CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions

Arezou Pakzad[1], Kumar Abhishek[1], and Ghassan Hamarneh[1]

School of Computing Science, Simon Fraser University, Canada
{arezou_pakzad, kabhishe, hamarneh}@sfu.ca

Abstract. While deep learning based approaches have demonstrated expert-level performance in dermatological diagnosis tasks, they have also been shown to exhibit biases toward certain demographic attributes, particularly skin types (e.g., light versus dark), a fairness concern that must be addressed. We propose CIRCLe, a skin color invariant deep representation learning method for improving fairness in skin lesion classification. CIRCLe is trained to classify images by utilizing a regularization loss that encourages images with the same diagnosis but different skin types to have similar latent representations. Through extensive evaluation and ablation studies, we demonstrate CIRCLe's superior performance over the state-of-the-art when evaluated on 16k+ images spanning 6 Fitzpatrick skin types and 114 diseases, using classification accuracy, equal opportunity difference (for light versus dark groups), and normalized accuracy range, a new measure we propose to assess fairness on multiple skin type groups. Our code is available at `https://github.com/arezou-pakzad/CIRCLe`.

Keywords: Fair AI · Skin Type Bias · Dermatology · Classification · Representation Learning.

1 Introduction

Owing to the advancements in deep learning (DL)-based data-driven learning paradigm, convolutional neural networks (CNNs) can be helpful decision support tools in healthcare. This is particularly true for dermatological applications where recent research has shown that DL-based models can reach the dermatologist-level classification accuracies for skin diseases [10, 20, 24] while doing so in a clinically interpretable manner [7, 38]. However, this data-driven learning paradigm that allows models to automatically learn meaningful representations from data leads DL models to mimic biases found in the data, i.e., biases in the data can propagate through the learning process and result in an inherently biased model, and consequently in a biased output.

Most public skin disease image datasets are acquired from demographics consisting primarily of fair-skinned people. However, skin conditions exhibit vast visual differences in manifestations across different skin types [56]. Lighter skinned populations suffer from over-diagnosis of melanoma [2] while darker skinned patients get diagnosed at later stages, leading to increased morbidity and mortality [4]. Despite this, darker skin is under-represented in most publicly available data sets [31, 36], reported studies [16], and in dermatology textbooks [3]. Kinyanjui et al. [31] performed an analysis on two popular benchmark dermatology datasets: ISIC 2018 Challenge dataset [13] and SD-198 dataset [53], to understand the skin type representations. They measured the individual typology angle (ITA), which measures the constitutive pigementation of skin images [43], to estimate the skin tone on these datasets, and found that the majority of the images in the two datasets ITA values between 34.8° and 48°, which are associated with lighter skin. This is consistent with the under-representation of darker skinned populations in these datasets. It has been shown that CNNs perform best at classifying skin conditions for skin types that are similar to those they were trained on [23]. Thus, the data imbalance across different skin types in the majority of the skin disease image datasets can manifest as racial biases in the DL models' predictions, leading to racial disparities [1]. However, despite these well-documented concerns, very little research has been directed towards evaluating these DL-based skin disease diagnosis models on diverse skin types, and therefore, their utility and reliability as disease screening tools remains untested.

Although research into algorithmic bias and fairness has been an active area of research, interest in fairness of machine learning algorithms in particular is fairly recent. Multiple studies have shown the inherent racial disparities in machine learning algorithms' decisions for a wide range of areas: pre-trial bail decisions [32], recidivism [5], healthcare [42], facial recognition [11], and college admissions [33]. Specific to healthcare applications, previous research has shown the effect of dataset biases on DL models' performance across genders and racial groups in cardiac MR imaging [47], chest X-rays [35, 50, 51], and skin disease imaging [23]. Recently, Groh et al. [23] showed that CNNs are the most accurate when classifying skin diseases manifesting on skin types similar to those they were trained on.

Learning domain invariant representations, a predominant approach in domain generalization [40], attempts to learn data distributions that are independent of the underlying domains, and therefore addresses the issue of training models on data from a set of source domains that can generalize well to previously unseen test domains. Domain invariant representation learning has been used in medical imaging for histopathology image analysis [34] and for learning domain-invariant shape priors in segmentation of prostrate MR and retinal fundus images [37]. On the other hand, previous works on fair classification and diagnosis of skin diseases have relied on skin type detection and debiasing [9] and classification model pruning [57].

One of the common definitions of algorithmic fairness for classification tasks, based on measuring statistical parity, aims to seek independence between the bias attribute (also known as the protected attribute; i.e., the skin type for our task) and the model's prediction (i.e., the skin disease prediction). Our proposed approach, **C**olor **I**nvariant **R**epresentation learning for unbiased **C**lassification of skin **Le**sions (**CIRCLe**), employs a color-invariant model that is trained to classify skin conditions independent of the underlying skin type. In this work, we aim to mitigate the skin type bias learnt by the CNNs and reduce the accuracy disparities across skin types. We address this problem by enforcing the feature representation to be invariant across different skin types. We adopt a domain-invariant representation learning method [41] and modify it to transform skin types from clinical skin images and propose a color-invariant skin condition classifier. In particular, we make the following contributions:

- To the best of our knowledge, this is the first work that uses skin type transformations and skin color-invariant disease classification to tackle the problem of skin type bias present in large scale clinical image datasets and how these biases permeate through the prediction models.
- We present a new state-of-the-art classification accuracy over 114 skin conditions and 6 Fitzpatrick skin types (FSTs) from the Fitzpatrick17K dataset. While previous works had either limited their analysis to a subset of diagnoses [9] or less granular FST labels [57], our proposed method achieves superior performance over a much larger set of diagnoses spanning over all the FST labels.
- We provide a comprehensive evaluation of our proposed method, CIRCLe, on 6 different CNN architectures, along with ablation studies to demonstrate the efficacy of the proposed domain regularization loss. Furthermore, we also assess the impact of varying the size and the FST distribution of the training dataset partitions on the generalization performance of the classification models.
- Finally, we propose a new fairness metric called Normalized Accuracy Range that, unlike several existing fairness metrics, works with multiple protected groups (6 different FSTs in our problem).

