Computer Methods and Programs in Biomedicine 245 (2024) 108044

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Understanding skin color bias in deep learning-based skin lesion segmentation

Marin Benčević [a,b,*], Marija Habijan [a], Irena Galić [a], Danilo Babin [c], Aleksandra Pižurica [b]

[a] J. J. Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Kneza Trpimira 2B, Osijek, 31000, Croatia

[b] Ghent University, Department of Telecommunications and Information Processing, TELIN-GAIM, St-Pietersnieuwstraat 41, Ghent, 9000, Belgium

[c] Ghent University, Department of Telecommunications and Information Processing, imec-TELIN-IPI, St-Pietersnieuwstraat 41, Ghent, 9000, Belgium

ARTICLE INFO

Keywords:
AI fairness
Dermatological image analysis
Deep neural networks
Skin lesion segmentation

ABSTRACT

Background: The field of dermatological image analysis using deep neural networks includes the semantic segmentation of skin lesions, pivotal for lesion analysis, pathology inference, and diagnoses. While biases in neural network-based dermatoscopic image classification against darker skin tones due to dataset imbalance and contrast disparities are acknowledged, a comprehensive exploration of skin color bias in lesion segmentation models is lacking. It is imperative to address and understand the biases in these models.

Methods: Our study comprehensively evaluates skin tone bias within prevalent neural networks for skin lesion segmentation. Since no information about skin color exists in widely used datasets, to quantify the bias we use three distinct skin color estimation methods: Fitzpatrick skin type estimation, Individual Typology Angle estimation as well as manual grouping of images by skin color. We assess bias across common models by training a variety of U-Net-based models on three widely-used datasets with 1758 different dermoscopic and clinical images. We also evaluate commonly suggested methods to mitigate bias.

Results: Our findings expose a significant and large correlation between segmentation performance and skin color, revealing consistent challenges in segmenting lesions for darker skin tones across diverse datasets. Using various methods of skin color quantification, we have found significant bias in skin lesion segmentation against darker-skinned individuals when evaluated both in and out-of-sample. We also find that commonly used methods for bias mitigation do not result in any significant reduction in bias.

Conclusions: Our findings suggest a pervasive bias in most published lesion segmentation methods, given our use of commonly employed neural network architectures and publicly available datasets. In light of our findings, we propose recommendations for unbiased dataset collection, labeling, and model development. This presents the first comprehensive evaluation of fairness in skin lesion segmentation.

1. Introduction

Dermatological image analysis using deep neural networks has emerged as an active research domain with a multitude of published papers. A subset of these studies focuses on semantic segmentation of lesion images, which plays a crucial role in facilitating lesion analysis, pathology inference, and the diagnosis of conditions like melanomas. Additionally, lesion segmentation is commonly employed as a preprocessing step in the evaluation of dermatological neural networks [1–3]. In this paper, we aim to answer the question of whether commonly used segmentation neural networks trained on publicly available datasets are biased against dark-skinned subjects.[1]

Numerous researchers have found evidence that deep neural networks for classifying dermatological images are biased against dark-skinned individuals due to dataset imbalance as well as lower contrast in images of darker-skinned individuals [1,4,5]. Unlike classification models, segmentation models directly evaluate each pixel, so it is not clear whether the same biases found in classification models also extend to segmentation. In addition, due to the direct relationship between inputs and outputs of segmentation models, encountered segmentation

* Corresponding author.
E-mail address: marin.bencevic@ferit.hr (M. Benčević).

[1] All code and data is available at github.com/marinbenc/lesion_segmentation_bias.

https://doi.org/10.1016/j.cmpb.2024.108044
Received 20 November 2023; Received in revised form 17 January 2024; Accepted 21 January 2024
Available online 24 January 2024

bias can serve as an indicator of underlying issues in the data itself, warranting a thorough examination of dataset collection and labeling practices.

However, despite the growing awareness of biases in classification models, there is no comprehensive investigation into skin color bias in lesion segmentation models. To this end, we present a thorough evaluation of skin tone bias within commonly used deep neural networks for skin lesion segmentation. The main contributions of our work are:

- Since widely used publicly available lesion segmentation datasets do not include skin color information, we present different methods to estimate skin color from clinical dermatological images: (1) Fitzpatrick skin type estimation using a neural network, (2) Individual Typology Angle estimation using image processing and (3) manual grouping of dark and light skinned images. By reaching the same conclusions using these different methods we ensure repeatability and reliability of the conclusions.
- We investigate bias by training commonly-used segmentation models with combinations of three different popular publicly available datasets, namely PH^2 [6], The Waterloo dataset [7], and The Dermofit dataset [8].
- We conduct an in-depth statistical evaluation of the segmentation performance for different estimated skin colors, drawing inspiration from the field of artificial intelligence fairness.
- To the best of our knowledge, this study represents the first comprehensive evaluation of bias in skin lesion segmentation models.

As will be evident in the rest of the paper, our findings reveal a significant and large correlation between segmentation performance and skin color, indicating that common neural networks consistently struggle with segmenting lesions in individuals with darker skin. This bias is evident both within and outside the training dataset, as well as across multiple publicly available datasets. In addition, the bias is evident both with estimated and manually determined skin tone quantification methods. To better understand the impact of this bias, we present a qualitative evaluation of biased predictions. We assess several commonly used preprocessing techniques aimed at reducing bias but find that they fail to significantly alleviate skin color bias. In light of these discoveries, several suggestions for future dataset collection, labeling, and model development are proposed, aiming to foster more equitable skin lesion segmentation models.

1.1. Related research

Skin color bias in neural network classifiers has been a subject of significant investigation [9,10]. Studies have explored disparities in model predictions based on skin color, using metrics such as Fitzpatrick skin type [11] (FP) and Individual Typology Angle [12] (ITA). In the context of dermatological images, Kinyanjui et al. [2] evaluated the distribution of ITA on two widely used dermoscopic image datasets but found limited correlations between accuracy and ITA. However, we show that the datasets they use for evaluation have inadequate representation of dark-skinned subjects to detect bias. Groh et al. [13] evaluated commonly used classifiers on a hand-labeled dataset, reporting lower accuracy for less-represented skin colors. However, no statistical analysis of the results is provided beyond reporting mean values for different skin types. Daneshjou et al. [4] curated a diverse dermatological image dataset with subjects of various skin colors and found worse model performance on dark-skinned subjects from models trained on widely used datasets, even after fine-tuning. Moreover, they observed inferior dermatologist diagnostic performance in dark-skinned subjects.

In attempts to address dermatological classifier bias, Bevan and Atapour-Abarghouei [1] employed common bias unlearning methods, such as Learning Not To Learn [14] and Turning a Blind Eye [15]. The impact of these methods on bias is not evaluated — their analysis only reported mean results across all subjects, neglecting the examination of differences between different skin colors.

There is a very small number of studies that focus on reducing skin lesion segmentation bias. Galdran et al. [5] proposed a technique for augmenting skin color in images to enhance dataset diversity, leading to improved segmentation and classification performance. Nonetheless, similar to [1], they solely presented mean results without a comprehensive statistical analysis of bias. Consequently, discerning whether the models became less biased or if the results generally improved while remaining biased towards light-skinned subjects is challenging.

