Addressing Bias and Data Scarcity in AI-Based Skin Disease Diagnosis with Non-Dermoscopic Images

Chiara Bellatreccia[1,*], Daniele Zama[2,**], Arianna Dondi[2], Luca Pierantoni[2], Andreozzi Laura[2], Iria Neri[2], Marcello Lanari[2], Andrea Borghesi[1] and Roberta Calegari[1]

[1]University of Bologna, Bologna, Italy
[2]IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

Abstract

AI-based diagnosis of skin diseases holds considerable promise for increasing healthcare accessibility, however, its effectiveness is currently limited by several challenges, including fairness. This study analyzes a real-world dataset collected from an Italian hospital, characterized by limited data availability, leading to poor diversity and representation—particularly evident in the scarcity of data for certain diseases and darker skin tones. Such limitations result in substantial classification biases. Additionally, the dataset includes non-dermoscopic, consumer-grade images that suffer from quality issues like inconsistent lighting and blurriness, complicating the training of fair and efficient AI models. Conventional strategies to mitigate these problems, such as synthesizing images for underrepresented groups, are hindered by the difficulty in accurately identifying skin tones from poor-quality images. Our research introduces a novel pipeline designed to enhance both the accuracy and fairness of skin disease diagnosis by addressing the challenges posed by real-world data. The proposed solution involves a two-stage approach: 1) data pre-processing and augmentation to obtain images that more accurately represent darker skin tones, generated through a state-of-the-art diffusion model; and 2) disease classification employing deep learning models. This methodology addresses data scarcity and improves fairness, with thorough validation of real-world data showing enhanced reliability and fairness in predictions across various skin diseases.

Keywords

AI Fairness, AI Ethics, Skin Disease Prediction

1. Introduction and Related Works

The application of Artificial Intelligence (AI) in dermatological disease prediction offers significant advancements in diagnostic processes, facilitating rapid and efficient disease identification. However, despite the promising capabilities of AI, issues such as bias and discrimination present substantial challenges, particularly when these systems are applied across diverse demographic groups. Significant efforts have been made to mitigate bias in dermatological AI applications without compromising the privacy or integrity of demographic data. For instance, the study by [1] introduces a method to ensure fairness by enhancing feature selection during the model training phase, purposely omitting sensitive demographic attributes. This technique relies on sophisticated feature entanglement strategies to focus solely on disease-relevant features, minimizing biases associated with non-disease attributes like skin tone. Moreover, the introduction of PatchAlign, as discussed in [2], marks a notable advancement in aligning skin condition image patches with corresponding clinical descriptions. Using a Masked Graph Optimal Transport (MGOT) algorithm effectively reduces noise and improves diagnostic accuracy and fairness across various skin tones by focusing on disease-relevant image regions. The work of [3] presents EDGEMIXUP, a preprocessing technique that alters image data to diminish bias by manipulating

AIMMES 2025 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies | co-located with EU Fairness Cluster Conference 2025, Barcelona, Spain
*Corresponding author.
chiara.bellatreccia@studio.unibo.it (C. Bellatreccia); daniele.zama@gmail.com (D. Zama); arianna.dondi@aosp.bo.it (A. Dondi); luca.pierantoni@aosp.bo.it (L. Pierantoni); laura.andreozzi4@unibo.it (A. Laura); iria.neri@aosp.bo.it (I. Neri); marcello.lanari@unibo.it (M. Lanari); andrea.borghesi3@unibo.it (A. Borghesi); roberta.calegari@unibo.it (R. Calegari)
0009-0001-0924-4760 (C. Bellatreccia); 0000-0002-9895-5942 (D. Zama); 0000-0002-7516-243X (A. Dondi); 0000-0001-8175-5488 (L. Pierantoni); 0000-0001-7755-9652 (A. Laura); 0000-0002-8552-0771 (I. Neri); 0000-0002-2586-314X (M. Lanari); 0000-0002-2298-2944 (A. Borghesi); 0000-0003-3794-2942 (R. Calegari)

colour saturation and integrating edge detection outputs. This method has shown efficacy in decreasing the performance disparity between different skin tones while maintaining overall diagnostic accuracy. Similarly, the FairSkin framework introduced in [4] leverages diffusion models to generate synthetic medical images that represent various skin tones equitably. Lastly, [5] and [6] propose innovative solutions to enhance fairness through structural model adjustments. The FairQuantize methodology employs weight quantization to adjust model performance across different demographics, and the channel pruning approach identifies and reduces bias by pruning channels that disproportionately affect specific demographic groups.

While the related works present innovative solutions for addressing bias and achieving acceptable accuracy in AI-based diagnostics for skin diseases, these solutions are still largely explorative and preliminary, rather than robust solutions to be applied in real-world scenarios. When applied to real-world scenarios, particularly employing non-dermoscopic images, they often yield unsatisfactory results [7]. When used in our specific scenario, the existing techniques still pose significant challenges that frequently lead to suboptimal outcomes if these techniques are applied in isolation [8]. It is worth emphasizing that the dataset used in our study introduces several unique challenges that must be responsibly addressed. The main challenges are related to *(1)* inherent dataset features (including its characteristics and variability), and (2) specific challenges related to the skewness of the available data, which significantly over-represent certain populations, thus inducing unfairness in the classification process (more details follow in the data description Section 2). Failing to meticulously study and address these issues within the development pipeline could lead to misdiagnoses, which in turn may exacerbate existing healthcare inequalities and result in adverse outcomes for affected patients. Such oversight highlights the critical need for rigorous evaluation and refinement of AI diagnostic tools to prevent potential harm and ensure their reliability and fairness across all populations.

