AI-Based Skin Disease Diagnosis with Non-Dermoscopic Images: Tackling Bias and Data Limitations

Chiara Bellatreccia
University of Bologna
chiara.bellatreccia@studio.unibo.it

Andrea Borghesi
University of Bologna
andrea.borghesi3@unibo.it

Arianna Dondi
Sant'Orsola Hospital of Bologna
arianna.dondi@aosp.bo.it

Roberta Calegari
University of Bologna
roberta.calegari@unibo.it

Daniele Zama
University of Bologna
daniele.zama@gmail.com

Luca Pierantoni
Sant'Orsola Hospital of Bologna
luca.pierantoni@aosp.bo.it

Laura Andreozzi
University of Bologna
laura.andreozzi4@unibo.it

Iria Neri
University of Bologna
iria.neri@aosp.bo.it

Marcello Lanari
University of Bologna
marcello.lanari@unibo.it

Abstract

AI-based skin disease diagnosis holds significant potential for improving healthcare equity but remains challenged by fairness concerns, particularly in underrepresented populations. This study addresses these issues using a real-world dataset from an Italian hospital, which suffers from limited diversity in skin tones and disease classes, as well as non-dermoscopic, low-quality images captured under inconsistent conditions. These factors contribute to classification bias and hinder existing fairness mitigation strategies. We propose a novel two-stage pipeline that combines (1) targeted data augmentation using DreamBooth fine-tuned Stable Diffusion to generate synthetic images for darker skin tones, and (2) disease classification using a Swin Transformer model. Our results show improved fairness metrics and balanced performance across skin tone groups, demonstrating the effectiveness of synthetic data in reducing dermatological AI bias.

Keywords: AI Fairness Compliance, AI Ethics, Skin Disease Prediction, AI Fairness

1. Introduction

Artificial Intelligence (AI) has increasingly demonstrated its transformative potential in the field of dermatology, particularly in the automated prediction and classification of skin diseases. AI-driven diagnostic tools can significantly accelerate the identification process, reduce clinician workload, and improve early detection rates—ultimately contributing to enhanced patient outcomes (Chiu et al., 2024; Yuan et al., 2022). However, despite these advantages, critical concerns remain regarding the fairness, reliability, and generalizability of such systems when deployed across demographically diverse populations (Gordon et al., 2024). Fairness in AI refers to the absence of bias or discrimination (Barocas & Selbst, 2016). Achieving fairness is challenging, as different types of bias must be identified and mitigated. The literature distinguishes several notions of fairness, including group fairness, individual fairness, and counterfactual fairness (Zafar et al., 2017). Group fairness requires equal or proportional treatment of different demographic groups. In this work, the groups under consideration are defined by skin color, an attribute that is frequently subject to bias in the classification of dermatological diseases. In particular, three mechanisms are especially relevant in our setting: (i) representation bias—under- or over-representation of specific skin tones; (ii) measurement bias—artifacts introduced during image acquisition, such as illumination, color cast/white balance, blur, and sensor noise; and (iii) label/process bias—label noise and task design choices (e.g., lesion cropping) that can amplify spurious cues. Our protocol directly addresses (i) through dataset construction and auditing, and explicitly acknowledges (ii)–(iii) as limitations on generalizability. The representation bias often originates from imbalanced training datasets, in which individuals with lighter skin tones are disproportionately represented. This imbalance frequently results from data collection processes conducted in geographic regions with predominantly Caucasian populations (Xu et al., 2024). When left unaddressed, such skewed distributions

URI: https://hdl.handle.net/10125/111817
978-0-9981331-9-5

lead to systematic underperformance of AI models for underrepresented groups, raising serious ethical concerns and potentially exacerbating existing disparities in dermatological healthcare delivery (Gordon et al., 2024; Zhang et al., 2024). These challenges are further amplified in real-world scenarios involving non-dermoscopic images and less frequently studied use cases such as pediatric dermatological conditions. The limitations are multifaceted. First, *image modality* represents a major issue: while melanoma-focused studies often utilize dermoscopic images that offer magnified and detailed lesion views, non-melanoma studies, such as ours, typically rely on clinical photographs. These images introduce significant variability due to lighting, angle, and resolution, which impairs classification consistency (Xu et al., 2024). Second, *disease complexity* is heightened in non-melanoma conditions, which tend to exhibit overlapping and less distinct visual features compared to the more structured patterns of melanomas, thus posing greater challenges for deep learning (DL) models (Corbin & Marques, 2023). Third, *dataset availability* is considerably limited, especially for non-melanoma diseases affecting pediatric patients. Most public datasets focus on adults, but pediatric skin conditions often differ, limiting the accuracy and generalizability of models trained solely on adult data.

In this work, we aim to address these limitations by proposing a data-centric pipeline that enhances fairness in dermatological disease classification under data scarcity conditions, without requiring modifications to model architecture. Our pipeline is designed to be model-agnostic and compatible with various DL classifiers, thereby ensuring flexibility while introducing fairness guarantees. Specifically, we *i)* employ the Individual Typology Angle (ITA) metric in combination with a Gaussian Mixture Model (GMM)-based thresholding method to categorize skin tones; *ii)* use DreamBooth fine-tuned Stable Diffusion to synthetically generate dermatological images for underrepresented skin tones; *iii)* rigorously curate the training data to mitigate the impact of lighting inconsistencies and image quality issues; and *iv)* demonstrate compatibility with state-of-the-art models, including the Swin Transformer (ST), to validate the effectiveness and generalizability of our approach. Our methodology tackles the underexplored need for fairness-aware AI in non-dermoscopic, real-world datasets, where many existing techniques fall short (Gordon et al., 2024; Xu et al., 2024). By emphasizing data preprocessing and augmentation over model architecture, we offer a complementary solution that integrates seamlessly into existing pipelines.

