EVALUATING THE IMPACT OF SKIN TONE REPRESENTATION ON OUT-OF-DISTRIBUTION DETECTION PERFORMANCE IN DERMATOLOGY

Assala Benmalek[⋆], *Celia Cintas*[†], *Girmaw Abebe Tadesse*[‡], *Roxana Daneshjou*[††], *Kush R. Varshney*[†], *Cherifi Dalila*[⋆]

[⋆]Institute of Electrical and Electronics Engineering – Boumerdes University, Boumerdes, Algeria
[†]IBM Research – Africa, Nairobi, Kenya
[†]IBM Research – T. J. Watson Research Center, Yorktown Heights, NY, USA
[‡]Microsoft AI for Good Research Lab, Nairobi, Kenya
[††]Stanford University, Stanford, CA, USA

ABSTRACT

Addressing representation issues in dermatological settings is crucial due to variations in how skin conditions manifest across skin tones, thereby providing competitive quality of care across different segments of the population. Although bias and fairness assessment in skin lesion classification has been an active research area, there is substantially less exploration of the implications of skin tone representations and Out-of-Distribution (OOD) detectors' performance. Current OOD methods detect samples from different hardware devices, clinical settings, or unknown disease samples. However, the absence of robustness analysis across skin tones questions whether these methods are fair detectors. As most skin datasets are reported to suffer from bias in skin tone distribution, this could lead to higher false positive rates in a particular skin tone. In this paper, we present a framework to evaluate OOD detectors across different skin tones and scenarios. We review and compare state-of-the-art OOD detectors across two categories of skin tones, FST I-IV (lighter tones) and FST V-VI (brown and darker tones), over samples collected from dermatoscopic and clinical protocols. Our experiments yield that in poorly performing OOD models, the representation gap measured between skin types is wider (from $\approx$ 10% to 30%) up for samples from darker skin tones. Compared to better performing models, skin type performance only differs for $\approx$ 2%. Furthermore, this work shows that understanding OOD methods' performance beyond average metrics is critical to developing more fair approaches. As we observe, models with a similar overall performance have a significant difference in the representation gap, impacting FST I-IV and FST V-VI differently. The code is publicly available at the repository[1].

***Index Terms*—** Algorithmic fairness, Skin tone representation, Out-of-distribution detection, Dermatology

1. INTRODUCTION

Skin diseases remain a global health challenge, with skin cancer being the most common cancer worldwide [1]. Following the recent success of Deep Learning (DL) in various computer vision problems, Convolutional Neural Networks (CNNs) have been employed for skin disease classification with improved performance. However, DL models have been shown to be prone to and exacerbate existing societal biases [2, 3]. Thus, as we observe increasing interest in DL for dermatology [4–6], it is imperative to address the transparency, robustness, and fairness of these solutions [7–11] in order to make them adopted clinically for positive societal impact. In dermatology, bias in representations of skin tones in academic materials [10, 12] and clinical care is becoming a primary concern. [12, 13] reports major disparities in dermatology when treating skin of color as common conditions often manifest differently on dark skin, and physicians are trained mostly to diagnose them on light skin. The growing practice of using machine learning algorithms to aid the diagnosis of skin diseases will further deepen the divide in patient care because these algorithms are trained with such imbalanced datasets [9], with an overwhelming majority of samples with light skin tones. Particularly, when we look at robustness, we are interested in the ability of the models to identify Out-of-Distribution (OOD) samples that differ from the training distribution. For example, OOD samples may come from new skin conditions, different collection protocols [14], or heterogeneous patient sub-populations. However, the fairness of these OOD detection methods has not been explored in the existing literature. OOD detectors need to guarantee equivalent detection capability across different sub-populations.