Our work builds upon existing state-of-the-art foundational efforts, with the goal of addressing additional limitations in real-world, highly imbalanced datasets. We explicitly consider both classification performance and fairness metrics in our analysis. There are a few methods in the literature that aim at improving the fairness of non-dermoscopic image disease classification through the refinement of sophisticated Deep Learning (DL) models ([9, 6, 5, 1]) – our approach is *orthogonal* as we do not focus on the classification model itself but rather propose a pipeline for image data pre-processing and data augmentation that can *complement* any existing DL model for classification of skin diseases. In particular, our pre-processing technique employs the Individual Typology Angle (ITA) metric along with a novel thresholding method based on a Gaussian Mixture Model to accurately measure the skin tone depicted in each image. For data augmentation, we propose a novel combination of stable diffusion with DreamBooth to address the challenge of data scarcity, which is particularly acute for darker skin tones. To the best of our knowledge, this is the first work to consider using DreamBooth for generating skin disease images for different skin shades. Our pre-processing method can be affected by issues such as poor lighting and image blurriness, which may distort the perceived skin tone. To counteract these problems, we carefully hand-pick the images used for training DreamBooth, ensuring that they represent the skin tones targeted for augmentation. This meticulous selection process is especially crucial as only three out of the nine diseases catalogued in our dataset have examples of 'dark' and 'brown' skin, necessitating precise and representative training data to enhance model fairness and accuracy. The final step is the training of DL models for skin disease classification using pre-processed and augmented data. In the current study, we opted for two of the most efficient models currently available, namely the Swin Transformer (ST) and the Convolutional Neural Network (CNN); potentially, other DL approaches could be plugged in, according to the available resources and desired outcomes. The overall pipeline is illustrated in Fig. 1 and consists of the previously discussed preprocessing steps, plus the comparison of enhanced results via data augmentation. Please note that the proposed pipeline requires co-design and co-creation phases (especially in the selection phase during the pre-processing), during which stakeholders (in this case, doctors) shall be involved to assist in the selection and validation processes.

The paper is organized as follows: Sec. 2 introduces the use case and the data that we targeted; Sec. 3 describes the data-preprocessing technique and Sec. 4 explains the data augmenting procedure; then,

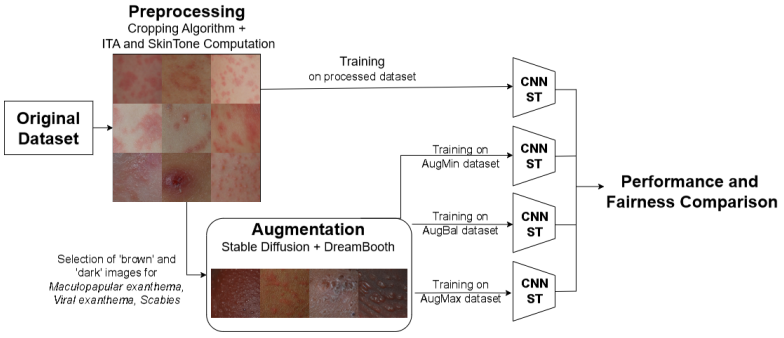

Figure 1: Diagram of our pipeline.

Sec. 5 shows the results of the whole pipeline (in terms of classification accuracy and fairness) once the classification models are inserted; finally, Sec. 6 concludes the paper.

2. Use Case and Dataset Description

The dataset consists of approximately 8,000 images of 273 pediatric patients at Sant'Orsola hospital in Bologna[1] representing nine possible skin diseases: *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*. The images were captured using consumer-grade cameras by the hospital's doctors, meaning they are **non-dermoscopic**. The dataset used in this study exhibits several critical characteristics that complicate the classification of skin diseases, primarily due to its inherent properties and the conditions under which the images were collected. Many of the images suffer from suboptimal lighting, causing skin tones to appear darker than they are, which not only complicates disease classification but also significantly affects the accuracy of skin tone assessments. This issue of misclassification is exacerbated by the high variability in the images' quality and size, as they were taken with different consumer-grade cameras. Such variability necessitates a comprehensive standardization process to make the dataset compatible for processing by neural networks. Additionally, the focus of the images is inconsistent—ranging from full-body shots to close-ups of specific affected areas. This variability presents further challenges in accurately identifying and analyzing disease-specific skin regions. Compounding these issues, some images exhibit blurriness, which diminishes the clarity and usefulness of the data for disease diagnosis. The dataset is further characterized by a notable scarcity of data, with only 273 patients represented, reducing the variability essential for a robust medical analysis. This limited data is particularly problematic for certain diseases, where only a few examples are available, skewing the class distribution and complicating the training of a reliable and generalizable model. These issues, coupled with the fact that the images were captured by medical professionals using consumer-grade cameras and are clinical rather than dermoscopic, introduce additional challenges. The photographs are prone to problems such as inconsistent lighting, blurriness, suboptimal angles, and other artefacts that negatively impact the quality and reliability of the data. These factors must be carefully managed to develop effective and accurate AI-based diagnostic tools. Moreover, the dataset presents significant challenges related to the representation of skin tones and disease classes. Predominantly, the dataset contains images of patients with lighter skin tones, which introduces a bias that complicates classification for less-represented skin tone categories. Additionally, some diseases are overrepresented in the dataset, leading to a class imbalance where the network is better at classifying certain illnesses over others. Although addressing class imbalance is not the primary focus of this work, it remains a critical aspect that influences the overall model performance.

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

3. Data preprocessing

The data preprocessing pipeline aims to standardize the dataset by generating uniformly sized image crops. The objective is to identify and extract regions of the images containing visible skin disease, using the binary mask associated with each image. The process follows a sliding-window approach and consists of several steps. Initially, the algorithm starts at the top-left corner and extracts a fixed-size patch measuring 256×256 pixels. Next, for each patch, a binary mask is used to calculate the disease coverage, defined as the ratio of positive labels (1) to negative labels (0) within the mask. This metric evaluates the presence of the skin disease based on contrast within the patch; patches are retained if the disease coverage exceeds a predefined threshold, indicating sufficient contrast, and discarded otherwise. The patch extraction process is repeated as the sliding window moves across the image in set steps, generating overlapping patches. To reduce redundancy, a non-maxima suppression procedure discards patches with lower disease coverage when overlap exceeds a threshold. Finally, patches exhibiting low contrast, such as those caused by poor illumination or blurriness, are removed to improve the overall quality of the dataset. This step ensures that only well-defined and informative regions are retained for further analysis. As expected, the preprocessing step reveals an inherent imbalance in the dataset across the different disease classes. In particular, certain diseases are underrepresented, with fewer than ten thousand examples available after preprocessing. This imbalance is anticipated to impact model performance, particularly for less-represented diseases.