Yuan et al. [16] propose EdgeMixup, a method for enhancing fairness in the segmentation and classification of clinical skin images for Lyme disease analysis. Inspired by the mixup augmentation technique, which linearly combines existing data samples to increase data diversity, EdgeMixup merges input images with random lesion boundary labels. A classifier network then selects the optimal boundary candidate, which is then refined iteratively by a segmentation network until metrics converge. This approach enables the classifier to focus more on lesion boundaries, reducing bias due to skin color variations. Their results show improved fairness in identifying Lyme disease in clinical images. Empirically, their results improve fairness for both segmentation and classification.

Recent research in medical image segmentation has highlighted the issue of fairness, albeit with limited studies. Puyol-Antón et al. [17] conducted an extensive analysis of fairness in cardiac MR segmentation, examining the influence of confounders and the effectiveness of dataset rebalancing. Their findings indicated a significant bias in imbalanced datasets. This bias is particularly apparent when considering intersectional unrepresented groups such as black women [18]. Further investigations into the impact of model selection on fairness in cardiac MR segmentation were explored by Lee et al. [19]. Ioannou et al. [20] reported that models trained on imbalanced datasets for brain MRI segmentation exhibited poorer performance for underrepresented subject groups. Meanwhile, Tian et al. [21] introduced the first extensive dataset for fair segmentation in medical imaging, specifically for optic disc and cup segmentation in ophthalmoscopy fundus images. Their research not only identified biases in existing approaches but also evaluated a novel loss function designed to mitigate segmentation bias.

Addressing the gap in medical image segmentation fairness research, our work provides a thorough evaluation of skin color bias in skin lesion segmentation neural networks across multiple widely used datasets. To the best of our knowledge, this represents the first thorough statistical analysis of fairness in commonly used skin lesion segmentation models.

1.1.1. Skin color estimation from images

Researchers have proposed various methods to estimate skin tones from dermatoscopic or clinical images. Kinyanjui et al. [22] utilized a neural network to segment and remove lesion regions from the image. The remaining pixels were then transformed to the CIELAB colorspace, and the ITA was estimated from the mean pixel value after removing outliers. In contrast, Bevan and Atapour-Abarghouei [1] estimated skin color by sampling mean colors from different patches of the image, selecting the lightest patch under the assumption that healthy skin is generally lighter than lesion regions. Note that this assumption does not hold with some lesion areas such as depigmentation, which could lead to higher errors in dark-skinned patients. Galdran et al. [5] used Shades of Gray color constancy to estimate the illuminant of an image. These approaches may encounter challenges related to varying lighting conditions as they rely heavily on the pixel values themselves.

A potentially more robust approach involves using a neural network to classify skin color into Fitzpatrick skin types. Groh et al. [13] employed this strategy by creating a dataset of clinical skin disease images labeled with the Fitzpatrick skin type. They then trained a neural network-based classifier on this dataset to estimate the Fitzpatrick skin type given an image. Both the human labelers and the neural network

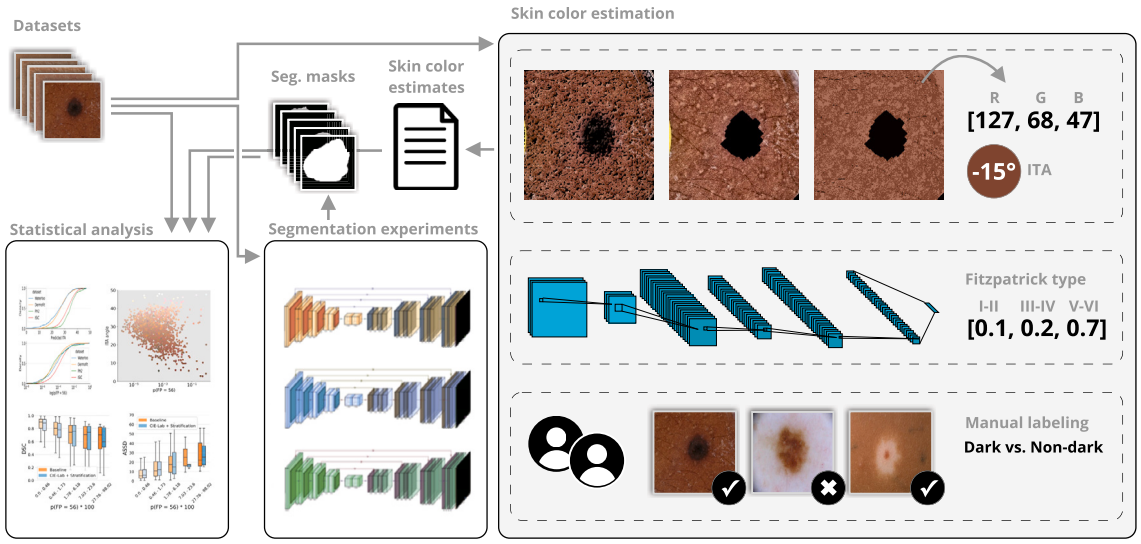

Fig. 1. A visual diagram of our methodology. First, the skin color of each image in the collected datasets is estimated using three different methods: (1) ITA estimation using image processing, (2) Fitzpatrick type estimation using a neural network, and (3) manually grouping images into dark vs. non-dark skin tones. Then, we train commonly used segmentation neural networks and obtain segmentation labels for in-sample and out-of-sample data. Finally, we perform a statistical analysis to investigate the relationship between the estimated skin color and segmentation quality.

achieved relatively low accuracy. Nonetheless, this approach still offers valuable insights when dealing with a sufficiently large sample size.

Kalb et al. [23] compare various methods of estimating skin tones from images and reach similar conclusions to us, namely that ISIC datasets [24] contain few dark-skinned images and are not a useful dataset for the evaluation of skin lesion analysis bias. Corbin and Marques [3] use explainable artificial intelligence methods to assess bias in skin lesion classification.

1.1.2. Skin lesion segmentation

Skin lesion segmentation is a widely studied problem with various proposed deep learning-based methods. Jha et al. [25] propose Double-U-Net, a network consisting of two connected U-Net-like networks with feature propagation between the two networks. In our previous work, we proposed using the polar transform [26] or image crops [27] as preprocessing steps with a neural network to predict the optimal preprocessing parameters.

There is also a class of methods focusing on the lesion boundary. Wang et al. [28] uses a transformer-based architecture with a boundary-wise attention gate to improve the capturing of local details. Lee et al. [29] predict boundary key points, which are used as residual attention to preserve boundary information in the network. They also use adversarial training, where an evaluator network is trained to predict whether the resulting segmentation maps are consistent with the boundary key points. Newer deep learning-based methods employ transformer-based architectures for this task [30,31].

2. Methods

The goal of our work is to evaluate the relationship between skin color and segmentation quality of commonly used neural networks, and our general methodology is explained visually in Fig. 1.

We select three different commonly used publicly available datasets of clinical and dermatoscopic skin lesion images with segmentation labels. These datasets do not include information about the patient skin color or ethnicity. Therefore, we employ three different methods to quantify or classify skin color in each image, as will be explained later in this section.

We then train commonly used image segmentation neural network architectures on the datasets. After calculating various segmentation metrics, both within and across datasets, we evaluate the relationships between the segmentation metrics and skin color estimates using statistical analysis. Finally, we also try to mitigate the effects of skin color bias by using preprocessing methods reported in skin lesion segmentation literature. The results of these methods are compared to the baseline model to evaluate whether the reported methods help mitigate the bias.

