK, and SCC, indicating reduced sensitivity for these conditions (Table 9). In contrast, lighter skin groups generally showed more consistent performance across diseases, suggesting that the model's learned feature representations may be biased toward lighter skin distributions present in the training data.

As shown in the underdiagnosis results (Table 3), darker skin tones (Groups 4–6) exhibited a higher underdiagnosis rate (0.17 [0.07–0.29]) whereas lighter tones were near zero. This pattern indicates that although the overall accuracy remained high, the model underperformed in identifying positive cases among darker skin tones. Such disparties likely stem from the under representation of these groups in the training data, resulting in limited feature diversity and reduce generalization. Addresing this imbalance through targeting data augmentation or domain adaptation may enhance recognition performance and mitigate diagnostic bias in future iterations of the models.

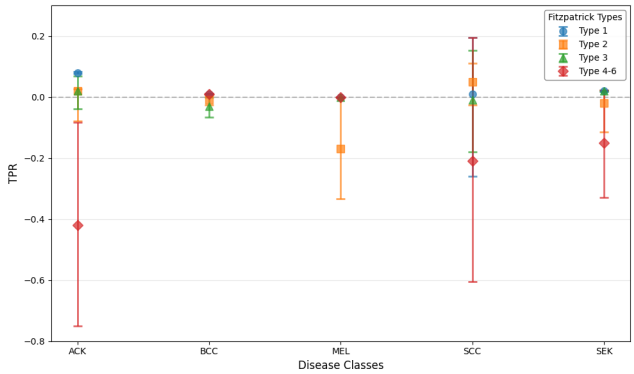

Figure 1: Per-disease TPR disparities across Fitzpatrick skin tone groups (1–6) on the PAD-UFES-20 test set. Each point represents the mean TPR disparity for a disease, with 95% confidence intervals indicated by vertical bars. Positive TPR disparities indicate more favorable sensitivity for that group, while negative disparities indicate reduced sensitivity. Darker skin tones (Groups 4–6) show generally lower TPRs for ACK, SEK, and SCC, suggesting that model sensitivity decreases for these lesion types in darker skin.

Table 2: Equalized Odds (EO) gaps across Fitzpatrick groups on the PAD-UFES-20 test set. Highest gap is highlighted in light red.

Disease	EO gap (95% CI)
ACK	0.50 (0.17–0.84)
BCC	0.10 (0.05–0.23)
MEL	0.17 (0.00–0.33)
SCC	0.25 (0.07–0.80)
SEK	0.18 (0.06–0.35)

Equalized Odds gaps (Table 2) show clear variation in fairness across skin tones. The largest EO gaps occurred for ACK (0.50 [0.17–0.84]) and SCC (0.25 [0.07–0.80]), indicating uneven error balance between lighter and darker groups for these lesion types. In contrast, BCC and MEL showed smaller EO gaps, suggesting more consistent behavior across skin tones.

Table 3: Underdiagnosis rates across Fitzpatrick groups on the PAD-UFES-20 test set. Highest rates are highlighted in light red.

Group	Underdiagnosis rate (95% CI)
1	0.00 (0.00–0.00)
2	0.02 (0.01–0.03)
3	0.01 (0.00–0.02)
4–6	0.17 (0.07–0.29)
Max–min gap	0.17 (0.07–0.29)

4.3 Fairness Evaluation across Demographics and Clinical Presentation of Lesions

4.3.1 Fairness Evaluation across Demographics

Across demographic groups (Table 10), age exhibited the strongest influence on model performance. Participants aged 55–64 showed the largest disparities in trur positive rate (TPR), particularly for SCC and MEL, while older adults (65–74 and 75+) demonstrated more stable results. Equalized Odds (EO) gaps (Table 4) were likewise most pronounced across age, especially for SEK (0.27 [0.07 to 0.60]) and MEL (0.20 [0.00 to 0.60]), suggesting less consistent model behavior across age ranges. In contrast, gender differences were minimal, with EO gaps generally below 0.07. Underdiagnosis rates (Table 5) followed a similar pattern, as variation across age (max to min gap 0.030 [0.015 to 0.057]) exceeded that observed for sex or clinical variables, confirming that age exerted the most prominent effect on fairness outcomes.