2 Method

2.1 Problem Definition

Given a dataset $\mathcal{D} = \{X, Y, Z\}$, consider x_i, y_i, z_i to be the input, the label, and the protected attribute for the i^{th} sample respectively, where we have M classes ($|Y| = M$) and N protected groups ($|Z| = N$). Let $\hat{y}_i$ denote the predicted label of sample i. Our goal is to train a classification model $f_\theta(\cdot)$ parametrized by θ that maps the input x_i to the final prediction $\hat{y}_i = f_\theta(x_i)$, such that (1) the prediction $\hat{y}_i$ is *invariant* to the protected attribute z_i and (2) the model's classification loss is minimized.

2.2 Feature Extractor and Classifier

In the representation learning framework, the prediction function $\hat{y}_i = f_\theta(x_i)$ is obtained as a composition $\hat{y}_i = \phi_C \circ \phi_E(x_i)$ of a feature extractor $r_i = \phi_E(x_i)$, where $r_i \in \mathbb{R}^p$ is a learned representation of data x_i, and a classifier $\hat{y}_i = \phi_C(r_i)$,

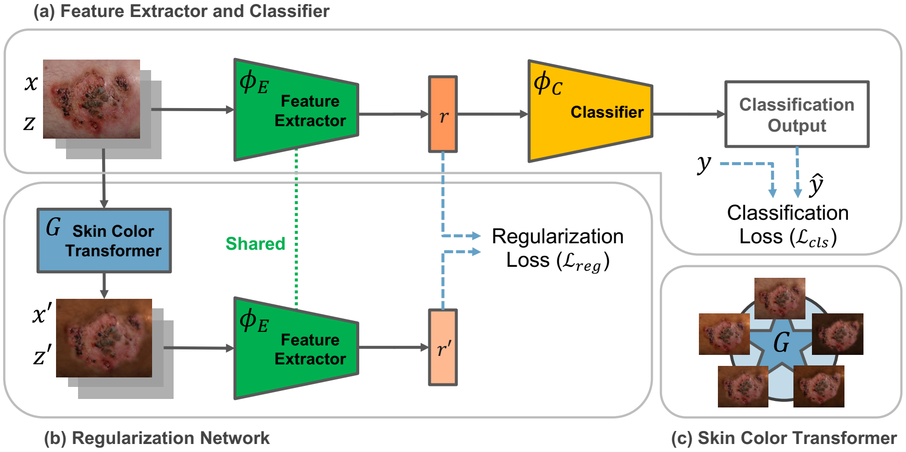

(a) Feature Extractor and Classifier
x
z
ϕ_E Feature Extractor
r
ϕ_C Classifier
Classification Output
y
$\hat{y}$
Classification Loss ($\mathcal{L}_{cls}$)
G Skin Color Transformer
Shared
Regularization Loss ($\mathcal{L}_{reg}$)
x'
z'
ϕ_E Feature Extractor
r'
G
(b) Regularization Network
(c) Skin Color Transformer

Fig. 1: Overview of CIRCLe. (a) The skin lesion image x with skin type z and diagnosis label y is passed through the feature extractor ϕ_E. The learned representation r goes through the classifier ϕ_C to obtain the predicted label $\hat{y}$. The classification loss enforces the correct classification objective. (b) The skin color transformer (G), transforms x with skin type z into x' with the new skin type z'. The generated image x' is fed into the feature extractor to get the representation r'. The regularization loss enforces r and r' to be similar. (c) The skin color transformer's schematic view with the possible transformed images, where one of the possible transformations is randomly chosen for generating x'.

predicting the label $\hat{y}_i$, given the representation r_i (Figure 1(a)). Thus, we aim to learn a feature representation r that is invariant to the protected attributes, and hypothesize that this will lead to better generalization for classification.

2.3 Regularization Network

Inspired by the method proposed by Nguyen et al. [41], we use a generative modelling framework to learn a function g that transforms the data distributions between skin types. To this end, we employ a method to synthesize a new image corresponding to a given input image with the subject's skin type in that image changed according to the desired Fitzpatrick skin type (FST) score. We call this model our Skin Color Transformer. After training the Skin Color Transformer model, we introduce an auxiliary loss term to our learning objective, whose aim is to enforce the domain invariance constraint. (Figure 1(b))

Skin Color Transformer. We learn the function G that performs image-to-image transformations between skin type domains. To this end, we use a Star Generative Adversarial Network (StarGAN) [12]. The goal of the StarGAN is to learn a unified network G (generator) that transforms the data density among multiple domains. In particular, the network $G(x, z, z')$ transforms an image x from skin type z to skin type z'. The generator's goal is to fool the discriminator D into classifying the transformed image as the destination skin type z'. In other words, the equilibrium state of StarGAN is when G successfully transforms the data density of the original skin type to that of the destination skin type. After training, we use $G(., z, z')$ as the Skin Color Transformer. This model takes the image x_i with skin type z_i as the input, along with a target skin type z_j and synthesizes a new image $z'_i = G(x_i, z_i, z_j)$ similar to x_i, only with the skin type of the image changed in accordance with z_j.