In this paper, we work towards quantifying and evaluating the detection disparity across skin tones in OOD detectors in different clinical scenarios. We are interested in answering questions such as: *how much does the skin tone representation of the In-Distribution Dataset (IDD) impact the OOD overall performance*? *do we observe changes in performance for different skin types? is the average performance of an OOD method a fair measurement across skin tones?* Specifically, our contributions are as follows:

- We propose an evaluation framework to assess OOD detectors under different skin types. We evaluate our approach across dermatoscopic and clinical datasets with different skin types under baselines and state-of-the-art OOD detectors.
- We create manually-labeled FST I-IV and FST V-VI for public benchmark dataset ISIC 2019 [1].
- From our experiments, we observe that in OOD models with poorly performance, the representation gap measured between skin types is wider ($\approx$ 20%). Compared to better performing models where skin type performance only differs for $\approx$ 2%.
- We argue that understanding performance OOD methods beyond average metrics is critical to ensure fair approaches. A clear example of this can be seen in Table 1, where IF ([15]), ODIN ([16]), and NN Softmax ([17]) have similar overall F_1 scores,

[1] https://github.com/assalaabnk/OOD-in-Dermatology

but when we observe performance by skin type, we see between $\approx 10\%$ to $\approx 30\%$ difference in the representation gap impacting FST I-IV and FST V-VI differently.

2. RELATED WORK

2.1. Algorithmic Fairness Studies

Several studies propose different ways to analyze skin tones; multiple approaches used individual typology angle (ITA) computed from pixel intensity values [9, 18]. The ITA values were then mapped to Fitzpatrick Skin Types (FST) [19]. This information is key to stratifying further studies regarding the algorithm fairness of classifiers. Rezk et al., [20] proposed data augmentation techniques to improve the diversity of skin tones at the training time of DL models. Moreover, the proximity of skin tones is found to play a significant impact on the classification performance as [18, 21] reported that skin condition classifiers trained on data from only two FSTs are most accurate on holdout images of the closest FSTs to the training data.

Although bias and fairness assessment in skin lesion classification has been an active research area [2, 9, 11, 20–23] to the best of the authors' knowledge, there is no research in understanding the impact of the lack of skin tone representation in OOD detectors.

2.2. Out-of-distribution Detection Methods

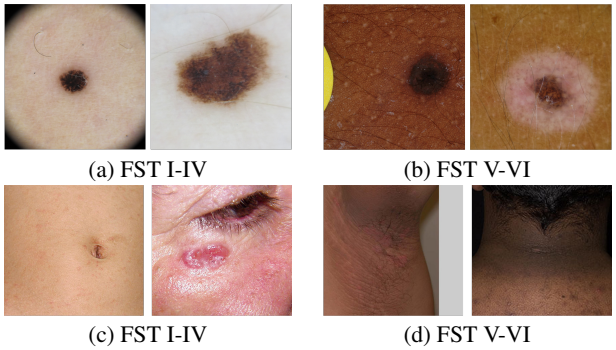
Existing OOD detection methods could be grouped into *ensemble methods* such as Isolation Forest (IF) [15], OneClassSVM [24] and *deep learning approaches* [14, 16, 17] based on the type of models employed. Xuan Li [25] used the IF approach on the features computed by a pre-trained CNN to detect OOD images of skin lesions, which is called DeepIF. ODIN [16] an NN Softmax methods [17] utilized CNNs trained for classification to build robust OOD detectors. [17] used temperature scaling in the last layer, and [16] extended this approach, adding small perturbations to the input to separate the softmax score distributions between in- and out-of-distribution images, allowing for more effective detection. Lastly, as autoencoders (AE) can model training data distribution, these neural networks are a common option for OOD detection. The majority of the methods discussed in the literature require the training data to consist of in-distribution examples only [26–28].

In this work, we aim to bridge the gap in analyzing the fairness of OOD detection methods by providing a model-agnostic evaluation framework that could be applied to either the ensemble or deep learning methods.

3. DATASETS

We validate the proposed framework using two datasets: ISIC 2019 [1] and Fitzpatrick 17k [29] for dermoscopic and clinical samples from different collection protocols. We stratify the samples from both datasets based on skin tones (FST I-IV and FST V-VI). See Figure 1 for reference examples for each skin tone across both datasets.