3.1. Skin Tone Detection via ITA

To accurately measure fairness across different skin tones, it is essential to correctly classify each image according to the skin tone it represents. This approach allows then for precise evaluations for each skin tone, identifying any performance differences, such as the presence of bias. Skin tone classification is commonly performed using the Individual Typology Angle (ITA), a metric first introduced by Chardon et al. in 1991 [10], and widely adopted in subsequent studies for its simplicity and effectiveness [11, 12, 13]. While this method has proven effective in controlled environments, such as dermoscopic datasets, it assumes uniform illumination and does not account for variations introduced by pathological changes in the skin or external artefacts. We propose a modified ITA computation method *tailored* to our dataset, which includes images of skin conditions captured under non-standardized conditions with consumer-grade cameras. To address challenges such as altered pigmentation in the affected skin, inconsistent illumination, and shadows, we exclude disease-affected regions from ITA computation using segmentation maps, ensuring only unaffected skin is analyzed.

Unlike prior works relying on fixed thresholds from dermoscopic datasets [12, 13, 14], we classify ITA values into skin tone categories using a Gaussian Mixture Model (GMM), which better handles dataset variability. This refined method provides a more accurate representation of skin tone, enabling a fairer evaluation of classification performance. The computation of the ITA must account for the fact that skin affected by disease often appears darker and reddish compared to healthy skin. To ensure reliable ITA values that represent the baseline skin tone, it is key to exclude disease-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 regions 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 resembles a Gaussian distribution with a longer tail extending towards lower values. Following the computation of ITA values, ranges are required to classify skin tone according to the Fitzpatrick scale [11], which categorizes skin into six types. Various thresholding schemes have been proposed to map ITA values to Fitzpatrick skin types [12, 13, 14]. However, these ranges were primarily designed for dermoscopic datasets, devoid of variability caused by illumination, angulation, or other artefacts. Given the non-dermoscopic nature of our dataset, these thresholds were deemed unsuitable. Instead, we assumed that images with similar skin tones exhibit similar ITA values within a reasonable range of variation. To classify the ITA values, we fitted the distribution using a Gaussian Mixture Model with six components, corresponding to the six skin tone categories in the Fitzpatrick scale. Each ITA

value was assigned to the Gaussian component that best represented its value. The resulting skin tone labels were categorized as *dark*, *brown*, *tan*, *intermediate*, *light*, and *very light*. Examples of the automatic skin tone classification are presented in Figure 2. While the ITA value is generally robust, shadows and poor illumination can lower the ITA value, resulting in a darker assigned skin tone. Nonetheless, darker images—whether due to actual skin tone or suboptimal lighting—were correctly assigned a lower ITA value, whereas lighter images were assigned higher ITA values. The distribution of skin tone labels across the dataset shows that the *dark* and *brown* skin tone categories are underrepresented, highlighting an imbalance in skin tone distribution within the dataset. Despite the robustness of the ITA calculation, this labelling process is not entirely accurate. Poor illumination or other artefacts cause the computed ITA value to deviate from the expected value for the true skin tone for a non-negligible number of images. Future work could address this limitation by incorporating advanced correction techniques for artefacts such as shadows and uneven lighting.

Figure 2: Examples of automatic skin tone classification based on the ITA values and the Gaussian mixture technique.

4. Synthetic Generation of Skin Disease Images

In this section, we describe the process used to generate synthetic images. We first explain how the DreamBooth model has been tailored to the specific use case, and its extreme imbalance. Then, we introduce three approaches for incorporating the synthetic images into training sets to be used for the downstream task of classifying the diseases via a DL model (described in Sec. 5).

4.1. Image Generation via the Combination of DreamBooth and Stable Diffusion

The dataset used in this study contains a limited number of examples of skin diseases affecting individuals with black skin, with at most 4 or 5 individuals with dark skin for each disease. The preprocessing pipeline described in Section 3 generates a large number of image crops from photographs of the same individual. However, using all these crops to train an image generation model would be redundant, as the crops originating from the same individual are highly similar to one another. Consequently, it is sufficient to select only a few representative crops (3 or 4) per individual with dark skin and construct a small, curated dataset comprising multiple individuals with dark skin exhibiting the specific disease of interest. This curated dataset can then be used to train an image generation model. One model well-suited for training with such a limited number of examples is DreamBooth, introduced by [15]. It is a fine-tuning technique for generative models, like diffusion-based ones, enabling the creation of high-quality, subject-specific images from just a few samples. This approach not only personalizes the model but also maintains its capacity to produce diverse and photorealistic outputs, making it ideal for scenarios constrained by data scarcity. In our work, DreamBooth was employed to fine-tune a pre-trained Stable Diffusion model, which was pre-trained on images of size 512×512. The fine-tuning process was divided into the following stages:

1. *Exploration of the Dataset* – A manual inspection of the dataset was conducted to identify the skin diseases for which images of 'dark' or 'brown' skin types were available. This exploration revealed that only three out of the nine diseases—*maculopapular exanthema, viral exanthema*, and *scabies*—contained images of individuals with 'dark' and 'brown' skin. Consequently, image generation was applied only to these three diseases.

2. *Construction of Mini-Datasets* – Mini-datasets were manually constructed for each of the three diseases, separately for 'brown' and 'dark' skin types. This process resulted in six datasets (two for each disease: one for 'brown' skin and one for 'dark' skin), with each dataset containing between 14 and 29 images.
3. *Fine-Tuning with DreamBooth* – For each of the six mini-datasets, the Stable Diffusion model was fine-tuned using the DreamBooth technique. A grid search was conducted to identify optimal hyperparameter configurations, exploring the following parameters:
    - *Learning rate*: Values of 5e-7, 2e-6, 5e-6, and 1e-5 were tested, to ensure adequate exploration of the parameter space, as the learning rate is a critical factor for convergence.
    - *Maximum training steps*: For mini-datasets with fewer than 15 images, values of 1000, 2000, 3000, and 4000 steps were tested. For mini-datasets with more than 15 images, values of 2000, 3000, 4000, and 5000 steps were tested. This choice was guided by a commonly applied rule of thumb in DreamBooth, which recommends fine-tuning with at least 100 training steps per image.
    - *Instance prompt*: The instance prompt in DreamBooth plays a key role in both the training and image generation phases. During training, a unique identifier (e.g., `<unique_ID>`) is included in the prompt alongside descriptive context (e.g., "human skin" or "a person with a skin condition") to associate the fine-tuned model with the specific features of the training images. This enables the model to learn how to reproduce those features while maintaining its broader generative capabilities. During image generation, the instance prompt is used to guide the model in synthesizing new images that reflect the characteristics of the fine-tuned training data. By combining or modifying the instance prompt with additional textual descriptions, keeping the `<unique_ID>` in the text, it is possible to control the specific details of the generated images, ensuring alignment with the desired output while retaining diversity and realism. In our case, both the prompt "`<unique_ID>`" and "`<unique_ID>` human skin" were evaluated. Including 'human skin' in the prompt was hypothesized to provide context and aid in accurately reproducing skin texture.