Domain Regularization Loss. In the training process of the disease classifier, for each input image x_i with skin type s_i, we randomly select another skin type $s_j \neq s_i$, and use the Skin Type Transformer to synthesize a new image $x'_i = G(x_i, s_i, s_j)$. After that, we obtain the latent representations $r_i = \phi_E(x_i)$, and $r'_i = \phi_E(x'_i)$ for the original image and the synthetic image respectively. Then we enforce the model to learn similar representations for r_i and r'_i by adding a regularization loss term to the overall loss function of the model:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} \tag{1}$$

where $\mathcal{L}_{cls}$ is the prediction loss of the network that predicts $\hat{y}_i$ given $r_i = \phi_E(x_i)$, and $\mathcal{L}_{reg}$ is the regularization loss. In this equation, $\lambda \in [0, 1]$ is a hyper-parameter controlling the trade-off between the classification and regularization losses. We define $\mathcal{L}_{reg}$ as the distance between the two representations r_i and r'_i to enforce the invariant condition. In our implementation, we use cross entropy as the classification loss $\mathcal{L}_{cls}$:

$$\mathcal{L}_{cls} = -\sum_{j=1}^{M} y_{ij} \, \log(\hat{y}_{ij}), \tag{2}$$

where y_{ij} is a binary indicator (0 or 1) if class label j is the correct classification for the sample i and $\hat{y}_{ij}$ is the predicted probability the sample i is of class j. The final predicted class $\hat{y}_i$ is calculated as

$$\hat{y}_i = \arg\max_j \; \hat{y}_{ij}. \tag{3}$$

We use squared error distance for computing the regularization loss $\mathcal{L}_{reg}$:

$$\mathcal{L}_{reg} = ||r_i - r'_i||_2^2. \tag{4}$$

3 Experiments

3.1 Dataset

We evaluate the performance of the proposed method on the Fitzpatrick17K dataset [23]. The Fitzpatrick17K dataset contains 16,577 clinical images with

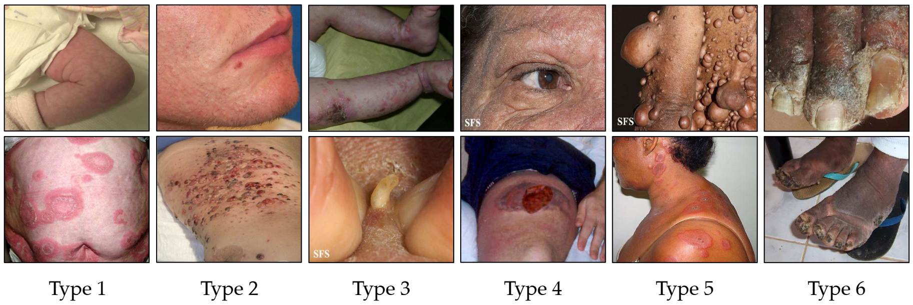
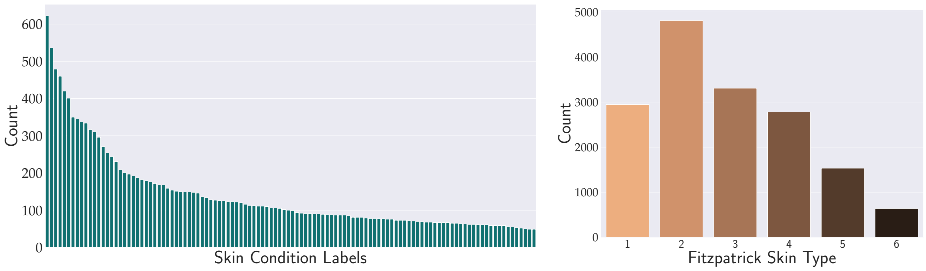


Fig. 2: Sample images of all six Fitzpatrick skin types (FSTs) from the Fitzpatrick17K dataset [23]. Notice the wide varieties in disease appearance, field of view, illumination, presence of imaging artifacts including non-standard background consistent with clinical images in the wild, and watermarks on some images.



(a)

(b)

Fig. 3: Visualizing the distribution of (a) the skin condition labels and (b) the Fitzpatrick skin type (FST) labels in the Fitzpatrick17K dataset. Notice that the number of images across different skin conditions is not uniformly distributed. Moreover, the number of images is considerably lower for darker skin types.

skin condition labels and skin type labels based on the Fitzpatrick scoring system [21]. The dataset includes 114 conditions with at least 53 images (and a maximum of 653 images) per skin condition, as shown in Figure 3 (a). The images in this dataset are annotated with (FST) labels by a team of non-dermatologist annotators. Figure 2 shows some sample images from this dataset along with their skin types. The Fitzpatrick labeling system is a six-point scale originally developed for classifying sun reactivity of skin and adjusting clinical medicine according to skin phenotype [21]. In this scale, the skin types are categorized in six levels from 1 to 6, from lightest to darkest skin types. Although Fitzpatrick labels are commonly used for categorizing skin types, we note that not all skin types are represented by the Fitzpatrick scale. [55].