ISIC 2019 Consists of $25,331$ dermoscopic images among eight diagnostic categories. Non-dermatologists trained on previous examples labeled skin images as FST I-IV and FST V-VI. The authors manually annotated the skin tone for this dataset, as this information was missing; the labels are available at the repository[2]. After carefully curating the labels, we have $25,327$ samples categorized as FST I-IV and 4 as FST V-VI.



(a) FST I-IV (b) FST V-VI

(c) FST I-IV (d) FST V-VI

Fig. 1. Examples for FST I-IV and FST V-VI skin types in both datasets. The samples in (a) and (b) belong to the ISIC 2019 dataset [1], while (c) and (d) to the Fitz17k dataset [21]. For the Fitz17k, we unified the skin tones into two labels, while ISIC 2019 was labeled manually by the authors with domain expert supervision; see more details in Section 3.

Fitzpatrick 17k Dataset [18, 21] contains $16,577$ clinical images with skin type labels based on the Fitzpatrick scoring system [29]. The images are sourced from two online open-source dermatology atlases. The images are annotated with Fitzpatrick skin type labels by a team of human annotators from Scale AI. 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. For this work, we grouped the six labels provided in [18, 21] into two classes, FST I-IV for lighter skin tones (13844) and FST V-VI (2168) for brown and darker tones. Figure 2 shows the variation of samples from both datasets.

4. METHODS

4.1. Proposed Evaluation Framework

Our fairness assessment has several steps (See Algorithm 1). Given skin image datasets ($\mathcal{D}_1$, $\mathcal{D}_2$),and its corresponding skin tone stratification $\mathcal{T}$. We train an Out-of-distribution (OOD) detector ($\mathcal{O}$) or use off-the-shelf when available. After the detector is trained, we extract the FST for training and evaluation samples (from $\mathcal{D}_i, i \in \{1, 2\}$) based on the stratification provided by $\mathcal{T}$. We do this by using $\mathcal{D}_1$ as IDD and $\mathcal{D}_2$ as OOD. This is to understand the impact of IDD datasets with predominantly one skin tone (FST I-IV). The evaluation pipeline then follows with training $\mathcal{O}$ and testing it with skin images stratified across skin tones ($\mathcal{T}_i$). The detection performance of $\mathcal{O}$ is then evaluated using F_1 score and Area Under Receiver operating characteristic (AUROC). The representation gap ($\mathcal{RG}$) score is then derived, which quantifies the divergence in the detection performance of $\mathcal{O}$ across skin tones. We hypothesize that fair OOD detectors will have a small $\mathcal{RG}$ score; this means that metrics will perform equally in both skin types. Finally, based on this score, we evaluate the fairness of all OOD models across different skin types and propose a ranking based on the aforementioned score.

4.2. OOD Models and Performance Metrics

We adopt Isolation Forest([15]) and OneClassSVM([24]) as baselines, and an AutoEncoder ([28]), Neural Network Softmax ([17]),

[2] https://github.com/assalaabnk/OOD-in-Dermatology

Table 1. OOD detection performance for samples from two skin tone categories: FST I-IV and FST V-VI. IDD: In-Distribution Dataset, OD: Out-of-distribution Dataset. IF: Isolation Forest. Bold is best in each column. (*) Results were obtained over the only four samples FST V-VI of ISIC 2019. (-): No Densenet model available trained on Fitz17k.