4.3.2 Fairness Evaluation across Clinical Presentation

Across clinical presentation features (Tables 4, 5 and 10), *SEK* showed the strongest disparities. TPR differences were greatest for bleeding lesions and medium-sized lesions, suggesting that *SEK* performance was particularly sensitive to variations in clinical appearance. Equalized Odds (EO) gaps showed a similar pattern, with the largest disparities for *SEK* across bleeding (0.28 [0.01–0.70]), hurt (0.14 [0.02–0.55]), and size-based groups (0.20 [0.01–0.60]).

This suggests that *SEK* predictions were most affected by lesion presentation compared to other diseases. Underdiagnosis rates showed smaller variation overall, though the greatest max–min gap occurred for painful lesions (0.019) and for non-elevated ones (0.031), indicating slightly higher missed-diagnosis risk in those subgroups.

Table 4: Equalized Odds (EO) gaps across demographic (age, sex) and clinical presentation (bleed, elevation, hurt, size) groups on PAD-UFES-20. Highest EO gap per attribute is highlighted in light red.

Disease	Age	Sex	Bleed	Elevation	Hurt	Size
ACK	0.08 (0.04–0.15)	0.02 (0.01–0.08)	0.03 (0.02–0.07)	0.01 (0.00–0.07)	0.03 (0.01–0.14)	0.02 (0.02–0.11)
BCC	0.07 (0.05–0.13)	0.03 (0.01–0.09)	0.08 (0.03–0.14)	0.02 (0.01–0.06)	0.12 (0.03–0.21)	0.04 (0.02–0.12)
MEL	0.20 (0.00–0.60)	0.03 (0.00–0.25)	0.09 (0.00–0.22)	0.07 (0.00–0.33)	0.12 (0.00–0.24)	0.20 (0.07–0.44)
SCC	0.17 (0.08–0.35)	0.05 (0.01–0.21)	0.04 (0.02–0.20)	0.03 (0.01–0.16)	0.01 (0.01–0.16)	0.13 (0.04–0.30)
SEK	0.27 (0.07–0.60)	0.07 (0.00–0.18)	0.28 (0.01–0.70)	0.02 (0.00–0.13)	0.14 (0.02–0.55)	0.20 (0.01–0.60)

Table 5: Underdiagnosis rate by demographics & clinical presentation. Groups with the highest underdiagnosis rate within each block are highlighted in light red.

Group	Rate (95% CI)	Group	Rate (95% CI)	Group	Rate (95% CI)
<55	0.029 (0.013–0.050)	Female	0.021 (0.008–0.033)	Small	0.014 (0.003–0.029)
55–64	0.034 (0.015–0.057)	Male	0.007 (0.002–0.015)	Medium	0.003 (0.000–0.009)
65–74	0.032 (0.014–0.053)	Max–min gap	0.013 (0.001–0.027)	Large	0.015 (0.003–0.029)
75+	0.004 (0.000–0.012)			Max–min gap	0.012 (0.003–0.029)
Max–min gap	0.030 (0.019–0.057)				
Elevation 0	0.031 (0.014–0.051)	Hurt 0	0.024 (0.015–0.035)	Bleed 0	0.020 (0.010–0.032)
Elevation 1	0.013 (0.006–0.022)	Hurt 1	0.005 (0.000–0.014)	Bleed 1	0.015 (0.003–0.030)
Max–min gap	0.018 (0.002–0.038)	Max–min gap	0.019 (0.005–0.033)	Max–min gap	0.005 (0.000–0.021)

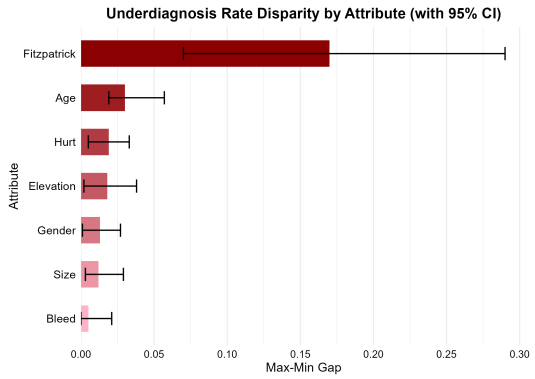

Figure 2: Max–min underdiagnosis disparity (95% CI) by attribute; Fitzpatrick skin tone exhibits the largest gap.