    A batch size of 1 was selected, as experiments revealed that smaller batch sizes promoted greater diversity when other hyperparameters were held constant.
4. *Model Selection* – For each of the six mini-datasets, the fine-tuned models with the most promising hyperparameter configurations were selected based on an empirical evaluation of the generated images. The evaluation prioritized diversity, accuracy, faithful representation of skin texture and colour, and fidelity to the real images in the mini-dataset.

In Figure 3 we can see a sample of the results obtained through the synthetic generation of images, in particular for images of the class 'brown' and 'dark'. We report the results for the three target diseases (scabies, viral exanthema, and maculopapular exanthema); for each disease, we show some real images and some generated ones. The synthetic images are extremely realistic and the skin tone matches the desired one.

4.2. Data Augmentation for Rare-colour Skin Images

After determining the required number of synthetic images for each of the three diseases and the two skin colours ("dark" and "brown"), we distributed this total among the fine-tuned models designated for each disease and skin colour. This approach ensured that the synthetic images were generated by models fine-tuned with various hyperparameter combinations, enhancing the diversity of the dataset. Models fine-tuned with different hyperparameters typically produce images with distinct characteristics, reflecting the variations in the mini-datasets used for training. Furthermore, to increase the diversity of the synthetic images, we frequently changed the generation seed.

To incorporate the synthetic images into the original training set—while keeping the test and validation sets comprised solely of real images—to provide more examples of diseases on darker skin tones, we followed and compared three distinct numerical approaches:

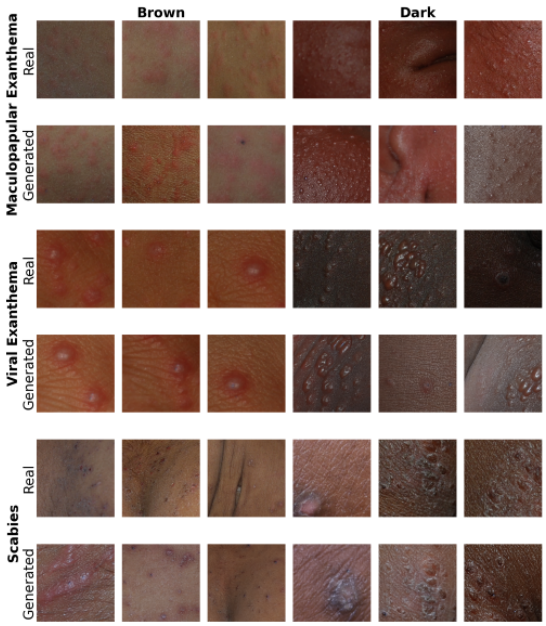

Figure 3: Real vs. generated images for each of the three target diseases using the DreamBooth technique.

1. AugMin– synthetic images of 'dark' and 'brown' skin are added to ensure that the total number of images (real + synthetic) for each disease matches the smallest image count among the other four skin colours ('very light,' 'light,' 'intermediate,' 'tan'). For example, if the 'tan' skin color has the fewest examples for scabies, with x images, then an equal number of synthetic images for 'dark' and 'brown' skin are added to reach a total of x images for these colors as well.
2. AugBalanced– synthetic images are added for 'dark' and 'brown' skin tones for each disease such that the number of images for each of these two colours represents approximately one-sixth (approximately 17%) of the total images for that disease. This strategy aims to achieve a more balanced distribution across all skin colours and diseases.
3. AugMax– similar to the first approach, but in this case, the total number of images (real + synthetic) for each disease and each of the two skin colours ('brown' and 'dark') was adjusted to match the largest number of images among the other four skin colours ('very light,' 'light,' 'intermediate,' 'tan') for that disease.

These three distinct approaches help evaluate the impact of including varying proportions of 'dark' and 'brown' skin images. This enables a constructive analysis of how representing underrepresented skin tones in the training set affects model performance and fairness outcomes. Comparing these proportions is particularly valuable for understanding the trade-off between fairness and performance and identifying the balance that optimizes equitable representation across skin tones with high predictive accuracy.

5. Disease Classification with DL Models

The dermatological classification task demands that the model captures complex features across different scales. For this reason, we have selected two of the best state-of-the-art models, the Convolutional Neural Network (CNN) and the Swin Transformer (ST). We start by measuring the performance of these models *without* data augmentation, to serve as a baseline. Since performance results offer limited insight into fairness concerns in the model's diagnostic outcomes, we evaluate the models using fairness metrics common in the literature and relevant to tasks like skin disease prediction [9], specifically reporting the Disparate Impact Ratio (*DI*) [16], Equalized Odds Ratio (*EOR*) [17], and Predictive Rate Ratio (*PRR*). Fairness metrics are often developed for binary outcome tasks, whereas the classification

Table 1
CNN accuracy and F1-score: disease aggregation.