In the Fitzpatrick17K dataset, there are significantly more images of light skin types than dark skin. There are 11,060 images of *light* skin types (FSTs 1, 2, and 3), and 4,949 images of *dark* skin types (FSTs 4, 5, and 6), as shown in Figure 3 (b).

3.2 Implementation Details

Dataset Construction Details. We randomly select 70%, 10%, and 20% of the images for the train, validation, and test splits, where the random selection is stratified on skin conditions. We repeat the experiments with five different random seeds for splitting the data. A series of transformations are applied to the training images which include: resize to 128 × 128 resolution, random rotations in $[-15^\circ, 15^\circ]$, and random horizontal flips. We also use ImageNet [18] training partition's mean and standard deviation values to normalize our images for training and evaluation.

Feature Extractor and Classifier's Details. We choose VGG-16 [52] pre-trained on ImageNet as our base network. We use the convolutional layers of VGG-16 as the feature extractor ϕ_E. We replace the VGG-16's fully-connected layers with a fully connected 256-to-114 layer as the classifier ϕ_C. We train the network for 100 epochs with plain stochastic gradient descent (SGD) using learning rate 1e-3, momentum 0.9, minibatch size 16, and weight decay 1e-3. We report the results for the epoch with the highest accuracy on the validation set.

StarGAN Details. StarGAN [12] implementation is taken from the authors' original source code with no significant modifications. We train StarGAN on the same train split used for training the classifier. As for the training configurations we use a minibatch size of 16. We train the StarGAN for 200,000 iterations and use the Adam [30] optimizer with a learning rate of 1e-4. For training the StarGAN's discriminator, we use cross entropy loss.

Model Training and Evaluation Setup. We use the PyTorch library [45] to implement our framework and train all our models on a workstation with AMD Ryzen 9 5950X processor, 32 GB of memory, and Nvidia GeForce RTX 3090 GPU with 24 GB of memory.

3.3 Metrics

We aim for an *accurate* and *fair* skin condition classifier. Therefore, we assess our method's performance using metrics for both accuracy and fairness. We use the well-known and commonly-used recall, F1-score, and accuracy metrics for evaluating our model's classification performance. For fairness, we use the equal opportunity difference (EOD) metric [25]. EOD measures the difference in true positive rates (TPR) for the two protected groups. Let TPR_z denote true positive rate of group z and $z \in \{0, 1\}$. Then EOD can be computed as:

$$EOD = |TPR_{z=0} - TPR_{z=1}|. \tag{5}$$

A value of 0 implies both protected groups have equal benefit. Given that the above metric (and other common fairness metrics in the literature [8, 19, 25]) are defined for two groups: privileged and under-privileged, w.r.t the protected attribute, we adopt the light (FSTs 1, 2, and 3) versus dark (FSTs 4, 5, and 6) as the two groups.

Additionally, to measure fairness in the model's accuracy for multiple groups of skin types, we assess the accuracy (ACC) disparities across all the six skin types by proposing the Normalized Accuracy Range (NAR) as follows:

$$NAR = \frac{ACC_{max} - ACC_{min}}{mean(ACC)}, \tag{6}$$

where ACC_{max} and ACC_{min} are the maximum and minimum accuracy achieved across skin types and $mean(ACC)$ is the mean accuracy across skin types, i.e.:

$$\begin{aligned} ACC_{max} &= max\{ACC_i : 1 \leq i \leq N\}, \\ ACC_{min} &= min\{ACC_i : 1 \leq i \leq N\}, \\ mean(ACC) &= \frac{1}{N} \sum_{i=1}^{N} ACC_i \end{aligned} \tag{7}$$

A perfectly fair performance of a model would result in equal accuracy across the different protected groups on a test set, i.e. $ACC_{max} = ACC_{min}$, leading to $NAR = 0$. As the accuracies across protected groups diverge, $ACC_{max} > ACC_{min}$, NAR will change even if the mean accuracy remains the same, thus indicating that the model's fairness is also changed. Moreover, NAR also takes into account the overall mean accuracy: this implies that in cases where the accuracies range ($ACC_{max} - ACC_{min}$) is the same, the model with the overall higher accuracy leads to a lower NAR, which is desirable. In our quantitative results, we report EOD for completeness; however, it is not an ideal measure, given it is restricted to only two protected groups whereas we have six. Therefore, we focus our attention on NAR.

3.4 Models

Baseline. For evaluating our method, we compare our results with the method proposed by Groh et al. [23], which has the current state-of-the-art performance on the Fitzpatrick17K dataset. We call their method the *Baseline*. To obtain a fair comparison, we use the same train and test sets they used.

Table 1: Comparing the model capacities and computational requirements of different backbones evaluated. For all the six backbones, we report the number of parameters and the number of multiply-add operations (**MulAddOps**). All numbers are in millions (**M**). Note how the six backbones encompass several architectural families and a large range of model capacities ($\sim$ 2M to $\sim$ 135M parameters) and computational requirements ($\sim$ 72M MulAddOps to $\sim$ 5136M MulAddOps).

	MobileNetV2	**MobileNetV3L**	**DenseNet-121**	**ResNet-18**	**ResNet-50**	**VGG-16**
Parameters (M)	2.55	4.53	7.22	11.31	24.03	135.31
MulAddOps (M)	98.16	72.51	925.45	592.32	1335.15	5136.16

Improved Baseline (Ours). In order to evaluate the effectiveness of the color-invariant representation learning process, we perform an ablation study, in which we remove the regularization loss $\mathcal{L}_{reg}$ from the learning objective of the model and train the classifier with only the classification objective. We call this model the *Improved Baseline.*

CIRCLe (Ours). The proposed model for unbiased skin condition classification, CIRCLe, is composed of two main components: the feature extractor and classifier, and the regularization network (Fig. 1).