4.4 Evaluation of Bias Mitigation Methods on Fitzpatrick Skin Tones

4.4.1 Importance Weighting (IW)

Importance weighting substantially reduced underdiagnosis for Fitzpatrick 4–6, decreasing rates from 0.17 at baseline to 0.07 (Table 7; $\Delta -0.10$), marking the largest reduction among all mitigation methods (max–min gap from 0.17 to 0.06; $\Delta -0.10$). Sensitivity (TPR) improvements were targeted and specific. The strongest gain occured for *SEK* in Groups 4-6 (Table 11; $\Delta +0.02$), while other disease classes remained largely stable Equalized Odds (EO) gaps shifted only marginally under importance weighting (Table 6), indicating that the improvement for darker skin tones was primarily driven by fewer missed detections rather than increased false positives.

4.4.2 Group-balanced resampling (RS): moderate improvement with trade-offs

Resampling also lowered underdiagnosis for Fitzpatrick 4–6 from 0.17 to 0.12 (Table 7; $\Delta -0.05$), reducing the max–min gap from 0.17 to 0.11 ($\Delta -0.06$). TPR values improved for several classes, most notably *ACK* ($\Delta +0.17$) and *SEK* ($\Delta +0.14$) (Table 11), but there are also regressions like *SCC* ($\Delta -0.20$). Equalized Odds (EO) gaps exhibited a mixed pattern of gains and degradations (Table 6), suggesting that resampling improved sensitivity for some diseases at the expense of fairness consistency across others.

4.4.3 Comparative summary

Across both strategies, importance weighting most effectively meets the clinical fairness objective for darker tones, achieving the greatest reduction in underdiagnosis and the largest shrinkage of inter-group gaps while maintaining stable performance across other disease classes (Figure 3). Group-balanced resampling provided moderate imporvements but with less consistency, enhancing sensitivity for certain classes in Groups 4–6 while degrading others (notably *SCC*).

Table 6: Equalized Odds (EO) gaps across fitzpatrick groups after after two bias-mitigation strategies on the PAD-UFES-20 test set. Orange = improved, red = worse, no color = unchanged.

Disease	**Importance Weighting**		**Group-Balanced Resampling**	
	EO gap (95% CI)	Δ	EO gap (95% CI)	Δ
BCC	0.50 (0.17–0.84)	0.00	0.27 (0.08–0.67)	−0.23
ACK	0.07 (0.04–0.18)	−0.04	0.09 (0.08–0.34)	−0.02
SEK	0.17 (0.00–0.33)	0.00	0.22 (0.06–0.44)	+0.06
SCC	0.23 (0.06–0.80)	height−0.02	0.42 (0.11–0.91)	+0.17
MEL	0.20 (0.02–0.60)	+0.02	0.20 (0.04–0.60)	+0.02

Table 7: Underdiagnosis rates across Fitzpatrick groups after two bias-mitigation strategies on the PAD-UFES-20 test set. Δ values indicate the change relative to the baseline model (Table 3). Orange = improved, red = worse, no color = unchanged.

Group	**Importance weighting**		**Group-balanced resampling**	
	Underdiagnosis Rate (95% CI)	Δ	Underdiagnosis Rate (95% CI)	Δ
1	0.01 (0.00–0.02)	+0.01	0.01 (0.00–0.04)	+0.01
2	0.02 (0.01–0.03)	0.00	0.02 (0.01–0.03)	0.00
3	0.01 (0.00–0.02)	0.00	0.04 (0.02–0.06)	+0.03
4–6	0.07 (0.00–0.17)	−0.10	0.12 (0.02–0.24)	−0.05
Max–min gap	0.06 (0.02–0.17)	−0.10	0.11 (0.03–0.22)	−0.06