	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.	71.3%	76.8%	0.73	0.78	76.5%	80.5%	0.77	0.80	**77.8%**	**82.6%**	**0.79**	**0.83**	76.5%	81.2%	0.78	**0.83**
MP ex.	57.7%	49.0%	0.65	0.55	**70.5%**	57.2%	**0.73**	0.61	69.3%	58.7%	**0.73**	0.63	68.5%	**59.0%**	**0.71**	0.63
MF ex.	71.5%	74.1%	0.74	0.78	75.2%	77.5%	0.79	0.82	**76.4%**	80.0%	**0.82**	**0.84**	73.3%	**80.7%**	0.79	**0.84**
PM ex.	63.8%	52.3%	0.69	0.59	69.9%	59.1%	0.73	0.64	**74.7%**	**62.6%**	**0.77**	0.66	71.2%	61.6%	0.75	**0.67**
V ex.	80.9%	76.8%	0.79	0.75	81.4%	78.8%	0.82	0.79	80.5%	79.4%	0.82	**0.80**	**83.2%**	**79.8%**	**0.83**	**0.80**
urticaria	85.9%	85.6%	0.84	0.83	88.4%	88.4%	0.85	0.86	**89.2%**	89.2%	**0.86**	**0.87**	88.9%	**89.4%**	**0.86**	**0.87**
pediculosis	62.4%	65.4%	0.68	0.69	57.7%	70.9%	0.63	0.74	**68.6%**	**74.8%**	0.71	**0.76**	67.5%	72.0%	**0.73**	**0.76**
scabies	74.1%	69.8%	0.75	0.72	79.1%	75.3%	0.79	0.77	**81.3%**	**77.6%**	**0.81**	**0.79**	78.4%	75.5%	**0.81**	0.78
chickenpox	54.3%	57.0%	0.61	0.61	50.9%	60.7%	0.59	0.66	52.0%	62.3%	0.61	0.69	**62.4%**	**65.5%**	**0.66**	**0.70**
All	78.2%	76.8%	0.72	0.70	81.1%	80.2%	0.75	0.74	82.0%	**81.5%**	**0.77**	**0.76**	**82.1%**	81.4%	**0.77**	**0.76**

Table 2
CNN Accuracy e F1-score: skin tones aggregations.

		No synthetic augmentation		AugMin		AugBalanced		AugMax	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Minority	dark	78.4%	0.74	81.6%	0.77	**82.7%**	**0.79**	81.9%	**0.79**
	brown	78.2%	0.71	81.0%	0.74	81.8%	**0.76**	**82.1%**	**0.76**
Majority	tan	75.5%	0.69	79.3%	0.73	**80.5%**	0.75	**80.5%**	**0.76**
	intermediate	75.1%	0.68	78.7%	0.73	**80.0%**	**0.75**	79.6%	**0.75**
	light	76.5%	0.68	79.7%	0.73	**81.0%**	**0.74**	80.9%	**0.74**
	very light	82.0%	0.75	84.9%	0.79	86.2%	**0.81**	**86.3%**	**0.81**
All		77.0%	0.77	80.3%	0.80	**81.6%**	**0.81**	81.5%	**0.81**

task in this study is multi-class and involves more than two demographic groups. To adapt these metrics to our setting, skin tones were aggregated 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 key considerations: *(1)* the observed underrepresentation of "dark" and "brown" skin tones in the dataset and *(2)* the adoption of a similar approach by [9]. This aggregation facilitates a meaningful application of fairness metrics while addressing the challenges posed by a multi-class, multi-group setting. Note that this aggregation did not require re-training, as the model is blind to the "skin tone" attribute during training. It affects only the evaluation process.

5.1. Convolutional Neural Network - No Data Augmentation

A deep CNN architecture comprising five convolutional layers, each with a kernel size of 3, was selected to address this. After each convolutional block, a MaxPooling layer with a kernel size of 2 is applied. A final fully connected layer outputs nine logits corresponding to the nine target classes. The CNN architecture was selected after a preliminary empirical evaluation, and its hyperparameters were hand-tuned. The dataset of cropped images was partitioned into training (60% of the samples), validation (20%), and test sets (20%). A hyperparameter tuning phase was conducted to optimize the batch size and learning rate. Six combinations of these parameters were evaluated, with the model trained for 5 epochs using stochastic gradient descent (SGD) with momentum as the optimizer. Optimal performance was achieved with a batch size of 128 and a learning rate of 0.01. For the final model training, the configuration included the following settings: batch size = 128, learning rate = 0.01, number of epochs = 15, and optimizer = SGD with momentum. Additionally, a cosine decay learning rate scheduler and an early stopping mechanism were employed to prevent unnecessary resource usage in cases of early overfitting.

Classification results. The model was evaluated using standard performance metrics, specifically F1 score and Accuracy. Results aggregated for each disease and for each skin tone are presented in the first column of respectively Table 1 and Table 2.

Accuracy results aggregated for skin tones indicate consistent performance across most skin tones, except for the "very light" category, which exhibited significantly higher accuracy. This discrepancy may be attributed to the higher overall quality and better illumination of "very light" samples, rendering them easier to classify. The model's performance lacks consistency when evaluated separately for each disease. Specifically, for approximately half of the diseases, the Accuracy and F1 score on the test set

Table 3
CNN DI, EOR and PRR: disease aggregation.

	No synthetic augmentation			AugMin			AugBalanced			AugMax		
	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR
DII ex.	**0.99**	0.78	0.93	0.97	**0.93**	**0.95**	0.98	0.83	0.94	**0.99**	0.76	0.94
MP ex.	**1.35**	0.85	1.18	1.44	0.81	1.23	1.46	0.85	1.18	1.44	**0.86**	**1.16**
MF ex.	1.32	0.55	**0.96**	1.26	0.68	0.97	1.21	**0.82**	0.95	**1.19**	0.71	0.91
PM ex.	**0.85**	0.63	1.22	**0.85**	**0.65**	1.18	0.82	0.56	1.19	0.83	0.61	**1.16**
V ex.	0.98	0.82	1.05	0.95	0.73	1.03	0.95	0.80	**1.01**	**0.99**	**0.86**	1.04
urticaria	0.93	0.86	**1.00**	**0.95**	**0.97**	**1.00**	**0.95**	**0.97**	**1.00**	0.94	0.93	0.99
pediculosis	0.74	0.68	**0.95**	0.73	0.81	0.81	**0.75**	**0.84**	0.92	0.74	0.73	0.94
scabies	1.46	0.67	1.06	1.44	0.68	1.05	1.43	0.69	1.05	**1.35**	**0.91**	**1.04**
chickenpox	0.76	0.70	**0.95**	0.71	0.75	0.84	0.73	0.83	0.83	**0.84**	**0.95**	**0.95**
All	1.04	0.73	1.03	0.93	0.78	**1.01**	**1.03**	**0.88**	**1.01**	**1.03**	0.81	**1.01**