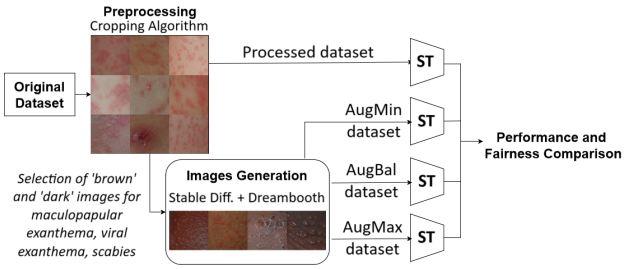

Figure 1. Our pipeline diagram.

2. Related Works

Several recent studies have proposed innovative strategies to address bias and enhance fairness in AI-based dermatological diagnostics. Chiu et al., 2024 introduced a fairness-preserving training methodology that improves feature selection by explicitly excluding sensitive demographic features. Their approach employs feature entanglement techniques to emphasize disease-relevant patterns while minimizing correlations with confounding attributes such as skin tone. Aayushman et al., 2024 proposed *PatchAlign*, which uses a Masked Graph Optimal Transport (MGOT) algorithm to align localized skin image patches with textual clinical descriptions. This enhances the relevance of diagnostic features and improves fairness across skin tones. Yuan et al., 2022 developed *EDGEMIXUP*, a preprocessing method that adjusts colour saturation and incorporates edge detection to balance model performance across skin tone groups, effectively reducing classification disparities. Zhang et al., 2024 introduced the *FairSkin* framework, leveraging diffusion models to generate synthetic dermatological images representing a wide range of skin tones. This addresses dataset imbalance and improves fairness. Munia and Imran, 2025 proposed *DermDiff*, a generative diffusion model designed to mitigate racial biases by generating representative dermoscopic images of underrepresented groups. Guo et al., 2024 proposed *FairQuantize*, a method that applies weight quantization to equalize model performance across demographic subgroups. Similarly, Kong et al., 2024 utilized channel pruning to eliminate feature channels that disproportionately affect specific demographic groups. Du et al., 2022 introduced *FairDisCo*, a disentangled contrastive learning framework that removes sensitive attribute information from learned representations to improve fairness in skin lesion classification. While these approaches mark significant progress, many remain experimental and show limited robustness when applied to real-world, non-dermoscopic datasets.

Xu et al., 2024 note that such methods often fail to generalize beyond controlled conditions. Gordon et al., 2024 further emphasizes that fairness interventions should not be used in isolation, but rather embedded within a broader, integrated development pipeline. Building on prior work, we adopt an orthogonal, model-agnostic perspective. By targeting upstream data transformation rather than model architectures, our approach enables fairness integration across diverse classifiers and deployment scenarios.

3. Dataset

The dataset comprises 8,000 non-dermoscopic clinical photographs from 273 pediatric patients at Sant'Orsola Hospital in Bologna[1]. It covers nine dermatological conditions: *drug-induced iatrogenic exanthema (DII ex.), maculopapular exanthema (MP ex.), morbilliform exanthema (MF ex.), polymorphous exanthema (PM ex.), viral exanthema (V ex.), urticaria, pediculosis, scabies*, and *chickenpox*. All images were captured by physicians using consumer-grade cameras. Demographic information was available for 200 patients, while data for the remaining individuals could not be retrieved because their images dated back several years and the corresponding records were no longer accessible. Table 1 reports descriptive statistics on patients' gender and age, indicating that most patients (76.7%) were assessed within the first five years of life. This dataset presents challenges for deep learning–based skin disease classification, primarily due to non-standardized image acquisition. Suboptimal lighting often distorts skin tones—typically making them appear darker—complicating disease classification and undermining skin tone detection methods crucial for fairness evaluation. Further complexity stems from heterogeneous image acquisition, with variations in quality, resolution, and focus—from close-up lesions to full-body shots. Inconsistencies, along with blurriness and noise, reduce diagnostic utility and hinder AI robustness. These artefacts highlight the limits of consumer-grade imaging in clinical settings and the need for careful preprocessing and standardization to ensure compatibility with deep learning models. Quantitatively, the dataset is limited to 273 patients, reducing variability across disease presentations and causing class imbalance, with several conditions underrepresented. While this study focuses on fairness by skin tone, skewed disease distributions still affect calibration and bias predictions toward majority classes. Equally critical is the predominance of lighter skin tones, reflecting local demographics but leading to underperformance for darker tones—a central concern in fairness-aware AI. Overall, limited size, acquisition inconsistency, class and demographic imbalance, and variable image quality pose major challenges to developing reliable, fair, and clinically useful models. Addressing these issues requires not only algorithmic advances but also careful preprocessing, augmentation, and fairness evaluation.

Age range (years)	Female (%)	Male (%)	Total (%)
0–1	14.7	27.9	42.6
2–5	14.0	20.1	34.1
6–10	3.9	11.6	15.5
11–18	3.0	4.8	7.8
Total	**35.6**	**64.4**	**100.0**

Table 1. Age distribution ($\leq$ 18 years) by gender, reported as percentages of the 200 patients with data.