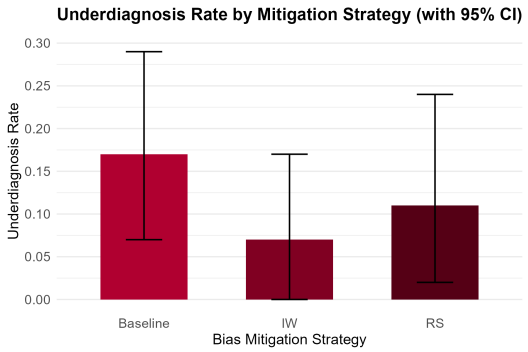

Figure 3: Comparison of underdiagnosis rates for Fitzpatrick group 4–6 across the baseline and bias mitigation strategies. Importance Weighting (IW) led to a noticeable reduction in underdiagnosis rate but did not completely eliminate it comapred to Group-balanced resampling (RS).

5 Discussion

Our results show that embeddings extracted from Google's Derm Foundation Model enable highly accurate classification of skin lesions, achieving a strong macro average AUC of 0.94 (95% CI 0.91–0.96) on the aggregated dataset. Combining data from multiple sources improved model performance across most disease categories, suggesting that greater diversity in training examples supports stronger generalization. The best results were achieved for Seborrheic Keratosis, indicating that the embeddings capture the distinctive color and texture characteristics of this lesion particularly well.

An important advantage of leveraging pre-encoded embeddings is their efficiency as it allows classifiers to be trained quickly and effectively, avoiding the need for large end-to-end models. This makes it much easier to deploy such systems in real clinical environments, where computational resources and time are limited. A model that performs this well could potentially assist dermatologists in diagnosis. Furthermore, it could serve as a valuable screening aid in underserved communities where access to dermatologists is limited.

Concurrently, fairness evaluations revealed that disparities persist across skin tones. The highest underdiagnosis rate was observed for Fitzpatrick Types 4-6, with a value of 0.17, while performance across age, sex, and clinical presentation features such as lesion size, bleeding, elevation, or pain showed only small differences (Figure 2). These findings suggest that although the embeddings are powerful, they still do not generalize equally across the full range of skin tones. This likely reflects bias in the data used to pretrain the foundation model rather than limitations in the downstream classifiers themselves.

Among the bias mitigation methods tested, importance weighting achieved the greatest improvement, reducing the underdiagnosis rate for darker skin tones by 0.10, decreasing it to 0.07 (Figure 3). The resulting trade offs for lighter tones were minimal, with a slight increase of 0.01 for Type one. While these results show that importance weighting can improve fairness, they also highlight that adjusting downstream classifiers alone is insufficient to eliminate representation bias present in the embeddings.

Overall, our findings support two main conclusions. First, foundation model embeddings offer a promising and scalable pproach for dermatology image classification, combining high diagnostic accuracy with low computational cost. Second, model fairness remains constrained by the limited diversity of available data. Improving representation across darker skin tones and developing training objectives that encourage skin tone invariant feature learning are essential steps toward building dermatology AI systems that are equitable, generalizable, and clinically ready for real-world deployment.

6 Limitations & Future Directions

A major limitation of this study is the lack of dermatology datasets that include Fitzpatrick skin tone labels, which limited our ability to analyze bias using only a small amount of available public data. The underrepresentation of darker skin tones remains a persistent challenge in dermatology AI, impacting both model training and fairness evaluation. We were also limited in the scope of our classification task to only five overlapping diagnostic categories between the two datasets. This restriction was necessary because PAD-UFES-20 is the only dataset among those we used that provides detailed metadata, including Fitzpatrick skin tone, demographics and clinical presentation information.

For future work, we plan to extend our analysis to include the DDI dataset, which contains skin lesion images annotated with Fitzpatrick skin tone ratings [13]. The DDI dataset was not included in the present study because it provides only Fitzpatrick skin tone ratings without accompanying metadata for demographics or lesion characteristics. After completing this broader bias analysis that incorporates demographic and clinical presentation variables, a follow-up study focused specifically on skin tone fairness using DDI will allow for a more targeted evaluation of skin tone representation. We also plan to conduct another study using the SCIN dataset [21], which, like DDI, includes Fitzpatrick skin tone ratings but covers dermatologic conditions that are not represented in PAD-UFES-20 or DERM12345. In addition, we hope to collaborate with private clinicians to collect more examples from individuals with darker skin tones. Building more diverse datasets is essential to ensuring that AI models in dermatology deliver equitable care across all patient populations. Future research should also focus on refining or retraining foundation models to improve generalization across the full spectrum of skin tones and disease types.