are higher for the Majority group than for the Minority group, contrary to the hypothesis that the Minority group would be systematically disadvantaged. Two potential explanations may account for this trend in traditional metrics: *(1)* factors such as poor illumination, body hair, or artefacts may lead to *misclassification of skin tone* in some images, causing certain images to be incorrectly categorized into the Minority group rather than the Majority group. This misclassification complicates the reliable assessment of bias; *(2)* a *lack of variability between the training set and the test set* for the "black" and "brown" skin categories (particularly for the former) may unintentionally inflate classification performance. For example, in the case of *"black"* skin tone and the disease *maculopapular exanthema*, we observe a remarkably higher *Accuracy* and *F1* score for the *Minority* group. Upon closer qualitative analysis, we find that the dataset includes only one individual with *"black"* skin tone and this disease. As described in Section 3, the cropping algorithm generates multiple image crops from a single individual, distributing them across the training, validation, and test sets. During training, the model learns to classify these crops effectively, and at test time, it encounters test crops highly similar to those seen during training. This results in an artificially inflated *Accuracy* for the *"black"* skin tone in this specific disease. We hypothesize that if the test set contained images of other individuals with *"black"* skin tone who were not seen during training, the model's performance would decrease significantly. In contrast, the Majority group likely benefits from greater diversity in the training data. The model is exposed to a wide variety of images from different individuals, enabling it to generalize better when presented with test crops from unseen Majority group individuals, resulting in more robust performance.

However, accuracy alone does not reveal the distribution of misclassifications across skin tones. As for fairness considerations, Table 3 summarizes the results, discussed in detail in the following.

The *DI* was calculated separately for each condition, where $\hat{Y} = 1$ represents the presence of the disease and $\hat{Y} = 0$ its absence. A value between 0.8 and 1.25 is generally considered fair. Values below 0.8 indicate unfairness against the minority group, whereas values above 1.25 suggest unfairness against the majority group. Notably, the model demonstrates *significant bias against the minority group* for diseases such as pediculosis and chickenpox, as evidenced by *DI* values below 0.8. This disparity may be attributed to the limited number of positive samples from the minority group for these conditions. In contrast, for diseases such as maculopapular rash, morbilliform rash, and scabies, there is a proportionally higher number of positive detections in the minority group compared to the majority group, resulting in *DI* values above 1.25. These observations highlight the varying degrees of fairness across different conditions and the impact of sample imbalances in fairness evaluations. As for *EOR*, in our experiments, *EOR* is computed for each disease, using the common division of skin tones into minority and majority groups. The results show that of the nine EOR values, only three, corresponding to *maculopapular exanthema*, *viral exanthema* and *urticaria* fall within the fairness range, indicating that the model's performance is not consistent across different demographic groups. In contrast to *DI* and *EOR* results, we observe that the *PRR* values are fair across all diseases. To understand this difference, it is important to note that the PPV (used to compute the *PRR*) relative to a group measures how often the model correctly predicts the positive class for that group. In this sense, PPV serves as a measure of the *quality* of predictions. On the other hand, the *DI* focuses on the probability of a positive prediction for each group, regardless of its correctness, making it a measure of *quantity*. Similarly, the *EOR* evaluates the True Positive Rate (recall) and the False Positive Rate, which also reflect the

Table 4
ST accuracy and F1-score: disease aggregation.

	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.	94.7%	96.8%	0.90	0.93	95.8%	94.8%	**0.94**	**0.94**	**96.3%**	**97.1%**	0.90	0.91	94.4%	96.2%	0.93	0.93
MP ex.	88.4%	88.6%	**0.84**	**0.84**	87.5%	**89.5%**	**0.84**	0.83	88.1%	87.4%	**0.84**	**0.84**	**91.8%**	86.5%	**0.84**	**0.84**
MF ex.	89.7%	**95.1%**	0.89	**0.92**	**96.4%**	93.9%	0.89	0.86	95.2%	93.9%	0.89	0.86	**96.4%**	91.8%	**0.91**	0.85
PM ex.	**91.7%**	87.1%	**0.92**	**0.89**	**91.7%**	**87.5%**	**0.92**	0.88	84.3%	84.4%	0.87	0.86	**91.7%**	86.5%	0.89	0.86
V ex.	95.1%	**95.6%**	0.90	**0.90**	95.9%	94.9%	0.90	**0.90**	95.1%	95.5%	0.88	0.88	**96.1%**	95.5%	**0.91**	0.89
urticaria	**91.9%**	**88.8%**	**0.93**	**0.93**	90.4%	88.1%	**0.93**	0.92	86.3%	84.7%	0.90	0.90	90.2%	86.9%	**0.93**	0.91
pediculosis	84.0%	88.2%	0.89	0.92	**86.6%**	**89.4%**	**0.90**	**0.93**	83.5%	89.1%	0.88	**0.93**	85.6%	87.3%	0.89	0.92
scabies	81.9%	**89.0%**	0.90	**0.93**	84.1%	87.6%	**0.91**	0.92	85.2%	88.2%	**0.91**	0.91	**85.7%**	87.8%	**0.91**	0.92
chickenpox	**88.4%**	91.0%	**0.89**	0.86	88.4%	**92.7%**	**0.89**	**0.87**	85.0%	88.6%	0.88	0.86	83.2%	87.7%	0.87	**0.87**
All	91.1%	**91.3%**	**0.90**	**0.90**	91.3%	90.7%	**0.90**	0.89	89.3%	89.4%	0.89	0.89	**91.4%**	90.1%	**0.90**	0.89

Table 5
ST DI, EOR and PRR: disease aggregation.

	No synthetic augmentation			AugMin			AugBalanced			AugMax		
	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR
DII ex.	1.01	0.72	0.98	1.01	0.87	**1.01**	**1.00**	**0.88**	**0.99**	0.94	0.70	0.98
MP ex.	1.36	0.74	**1.00**	**1.28**	**0.93**	0.98	1.41	0.65	1.01	1.52	0.56	1.06
MF ex.	1.21	0.76	0.94	**1.17**	0.91	0.98	1.22	**0.95**	**1.01**	1.21	0.77	1.05
PM ex.	0.81	0.73	1.05	0.79	0.54	1.05	0.76	0.58	**1.00**	**0.83**	**0.79**	1.06
V ex.	0.96	0.91	0.99	**0.98**	0.91	1.01	0.97	**0.95**	**1.00**	0.95	0.76	1.01
urticaria	**0.99**	0.74	1.03	0.98	0.78	1.03	0.97	0.75	**1.02**	0.97	**0.93**	1.04
pediculosis	0.78	0.61	0.95	0.78	**0.88**	0.97	0.76	0.76	0.94	**0.82**	0.43	**0.98**
scabies	**1.23**	0.10	0.92	1.28	0.40	0.96	1.28	0.45	0.96	1.32	**0.91**	**0.98**
chickenpox	0.72	0.35	**0.97**	0.71	0.40	0.95	0.72	0.35	0.96	**0.73**	**0.42**	0.95
All	1.01	0.63	0.98	**1.00**	**0.74**	0.99	1.01	0.70	**1.00**	1.03	0.70	1.01