4. Data Preprocessing

The goal of the preprocessing pipeline is to standardize the dataset by generating uniformly sized image crops that contain diagnostically relevant skin regions. This step is essential to mitigate the variability in image resolution and composition discussed previously and to prepare the dataset for effective training of deep learning models. Specifically, the pipeline focuses on extracting well-localized patches that contain visible manifestations of dermatological conditions, using binary masks that delineate areas affected by disease[2]. The preprocessing procedure follows a sliding-window approach and consists of multiple stages. Initially, the algorithm begins in the top-left corner of the image and extracts a patch of fixed size—specifically, 256×256 pixels. For each extracted patch, the corresponding binary mask is used to compute the disease coverage, defined as the proportion of positive (value = 1) to negative (value = 0) pixels within the mask. This metric quantifies the presence of visible disease in the patch and serves as a filtering criterion: patches are retained only if their disease coverage exceeds a predefined threshold, ensuring sufficient pathological content and visual contrast. Patches falling below this threshold are discarded as non-informative. The sliding window then progresses across the image at fixed intervals, generating overlapping patches. To reduce redundancy and promote diversity in the retained data, a non-maxima suppression step is applied. When overlapping patches exceed a predefined Intersection over Union (IoU) threshold, only the patch with the highest disease coverage is retained, while the others

[1] https://www.aosp.bo.it/

[2] The code for this project is publicly available and anonymized for review purposes at https://github.com/AnonymousUser2283/BiasMitigationProject/tree/main.

are discarded. This ensures that only the most informative regions are preserved, avoiding duplication of similar content. To further enhance the quality of the preprocessed dataset, a final filtering step removes patches exhibiting low visual contrast—typically caused by poor lighting, blur, or underexposure. These artifacts negatively affect the reliability of downstream analysis and classification tasks. As such, low-contrast patches are systematically excluded, ensuring that the final dataset comprises only high-quality, diagnostically relevant image regions. As anticipated, the application of this preprocessing pipeline reveals a significant class imbalance across the nine disease categories. Figure 2 illustrates the distribution of the retained patches per disease. Certain conditions are notably underrepresented, with fewer than ten thousand patches available post-processing. While class imbalance is not the primary focus of this study, it remains an important confounding factor that may adversely impact the model's ability to generalize, particularly in accurately detecting less-represented diseases.

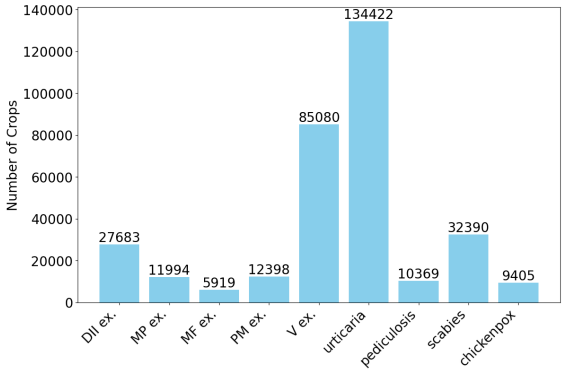

Figure 2. Crop distribution by disease.

To accurately measure fairness across different skin tones, it is essential to correctly classify each image according to the skin tone it represents. This enables precise evaluations for each skin tone, allowing for the identification of potential performance disparities, such as the presence of bias. Skin tone classification is commonly performed using the ITA, a metric first introduced by Chardon et al., 1991, and widely adopted in subsequent studies for its simplicity and effectiveness (Groh et al., 2021; Gupta & Sharma, 2019; Kinyanjui et al., 2020). While this method has proven effective in controlled environments – such as those involving dermoscopic datasets – it assumes uniform illumination and does not account for variations introduced by pathological changes in the skin or by external artefacts. We therefore propose a modified ITA computation method, *tailored* to our dataset, which consists of clinical images of skin conditions captured under non-standardized conditions using consumer-grade cameras. To mitigate challenges such as altered pigmentation in diseased areas, inconsistent lighting, and shadows, we exclude disease-affected regions from the ITA computation using segmentation maps, ensuring that only unaffected skin is analyzed. Unlike prior works that rely on fixed thresholds derived from dermoscopic datasets (Charlton et al., 2020; Groh et al., 2021; Kinyanjui et al., 2020), we classify ITA values into skin tone categories using a GMM, which is better suited to handle the variability present in our dataset. This refined method yields a more accurate representation of skin tone, enabling fairer evaluations of classification performance. The computation of ITA must account for the fact that skin affected by disease often appears darker and redder than healthy skin. To obtain reliable ITA values that reflect the baseline skin tone, it is critical to exclude affected regions from the calculation. This was achieved by applying a bitwise `and` operation between the original image crop and its corresponding segmentation mask, replacing disease-affected areas with black pixels. The ITA value was then computed exclusively for the non-black pixels in the crop. The resulting distribution of ITA values closely approximates a Gaussian distribution with a longer tail toward lower values. Following ITA computation, it is necessary to map these values to discrete skin tone categories according to the Fitzpatrick scale (Gupta & Sharma, 2019), which defines six skin types. Several thresholding schemes have been proposed to map ITA values to Fitzpatrick types (Charlton et al., 2020; Groh et al., 2021; Kinyanjui et al., 2020). However, these thresholds were primarily designed for dermoscopic datasets and do not account for variability due to illumination, viewing angles, or other artefacts. Given the non-dermoscopic nature of our dataset, we found these thresholds to be unsuitable. Instead, we assume that images with similar skin tones exhibit ITA values within reasonably consistent ranges. To classify the ITA values, we fit the distribution using a Gaussian Mixture Model with six components, corresponding to the six categories in the Fitzpatrick scale. Each ITA value is assigned to the Gaussian component that best represents it. The resulting skin tone labels, the distribution of which is shown in Figure 3, are categorized as *dark*, *brown*, *tan*, *intermediate*, *light*, and *very light*. Examples of automatic skin tone classification are presented in Figure 4. While the ITA value is generally robust, shadows and poor illumination can reduce ITA estimates, leading to the assignment of a darker skin tone. Nevertheless, darker images—whether due to