7 Conclusion

Overall, embeddings extracted from Google's Derm Foundation Model demonstrate strong performance for skin lesion classification tasks and generally has consistent results across age, sex, and clinical presentation groups, with only minor variations ($\approx$3% underdiagnosis gap in the most affected group). However, our fairness analyses reveal that these embeddings do not generalize equally across skin tones particularly for darker skin tone (Fitzpatrick group 4-6). This suggests that the underlying representation space itself encodes skin tone dependent differences, likely reflecting the limited diversity of the pretraining data. To address this, future work should prioritize expanding on training data to include individuals with darker skin tones and developing methods that pronote learning of skin tone invariant features, thereby ensuring more accurate and equitable performance across all groups.

References

[1] Haiwen Gui, Jesutofunmi A Omiye, Crystal T Chang, and Roxana Daneshjou. The promises and perils of foundation models in dermatology. *Journal of Investigative Dermatology*, 144(7): 1440–1448, 2024.

[2] Atilla P. Kiraly, Sebastien Baur, Kenneth Philbrick, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Nick George, Fayaz Jamil, Jing Tang, Kai Bailey, Faruk Ahmed, Akshay Goel, Abbi Ward, Lin Yang, Andrew Sellergren, Yossi Matias, Avinatan Hassidim, Shravya Shetty, Daniel Golden, Shekoofeh Azizi, David F. Steiner, Yun Liu, Tim Thelin, Rory Pilgrim, and Can Kirmizibayrak. Health ai developer foundations, 2024. URL `https://arxiv.org/abs/2411.15128`.

[3] Chloe Sales and Sarah J Coates. Applications of artificial intelligence for high-burden, underserved skin diseases in global settings: a review. *Current Dermatology Reports*, 14(1):14, 2025.

[4] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, pages 1–12, 2025.

[5] Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer methods and programs in biomedicine*, 245:108044, 2024.

[6] Anurag Vaidya, Richard J Chen, Drew FK Williamson, Andrew H Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C Dyer, Ming Y Lu, Jana Lipkova, et al. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30(4):1174–1190, 2024.

[7] Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023.

[8] Burak Kocak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and interventional radiology*, 31(2):75, 2025.

[9] Google Developers. Derm foundation model. `https://developers.google.com/health-ai-developer-foundations/derm-foundation`, 2025. Last updated 2025-02-11 UTC.

[10] Kai Sun, Siyan Xue, Fuchun Sun, Haoran Sun, Yu Luo, Ling Wang, Siyuan Wang, Na Guo, Lei Liu, Tian Zhao, Xinzhou Wang, Lei Yang, Shuo Jin, Jun Yan, and Jiahong Dong. Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions, 2024. URL `https://arxiv.org/abs/2412.02621`.

[11] Mohan Timilsina, Samuele Buosi, Muhammad Asif Razzaq, Rafiqul Haque, Conor Judge, and Edward Curry. Harmonizing foundation models in healthcare: A comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. *Computers in Biology and Medicine*, 189:109925, 2025.

[12] Jingkai Xu, De Cheng, Xiangqian Zhao, Jungang Yang, Zilong Wang, Xinyang Jiang, Xufang Luo, Lili Chen, Xiaoli Ning, Chengxu Li, Xinzhu Zhou, Xuejiao Song, Ang Li, Qingyue Xia, Zhou Zhuang, Hongfei Ouyang, Ke Xue, Yujun Sheng, Rusong Meng, Feng Xu, Xi Yang, Weimin Ma, Yusheng Lee, Dongsheng Li, Xinbo Gao, Jianming Liang, Lili Qiu, Nannan Wang, Xianbo Zuo, and Cui Yong. Dermino: Hybrid pretraining for a versatile dermatology foundation model, 2025. URL `https://arxiv.org/abs/2508.12190`.

[13] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

[14] Muhammad Osama Khan, Muhammad Muneeb Afzal, Shujaat Mirza, and Yi Fang. How fair are medical imaging foundation models? In *Machine Learning for Health (ML4H)*, pages 217–231. PMLR, 2023.