2.1. Skin tone extraction methods

The two most commonly used ways to quantify or classify skin color are the Fitzpatrick skin type or Individual Typology Angle [9,13,22]. The Fitzpatrick scale categorizes skin types into six groups based on UV response, ranging from type I (palest, never tans, always burns) to type VI (darkest, never burns). While subjectively evaluating skin types from images poses challenges [13], it can be effective with large sample sizes to evaluate bias [9].

ITA, being a more objective measure based on colorimetry, quantifies the skin's constitutive pigmentation [12]. Higher ITA values correspond to lighter skin. ITA can also be estimated from images using the CIELAB colorspace:

$$\mathrm{ITA}(L*, b*) = \arctan\left(\frac{L* - 50}{b*}\right) \cdot \frac{180}{\pi}, \tag{1}$$

where $L*$ and $b*$ are the lightness and blue-yellow opponents of the CIELAB colorspace, respectively. However, this estimate is highly dependent on lighting conditions and image contents. Estimated ITA from images can only be used as an indication of relative skin darkness within the dataset and not as an objective measure.

Since the publicly available datasets do not include skin color information, we use three distinct methods to estimate skin color: (1) ITA estimation using image processing, (2) Fitzpatrick skin type estimation using a neural network, and (3) manually labeling dark vs. non-dark skinned images.

In addition to using the Fitzpatrick type classifier directly to classify the subjects, we also use a proxy value for skin darkness $p(\mathrm{FP} = \text{V-VI})$. This is the probability, as predicted by the Fitzpatrick type classifier neural network, that the image belongs to Fitzpatrick types V or VI. We use this for calculating correlations and visualizing results, as it is a numerical instead of a categorical measure.

As can be seen in Fig. 2, while there is an overlap between the ITA and FP estimations, they have different strengths and weaknesses.

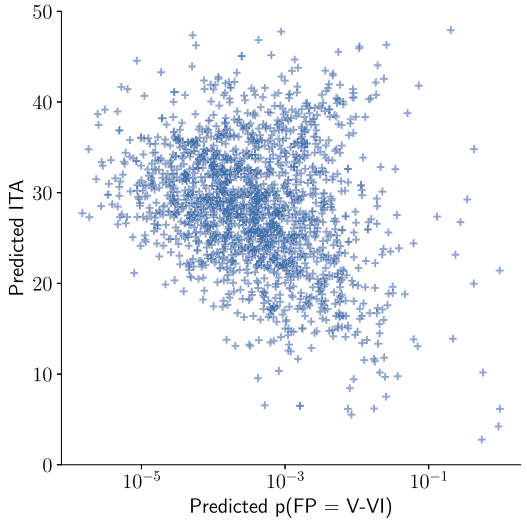

Fig. 2. A scatterplot of estimated ITA angle and the predicted probability of class "V-VI" by the FP estimation neural network classifier (p(FP = V-VI)). A higher probability corresponds to darker skin, while a higher ITA corresponds to lighter skin.

Namely, the ITA estimation is prone to errors but is a white box method for which the limitations and biases can be understood ahead of time. The FP estimation, being a neural network-based method, potentially offers better accuracy but is a black box method and therefore could include unknown biases. Finally, manual labeling offers the most reliable estimation but is coarse as we only use binary dark or non-dark labels to increase the reliability of the labels. We hope that by reaching the same conclusion using these distinct methods we can make more robust claims about the relationship between segmentation quality and skin color.

2.1.1. Fitzpatrick type estimation

To derive a Fitzpatrick type for each image, we trained a VGG16-based network to classify skin color into three Fitzpatrick type classes: I-II, III-IV, and V-VI. We aggregate the six Fitzpatrick types into three classes to account for off-by-one errors which are prevalent in the automatic estimation of skin color [32,4] as well as to reduce the effect of class imbalance in types V and VI.

The network was initialized using weights trained on the ImageNet dataset [33]. We then pre-train the network using the Fitzpatrick-17k dataset [13,32], which contains clinical skin disease images. Since the dataset is comprised of various skin condition images we further fine-tune the model on only skin lesion image datasets including the Diverse Dermatology Images [4] and PAD-UFES-20 [34]. Augmentation techniques were applied to increase out-of-sample robustness, including random scaling, translation, rotation, as well as horizontal and vertical flipping. Altogether, the network is pre-trained on 16,577 images and fine-tuned on 2,954 images from 1,943 subjects.

To address class imbalance, we used cross entropy loss with increased error weight for less-represented classes ($w_c = 1/|C|$ where $|C|$ is the number of samples of class C in the training dataset). 5-fold cross-validation was employed for model evaluation during training. During inference on segmentation datasets, an ensemble of the five folds was used for skin type prediction through majority voting.

The network achieves a balanced accuracy of 57% when evaluated using 5-fold cross-validation on the PAD-UFES-20 and Diverse datasets. Note that this performance is in line with human performance on the task of Fitzpatrick skin type estimation [13]. We also explored deeper backbone architectures like ResNet18 and ResNet34 [35], but validation results showed no improvement in accuracy. Therefore, we opted for the smaller VGG16 backbone.

2.1.2. Individual typology angle estimation

To perform the ITA estimation, we first extract skin pixels on the image without the lesion, hair, and other artifacts. We do this by preprocessing the image using contrast-limited adaptive histogram equalization of the $L*$ channel in the CIELAB colorspace and artifact removal using Dullrazor [36]. The HSV colorspace is then employed to obtain a mask of skin pixels by using Otsu thresholding of the value channel. The resulting mask is morphologically expanded and used to mask the image, leaving behind healthy skin areas with background, hairs, lesions, and pigmentations removed.

After extracting the skin region, we convert the image back to the CIELAB colorspace and perform k-means clustering on all skin pixel value vectors $(L_i, a*_i, b*_i)$. The optimal value of k is automatically determined for each image following [37], and we identify the most populated cluster as the estimated skin color of the subject.

A sample of the predicted dominant skin colors and ITA angles can be seen in Fig. 3. Visually, there is a high agreement between predicted and actual skin colors.

It is important to emphasize that while ITA and Fitzpatrick types may show correlation, they are not interchangeable. The ITA calculation serves as an estimate and is only intended to represent relative skin color lightness among images in the dataset. Its absolute value may not directly correspond to the actual ITA of the individual subjects.

2.1.3. Manual grouping of light- and dark-skinned subjects

Finally, human raters also manually labeled each image in the evaluation datasets as either belonging to dark-skinned (FP V or VI) or non-dark-skinned classes. These labels were assigned by two computer scientists, and disagreements were solved by preferring the non-dark-skinned class. This ensures conservative classification and minimizes the risk of false-positive dark-skinned images, thus avoiding overestimating bias. We note that the raters are not medical experts, however, the methodology of binary classification was chosen to ensure that the labeling is both easily manageable by individuals without specialized training and minimizes the potential for false-positive identifications.

2.2. Training the segmentation models

To evaluate bias in neural networks we train various U-Net-based models using a ResNet-18 encoder, a widely used architecture in lesion segmentation [24–26]. To ensure representative results, we initialize each model using weights obtained by training on the ISIC 2018 challenge Task 1 data [24,38] consisting of 3594 multi-source dermatoscopic images. Then, we train the models using 5-fold cross-validation on a combination of the PH^2 [6], Dermofit [8] and Waterloo datasets [7]. Additionally, we assess out-of-sample performance by employing a leave-one-dataset-out approach, training models on two datasets and evaluating them on the left-out dataset.