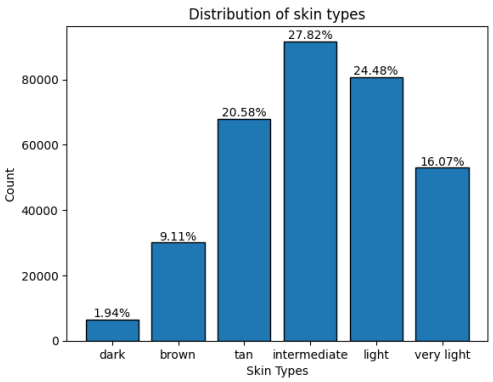

Figure 3. Distribution of the corresponding skin tones, obtained with the Gaussian Mixture technique.

actual skin tone or lighting artefacts—were typically assigned lower ITA values, while lighter images received higher values. The overall distribution of skin tone labels highlights the underrepresentation of the *dark* and *brown* categories and reflects an imbalance in skin tone representation within the dataset. Although the modified ITA calculation improves accuracy, labeling remains imperfect. Poor lighting and image artefacts can distort ITA values, misrepresenting skin tone in many cases. Future work could mitigate this by applying advanced corrections for shadows, glare, and uneven illumination.

Figure 4. Examples of skin tone classification using ITA and GMM.

5. Synthetic Images Generation

5.1. DreamBooth and Stable Diffusion

The dataset used in this study contains a limited number of examples of skin diseases affecting individuals with dark skin, with at most four or five individuals per disease identified as having ‘dark’ skin. The preprocessing pipeline described in Section 4 generates a large number of image crops per individual. However, using all of these crops to train an image generation model would be redundant, as the crops derived from the same individual are highly similar. Consequently, it is sufficient to select only a few representative crops (typically 3 or 4) per individual with dark skin and construct a small, curated dataset comprising several individuals with the same skin tone affected by a specific disease. This curated dataset is then used to train an image generation model. In this step, we adopted a co-creation and co-design approach. A multidisciplinary team of computer scientists and medical doctors, with expertise in dermatological disease prediction and active clinical roles at an Italian hospital, collaborated to ensure both technical and clinical accuracy in image selection.

One model particularly well-suited for training with such limited data is DreamBooth, introduced by Ruiz et al., 2023. DreamBooth is a fine-tuning approach for diffusion-based generative models, designed to produce high-quality, subject-specific images using only a few examples. The method enables personalization of a pre-trained generative model while preserving its ability to produce diverse and photorealistic outputs, making it ideal for scenarios characterized by extreme data scarcity. In our work, DreamBooth was used to fine-tune a pre-trained Stable Diffusion model (specifically the version provided by RunwayML), originally trained on images of size 512×512. The fine-tuning process was organized into the following phases:

Exploration of the Dataset A manual inspection of the dataset was conducted to identify which skin diseases included images of individuals with ‘dark’ or ‘brown’ skin. This analysis revealed that only three out of the nine diseases—*maculopapular exanthema*, *viral exanthema*, and *scabies*—contained examples of individuals with darker skin tones. As a result, image generation was limited to these three diseases.

Construction of Mini-Datasets For each of the three selected diseases, two mini-datasets were manually constructed—one for ‘brown’ skin and one for ‘dark’ skin. This resulted in a total of six distinct datasets, each containing between 14 and 29 images. The careful curation ensured that the selected samples were representative and of sufficient quality for fine-tuning.

Fine-Tuning with DreamBooth Each of the six mini-datasets was used to fine-tune the Stable Diffusion model using the DreamBooth technique. A grid search was conducted to explore optimal hyperparameter configurations, testing the following parameters: learning rate values of 5e-7, 2e-6, 5e-6, and 1e-5 were evaluated due to the learning rate’s critical role in convergence. For mini-datasets with fewer than 15 images, maximum training steps of 1000, 2000, 3000, and 4000 were tested; for datasets with more than 15 images, the tested range extended to 5000 steps. The instance prompt in DreamBooth is essential during both training and generation. During fine-tuning, a unique identifier (e.g., ”<unique_ID>”) was included in the prompt alongside contextual terms such as “human

skin" or "a person with a skin condition" to associate the model with features from the training images. During image generation, this prompt was reused or extended to control the characteristics of the synthetic outputs. A batch size of 1 was adopted, as preliminary experiments showed that smaller batch sizes improved diversity under fixed conditions.

Model Selection For each of the six fine-tuned models, the most promising configuration was selected based on empirical evaluation of generated images. Evaluation criteria included visual diversity, color and texture accuracy, fidelity to the original samples, and realism of the synthetic skin representation. Figure 5 shows representative DreamBooth-generated images for scabies, viral exanthema, and maculopapular exanthema, alongside real examples. The synthetic images are visually realistic and accurately reflect the targeted dark and brown skin tones.