[15] Dilermando Queiroz, Anderson Carlos, André Anjos, and Lilian Berton. Fair foundation models for medical image analysis: Challenges and perspectives, 2025. URL `https://arxiv.org/abs/2502.16841`.

[16] Google Cloud. Derm foundation — vertex ai model garden. `https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/derm-foundation`, 2025. Accessed: 2025-10-15.

[17] Andre GC Pacheco and Renato A Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545, 2020.

[18] Abdurrahim Yilmaz, Sirin Pekcan Yasar, Gulsum Gencoglan, and Burak Temelkuran. Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data*, 11(1):1302, 2024.

[19] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[20] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[21] Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615–e2446615, Nov 2024.

A Technical Appendices and Supplementary Material

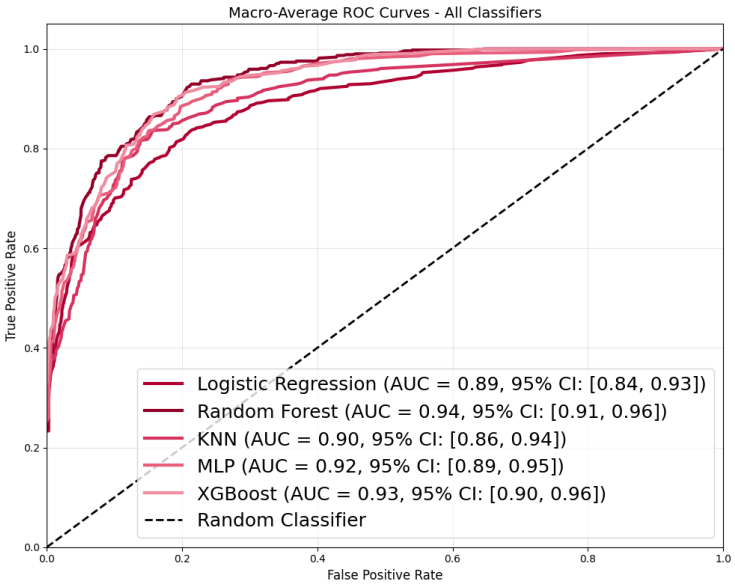

Figure 4: ROC curves of our best performing models on the aggregated dataset, showing Macro-AUC performance.

Table 8: Hyperparameter settings for all classifiers used in the training and evaluation pipeline.

Model	Hyperparameters
Logistic Regression	`max_iter=1000`, `class_weight="balanced"`, `multi_class="multinomial"`, `random_state=42`
Random Forest	`n_estimators=300`, `class_weight="balanced"`, `random_state=42`
K-Nearest Neighbors (KNN)	`n_neighbors=15`, `weights="distance"`, features standardized with `StandardScaler()`
Multilayer Perceptron (MLP)	`hidden_layer_sizes=(128, 64)`, `max_iter=500`, `random_state=42`, `early_stopping=True`, features standardized with `StandardScaler()`
XGBoost	`objective="multi:softprob"`, `eval_metric="mlogloss"`, `random_state=42`

B Fairness Metrics

B.1 TPR Disparities

Table 9: TPR disparities across fitzpatrick skin tone (fst) groups for each disease with 95% confidence intervals (CI). The group with the largest disparity for each disease class is highlighted in light red.

Group	BCC	ACK	SEK	SCC	MEL
FST 1	0.01 (0.01–0.01)	0.08 (0.08–0.08)	0.02 (0.02–0.02)	0.01 (-0.26–0.20)	0.00 (0.00–0.00)
FST 2	-0.01 (-0.03–0.01)	-0.02 (-0.08–0.03)	-0.02 (-0.11–0.02)	0.05 (-0.03–0.11)	-0.17 (-0.33–0.00)
FST 3	-0.03 (-0.07–0.00)	0.02 (-0.04–0.07)	0.02 (0.02–0.02)	-0.01 (-0.18–0.15)	0.00 (0.00–0.00)
FST 4–6	0.01 (0.01–0.01)	-0.42 (-0.75—0.08)	-0.15 (-0.33–0.02)	-0.21 (-0.61—0.20)	0.00 (0.00–0.00)

Table 10: TPR disparities across demographic and clinical groups with 95% confidence intervals (CI). Greatest disparities (point estimates) are highlighted in light red.