Table 6
ST Accuracy and F1-score: skin tones aggregation

		No synthetic augmentation		AugMin		AugBalanced		AugMax	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Minority	dark	90.8%	0.91	**91.5%**	**0.92**	89.9%	0.90	**91.5%**	0.91
	brown	91.2%	**0.91**	91.3%	0.90	89.2%	0.88	**91.4%**	0.90
Majority	tan	**91.9%**	**0.92**	91.7%	0.90	90.3%	0.89	91.7%	0.90
	intermediate	**91.7%**	**0.92**	91.0%	0.90	90.2%	0.89	90.5%	0.89
	light	**90.3%**	**0.90**	89.4%	0.87	88.2%	0.87	88.1%	0.86
	very light	**91.5%**	**0.92**	91.1%	0.90	88.5%	0.89	90.1%	0.89
All		**91.3%**	0.90	90.8%	**0.91**	89.4%	0.89	90.2%	0.90

quantity of True Positives and False Positives identified by the model. Therefore, while *PRR* compares the model's precision across groups, the other metrics (*DI* and *EOR*) assess the distribution and balance of predictions among groups.

5.2. Swin Transformer - No Data Augmentation

As the second model for our study, we utilized a Swin Transformer [18]. During training, the same dataset partitioning used for the CNN was adopted: 60% for training, 20% for validation, and 20% for testing. To effectively capture the skin texture and disease-specific characteristics, a model variant pre-trained on ImageNet-1k and later fine-tuned on a skin cancer dataset [2] was employed. The last three stages of the ST were fine-tuned, while the weights of the first stage were frozen, resulting in a total of 26 million trainable parameters. To ensure convergence and fully explore the weight space, hyperparameter tuning focused on learning rate values, specifically 1e-5, 1e-4, 1e-3, and 1e-2. While lower learning rates ensured good convergence, they often led the model to converge to local minima, resulting in low accuracy and F1 scores. Consequently, a learning rate of 1e-2 was selected, enabling larger training steps. To stabilize training in the later epochs, the learning rate was reduced by a factor of 100 after nine epochs, based on the observed loss trends.

Classification results. The final performance results are shown in the first column of Tables 4, 5 and 6 aggregated by disease and skin tones respectively.

First, we notice a significant performance improvement compared to the CNN, likely due to the remarkably higher capacity of the ST (~26 million trainable parameters versus the ~4 million of the

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

CNN). Moreover, it is evident that, while for the CNN the *Accuracy* and *F1* scores vary significantly between demographic groups depending on the disease, the ST shows more consistent Accuracy and F1-score values. The *DI* values of the ST indicate that, on average, it demonstrates greater fairness compared to the CNN. Specifically, the Swin Transformer exhibits only three out of nine instances of unfair values for the DI metrics (Table 5), in contrast to the CNN, which presents five out of nine instances of unfair values for the same metrics. On the other hand, *EOR* values are systematically lower in the Swin Transformer results compared to those of the CNN: of the nine *EOR* values, only one—specifically the one corresponding to *viral exanthema*—falls within the fairness range, once again indicating that the model's predictions are strongly influenced by the *skin tone* attribute. As for the *PRR* values, they remain within acceptable limits, indicating that the model's precision is comparable for both the Minority and Majority categories. However, as already stated at the end of Section 5.1, this does not imply a fair prediction across the various skin tones, as evidenced by the values of the other metrics.

5.3. Convolutional Neural Network - With Data Augmentation

The addition of synthetic images generally resulted in a significant performance improvement across all diseases[3], as evidenced by the values of Table 1. This effect may be attributed to the regularizing impact of these new data on the dataset, which benefited all diseases. Furthermore, Accuracy and F1-score also improved across individual skin tones, including both darker and lighter tones, for which no synthetic images were generated. Overall, except for the 'very light' skin tone, the addition of synthetic data helped equalize performance across skin tones, raising metrics for lighter tones (which previously had lower Accuracy and F1 scores compared to 'dark' and 'brown' tones) more than it did for darker tones. Regarding fairness metrics, synthetic data also benefited diseases for which no synthetic images were generated. We provide now a more detailed analysis of each augmentation technique.

AugMin adds the fewest synthetic images and provides the smallest improvement in terms of both traditional and fairness metrics, suggesting that more synthetic data could be beneficial. As shown in Table 1, the Accuracy and F1-score improvements for individual diseases sometimes narrowed the performance gap between Minority and Majority groups (e.g., in the case of *drug-induced iatrogenic exanthema, morbilliform exanthema, polymorphous exanthema, viral exanthema, urticaria*, and *scabies*), while in other cases, the performance gap widened. Regarding fairness metrics (Table 3), no significant improvement was observed for the three diseases targeted with synthetic images in 'dark' and 'brown' tones, and in some cases, a decline was noted. Interestingly, however, certain diseases for which no synthetic images were generated showed counterintuitive fairness improvements. In summary, this approach appears to function as a regularizer that enhances overall performance and improves the homogeneity of model performance across skin tones. However, it is not effective in improving classification fairness, particularly for the targeted diseases (i.e., *maculopapular exanthema, viral exanthema*, and *scabies*). AugBalanced outperformed the previous approach in terms of both Accuracy and F1-score. In this case, fairness outcomes for the three targeted diseases were very similar to those observed without synthetic data. However, for most other diseases, fairness appeared to improve, particularly for *EOR* values. Overall, this demonstrates that a greater presence of synthetic data has a stronger regularizing effect on performance, benefiting nearly all diseases and all skin tones. AugMax yielded the best trade-off between fairness and performance. Accuracy and F1-score metrics remained higher than the model trained on the original dataset, while the average fairness metrics for each disease fell within ranges considered fair. Significant improvements were observed for both *DI* and *EOR* values in two of the three diseases for which synthetic images were generated (i.e., *viral exanthema* and *scabies*). Additionally, for most other diseases, fairness metrics also improved. For the CNN, generating synthetic images for the three targeted diseases and incorporating them into the dataset proved beneficial for both overall model performance and classification fairness. The more synthetic images, the merrier: AugMax achieves the best trade-off between fairness and performance.