	Datasets		AUROC ↑			F_1 ↑			$\mathcal{RG}$ ↓
Methods	IDD	OD	FST I-IV	FST V-VI	All	FST I-IV	FST V-VI	All	
OneSVM [24]	ISIC 2019	Fitz17k	0.52 ± 0.011	0.53 ± 0.033	0.51 ± 0.011	0.67 ± 0.014	0.70 ± 0.015	0.64 ± 0.008	0.03
IF [15]	ISIC 2019	Fitz17k	0.53 ± 0.004	0.47 ± 0.013	0.52 ± 0.012	0.80 ± 0.009	0.89 ± 0.004	0.84 ± 0.014	0.09
AE [28]	ISIC 2019	Fitz17k	$\mathbf{0.97 \pm 0.005}$	$\mathbf{0.98 \pm 0.007}$	$\mathbf{0.97 \pm 0.005}$	$\mathbf{0.92 \pm 0.010}$	$\mathbf{0.93 \pm 0.013}$	$\mathbf{0.91 \pm 0.012}$	**0.02**
ODIN [16]	ISIC 2019	Fitz17k	0.67 ± 0.006	0.55 ± 0.006	0.64 ± 0.003	0.84 ± 0.001	0.50 ± 0.009	0.84 ± 0.001	0.34
NN Softmax [17]	ISIC 2019	Fitz17k	0.88 ± 0.002	0.84 ± 0.005	0.87 ± 0.001	0.85 ± 0.004	0.75 ± 0.014	0.84 ± 0.003	0.1
OneSVM [24]	Fitz17k	ISIC 2019	0.51 ± 0.014	0.27 ± 0.004(*)	0.51 ± 0.019	0.66 ± 0.028	0.72 ± 0.016(*)	0.66 ± 0.007	0.06
IF [15]	Fitz17k	ISIC 2019	0.57 ± 0.015	0.44 ± 0.000(*)	0.42 ± 0.008	0.85 ± 0.002	$\mathbf{0.94 \pm 0.004}$(*)	0.86 ± 0.005	0.09
AE [28]	Fitz17k	ISIC 2019	$\mathbf{0.93 \pm 0.002}$	$\mathbf{0.95 \pm 0.003}$(*)	$\mathbf{0.93 \pm 0.001}$	$\mathbf{0.89 \pm 0.002}$	0.84 ± 0.002(*)	$\mathbf{0.89 \pm 0.001}$	**0.05**
ODIN [16]	Fitz17k	ISIC 2019	-	-	-	-	-	-	-
NN Softmax [17]	Fitz17k	ISIC 2019	-	-	-	-	-	-	-

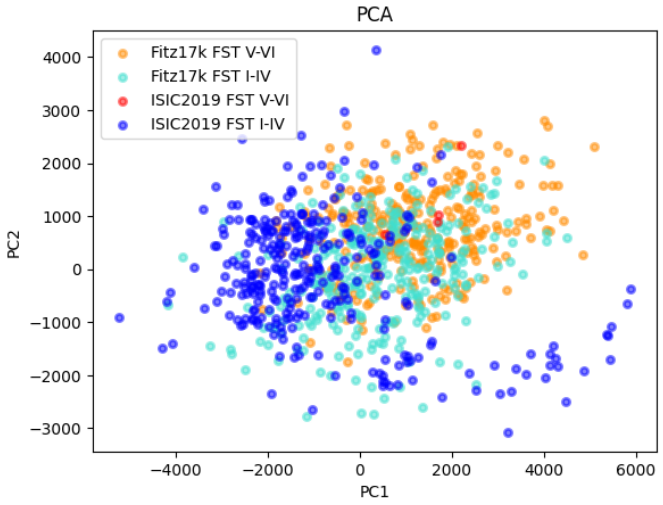


Fig. 2. Principal Component Analysis for a subset of samples from ISIC 2019 [1], and samples from Fitz17k [21] for lighter tones (FST I-IV) and darker tones (FST V-VI).

and ODIN [16] as state-of-the-art OOD methods. All models, training, and testing procedures, as well as hyperparameter setup, can be replicated following the repository code.

IF and OneClassSVM: The IF was configured with 300 estimators and a contamination of 0.1; The OneSVM was configured with $\nu = 0.01$, and $\gamma = 0.0001$. The parameters for both traditional methods were obtained via grid search.

NN Softmax and ODIN: Following [14], we used the DenseNet pre-trained model for diagnosis classification in ISIC-2019. In order to get the optimal parameters for both OOD models, we employ a grid search by testing a variation of parameters consisting of temperature scaling ($\tau = 200$) and magnitudes of perturbation ($\epsilon = 0.0002$) for ODIN. The threshold in these models is called optimal delta ($\delta = 0.996$ for NN Softmax and $\delta = 0.179$ for ODIN. Scores below δ are considered OOD samples, while scores above the threshold are considered in-distribution samples [14, 16, 17].