Group	BCC	ACK	SEK	SCC	MEL
Age disparities					
<55	-0.02 (-0.08–0.02)	0.02 (-0.01–0.05)	-0.16 (-0.47–0.16)	0.05 (0.00–0.10)	0.10 (0.00–0.20)
55–64	-0.01 (-0.05–0.03)	-0.03 (-0.10–0.03)	-0.01 (-0.10–0.08)	-0.12 (-0.30–0.05)	-0.10 (-0.30–0.10)
65–74	0.01 (-0.03–0.05)	-0.02 (-0.09–0.04)	0.04 (0.00–0.08)	0.03 (-0.03–0.08)	0.00 (0.00–0.00)
75+	0.01 (-0.02–0.05)	0.05 (0.02–0.09)	0.01 (-0.04–0.06)	-0.03 (-0.11–0.06)	0.00 (0.00–0.00)
Gender disparities					
Female	-0.01 (-0.03–0.00)	-0.01 (-0.04–0.02)	-0.04 (-0.09–0.00)	-0.03 (-0.10–0.04)	0.02 (-0.06–0.13)
Male	0.01 (-0.00–0.03)	0.01 (-0.02–0.04)	0.04 (0.00–0.09)	0.03 (-0.04–0.10)	-0.02 (-0.13–0.06)
Bleed disparities					
Bleed 0	-0.02 (-0.04–0.00)	-0.01 (-0.04–0.01)	0.14 (-0.01–0.29)	0.02 (-0.05–0.10)	-0.04 (-0.11–0.00)
Bleed 1	0.02 (-0.00–0.04)	0.01 (-0.01–0.04)	-0.14 (-0.29–0.01)	-0.02 (-0.10–0.05)	0.04 (0.00–0.11)
Elevation disparities					
Elevation 0	0.00 (-0.01–0.02)	0.01 (-0.02–0.03)	-0.01 (-0.07–0.03)	-0.01 (-0.08–0.05)	-0.03 (-0.17–0.11)
Elevation 1	-0.00 (-0.02–0.01)	-0.01 (-0.03–0.02)	0.01 (-0.03–0.07)	0.01 (-0.05–0.08)	0.03 (-0.11–0.17)
Hurt disparities					
Hurt 0	-0.01 (-0.02–0.01)	0.02 (-0.03–0.07)	0.00 (0.00–0.00)	-0.00 (-0.07–0.08)	0.00 (0.00–0.00)
Hurt 1	0.01 (-0.01–0.02)	-0.02 (-0.07–0.03)	0.00 (0.00–0.00)	0.00 (-0.08–0.07)	0.00 (0.00–0.00)
Size disparities					
Small	-0.02 (-0.05–0.01)	-0.01 (-0.07–0.03)	0.07 (0.00–0.21)	0.13 (0.00–0.22)	0.11 (0.00–0.22)
Medium	0.00 (-0.02–0.02)	0.01 (-0.04–0.06)	0.00 (0.00–0.00)	0.00 (-0.16–0.00)	0.00 (-0.27–0.00)
Large	0.01 (-0.01–0.03)	0.00 (-0.06–0.05)	-0.07 (-0.21–0.00)	-0.01 (-0.15–0.01)	-0.02 (-0.27–0.00)

Table 11: TPR disparities across Fitzpatrick groups for each disease after bias-mitigation strategies. Orange = improved (more positive), Red = worse (more negative), no color = unchanged, compared to Table 9. 95% confidence intervals (CI) are shown in parentheses.