[3]Including those for which no synthetic data was added

5.4. Swin Transformer - With Data Augmentation

The ST model has already demonstrated very high accuracy and F1-score values across all diseases and skin tones, although *EOR* values were notably problematic, especially for the latter diseases. With augmentation, AugMin improved accuracy and F1-scores for several diseases in minority categories (i.e., 'dark' and 'brown' skin tones), although at a slight cost to the majority category. Overall, the classification performance remained comparable to the model trained on the original dataset across all skin tones and diseases. In terms of fairness, AugMin led to significant improvements, particularly in *EOR* values; moreover, *DI* values improved for two of the three target diseases, and *PRR* values also showed improvement. Compared to other approaches, AugBalanced reduces accuracy and F1-score values. However, it shows notable improvements in *EOR* values, with six out of nine improving, although at the expense of two deteriorating compared to the model trained on the original dataset. *DI* values worsened, while *PRR* values improved. On the other hand, AugMax maintains good accuracy and F1-score values, though distributed differently compared to AugMin. In terms of fairness metrics, this approach performs well for *DI* and *PRR* values, though it does not substantially improve over the model trained on the original dataset, except for *pediculosis*. However, it is less effective for *EOR* values, except for *urticaria* and *scabies*, where it achieves fairness range values. In conclusion, for the ST model, AugMin proved to be the most effective.

This outcome can be attributed to at least two factors: (1) Although the ST model has significantly more trainable parameters than the CNN, it is pre-trained, which makes it inherently more resistant to substantial changes. In contrast, the CNN is trained entirely from scratch, providing greater flexibility for performance improvements. This explains why the impact of synthetic images is more pronounced with the CNN; (2) the ST had already achieved high accuracy and F1-score during initial training without synthetic data, showing fewer signs of overfitting compared to the CNN. Therefore, synthetic images did not produce the same regularizing effect on the ST as on the CNN, where there was more room for improvement. This also clarifies why the ST benefited more from an approach involving fewer synthetic data: adding more data likely pushed the model beyond its 'saturation point,' thus limiting the desired improvements in fairness, although it still managed to deliver better fairness results than the same model trained on the original dataset.

6. Conclusion

This study demonstrated the effectiveness of using advanced image generation techniques, like DreamBooth combined with stable diffusion, to enhance the representation of underrepresented skin tones in medical datasets. Our methods significantly improved fairness metrics, balancing performance and fairness effectively. Incorporating synthetic images, especially in the training sets for diseases affecting 'dark' and 'brown' skin tones, addressed data scarcity issues and reduced bias in medical image analysis. The comparison of different data augmentation strategies (AugMin, AugBalanced, AugMax) helped us understand the trade-offs between dataset diversity and predictive accuracy. While the CNN showed more significant improvements due to its flexibility, the pre-trained nature of the ST limited its adaptability to synthetic data enhancements. However, both models benefited from our approach, underscoring the potential of synthetic data to improve diagnostic tools across diverse skin tones. We also noticed that although synthetic images are produced only for specific diseases, the experimental results demonstrate enhanced performance across all nine diseases catalogued in our study. Our findings advocate for the continued use of synthetic data augmentation to enhance fairness and performance in dermatological AI applications, paving the way for more equitable healthcare solutions.

A potential extension of this work could involve generating images of diseases on dark skin by adapting images of diseases from lighter skin tones. This approach would create more examples of diseases of dark skin, which are currently underrepresented. However, this method must be approached with caution, as the texture and appearance of dermatological conditions can vary significantly between different skin tones, potentially affecting the scientific accuracy of the generated images. This careful

consideration is essential to ensure the development of AI-based diagnostic tools that are both effective and equitable across diverse populations.

References

[1] C.-H. Chiu, Y.-J. Chen, Y. Wu, Y. Shi, T.-Y. Ho, Achieve fairness without demographics for dermatological disease diagnosis, Medical Image Analysis 95 (2024) 103188. URL: https://www.sciencedirect.com/science/article/pii/S1361841524001130. doi:https://doi.org/10.1016/j.media.2024.103188.

[2] Aayushman, H. Gaddey, V. Mittal, M. Chawla, G. R. Gupta, Fair and accurate skin disease image classification by alignment with clinical labels, 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, Springer Nature Switzerland, Cham, 2024, pp. 394–404.

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

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

[5] Y. Guo, Z. Jia, J. Hu, Y. Shi, FairQuantize: Achieving Fairness Through Weight Quantization for Dermatological Disease Diagnosis , in: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, volume LNCS 15010, Springer Nature Switzerland, 2024.

[6] Q. Kong, C.-H. Chiu, D. Zeng, Y.-J. Chen, T.-Y. Ho, J. Hu, Y. Shi, 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, Springer Nature Switzerland, Cham, 2024, pp. 24–34.

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

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

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

[10] A. Chardon, I. Cretois, C. Hourseau, Skin colour typology and suntanning pathways, International Journal of Cosmetic Science 13 (1991). URL: https://api.semanticscholar.org/CorpusID:25650931.

[11] V. Gupta, V. K. Sharma, Skin typing: Fitzpatrick grading and others, Clinics in Dermatology 37 (2019) 430–436. URL: https://www.sciencedirect.com/science/article/pii/S0738081X1930121X. doi:https://doi.org/10.1016/j.clindermatol.2019.07.010, the Color of Skin.

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

[13] N. M. Kinyanjui, T. Odonga, et al., Fairness of classifiers across skin tones in dermatology, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham, 2020, pp. 320–329.

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

[15] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-

image diffusion models for subject-driven generation, 2023. URL: https://arxiv.org/abs/2208.12242. arXiv:2208.12242.

[16] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, 2015. URL: https://arxiv.org/abs/1412.3756. arXiv:1412.3756.

[17] A. Agarwal, A. Beygelzimer, et al., A reductions approach to fair classification, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 60–69. URL: https://proceedings.mlr.press/v80/agarwal18a.html.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows (2021) 10012–10022.