Multiple Backbones. To demonstrate the efficacy of our method, we present evaluation with several other backbone architectures in addition to VGG-16 [52] used by Groh et al. [23]. In particular, we use MobileNetV2 [49], MobileNetV3-Large (referred to as MobileNetV3L hereafter) [27], DenseNet-121 [28], ResNet-18 [26], and ResNet-50 [26], thus covering a wide range of CNN architecture families and a considerable variety in model capacities, i.e. from 2.55 million parameters in MobileNetV2 to 135.31 million parameters in VGG-16 (Table 1).

For all the models, we perform an ablation study to evaluate if adding the regularization loss $\mathcal{L}_{reg}$ helps improve the performance.

4 Results and Analysis

4.1 Classification and Fairness Performance.

Table 2 shows the accuracy and fairness results for the proposed method in comparison with the baseline. From the table, we can see that our Improved Baseline method recognizably outperforms the baseline method in accuracy and fairness. By using a powerful backbone and a better and longer training process, we more than doubled the classification accuracy on the Fitzpatrick17K dataset for all the skin types. This indicates that the choice of the base classifier and training settings plays a significant role in achieving higher accuracy rates on the

Table 2: Classification performance and fairness of CIRCLe for classifying 114 skin conditions across skin types as assessed by the mean (std. dev.) of the metrics described in Section 3.3. We compute the overall accuracy based on the micro average accuracy across all skin types. Values in bold indicate the best results. CIRCLe yields the best performance while also improving fairness.

Model	**Recall**	**F1-score**	**Accuracy**							**EOD ↓**	**NAR ↓**
			Overall	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**		
Baseline	0.251	0.193	0.202	0.158	0.169	0.222	0.241	0.289	0.155	0.309	0.652
Improved	0.444	0.441	0.471	0.358	0.408	0.506	0.572	0.604	0.507	0.261	0.512
Baseline (Ours)	(0.007)	(0.009)	(0.004)	(0.026)	(0.014)	(0.023)	(0.022)	(0.029)	(0.027)	(0.028)	(0.078)
CIRCLe	**0.459**	**0.459**	**0.488**	**0.379**	**0.423**	**0.528**	**0.592**	**0.617**	**0.512**	**0.252**	**0.474**
(Ours)	(0.003)	(0.003)	(0.005)	(0.019)	(0.011)	(0.024)	(0.022)	(0.021)	(0.043)	(0.031)	(0.047)

Fitzpatrick17K dataset. Moreover, we can see that CIRCLe further improves the performance of our Improved Baseline across all the skin types, as well as the overall accuracy. This significant improvement demonstrates the effectiveness of the color-invariant representation learning method in increasing the model's generalizability. This observation shows that when the model is constrained to learn similar representations from different skin types that the skin condition appears on, it can learn richer features from the disease information in the image, and its overall performance improves. In addition, CIRCLe shows improved fairness scores (lower EOD and lower NAR), which indicates that the model is less biased. To the best of our knowledge, we set a new state-of-the-art performance on the Fitzpatrick17K dataset for the task of classifying the 114 skin conditions.

Different model architectures may show different disparities across protected groups [46].

We can see in Table 3 that the color-invariant representation learning (i.e. with the regularization loss $\mathcal{L}_{reg}$ activated) significantly improves the accuracy and fairness results in different model architecture choices across skin types, which indicates the effectiveness of the proposed method independently from the backbone choice and its capacity. We can see that while the regularization loss does not necessarily improve the EOD for all the backbones, EOD is not the ideal measure of fairness for our task since as explained in Section 3.3, it can only be applied to a lighter-versus-darker skin tone fairness assessment. However, employing the regularization loss does improve the NAR for all the backbone architectures.

4.2 Domain Adaptation Performance

For evaluating the model's performance on adapting to unseen domains, we perform a "two-to-other" experiment, where we train the model on all the images from two FST domains and test it on all the other FST domains. Table 4 shows the performance of our model for this experiment. CIRCLe recognizably improves the domain adaptation performance in comparison with the Baseline and

Table 3: Evaluating the classification performance improvement contribution of the regularization loss $\mathcal{L}_{reg}$ with multiple different feature extractor backbones. Reported values are the mean (std. dev.) of the metrics described in Section 3.3. Best values for each backbone are presented in bold. EOD reported (for two groups of light and dark FSTs) for completeness but evaluation over all the 6 FSTs uses NAR (see text for details). Observe that $\mathcal{L}_{reg}$ improves the classification accuracy and the fairness metric NAR for all backbones.