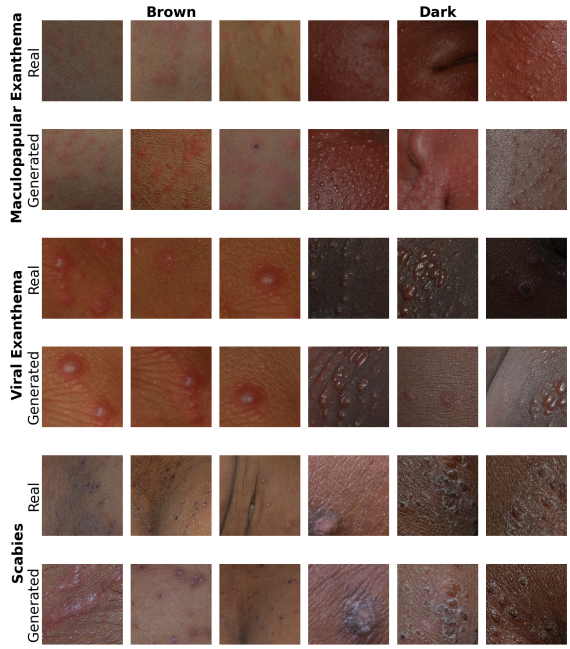

Figure 5. Real vs. generated images for diseases.

5.2. Data Augmentation

To address the underrepresentation of darker skin tones, we generated synthetic images using the previously selected DreamBooth models. The number of synthetic samples required for each disease and skin tone category ('dark' and 'brown') was determined based on three augmentation strategies, described below. In all cases, the generated images were added to the training set, while the validation and test sets remained composed entirely of real images. To promote diversity among synthetic images, we distributed generation tasks across models fine-tuned with different hyperparameter configurations. Each model contributes distinct characteristics, and varying the random seed during image generation further increases diversity.

AUGMIN Synthetic images of 'dark' and 'brown' skin are added so that the total number of images (real + synthetic) for each disease and for each of the two latter skin tones matches the smallest number of images available among the other four skin tones ('very light', 'light', 'intermediate', 'tan') for that specific disease.

AUGBALANCED Synthetic images are added such that 'dark' and 'brown' skin tones each represent approximately one-sixth (approximately 17%) of the total image count for each disease. This approach seeks to achieve a more balanced distribution across all skin tones, addressing fairness without overcompensating.

AUGMAX The total number of images (real + synthetic) for 'dark' and 'brown' skin is increased to match the largest image count among the other four skin tones for each disease. This strategy aims to provide maximum representation for underrepresented tones.

These three augmentation strategies allow systematic evaluation of how adding synthetic images of rare skin tones impacts classification performance and fairness, highlighting trade-offs between diversity, data volume, and predictive accuracy.

6. Disease Classification

The dermatological classification task requires the model to capture complex features across multiple spatial scales. For this reason, we selected the ST, one of the most advanced and effective state-of-the-art architectures for visual recognition tasks. We begin by evaluating the performance of this model *without* data augmentation, which serves as a baseline for subsequent comparisons. However, raw performance metrics alone offer limited insight into fairness-related concerns in diagnostic outcomes. To address this, we complement the evaluation with fairness metrics commonly used in the literature, particularly in the context of medical imaging and skin disease prediction (Corbin & Marques, 2023). Specifically, we report the Disparate Impact Ratio (*DI*) (Feldman et al., 2015), Equalized Odds Ratio (*EOR*) (Agarwal et al., 2018), and Predictive Rate Ratio (*PRR*). A DI outside the fairness threshold (0.80–1.25) suggests that patients with certain skin tones may receive fewer positive disease detections, increasing the risk of delayed or missed diagnoses. An imbalanced EOR (i.e., below 0.8) indicates unequal error rates: higher false negatives in darker tones exacerbate underdiagnosis, whereas higher false positives can

lead to unnecessary procedures. A PRR below the fairness threshold (i.e., under 0.8) shows that precision for darker skin tones is lower than for lighter ones. Importantly, groups can display similar precision (fair PRR) while still exhibiting unequal error rates (unfair EOR). This divergence illustrates how different fairness metrics capture complementary but distinct clinical risks. Because these fairness metrics are typically defined for binary classification tasks and binary demographic groups, we adapt them to our multi-class and multi-group setting by aggregating skin tones into two broader categories: a minority group ('dark' and 'brown' skin tones) and a majority group ('tan', 'intermediate', 'light', and 'very light' skin tones). This grouping is informed by two main considerations: *(i)* the clear underrepresentation of 'dark' and 'brown' tones in the dataset; and *(ii)* the precedent set by similar approaches in prior work (Corbin & Marques, 2023). This aggregation enables meaningful fairness evaluation in multi-class classification, affecting only assessment; the model itself remains agnostic to skin tone, which is not used in training.

6.1. ST - No Augmentation

During training, we adopted the following dataset partitioning: 60% for training, 20% for validation, and 20% for testing, using 5-fold cross-validation to obtain robust estimates. To effectively capture both skin texture and disease-specific features, we employed a variant of the ST pre-trained on ImageNet-1k and subsequently fine-tuned on a skin cancer dataset[3]. During fine-tuning, the first stage of the ST was frozen while the last three stages were updated, resulting in a model with approximately 26 million trainable parameters. Because the frozen layers mainly encode generic low-level features (e.g., edges, textures), task-specific representations and decision boundaries were learned directly from our dermatology dataset. Therefore, the bias patterns we observed likely stem from our dataset distribution rather than inherited effects from pretraining. To ensure convergence and optimize performance, hyperparameter tuning focused on learning rate selection. While lower learning rates promoted convergence, they often led to local minima with suboptimal accuracy and F1-scores. Consequently, we selected a learning rate of 1e-2 to allow for larger gradient steps during training. To stabilize training in later epochs, the learning rate was reduced by a factor of 100 after nine epochs, based on validation loss trends.