2.3. Datasets used for bias evaluation

We evaluate bias on three publicly available datasets:

1. PH^2 [6], a set of 200 dermatoscopic images.
2. The Waterloo dataset [7], consisting of 191 photographs of skin lesions from two different databases.
3. The Dermofit dataset [8] consisting of 1300 dermatoscopic images with internal color standards.

In Fig. 4, we present the distribution of the predicted Individual Typology Angle (ITA) and the probability of belonging to the dark-skinned class. Notably, the ISIC dataset demonstrates the least diversity among the four datasets. Therefore, we only use the ISIC dataset for model initialization and not for bias evaluation.

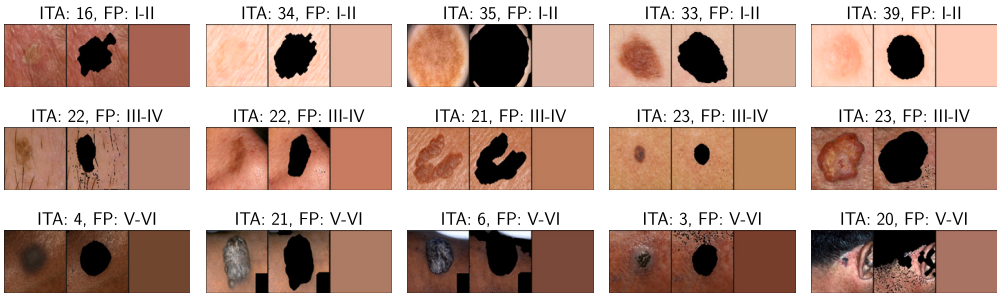

Fig. 3. Examples of ITA estimation. Each image shows, from left to right, the original image, the image-processing derived skin region used for ITA estimation, and the estimated dominant skin color. Above each image, the estimated ITA angle is shown together with the Fitzpatrick type prediction for that image, labeled FP.

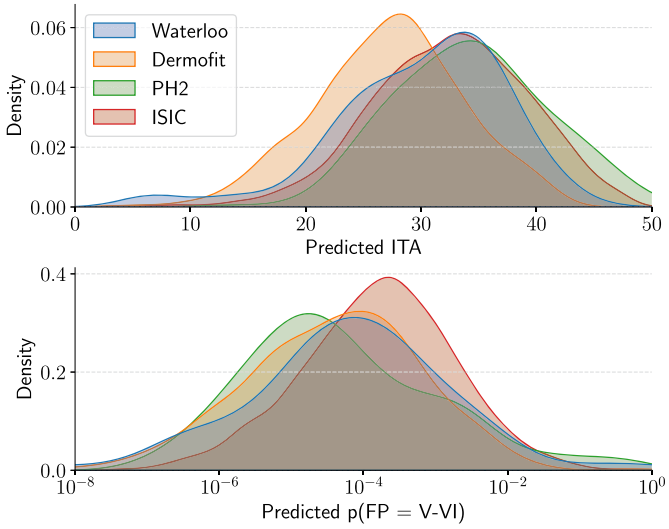

Fig. 4. A kernel density estimate plot of the skin color estimation results. The estimated ITA is shown on top. Lower ITA values correspond with darker skin. Below, the predicted probability of the dark-skinned class p(FP = V-VI) comes from the predictions of the Fitzpatrick type neural network classifier. Lower probability values correspond with darker skin. Note that Waterloo and Dermofit datasets contain more dark-skinned images according to the estimated ITA, while ISIC contains few dark-skinned images according to both estimates.

2.4. Statistical evaluation of bias

We use several segmentation metrics including the Dice similarity coefficient (DSC), Housdorff distance (HD), and average symmetric surface distance (ASSD). DSC is a measure of both sensitivity and precision and thus provides a comprehensive evaluation of the similarity between predicted and ground truth segmentation masks. On the other hand, HD and ASSD focus on boundary quality, with HD measuring the maximum distance between the predicted and ground truth boundary, while ASSD represents the average boundary distance. Therefore, these three metrics capture different problems in lesion boundary segmentation.

We use several measures to quantify skin darkness. Firstly, we use the estimated Fitzpatrick type as a categorical variable, comparing the distributions of segmentation metrics between the light (I-II), medium (III-IV), and dark (V-VI) skin color categories. Secondly, we use the Fitzpatrick type classifier's predicted probability of the dark-skinned class (p(FP = V-VI)) as a continuous proxy for skin darkness, examining correlations between segmentation metrics and p(FP = V-VI). Similarly, we employ the estimated ITA as a continuous variable to evaluate correlations with segmentation metrics. Finally, we also use the manual grouping as a categorical variable to verify if there are significant differences in segmentation metrics between dark and non-dark subjects.

Table 1
In-sample results for baseline lesion segmentation on different skin types (I-II – light skin, III-IV – medium skin, V-VI – dark skin) as classified by the Fitzpatrick type classifier. One-way ANOVA statistics and p-values are shown for each metric.

Skin Type	N	DSC	HD	ASSD
I-II	1587	0.904 ± 0.092	22.179 ± 17.285	6.190 ± 6.453
III-IV	166	0.878 ± 0.111	20.854 ± 18.684	6.494 ± 7.826
V-VI	5	0.744 ± 0.321	37.342 ± 39.039	10.814 ± 9.758
ANOVA	-	$F = 18.842$ $p < 0.0001$	$F = 3.127$ $p = 0.044$	$F = 1.703$ $p = 0.182$

To ensure normality for the ANOVA, Tukey's honestly significant difference (HSD) and t-tests, we use $log(1 - DSC)$, $log(HD)$, and $log(ASSD)$ for DSC, HD, and ASSD, respectively.

3. Results

3.1. In-sample bias quantification results

The in-sample evaluation was performed by training a model on all three datasets and evaluating using a held-out test set sampled from all three datasets. When using the Fitzpatrick type classifier to evaluate in-sample bias, we observe a large difference in both the mean and standard deviation of DSC, HD, and ASSD for light (FP I or II), medium (FP III or IV) and dark (FP V or VI) individuals. This difference is presented in Table 1. A one-way ANOVA confirms a statistically significant difference between the mean DSC of the groups ($F(2, 1755) = 18.842, p < 0.0001$). Subsequent Tukey's HSD test indicates that the mean DSC of dark-skinned individuals was significantly lower than that of light-skinned ($p < 0.0001, 95\%$ CI= $[0.192, 0.479]$) as well as medium-skinned ($p < 0.0001, 95\%$ CI = $[0.158, 0.449]$) individuals. In other words, in terms of DSC, the in-sample segmentation is worse for dark individuals than for other groups.

Further evidence of bias arises when using the Fitzpatrick classifier probability of belonging to the dark-skinned class (p(FP = V-VI)) and the estimated ITA as two continuous proxies for skin color darkness. Using Spearman's rank correlation, we find a significant negative correlation between DSC and p(FP = V-VI) ($r(1756) = -0.320, p < 0.0001$) as well as a positive correlation between DSC and ITA ($r(1756) = 0.283, p < 0.0001$). This is presented visually in Fig. 5. Higher ITA corresponds to lighter skin, so both of these results strongly indicate that segmentation is worse for individuals of darker skin colors.