Model	$\mathcal{L}_{reg}$	**Recall**	**F1-score**	**Accuracy**							**EOD ↓**	**NAR ↓**
				Overall	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**		
MobileNetV2	✗	0.375	0.365	0.398	0.313	**0.364**	0.409	0.503	0.491	0.333	0.280	0.472
	✓	**0.404**	**0.397**	**0.434**	**0.354**	0.357	**0.471**	**0.559**	**0.544**	**0.421**	0.258	**0.455**
MobileNetV3L	✗	**0.427**	0.403	0.438	0.357	0.388	0.449	0.543	**0.560**	0.413	0.271	0.449
	✓	0.425	**0.412**	**0.451**	**0.369**	**0.400**	**0.464**	**0.565**	0.550	**0.444**	0.275	**0.420**
DenseNet-121	✗	0.425	0.416	0.451	0.393	0.397	0.452	**0.565**	0.522	**0.500**	0.278	0.364
	✓	**0.441**	**0.430**	**0.462**	**0.413**	**0.406**	**0.473**	0.561	**0.550**	0.452	0.294	**0.324**
ResNet-18	✗	0.391	0.381	0.417	0.355	0.353	0.431	0.538	0.516	**0.389**	0.263	0.430
	✓	**0.416**	**0.410**	**0.436**	**0.367**	**0.380**	**0.458**	**0.543**	**0.538**	**0.389**	0.282	**0.395**
ResNet-50	✗	0.390	0.382	0.416	0.337	0.363	0.422	0.549	0.506	0.389	0.257	0.497
	✓	**0.440**	**0.429**	**0.466**	**0.384**	**0.402**	**0.502**	**0.580**	**0.569**	**0.421**	0.283	**0.411**

Table 4: Classification performance measured by micro average accuracy when trained and evaluated on holdout sets composed of different Fitzpatrick skin types (FSTs). For example, "FST3-6" denotes that the model was trained on images only from FSTs 1 and 2 and evaluated on FSTs 3, 4, 5, and 6. CIRCLe achieves higher classification accuracies than Baseline (Groh et al. [23]) and Improved Baseline (also ours) for all holdout partitions and for all skin types.

Holdout Partition	**Method**	**Overall**	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**Type 5**	**Type 6**
FST3-6	Baseline	0.138	-	-	0.159	0.142	0.101	0.090
	Improved Baseline	0.249	-	-	0.308	0.246	0.185	0.113
	CIRCLe	**0.260**	-	-	**0.327**	**0.250**	**0.193**	**0.115**
FST12 and FST56	Baseline	0.134	0.100	0.130	-	-	0.211	0.121
	Improved Baseline	0.272	0.181	0.274	-	-	0.453	0.227
	CIRCLe	**0.285**	**0.199**	**0.285**	-	-	**0.469**	**0.233**
FST1-4	Baseline	0.077	0.044	0.055	0.091	0.129	-	-
	Improved Baseline	0.152	0.078	0.111	0.167	0.280	-	-
	CIRCLe	**0.163**	**0.095**	**0.121**	**0.177**	**0.293**	-	-

Improved Baseline, demonstrating the effectiveness of the proposed method in learning a color-invariant representation.

4.3 Classification Performance Relation with Training Size

As CIRCLe's performance improvement and effectiveness in comparison with the baselines is established in Section 4.1, we further analyze the relation of

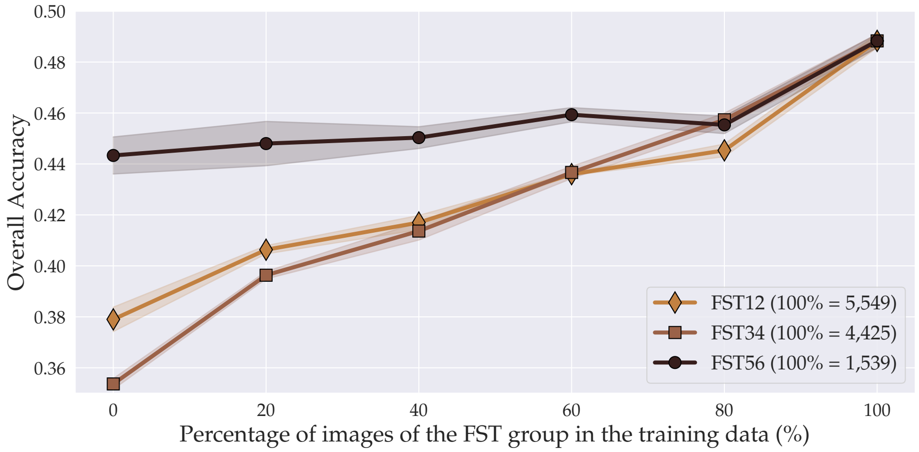


Fig. 4: Classification performance of CIRCLe on the test set as the number of training images of the FST groups increases. Each FST group line plot indicates the series of experiments in which the percentage of number of training images of that FST group changes as the rest of the training images remain idle. The rightmost point in the plot, with 100%, is identical for all the FST groups, which is the overall accuracy achieved by CIRCLe in Table 2. The std. dev. error band, illustrated in the figure, is computed by repetition of experiments with three different random seeds.

Table 5: Total number of training images for each experiment illustrated in Figure 4. Note that the test set for all these experiments is the original test split with 3,205 images (20% of the Fitzpatrick17K dataset images), and the number of training images for experiments with 100% of each FST group is the same for all three groups, and is equal to the original train split with 11,934 images (70% of the Fitzpatrick17K dataset images).