Classification results. Final performance metrics are reported in the first column of Tables 2, 3, and 4, aggregated by disease class and skin tone group. The ST demonstrated strong overall performance, both Accuracy and F1-score were consistent across demographic groups and disease categories. The *DI* values suggest that the ST achieves relatively high fairness: only 3 out of 9 conditions fall outside the accepted fairness range (Table 3). However, *EOR* values are consistently lower; of the 9 reported values, only one—corresponding to *viral exanthema*—falls within the fairness threshold. This indicates that the model's recall and error rates are still influenced by the *skin tone* attribute. In contrast, *PRR* values remain within acceptable bounds, implying that precision is relatively balanced between minority and majority groups. Nonetheless, as previously discussed, acceptable *PRR* values do not guarantee fairness, as evidenced by disparities reflected in the *DI* and *EOR* metrics.

6.2. ST - Data Augmentation

The ST model already demonstrated high Accuracy and F1-score values across all diseases and skin tone groups, although the *EOR* values were notably problematic—particularly for the latter diseases. When applying augmentation, AUGMIN improved both Accuracy and F1-score metrics for several diseases within minority skin tone categories (i.e., 'dark' and 'brown'), albeit with a slight performance trade-off for the majority group. Overall, classification performance remained comparable to the baseline model trained on the original dataset, across both skin tones and diseases. In terms of fairness, AUGMIN yielded significant improvements, particularly for *EOR* values. Additionally, *DI* values improved in two of the three target diseases, and *PRR* values also showed gains. In comparison, AUGBALANCED led to a reduction in Accuracy and F1-scores but produced notable improvements in *EOR*, with six out of nine values showing enhancement—although three deteriorated relative to the baseline. However, *DI* values worsened, while *PRR* values again improved. On the other hand, AUGMAX maintained strong Accuracy and F1-score values, though with a different distribution across classes compared to AUGMIN. In terms of fairness, AUGMAX performed well for *DI* and *PRR*, but it did not yield substantial improvements over the original model—except in the case of *pediculosis*. Its impact on *EOR* was limited, except for *urticaria* and *scabies*, where *EOR* values fell within the fairness threshold. In conclusion, AUGMIN proved to be the most effective augmentation strategy overall. This outcome may be attributed to the fact that the ST model

[3]https://huggingface.co/gianlab/swin-tiny-patch4-window7-224-finetuned-skin-cancer

	No synthetic augmentation				AUGMIN				AUGBALANCED				AUGMAX			
	Accuracy		F1 score		Accuracy		F1 score		Accuracy		F1 score		Accuracy		F1 score	
	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj
DII ex.	95.4%	95.6%	0.90	0.92	96.0%	95.1%	**0.95**	**0.93**	**96.9%**	**96.2%**	0.90	0.90	93.3%	92.8%	0.94	**0.93**
MP ex.	88.6%	89.0%	**0.86**	0.82	87.0%	**90.1%**	0.82	0.84	87.2%	88.9%	**0.82**	0.83	**90.2%**	87.1%	**0.85**	**0.85**
MF ex.	90.8%	**93.9%**	0.87	**0.93**	95.1%	92.9%	**0.90**	0.82	95.2%	93.0%	0.86	0.86	**96.0%**	91.9%	0.84	0.88
PM ex.	91.0%	**88.7%**	0.90	**0.90**	91.9%	88.0%	**0.93**	0.86	85.0%	81.4%	0.89	0.85	**92.0%**	84.3%	0.90	0.88
V ex.	93.7%	**96.2%**	0.93	0.90	**96.4%**	93.2%	0.91	0.90	95.9%	95.6%	0.86	**0.92**	94.1%	94.7%	**0.88**	0.91
urticaria	**91.3%**	86.5%	0.91	**0.94**	87.2%	**91.1%**	**0.93**	0.93	86.5%	85.1%	0.90	0.91	89.6%	87.0%	0.92	0.90
pediculosis	83.5%	88.8%	**0.91**	0.93	85.9%	**89.1%**	0.89	0.94	83.9%	90.8%	0.86	**0.95**	**86.1%**	87.9%	0.88	0.91
scabies	82.7%	**89.3%**	0.91	**0.96**	83.9%	87.2%	**0.93**	0.90	**85.9%**	89.0%	**0.90**	0.91	83.7%	88.9%	**0.89**	0.92
chickenpox	88.4%	90.7%	**0.89**	0.83	**88.9%**	**92.1%**	0.88	**0.90**	84.1%	89.2%	0.86	0.85	83.6%	87.9%	0.88	0.85
All	91.2%	**91.3%**	**0.91**	**0.90**	91.0%	90.3%	0.90	0.88	89.5%	89.1%	0.90	0.89	**91.3%**	90.0%	0.90	**0.90**

Table 2. ST accuracy and F1-score: disease aggregation. Results obtained with 5-fold Cross Validation.

	No synthetic augmentation			AugMin			AugBalanced			AugMax		
	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR
DII ex.	1.03	0.74	**0.96**	**1.00**	**0.89**	1.05	**1.00**	0.86	**0.94**	0.98	0.71	0.94
MP ex.	1.32	0.72	1.03	**1.25**	**0.90**	**1.00**	1.39	0.65	0.98	1.49	0.54	1.02
MF ex.	1.19	0.78	0.95	**1.17**	**0.94**	0.96	1.18	0.91	**1.00**	1.24	0.73	1.05
PM ex.	**0.80**	0.76	1.03	0.77	0.56	**1.02**	0.72	0.59	**1.04**	**0.79**	**0.83**	1.03
V ex.	0.94	0.88	**1.00**	0.96	0.88	1.03	0.96	**0.92**	1.01	**0.97**	0.73	**1.00**
urticaria	0.96	0.78	**1.00**	1.04	0.79	1.02	**0.99**	0.80	1.03	**0.99**	**0.91**	1.02
pediculosis	0.77	0.65	0.96	0.79	**0.92**	0.96	0.80	0.73	0.97	**0.83**	0.47	**0.99**
scabies	**1.24**	0.18	0.86	1.30	0.49	0.95	1.26	0.49	0.95	1.33	**0.89**	**0.96**
chickenpox	**0.77**	0.39	**0.99**	0.74	0.44	0.94	0.69	0.33	0.96	0.75	**0.46**	**0.99**
All	**1.00**	0.65	0.98	**1.00**	**0.76**	0.99	**1.00**	0.70	0.99	1.04	0.70	**1.00**

Table 3. ST DI, EOR and PRR: disease aggregation. Results obtained with 5-fold Cross Validation.