AE: We trained the AE [28] on ISIC 2019 and Fitz17k datasets. Applied early stopping with patience of 5 steps to get the best-performing model and avoid overfitting; the training halted after 22 epochs for ISIC 2019 and 33 epochs for Fitz17k. The threshold was calculated using Brent's method to find the root of the reconstruction error distributions for IDD and OD samples.

Algorithm 1: Proposed framework pseudo-code for evaluating OOD detectors and computing the representation gap across skin tones.

input : Dataset: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_n\}$
input : Skin Types (FST): $\mathcal{T} = [f_1, f_2, \cdots, f_m, \cdots, f_M]$
output: Representation gap: $\mathcal{RG}$,

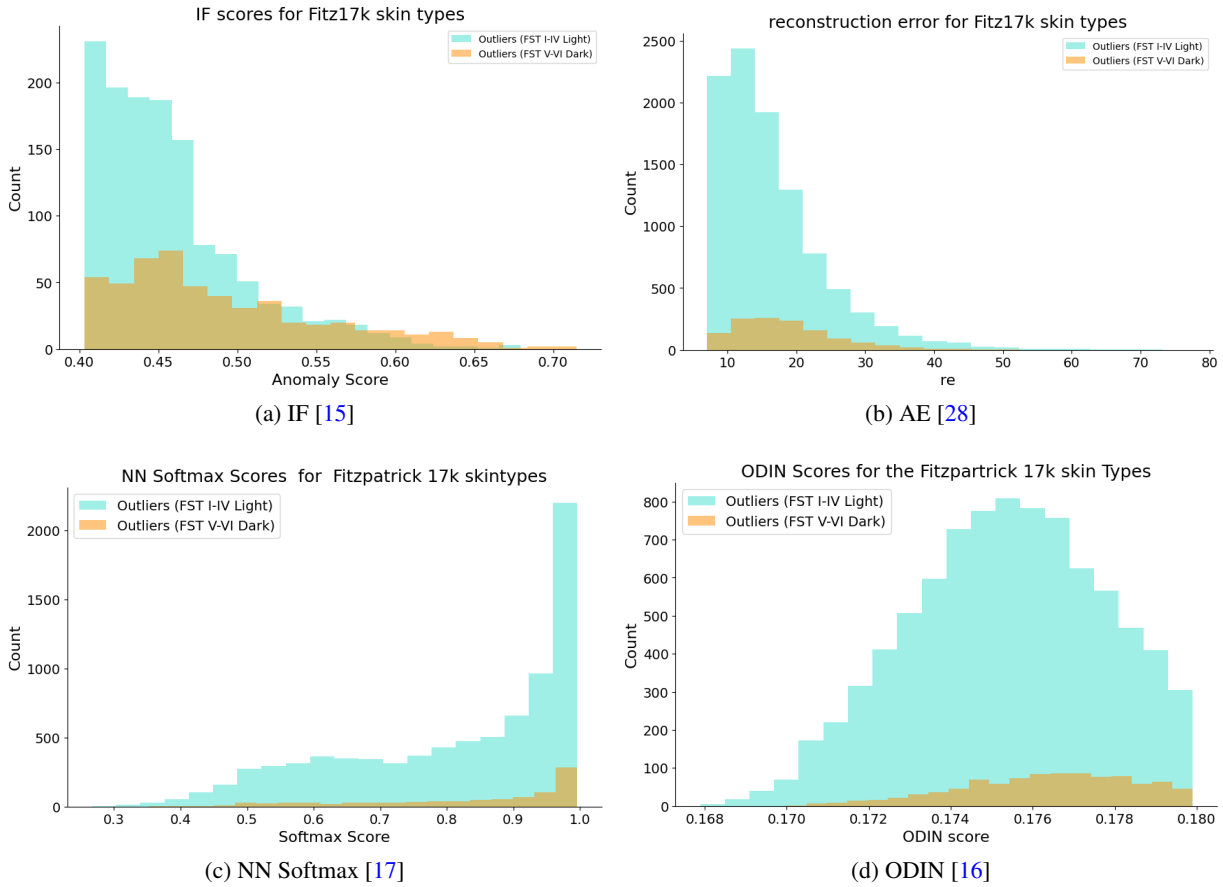
for $\mathcal{D}_i$ *in* $\{\mathcal{D}_1, \cdots, \mathcal{D}_n\}$ **do**
  for *model in [IF, AE, ODIN,* $\cdots$*]* **do**
    $\mathcal{O} \leftarrow SelectOOD(type = model)$;
    $\mathcal{O}_\theta \leftarrow TrainOOD(\mathcal{O}, \mathcal{D}_i)$;
    for f_m *in* $\mathcal{T}$ **do**
      $\mathcal{T} \leftarrow ExtractFST(FST = f_m)$;
    $\mathcal{S}_i \leftarrow Stratify(\mathcal{D}_i, \mathcal{T}_i)$;
    $\mathcal{F}_1^i, AUROC_i \leftarrow Evaluate(\mathcal{O}, \mathcal{S}_i)$;
  $\mathcal{RG} \leftarrow Rank(\mathcal{F}_1^i, AUROC_i, \mathcal{T}_i)$;
return $\mathcal{RG}$;