	0%	**20%**	**40%**	**60%**	**80%**
FST12	5,964	7,073	8,183	9,293	10,403
FST34	7,088	7,973	8,858	9,743	10,628
FST56	9,974	1,0281	10,589	10,897	11,205

CIRCLe's classification performance with the percentage of images of the FST groups in the training data. To this end, we consider the FST groups of light skin types (FSTs 1 and 2) with 5,549 images, medium skin types (FSTs 3 and 4) with 4,425 images, and dark skin types (FSTs 5 and 6) with 1,539 images in the training set. For each FST group, we gradually increase the number of images of that group in the training set, while the number of training images in other groups remains unchanged, and report the model's overall accuracy on the test set. The total number of training images for each of these experiments is provided in Table 5. As we can see in Figure 4, as the number of training images in a certain FST group increases, the overall performance improves, which is expected since DL-based models generalize better with larger training datasets. However, we can see that for the least populated FST group, i.e., dark skin types (FST56) with 13% of the training data, our method demonstrates a more robust performance across experiments, and even with 0% training data of FST56, it achieves a relatively high classification accuracy of 0.443. In addition, note that in these experiments, FST groups with lower number of images in the dataset, would have a larger number of total training images, since removing a percentage of them from the training images will leave a larger portion of images available for training (see Table 5). This indicates that when the number of training images is large enough, even if images of a certain skin type are not available, or are very limited, our model can perform well overall. This observation signifies our method's ability to effectively utilize the disease-related features in the images from the training set, independently from their skin types, as well as the ability to generalize well to minority groups in the training set.

5 Discussion and Future Work

In order to develop fair and accurate DL-based data-driven diagnosis methods in demotology, we need annotated datasets that include a diversity of skin types and a range of skin conditions. However, only a few publicly available datasets satisfy these criteria. Out of all the datasets identified by the Seventh ISIC Skin Image Analysis Workshop at ECCV 2022 (derm7pt [29], Dermofit Image Library [6], Diverse Dermatology Images (DDI) [17], Fitzpatrick17K [23], ISIC 2018 [13], ISIC 2019 [14, 15, 54], ISIC 2020 [48], MED-NODE [22], PAD-UFES-20 [44], PH2 [39], SD-128 [53], SD-198 [53], SD-260 [58]), only three datasets contain Fitzpatrick skin type labels: Fitzpatrick17K with 16,577, DDI with 656, and PAD-UFES-20 with 2,298 clinical images. The Fitzpatrick17K dataset is the only dataset out of these three which covers all the 6 different skin types (with over 600 images per skin type) and contains more than 10K images, suitable for training high-capacity DL-based networks and our GAN-based color transformer. It also contains samples from 114 different skin conditions, which is the largest number compared to the other two. For these reasons, in this work, we used the Fitzpatrick17K dataset for training and evaluating our proposed method. However, skin conditions in the Fitzpatrick17K dataset images are not verified by dermotologists and skin types in this dataset are annotated by non-dermatologists. Also, the patient images captured in the clinical settings exhibit various lighting conditions and perspectives. During our experiments, we found many erroneous and wrongly labeled images in the Fitzpatrick17K dataset, which could affect the training process. Fig. 5 shows some erroneous images in the Fitzpatrick17K dataset. Therefore, one possible future work can be cleaning the Fitzpatrick17K dataset and verifying its skin conditions and skin types by dermatologists.

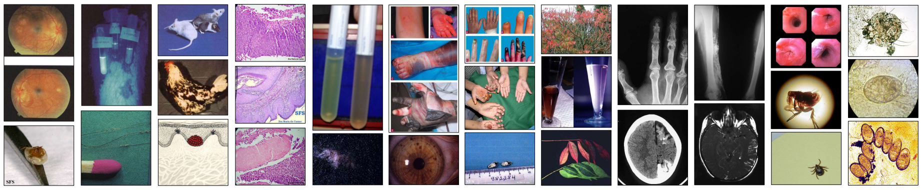


Fig. 5: Sample erroneous images from the Fitzpatrick17K dataset that are not clinical images of skin conditions, but are included in the dataset and are wrongly labeled with skin conditions.

In addition, as we can see in Section 4.3 and Figure 4, the number of training images plays a significant role in the model's performance across different skin types. Although in this paper we proposed a method for improving the skin condition classifier's fairness and generalizability, the importance of obtaining large and diverse datasets must not be neglected. Mitigating bias in AI diagnosis tools in the algorithm stage, as we proposed, can be effective and is particularly essential for the currently developed models, however, future research at the intersection of dermatology and computer vision should have specific focus on adding more diverse and annotated images to existing databases.

6 Conclusion

In this work, we proposed CIRCLe, a method based on domain invariant representation learning, for mitigating skin type bias in clinical image classification. Using a domain-invariant representation learning approach and training a color-invariant model, CIRCLe improved the accuracy for skin disease classification across different skin types for the Fitzpatrick17K dataset and set a new state-of-the-art performance on the classification of the 114 skin conditions. We also proposed a new fairness metric Normalized Accuracy Range for assessing fairness of classification in the presence of multiple protected groups, and showed that CIRCLe improves fairness of classification. Additionally, we presented an extensive evaluation over multiple CNN backbones as well as experiments to analyze CIRCLe's domain adaptation performance and the effect of varying the number of training images of different FST groups on its performance.

Acknowledgements. We would like to thank lab members Jeremy Kawahara and Ashish Sinha for their helpful discussions and comments on this work. We would also like to thank the reviewers for their valuable feedback that helped in improving this work. This project was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), and its computational resources were provided by NVIDIA and Compute Canada (computecanada.ca).