In addition, we evaluate the impact of ISIC pre-training on the results. When omitting the pre-training process, we see almost no change in bias evaluation results. The resulting Spearman's rank correlation between DSC and p(FP = V-VI) decreases from -0.320 to -0.319 while the correlation between DSC and ITA increases from 0.283 to 0.288, when omitting ISIC pre-training.

However, it is important to note that we do not find a significant in-sample difference between manually classified dark and non-dark

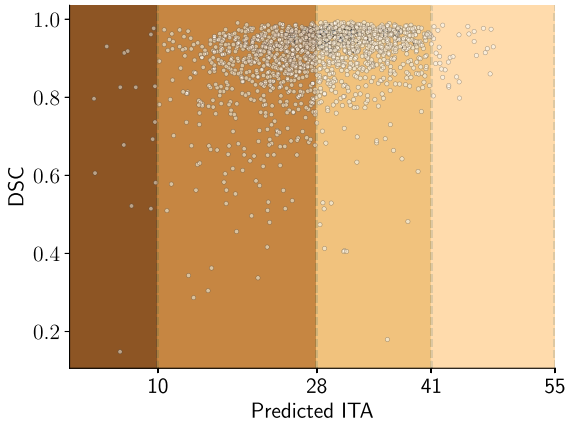

Fig. 5. A scatter plot of estimated ITA and the DSC scores of each test image result. The background colors exemplify a skin color from the ITA range of that region.

Table 2
Out-of-sample results for baseline lesion segmentation on different skin types (I-II – light skin, III-IV – medium skin, V-VI – dark skin) as classified by the Fitzpatrick type classifier. One-way ANOVA statistics and p-values are shown for each metric.

Skin Type	N	DSC	HD	ASSD
I-II	1587	0.843 ± 0.150	30.141 ± 23.379	9.685 ± 10.122
III-IV	166	0.770 ± 0.197	33.772 ± 27.507	12.399 ± 13.450
V-VI	5	0.532 ± 0.293	65.052 ± 39.922	30.098 ± 19.174
ANOVA	-	$F = 25.810$ $p < 0.0001$	$F = 3.477$ $p = 0.031$	$F = 8.014$ $p < 0.0001$

Table 3
Results of the Tukey HSD test on out-of-sample evaluation using the predicted Fitzpatrick types (I-II – light skin, III-IV – medium skin, V-VI – dark skin). Only significant differences ($p < 0.01$) are shown.

Skin Types	p	95% CI
DSC		
I-II vs V-VI	< 0.0001	[0.342, 0.916]
I-II vs III-IV	< 0.0001	[0.062, 0.166]
III-IV vs V-VI	< 0.0001	[0.224, 0.806]
ASSD		
I-II vs V-VI	0.003	[-2.365, -0.408]

Table 4
Spearman rank correlations between segmentation metrics and predicted skin color probabilities for the darkest class (p(FP = V-VI)) and estimated ITA for different experiments. All values are statistically significant with $p < 0.0001$.

	In-sample results			Out-of-sample results		
Attribute	DSC	HD	ASSD	DSC	HD	ASSD
Baseline						
p(FP = V-VI)	-0.320	0.268	0.292	-0.357	0.330	0.358
Pred. ITA	0.283	-0.209	-0.272	0.271	-0.258	-0.275
Stratified sampling						
p(FP = V-VI)	-0.311	0.272	0.289	-0.347	0.320	0.352
Pred. ITA	0.280	-0.211	-0.273	0.230	-0.238	-0.234
CIELAB & strat. sampling						
p(FP = V-VI)	-0.309	0.283	0.290	-0.346	0.321	0.348
Pred. ITA	0.269	-0.207	-0.261	0.203	-0.205	-0.197

groups. We deliberately included only high-certainty images in the dark class to reduce false positives, which might result in an underestimation of bias when using manual classification. As described later in this section, this test yields a statistically significant difference in the out-of-sample evaluation.

3.2. Out-of-sample bias quantification results

Given that segmentation models are often deployed in diverse settings and tasked with segmenting images outside their initial training domain, assessing out-of-sample performance becomes crucial in evaluating the fairness and safety of a model. To address this, we adopt a leave-one-dataset-out training procedure to comprehensively evaluate the model's performance beyond its training data.

Out-of-sample, the models are worse both in terms of segmentation metrics as well as bias, as is evident in Table 2. The ANOVA analysis of the metrics reveals significant differences between the DSCs of the three groups ($F(2, 1755) = 25.810, p < 0.0001$) as well as ASSD ($F(2, 1755) = 8.014, p < 0.0001$) that are notably larger than the in-sample differences. Subsequent testing using Tukey HSD confirms significant differences between the DSCs of all groups as well as the ASSD between the light and dark groups. The results of the Tukey HSD test are presented in Table 3. This increase in out-of-sample bias also persists in the baseline models when evaluated using estimated ITA as well as p(FP = V-VI), as shown in Table 4.

Finally, a one-tailed independent samples Welch's t-test indicates the DSC scores of manually labeled dark-skinned subjects ($M = 0.75, SD = 0.22$) and non-dark-skinned ($M = 0.84, SD = 0.16$) are statistically different ($t(1756) = 2.14, p = 0.020$).

3.3. Evaluation of methods for skin color bias correction

Various methods, such as stratified sampling or utilizing the CIELAB colorspace, are commonly employed to reduce bias in lesion segmentation models. While these methods are intuitively expected to mitigate bias, their quantitative impact remains largely unexplored. To address this gap, we conduct a comparative study involving three different preprocessing procedures: (1) RGB images with minimal pre-processing (Baseline), (2) stratified sampling of RGB images, and (3) stratified sampling and conversion to the CIELAB colorspace. To better understand if these steps mitigate bias, we evaluate the effects of the preprocessing steps in out-of-sample data.

The stratified sampling was implemented using the Fitzpatrick type classification results. Given an image I of class $C \in \{\text{I-II, III-IV, V-VI}\}$, the probability of selecting the image during batch sampling is proportional to $1/|C|$, where $|C|$ is the number of samples belonging to class C. The samples are drawn with replacement.

For the out-of-sample segmentation masks, independent samples t-test revealed a significant difference ($t(1756) = 3.57, p = 0.0004$) between the baseline DSCs ($M = 0.835, SD = 0.16$) and those obtained from the stratified CIELAB model ($M = 0.815, 0.17$). Although both stratified sampling on its own as well as with CIELAB conversion slightly reduced the correlation between p(FP = V-VI) and DSC, the bias persisted without significant reduction, as can be seen in Fig. 6.

A one-way ANOVA of the out-of-sample results of the model using stratification and CIELAB still revealed significant differences for DSC ($F(2, 1755) = 15.605, p < 0.0001$), HD ($F(2, 1755) = 5.453, p = 0.004$) and ASSD ($F(2, 1755) = 6.334, p = 0.002$) between the light, medium and dark skin color groups.

The results are slightly more promising when evaluating bias using predicted ITA. There is a slight reduction in Spearman's rank correlation between ITA and out-of-sample DSC from 0.283 ($p < 0.0001$) to 0.269 ($p < 0.0001$), as shown in Table 4. However, as indicated by other results, this small reduction in correlation does not translate into better segmentation results for less-represented subjects.