		No synthetic augmentation		AugMin		AugBalanced		AugMax	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Minority	dark	90.2%	0.90	**91.9%**	0.92	90.7%	0.92	90.7%	**0.93**
	brown	90.9%	**0.92**	91.0%	0.90	88.9%	0.89	**91.5%**	0.91
Majority	tan	**91.3%**	0.91	91.1%	0.90	90.8%	0.88	90.7%	**0.92**
	intermediate	89.8%	**0.92**	**91.1%**	0.91	89.2%	0.88	90.0%	0.90
	light	89.5%	**0.92**	**89.7%**	0.86	89.0%	0.89	89.0%	0.88
	very light	**91.6%**	**0.93**	90.2%	0.90	89.0%	0.89	90.8%	0.90
All		**90.9%**	**0.91**	90.3%	**0.91**	89.8%	0.89	90.3%	**0.91**

Table 4. ST Accuracy and F1-score: skin tones aggregation. Results obtained with 5-fold Cross Validation.

is pre-trained, making it more resistant to substantial changes in the training data. This explains why the more conservative augmentation strategy, AUGMIN, was the most effective. It also clarifies why the model benefited more from a moderate increase in synthetic data: introducing excessive augmentation likely pushed the model beyond its performance "saturation point," thereby limiting gains in fairness. Nevertheless, all augmentation strategies—particularly AUGMIN—led to fairness improvements over the original baseline model.

7. Addressing Data Leakage

Data leakage is a well-known challenge in medical machine learning, often caused by improper dataset partitioning that places data from the same patient across training, validation, and test sets (Bussola et al., 2020). This can inflate performance metrics and misrepresent model generalizability. Prior studies in dermatology, radiology, and brain MRI analysis (Bussola et al., 2020; Mohanakumar et al., 2024; Rumala, 2023) show that inappropriate splitting or inclusion of augmented samples leads models to learn subject-specific features or spurious artifacts rather than diagnostic signals, with reported score inflation of up to 41%. These findings highlight the necessity of subject-level separation for reliable evaluation, especially in high-stakes fields such as dermatology and oncology.

In this study, we were aware of the risk of data leakage, but the limited number of patients—especially in underrepresented skin tone groups—made patient-level splits infeasible. Instead, a crop-wise split was used, allowing patches from the same patient to appear across subsets, which introduces leakage and limits the reliability of performance metrics. Attempts at a non-stratified patient-wise split showed high variability due to small sample sizes, particularly for certain conditions, echoing challenges reported in Ghorbani et al., 2019. Despite these limitations, the study remains valuable for assessing and improving fairness in dermatological AI. By augmenting the dataset with synthetic images of brown and dark skin and retraining the model, we consistently reduced performance disparities between skin tone groups. This demonstrates the potential of targeted augmentation to

mitigate bias, even if absolute performance metrics are not fully reliable. Future work will focus on validating this approach on external datasets with more diverse and balanced patient populations, to confirm both the generalizability and clinical impact of the fairness improvements observed here.

8. Conclusion

This study shows that advanced image generation techniques—specifically DreamBooth with Stable Diffusion—can enhance representation of underrepresented skin tones in dermatology datasets. Adding synthetic images for 'dark' and 'brown' tones improved fairness metrics without reducing accuracy. Among the tested strategies, AUGMIN was most effective, reducing disparities while maintaining performance. Although synthetic data covered only three diseases, gains extended across all nine, underscoring the broader value of targeted augmentation. Overall, our findings support synthetic data as a practical means to improve fairness in dermatological AI and foster more equitable diagnostics. As noted in Section 7, the crop-wise split represents a limitation that reduces the interpretability of fairness metrics. Accordingly, the results should be interpreted with caution. While the observed improvements suggest that synthetic augmentation can modestly mitigate bias in this dataset and affect both accuracy and group-fairness metrics, these findings do not justify claims of clinical readiness. Generalizability remains constrained by several factors: (i) single-center data acquisition, (ii) potential measurement bias from lighting, blur, or device variability, and (iii) reliance on crop-wise rather than patient-wise splitting. Performance on entirely unseen patients or in different clinical settings cannot be definitively inferred, and both accuracy and disparities are expected to vary under domain shift (e.g., new devices, care settings, or prevalence). Patient-wise splits and external validation are therefore essential to establish real-world generalizability. Our results should thus be viewed as providing directional evidence of mechanisms rather than definitive estimates of clinical performance. A promising avenue for future work is translating light-skinned images into darker tones to achieve a sufficiently balanced distribution for patient-wise splits. However, this approach requires caution: disease presentation varies across skin tones, and the clinical realism of translated images must be preserved. Another potential improvement involves quantifying image defects (e.g., using blur indices or gray-world metrics) such as illumination, color cast, and sensor noise. Such analysis would allow for a more comprehensive understanding of model performance across images of varying quality and support unsupervised annotation of skin tone, clarifying potential error sources linked to non-dermoscopic imaging. Finally, validation of generated images by qualified medical personnel constitutes an important follow-up. The creation of synthetic dermatological images raises ethical considerations, including (i) potential artifact introduction, (ii) reinforcement of bias, and (iii) misuse of synthetic data. These risks highlight the need for human–AI co-creation, where expert validation ensures clinical realism, guides generation toward bias reduction and fairness, and establishes safeguards for trustworthy and responsible use. In this study, images underwent qualitative review by computer scientists; however, our team is currently developing a web application to enable validation by board-certified dermatologists, ensuring that synthetic data remain clinically realistic and reliable.