To measure performance across all models, we employ $AUROC$, F_1-score (F_1), and a RG score. (RG) is the difference in the performance of an OOD detector under different skin types compared to overall performance.

5. RESULTS

Table 1 shows the OOD detection performance for samples of different skin types, FST I-IV and FST V-VI, across traditional and deep learning-based OOD methods. We can observe that in poorly general performance models such as IF and ODIN (AUROC 0.52, and 0.64), the representation gap measured between skin types is wider (0.09 and 0.3 respectively). Compared to a AE([28]) (AUROC 0.97), skin type performance only differs for 0.02.

Similar behavior can be seen in Figure 3 when we observe the different anomalous scores assigned by each method to both skin categories. We can observe that IF assigned higher abnormal scores (≥ 0.5) to $\approx 16\%$ of true outliers from FST I-IV skin type, while $\approx 36\%$ of outliers in FST V-VI were assigned. Similar behavior can be seen in the scores generated by ODIN. In comparison, the AE ($t_0 = 7$ and $t_1 = 8$) assigned $\approx 91\%$ above the threshold for FST I-IV and $\approx 98\%$ for FST V-VI, while we see a reduction in the gap, from 20% to 6.9% between scores. Figure 3(b) also shows a more concentrated set of scores within a small range for all samples across skin types compared to the rest of the score distributions. Un-



(a) IF [15]

(b) AE [28]

(c) NN Softmax [17]

(d) ODIN [16]

Fig. 3. Abnormal scores distributions for all OOD methods stratified by skin tone. We can observe in (a) that IF assigned higher abnormal scores (≥ 0.5) to $\approx 16\%$ of true outliers from FST I-IV skin type and $\approx 36\%$ of outliers in FST V-VI. In (b), the AE assigned $\approx 91\%$ above threshold ($t_0 = 7$ and $t_1 = 8$) for FST I-IV and $\approx 98\%$ for FST V-VI. We see a reduction in the gap, from 20% to 6.9% between scores compared to IF. (c) NN Softmax assigned below threshold (optimal delta 0.996) $\approx 70\%$ of FST V-VI samples of true outliers and $\approx 75\%$ for FST I-IV samples, reducing even further the gap between skin types. (d) ODIN assigned bellow threshold (optimal $\delta = 0.179$) $\approx 28\%$ of FST V-VI samples of true outliers and $\approx 84\%$ for FST I-IV samples.

derstanding the performance of OOD methods beyond average metrics is critical to understanding potential blind spots and developing more fair approaches. A clear example of this can be seen in Table 1, where IF, ODIN and NN Softmax have similar overall F_1 scores, but when we observe performance by skin type, we see a ≈ 0.2 difference in the representation gap in both methods impacting FST I-IV and FST V-VI differently in each approach. This instability of performance may be partially because the Densenet used for NN softmax and ODIN is trained on a dataset that heavily lacks samples of Dark skin tones. For instance, FST V-VI (brown and dark-skinned samples) constitute only 13.5% of Fitz17k and less than 0.01% of ISIC-2019. This could also encourage OOD detectors to classify them to be out of distribution easily.