In addition, when examining subjects manually grouped into dark or non-dark groups, a one-tailed Welch's independent samples t-test indi-

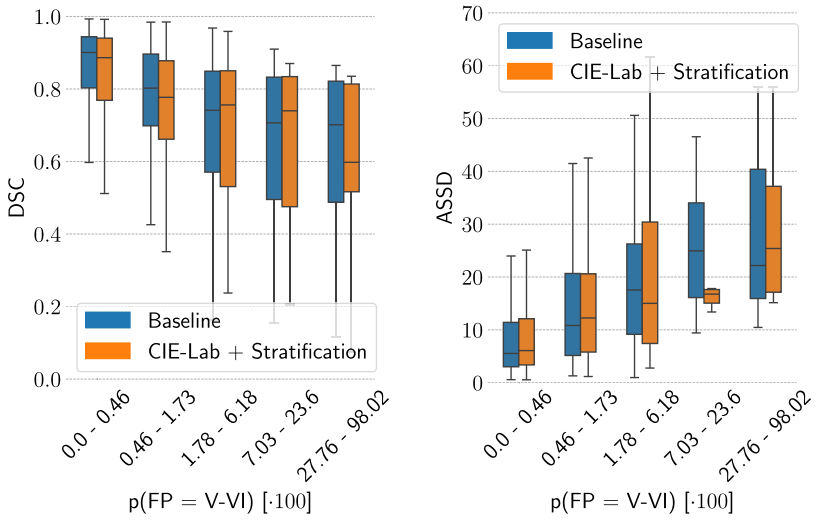

Fig. 6. A box plot of out-of-sample DSC and ASSD scores for groups binned by the neural Fitzpatrick skin type classifier probability of belonging to the dark-skinned class ($p(FP = V - VI)$) for two different preprocessing procedures.

Table 5
A comparison of segmentation results of subjects manually grouped into dark-skinned ($n_d = 23$) and the rest of the dataset ($n_r = 1735$). p values of a one-sided Welch's t-test with 10,000 permutations are reported. Significant results ($p < 0.05$) are marked with an asterisk.

Skin Type	DSC	
Baseline in-sample		
Dark	0.84 ± 0.22	$t = 1.31$
Rest	0.90 ± 0.09	$p = 0.0953$
CIELAB & strat. sampling in-sample		
Dark	0.85 ± 0.18	$t = 1.31$
Rest	0.90 ± 0.09	$p = 0.0968$
Baseline out-of-sample		
Dark	0.75 ± 0.22*	$t = 2.14$
Rest	0.84 ± 0.16*	$p = 0.0200$
CIELAB & strat. sampling out-of-sample		
Dark	0.73 ± 0.24*	$t = 2.13$
Rest	0.82 ± 0.17*	$p = 0.0225$

cated a significant difference between the groups for both the baseline and stratified CIELAB models, as presented in Table 5.

3.4. Qualitative assessment

Examples of out-of-sample predictions on manually labeled dark-skinned subjects are presented in Fig. 7. Qualitatively, the predicted lesion border follows the ground truth border better on light-skinned subjects. On dark-skinned subjects, the border exhibits both false positives and false negatives.

A significant challenge for the model lies in accurately identifying depigmented regions, as evident in Fig. 7(B), (D), and (E). The predicted area in Fig. 7(D) contains a sizable area that appears depigmented but was not labeled as such by the dermatologist. Conversely, the ground truth border of Fig. 7(E) includes the depigmented area, however, the predicted border encompasses an even larger area of surrounding skin. Depigmented areas with gradual borders pose difficulties for both manual labelers and fully automatic models in defining precise borders. Consequently, these examples might inadvertently influence the model to incorporate the surrounding skin of the lesion, even when no depigmentation occurs, as observed in Fig. 7(H). Although this issue is also present in light-skinned subjects, the underrepresentation of dark-skinned subjects accentuates its prominence in these images.

4. Discussion

We have found a significant bias against dark-skinned subjects in commonly used lesion segmentation methods. Furthermore, qualitative assessment has revealed both under and over-segmentation of skin lesions for dark-skinned subjects, especially in areas of depigmentation or low-contrast lesion borders.

To address this issue, a potential approach is to adopt more descriptive labels. Instead of solely labeling binary lesion/non-lesion areas, incorporating labels for differently colored areas such as white globules, yellow or orange areas, black lacunae, blue-gray areas, as well as structures like hypopigmented areas, structureless areas, and blue-white veils would be valuable. By providing more detailed labels, the model can learn border contrasts, shapes, smoothness, and other relevant features for each distinct structure. This, in turn, would enhance the accuracy of segmentation for areas such as depigmented skin.

In addition, considering the challenges posed by gradual lesion borders, it may be beneficial to incorporate methods that take varying degrees of lesion border contrast into account. The inherent difficulty in precisely defining gradual borders suggests that introducing fuzzy labels could facilitate the development of new models that better capture the ambiguity associated with gradual borders and make more nuanced predictions.

These approaches could also be extended to post-processing techniques. Instead of binarizing predictions into lesion and non-lesion areas, post-processing methods could be employed to treat smooth borders probabilistically. This would lead to more refined and reliable segmentation results, particularly for cases with ambiguous or gradual lesion borders.

Lastly, the existing datasets [24,7] have been compiled from various sources and annotated by different dermatologists without standardized guidelines. Consequently, there are instances where depigmented areas and surrounding lesion tissue are inconsistently incorporated into the lesion area. Additionally, there is variability even within one dataset in annotation methods, with some images manually labeled using polygons while others employ semi-automatic pixel-level labels [24]. All of these issues lead to a large degree of intra- and inter-observer variability within the datasets [39].

Another key reason for this bias is the lack of diverse publicly available datasets. We have shown that dark skin is severely underrepresented in widely used datasets including ISIC 2018 [24], PH^2 [6],