References

1. Adamson, A.S., Smith, A.: Machine learning and health care disparities in dermatology. JAMA Dermatology **154**(11), 1247–1248 (2018) 2
2. Adamson, A.S., Suarez, E.A., Welch, H.G.: Estimating overdiagnosis of melanoma using trends among black and white patients in the US. JAMA Dermatology **158**(4), 426–431 (2022) 2
3. Adelekun, A., Onyekaba, G., Lipoff, J.B.: Skin color in dermatology textbooks: An updated evaluation and analysis. Journal of the American Academy of Dermatology **84**(1), 194–196 (2021) 2
4. Agbai, O.N., Buster, K., Sanchez, M., Hernandez, C., Kundu, R.V., Chiu, M., Roberts, W.E., Draelos, Z.D., Bhushan, R., Taylor, S.C., Lim, H.W.: Skin cancer and photoprotection in people of color: A review and recommendations for physicians and the public. Journal of the American Academy of Dermatology **70**(4), 748–762 (2014) 2
5. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, pp. 254–264 (2016) 2
6. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis, pp. 63–86 (2013) 13
7. Barata, C., Marques, J.S., Emre Celebi, M.: Deep attention model for the hierarchical diagnosis of skin lesions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2757–2765 (2019) 1
8. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019) 8
9. Bevan, P.J., Atapour-Abarghouei, A.: Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. arXiv preprint arXiv:2202.02832 (2022) 2, 3
10. Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Fröhling, S., et al.: A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. European Journal of Cancer **111**, 148–154 (2019) 1
11. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 77–91 (2018) 2
12. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018) 4, 7
13. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019) 2, 13
14. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on

Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging. pp. 168–172 (2018) 13

15. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019) 13
16. Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V., Zou, J.: Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. JAMA Dermatology **157**(11), 1362–1369 (2021) 2
17. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances **8**(31), eabq6147 (2022) 13
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009) 7
19. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012) 8
20. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017) 1
21. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology **124**(6), 869–871 (1988) 6
22. Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N.: MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Systems with Applications **42**(19), 6578–6585 (2015) 13
23. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021) 2, 5, 6, 8, 9, 11, 13
24. Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology **29**(8), 1836–1842 (2018) 1
25. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems **29**, 3323–3331 (2016) 7, 8
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 9
27. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019) 9
28. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017) 9

29. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. IEEE Journal of Biomedical and Health Informatics **23**(2), 538–546 (2019) 13
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 7
31. Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C., Panda, R., Sattigeri, P., Varshney, K.R.: Fairness of classifiers across skin tones in dermatology. In: Medical Image Computing and Computer-Assisted Intervention. pp. 320–329 (2020) 2
32. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. The Quarterly Journal of Economics **133**(1), 237–293 (2018) 2
33. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. In: American Economic Association Papers and Proceedings. vol. 108, pp. 22–27 (2018) 2
34. Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Veta, M.: Learning domain-invariant representations of histological images. Frontiers in Medicine **6**, 162 (2019) 2
35. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020) 2
36. Lester, J., Jia, J., Zhang, L., Okoye, G., Linos, E.: Absence of images of skin of colour in publications of COVID-19 skin manifestations. The British Journal of Dermatology **183**(3), 593–595 (2020) 2
37. Liu, Q., Chen, C., Dou, Q., Heng, P.A.: Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. Proceedings of the AAAI Conference on Artificial Intelligence **36**(2), 1756–1764 (2022) 2
38. Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C., Coz, D.: A deep learning system for differential diagnosis of skin diseases. Nature Medicine **26**(6), 900–908 (2020) 1
39. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: PH^2—A Dermoscopic Image Database for Research and Benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5437–5440 (2013) 13
40. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International Conference on International Conference on Machine Learning. pp. 10–18 (2013) 2
41. Nguyen, A.T., Tran, T., Gal, Y., Baydin, A.G.: Domain invariant representation learning with domain density transformations. Advances in Neural Information Processing Systems **34**, 5264–5275 (2021) 3, 4
42. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019) 2
43. Osto, M., Hamzavi, I.H., Lim, H.W., Kohli, I.: Individual typology angle and Fitzpatrick skin phototypes are not equivalent in photodermatology. Photochemistry and Photobiology **98**(1), 127–129 (2022) 2
44. Pacheco, A.G., Lima, G.R., Salomão, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves Jr, F.C., Esgario, J.G., Simora, A.C., Castro, P.B., et al.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief **32**, 106221 (2020) 13

45. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035 (2019) 7
46. Prince, S.: Tutorial #1: Bias and fairness in AI (2019), https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/. Accessed: 2022-04-14 10
47. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 413–423 (2021) 2
48. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data **8**(1), 1–8 (2021) 13
49. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018) 9
50. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: Biocomputing 2021: proceedings of the Pacific symposium. pp. 232–243 (2020) 2
51. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**(12), 2176–2182 (2021) 2
52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 7, 9
53. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: European Conference on Computer Vision. pp. 206–222. Springer (2016) 2, 13
54. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 1–9 (2018) 13
55. Ware, O.R., Dawson, J.E., Shinohara, M.M., Taylor, S.C.: Racial limitations of Fitzpatrick skin type. Cutis **105**(2), 77—80 (2020) 6
56. Weiss, E.B.: Brown skin matters, https://brownskinmatters.com/. Accessed: 2022-06-23 1
57. Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J.: FairPrune: Achieving fairness through pruning for dermatological disease diagnosis. arXiv preprint arXiv:2203.02110 (2022) 2, 3
58. Yang, J., Wu, X., Liang, J., Sun, X., Cheng, M.M., Rosin, P.L., Wang, L.: Self-paced balance learning for clinical skin disease recognition. IEEE Transactions on Neural Networks and Learning Systems **31**(8), 2832–2846 (2019) 13