References

Aayushman, Gaddey, H., Mittal, V., Chawla, M., & Gupta, G. R. (2024). Fair and accurate skin disease image classification by alignment with clinical labels. In M. G. e. a. Linguraru (Ed.), *Medical image computing and computer assisted intervention* (pp. 394–404). Springer Nature Switzerland.

Agarwal, A., Beygelzimer, A., & et al. (2018, October). A reductions approach to fair classification. In J. Dy & A. Krause (Eds.), *Proc. of the 35th international conference on machine learning* (pp. 60–69, Vol. 80). PMLR.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671. https://api.semanticscholar.org/CorpusID:143133374

Bussola, N., Marcolini, A., Maggio, V., Jurman, G., & Furlanello, C. (2020). Ai slipping on tiles: Data leakage in digital pathology. https://arxiv.org/abs/1909.06539

Chardon, A., Cretois, I., & Hourseau, C. (1991). Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, *13*.

Charlton, M., Stanley, S. A., Whitman, Z., Wenn, V., Coats, T. J., Sims, M., & Thompson, J. P. (2020). The effect of constitutive pigmentation on the measured emissivity of human skin. *PLOS ONE*, *15*(11), 1–9. https://doi.org/10.1371/journal.pone.0241843

Chiu, C.-H., Chen, Y.-J., Wu, Y., Shi, Y., & Ho, T.-Y. (2024). Achieve fairness without demographics for dermatological disease diagnosis. *Medical Image Analysis*, *95*, 103188. https://doi.org/https://doi.org/10.1016/j.media.2024.103188

Corbin, A., & Marques, O. (2023). Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques. *IEEE Access*, *11*, 78339–78352. https://doi.org/10.1109/ACCESS.2023.3289320

Du, S., Hers, B., Bayasi, N., Hamarneh, G., & Garbi, R. (2022). Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. *European Conference on Computer Vision*, 185–202.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. https://arxiv.org/abs/1412.3756

Ghorbani, A., Natarajan, V., Coz, D., & Liu, Y. (2019). Dermgan: Synthetic generation of clinical skin images with pathology. *ArXiv*, *abs/1911.08716*.

Gordon, E. R., Trager, M. H., Kontos, D., Weng, C., Geskin, L. J., Dugdale, L. S., & Samie, F. H. (2024). Ethical considerations for artificial intelligence in dermatology: A scoping review. *British Journal of Dermatology*, ljae040.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1820–1828. https://doi.org/10.1109/CVPRW53098.2021.00201

Guo, Y., Jia, Z., Hu, J., & Shi, Y. (2024). FairQuantize: Achieving Fairness Through Weight Quantization for Dermatological Disease Diagnosis. *Proc. of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, *LNCS 15010*.

Gupta, V., & Sharma, V. K. (2019). Skin typing: Fitzpatrick grading and others [The Color of Skin]. *Clinics in Dermatology*, *37*(5), 430–436. https://doi.org/https://doi.org/10.1016/j.clindermatol.2019.07.010

Kinyanjui, N. M., Odonga, T., & et al. (2020). Fairness of classifiers across skin tones in dermatology. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 320–329.

Kong, Q., Chiu, C.-H., Zeng, D., Chen, Y.-J., Ho, T.-Y., Hu, J., & Shi, Y. (2024). Achieving fairness through channel pruning for dermatological disease diagnosis. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, & J. A. Schnabel (Eds.), *Medical image computing and computer assisted intervention – miccai 2024* (pp. 24–34). Springer Nature Switzerland.

Mohanakumar, M., Amalraj, C., & Upeksha, P. (2024). Few-shot melanoma stage classification with siamese networks and resnet encoders: A focus on data leakage prevention, 1–6. https://doi.org/10.1109/ICITR64794.2024.10857790

Munia, N., & Imran, A.-A.-Z. (2025). Dermdiff: Generative diffusion model for mitigating racial biases in dermatology diagnosis. *arXiv preprint arXiv:2503.17536*.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. https://arxiv.org/abs/2208.12242

Rumala, D. J. (2023). How you split matters: Data leakage and subject characteristics studies in longitudinal brain mri analysis. In *Clinical image-based procedures, fairness of ai in medical imaging, and ethical and philosophical issues in medical imaging* (pp. 235–245). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45249-9_23

Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., & Zhou, S. K. (2024). Addressing fairness issues in deep learning-based medical image analysis: A systematic review. *npj Digital Medicine*, *7*(1), 286.

Yuan, H., Hadzic, A., Paul, W., de Flores, D. V., Mathew, P., Aucott, J., Cao, Y., & Burlina, P. (2022). Edgemixup: Improving fairness for skin disease classification and segmentation. *arXiv preprint arXiv:2202.13883*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proc. of the 26th International Conference on World Wide Web*, 1171–1180. https://doi.org/10.1145/3038912.3052660

Zhang, R., Yao, Y., Tan, Z., Li, Z., Wang, P., Hu, J., Liu, S., & Chen, T. (2024). Fairskin: Fair diffusion for skin disease image generation. *arXiv preprint arXiv:2410.22551*.