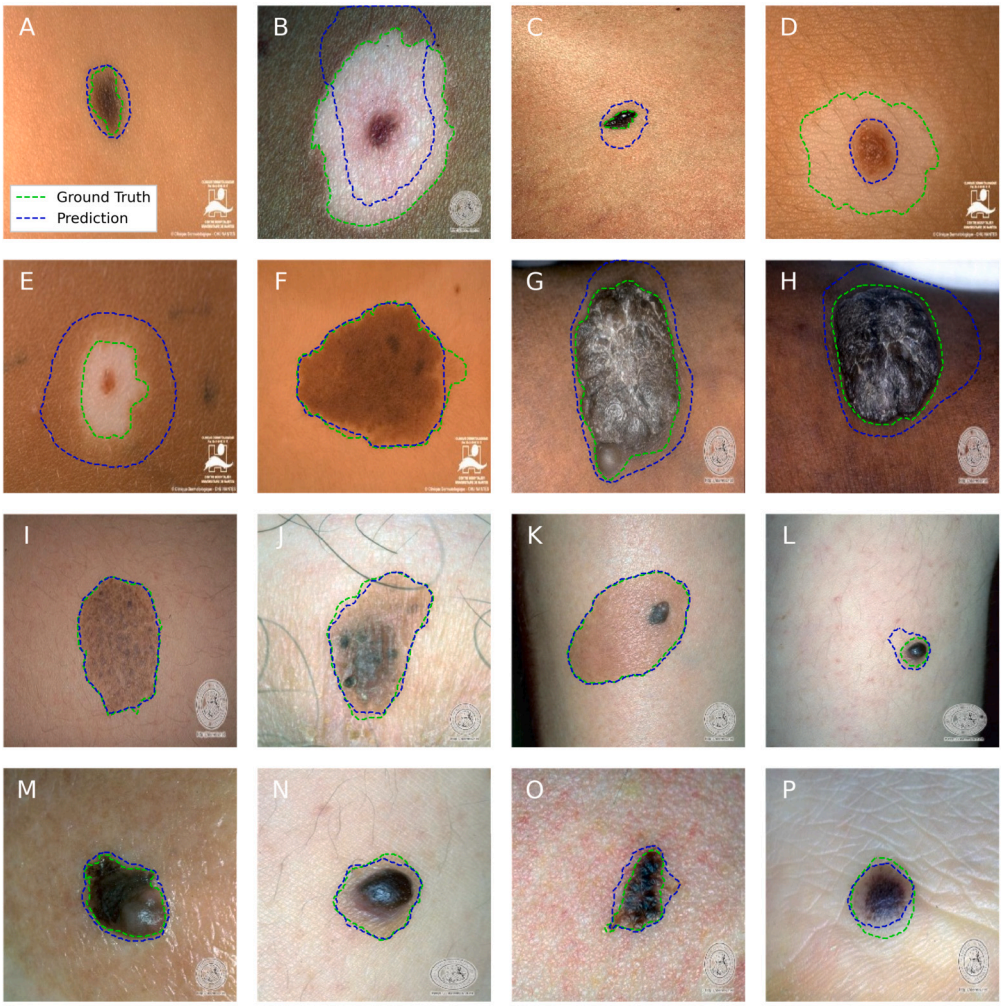

Fig. 7. Examples of model prediction of the out-of-sample baseline RGB model. The top two rows (A-H) show dark-skinned subjects as determined by manual labeling, while the bottom two rows (I-P) show non-dark-skinned subjects.

Dermofit [8] and Waterloo [7]. It is hard to claim that the segmentation quality of any network trained on these datasets generalizes to dark-skinned populations. Furthermore, with the exception of [16], the absence of information regarding skin color, race, or ethnicity within these datasets poses challenges for tracking and evaluating fairness, inadvertently incentivizing mean results as the primary focus.

Additionally, the bias observed may be attributed to characteristics inherent in the images and labels themselves. Most lesions are harder to segment in dark-skinned individuals due to a lower amount of contrast between surrounding tissue. This could lead to noisier segmentation labels as well as more challenging automatic segmentation. To identify the source of bias in image segmentation, further studies are necessary. These studies can include the use of balanced datasets to assess the impact of dataset imbalances and the development of image complexity metrics to evaluate how segmentation results are affected by image complexity and the lack of contrast in darker-skinned images.

Despite employing stratified sampling and the CIELAB colorspace, our attempts to mitigate bias did not yield significant improvements. In fact, the mean results worsened when evaluated out-of-sample. This highlights the importance of not solely relying on reporting the mean in-sample result for lesion segmentation, as it may not accurately reflect the model's performance in real-world scenarios. More generally, it could be that these simple preprocessing procedures do not have sufficient power to mitigate real-world domain shifts, as shown in [40].

We summarize our results in the following suggestions for future dataset curation and lesion border segmentation research:

- The development of publicly available datasets, standards, and challenges is critical for lesion segmentation across varied patient demographics, incorporating detailed demographic data. Future datasets and challenges in this field must contain data about the patient's skin color to enhance assessments of algorithmic fairness.
- These datasets should follow standardized guidelines and annotation methods to ensure consistent and reliable lesion area inclusion, especially for areas such as depigmented regions. The datasets should report on skin color bias present in the annotations themselves. Otherwise, biases present in the data could propagate to neural networks trained on those datasets.
- A shift from binary labels to labeling regions such as, among others, white globules, yellow or orange areas, hypopigmented areas, or blue-white veils, would allow models to learn and predict borders more accurately and fairly.
- Addressing the issue of ambiguous lesion margins requires the integration of probabilistic labels and uncertainty metrics within model predictions.
- Published skin lesion segmentation methods should employ robust validation procedures in terms of fairness and report outcomes beyond mean results averaged across all skin colors.
- Since light-skinned populations are overrepresented in existing datasets, benchmarks and challenges such as [24] should focus on a fair evaluation across subject populations as the primary metric, instead of mean segmentation results.

5. Conclusion

We have used several methods of estimating skin color from dermatological images including a Fitzpatrick type classifier, image processing-based ITA estimation, and manual grouping. In all cases, we have found significant bias in skin lesion segmentation against darker-skinned individuals when evaluated both in and out-of-sample. These findings indicate a pervasive bias in most published lesion segmentation methods, given our use of commonly employed neural network architectures and publicly available datasets.

It is important to acknowledge a limitation of our study, namely the utilization of algorithmic estimates of skin color from relatively small skin areas, which inherently introduces error. However, we have deliberately erred on the side of underestimating bias to ensure the robustness of our results, which consistently reveal the presence of bias according to different skin color estimation methods and even when accounting for errors in estimated skin colors.

In essence, we conclude that, while producing impressive mean results, existing lesion segmentation models are inadequate for practical deployment in real-world scenarios involving diverse patient populations. There is a pressing need for public datasets, benchmarks, and challenges for lesion segmentation on diverse patients. Such efforts, coupled with rigorous validation and inclusive performance metrics, are imperative for achieving equitable and accurate lesion segmentation across all patient demographics.

Ethical approval

No human or animal subjects have been part of this study.

CRediT authorship contribution statement

Marin Benčević: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marija Habijan:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Irena Galić:** Writing – review & editing, Supervision, Resources, Project administration. **Danilo Babin:** Writing – review & editing, Validation, Supervision, Resources. **Aleksandra Pižurica:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work has been supported in part by the Croatian Science Foundation under Project UIP-2017-05-4968, as well as the Faculty of Electrical Engineering, Computer Science and Information Technology Osijek grant “IZIP 2023”. This research has been partially supported by the Flanders AI Research Programme grant no. 174B09119.

References

[1] P.J. Bevan, A. Atapour-Abarghouei, Detecting melanoma fairly: skin tone detection and debiasing for skin lesion classification, arXiv:2202.02832, 2022.

[2] N.M. Kinyanjui, T. Odonga, C. Cintas, N.C.F. Codella, R. Panda, P. Sattigeri, K.R. Varshney, Fairness of classifiers across skin tones in dermatology, in: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, vol. 12266, Springer International Publishing, Cham, 2020, pp. 320–329.

[3] 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, https://doi.org/10.1109/ACCESS.2023.3289320.

[4] R. Daneshjou, K. Vodrahalli, R.A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S.M. Swetter, E.E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J.A.C. Allerup, U. Okata-Karigane, J. Zou, A.S. Chiou, Disparities in dermatology AI performance on a diverse, curated clinical image set, Sci. Adv. 8 (2022) eabq6147, https://doi.org/10.1126/sciadv.abq6147.

[5] A. Galdran, A. Alvarez-Gila, M.I. Meyer, C.L. Saratxaga, T. Araújo, E. Garrote, G. Aresta, P. Costa, A.M. Mendonça, A. Campilho, Data-driven color augmentation techniques for deep skin image analysis, arXiv:1703.03702, 2017.

[6] T. Mendonca, P.M. Ferreira, J.S. Marques, A.R.S. Marcal, J. Rozeira, 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 (EMBC), IEEE, Osaka, 2013, pp. 5437–5440.