Group	BCC	ACK	SEK	SCC	MEL
Importance Weighting (IW)					
FST 1	-0.00 (-0.04–0.02)	0.08 (0.08–0.08)	0.02 (0.02–0.02)	0.01 (-0.26–0.20)	0.00 (0.00–0.00)
FST 2	0.00 (-0.01–0.02)	-0.02 (-0.08–0.03)	-0.02 (-0.11–0.02)	0.03 (-0.06–0.10)	-0.17 (-0.33–0.00)
FST 3	-0.03 (-0.06–0.00)	0.02 (-0.04–0.07)	-0.18 (-0.58–0.02)	-0.01 (-0.18–0.15)	0.00 (0.00–0.00)
FST 4–6	0.02 (0.02–0.02)	-0.42 (-0.75—0.08)	0.02 (0.02–0.02)	-0.21 (-0.61–0.20)	0.00 (0.00–0.00)
Group-Balanced Resampling (RS)					
FST 1	0.03 (-0.02–0.07)	0.01 (-0.09–0.09)	0.05 (0.05–0.05)	0.01 (-0.26–0.20)	0.00 (0.00–0.00)
FST 2	0.05 (0.03–0.07)	-0.01 (-0.07–0.04)	0.01 (-0.08–0.05)	0.02 (-0.07–0.09)	-0.22 (-0.44—0.06)
FST 3	-0.03 (-0.08–0.02)	0.03 (-0.03–0.07)	-0.15 (-0.55–0.05)	-0.01 (-0.18–0.15)	0.00 (0.00–0.00)
FST 4–6	-0.04 (-0.29–0.09)	-0.25 (-0.58–0.09)	-0.01 (-0.12–0.05)	-0.41 (-0.81—0.01)	0.00 (0.00–0.00)

C Dataset Distribution

Table 12: Distribution of images per disease class across the DERM12345 and PAD-UFES-20 datasets.

Disease Class	**DERM12345**	**PAD-UFES-20**	**Total**
Actinic Keratosis (ACK)	58	730	788
Basal Cell Carcinoma (BCC)	423	845	1268
Malignant Melanoma (MEL)	52	400	452
Squamous Cell Carcinoma (SCC)	266	192	458
Seborrheic Keratosis (SEK)	607	235	842
Total Images	**1406**	**2402**	**3808**

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the goals, methods, and findings of the paper which focus on evaluating Derm Foundation embeddings for accuracy and fairness, and these claims are consistently supported by the experiments and discussion, accurately reflecting the study's scope and contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper includes a dedicated “Limitations & Future Directions” section that clearly discusses constraints such as limited availability of diverse dermatology datasets, underrepresentation of darker skin tones, and restricted diagnostic categories, outlining how these factors affect fairness analysis and future research directions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results or formal proofs, as it is primarily an empirical study focused on experimental evaluation of model performance and fairness.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of datasets, data splits, classifiers, hyperparameters, evaluation metrics, and experimental procedures, ensuring that the main results can be reproduced even without direct access to the original code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper does not provide open access to the code or replication scripts, though it uses publicly available datasets and describes experimental settings in sufficient detail to allow independent reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper clearly specifies all training and testing details, including dataset splits, classifier types, hyperparameter settings, and evaluation procedures, allowing readers to fully understand how the results were obtained.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports 95% confidence intervals for all major performance and fairness metrics using bootstrap resampling, providing appropriate measures of variability and statistical significance for the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper does not specify details about computational resources such as hardware type, memory, or runtime, although the experiments are lightweight and based on pre-encoded embeddings that can be reproduced on standard hardware.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research adheres to the NeurIPS Code of Ethics, as it uses publicly available datasets, maintains participant anonymity, and focuses on fairness and equity in medical AI without involving any harmful or unethical data collection or practices.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the positive societal impact of improving accessibility and fairness in dermatologic AI while acknowledging potential risks of bias and unequal performance across skin tones, emphasizing the importance of equitable deployment in clinical settings.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any new models or datasets with potential misuse risks, and all experiments are conducted using publicly available, ethically sourced datasets, so additional safeguards were not required.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits all original creators of datasets and models, including PAD-UFES-20, DERM12345, and Google's Derm Foundation Model, with appropriate citations and respect for their licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce any new datasets, models, or code assets, as it relies entirely on publicly available resources for all experiments and analyses.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any crowdsourcing or direct research with human subjects, as all analyses are performed on existing, publicly available dermatology datasets.

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve direct interaction with human subjects, and all datasets used are publicly available and de-identified; therefore, IRB approval was not required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not use large language models as part of its core methodology; any AI tools were used solely for writing or formatting assistance and did not influence the scientific content or analysis.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.