[7] J. Glaister, R. Amelard, A. Wong, D.A. Clausi, MSIM: multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis, IEEE Trans. Biomed. Eng. 60 (2013) 1873–1883, https://doi.org/10.1109/TBME.2013.2244596.

[8] L. Ballerini, R.B. Fisher, B. Aldridge, J. Rees, A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions, in: M.E. Celebi, G. Schaefer (Eds.), Color Medical Image Analysis, vol. 6, Springer, Netherlands, Dordrecht, 2013, pp. 63–86.

[9] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, in: S.A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in: Proceedings of Machine Learning Research, PMLR, vol. 81, 2018, pp. 77–91.

[10] M. Merler, N. Ratha, R.S. Feris, J.R. Smith, Diversity in faces, arXiv:1901.10436, 2019.

[11] T.B. Fitzpatrick, The validity and practicality of sun-reactive skin types I through VI, Arch. Dermatol. 124 (1988) 869, https://doi.org/10.1001/archderm.1988.01670060015008.

[12] L.G. Farkas, M.J. Katic, C.R. Forrest, International anthropometric study of facial morphology in various ethnic groups/races, J. Craniofac. Surg. 16 (2005) 615–646, https://doi.org/10.1097/01.scs.0000171847.58031.9e.

[13] 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, Nashville, TN, USA, 2021, pp. 1820–1828.

[14] B. Kim, H. Kim, K. Kim, S. Kim, J. Kim, Learning not to learn: training deep neural networks with biased data, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[15] M. Alvi, A. Zisserman, C. Nellåker, Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings, in: L. Leal-Taixé, S. Roth (Eds.), Computer Vision – ECCV 2018 Workshops, vol. 11129, Springer International Publishing, Cham, 2019, pp. 556–572.

[16] H. Yuan, J. Aucott, A. Hadzic, W. Paul, M. Villegas de Flores, P. Mathew, P. Burlina, Y. Cao, EdgeMixup: embarrassingly simple data alteration to improve lyme disease lesion segmentation and diagnosis fairness, Springer Nature, Switzerland, 2023, pp. 374–384.

[17] E. Puyol-Antón, B. Ruijsink, J. Mariscal Harana, S.K. Piechnik, S. Neubauer, S.E. Petersen, R. Razavi, P. Chowienczyk, A.P. King, Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation, Front. Cardiovasc. Med. 9 (2022) 859310, https://doi.org/10.3389/fcvm.2022.859310.

[18] T. Lee, E. Puyol-Antón, B. Ruijsink, M. Shi, A.P. King, A systematic study of race and sex bias in CNN-based cardiac MR segmentation, in: O. Camara, E. Puyol-Antón, C. Qin, M. Sermesant, A. Suinesiaputra, S. Wang, A. Young (Eds.), Statistical Atlases and Computational Models of the Heart, in: Regular and CMRxMotion Challenge Papers, vol. 13593, Springer Nature, Switzerland, Cham, 2022, pp. 233–244.

[19] T. Lee, E. Puyol-Antón, B. Ruijsink, K. Aitcheson, M. Shi, A.P. King, An investigation into the impact of deep learning model choice on sex and race bias in cardiac MR segmentation, Springer Nature, Switzerland, 2023, pp. 215–224.

[20] S. Ioannou, H. Chockler, A. Hammers, A.P. King, A study of demographic bias in CNN-based brain MR segmentation, Springer Nature Switzerland, 2022, pp. 13–22.

[21] Y. Tian, M. Shi, Y. Luo, A. Kouhana, T. Elze, M. Wang, Fairseg: a large-scale medical image segmentation dataset for fairness learning with fair error-bound scaling, arXiv:2311.02189, 2023.

[22] N.M. Kinyanjui, T. Odonga, C. Cintas, N.C.F. Codella, R. Panda, P. Sattigeri, K.R. Varshney, Estimating skin tone and effects on classification performance in dermatology datasets, arXiv:1910.13268, 2019.

[23] T. Kalb, K. Kushibar, C. Cintas, K. Lekadir, O. Diaz, R. Osuala, Revisiting skin tone fairness in dermatological lesion classification, Springer Nature, Switzerland, 2023, pp. 246–255.

[24] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC), arXiv:1902.03368, 2019.

[25] D. Jha, M.A. Riegler, D. Johansen, P. Halvorsen, H.D. Johansen, Doubleu-net: a deep convolutional neural network for medical image segmentation, in: 2020 IEEE

33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE Computer Society, Los Alamitos, CA, USA, 2020, pp. 558–564.

[26] M. Benčević, I. Galić, M. Habijan, D. Babin, Training on polar image transformations improves biomedical image segmentation, IEEE Access 9 (2021) 133365–133375, https://doi.org/10.1109/ACCESS.2021.3116265.

[27] M. Benčević, Y. Qiu, I. Galić, A. Pižurica, Segment-then-segment: context-preserving crop-based segmentation for large biomedical images, Sensors 23 (2023), https://doi.org/10.3390/s23020633.

[28] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion segmentation, in: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 206–216.

[29] M.C.H. Lee, O. Oktay, A. Schuh, M. Schaap, B. Glocker, Image-and-spatial transformer networks for structure-guided image registration, in: D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 337–345.

[30] B. Chen, Y. Liu, Z. Zhang, G. Lu, A.W.K. Kong, Transattunet: multi-level attention-guided u-net with transformer for medical image segmentation, in: IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, pp. 1–14.

[31] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E.K. Aghdam, J. Cohen-Adad, D. Merhof, Hiformer: hierarchical multi-scale representations using transformers for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6202–6212.

[32] M. Groh, C. Harris, R. Daneshjou, O. Badri, A. Koochek, Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm, arXiv preprint, arXiv:2207.02942, 2022.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, 2009, pp. 248–255.

[34] A.G. Pacheco, G.R. Lima, A.S. Salomão, B. Krohling, I.P. Biral, G.G. De Angelo, F.C. Alves Jr, J.G. Esgario, A.C. Simora, P.B. Castro, F.B. Rodrigues, P.H. Frasson, R.A. Krohling, H. Knidel, M.C. Santos, R.B. Do Espírito Santo, T.L. Macedo, T.R. Canuto, L.F. De, Barros, PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones, Data Brief 32 (2020) 106221, https://doi.org/10.1016/j.dib.2020.106221.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.

[36] T. Lee, V. Ng, R. Gallagher, A. Coldman, D. McLean, Dullrazor®: a software approach to hair removal from images, Comput. Biol. Med. 27 (1997) 533–543, https://doi.org/10.1016/S0010-4825(97)00020-6.

[37] V. Satopa, J. Albrecht, D. Irwin, B. Raghavan, Finding a “Kneedle” in a Haystack: detecting knee points in system behavior, https://doi.org/10.5281/zenodo.6496266, 2023.

[38] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) 180161, https://doi.org/10.1038/sdata.2018.161.

[39] P. Tschandl, C. Sinz, H. Kittler, Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation, Comput. Biol. Med. 104 (2019) 111–116, https://doi.org/10.1016/j.compbiomed.2018.11.010.

[40] J. Schrouff, N. Harris, O. Koyejo, I. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen, A. Dieng, Y. Liu, V. Natarajan, A. Karthikesalingam, K. Heller, S. Chiappa, A. D’Amour, Diagnosing failures of fairness transfer across distribution shift in real-world medical settings, arXiv:2202.01034, 2022.