6. CONCLUSION

We propose an evaluation framework to assess the impact of skin type representation in OOD detectors. We stratify OOD samples based on skin tone and observe imbalanced detection performance for FST V-VI samples, where the samples from darker skin tones are detected as OOD with higher performance in most cases. We showcase the importance of quantifying the representation gap, as existing OOD models with similar overall performance diverge differently on skin types. This information should be taken into account when deciding the type of OOD method to implement in the robustness pipeline. Furthermore, we provided labeled samples for ISIC-2019, and we highlighted the need for more diverse dermatoscopic datasets,while clinical datasets, such as Fitz17k, yield more representation across both labels.

Future work aims to understand further the impact of different proportions of skin types and potential interventions that can be done during training to reduce the representation gap that we observed in this study.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research uses human subject data made available in open access by the corresponding authors (ISIC 2019 [1] and Fitzpatrick 17k Dataset [21]) licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, and ethical approval was not required. No funding was received for this study.

8. REFERENCES

[1] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)," arXiv:1710.05006, 2017.

[2] Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu, "FairPrune: Achieving Fairness Through Pruning for Dermatological Disease Diagnosis," in *MICCAI*, Cham, 2022, Springer, pp. 743–753.

[3] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti, "On hate scaling laws for data-swamps," *arXiv preprint arXiv:2306.13141*, 2023.

[4] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[5] Arieh Gomolin, Elena Netchiporouk, Robert Gniadecki, and Ivan V Litvinov, "Artificial intelligence applications in dermatology: Where do we stand?," *Front. Med.*, vol. 7, 2020.

[6] Pengyi Zhang, Yunxin Zhong, and Xiaoqiong Li, "Melanet: A deep dense attention network for melanoma detection in dermoscopy images," 2019.

[7] Adewole S Adamson and Avery Smith, "Machine learning and health care disparities in dermatology," *JAMA Derm.*, vol. 154, no. 11, pp. 1247–1248, 2018.

[8] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," arXiv:2001.08103, 2020.

[9] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney, "Fairness of classifiers across skin tones in dermatology," in *Proc. Int. Conf. Med. Image Comp. Comp.-Assist. Interv.*, 2020, pp. 320–329.

[10] Girmaw Abebe Tadesse, Celia Cintas, Kush R Varshney, Peter Staar, Chinyere Agunwa, Skyler Speakman, Justin Jia, Elizabeth E Bailey, Ademide Adelekun, Jules B Lipoff, et al., "Skin tone analysis for representation in educational materials using machine learning," *npj Digital Medicine*, vol. 6, no. 1, pp. 151, 2023.

[11] Thorsten Kalb, Kaisar Kushibar, Celia Cintas, Karim Lekadir, Oliver Diaz, and Richard Osuala, "Revisiting skin tone fairness in dermatological lesion classification," in *Workshop on Clinical Image-Based Procedures*. Springer, 2023, pp. 246–255.

[12] Usha Lee Mcfarling, "Dermatology faces a reckoning: Lack of darker skin in textbooks and journals harms care for patients of color," July 2020.

[13] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff, "Skin color in dermatology textbooks: an updated evaluation and analysis," *Journal of the American Academy of Dermatology*, vol. 84, no. 1, pp. 194–196, 2021.

[14] Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush Varshney, "Out-of-distribution detection in dermatology using input perturbation and subset scanning," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–4.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[16] Shiyu Liang, Yixuan Li, and R. Srikant, "Principled detection of out-of-distribution examples in neural networks," arXiv:1706.02690, 2017.

[17] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[18] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri, "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*, 2021, pp. 1820–1828.

[19] Marcus Wilkes, Caradee Y. Wright, Johan L. du Plessis, and Anthony Reeder, "Fitzpatrick skin type, individual typology angle, and melanin index in an African population," *JAMA Dermatol.*, vol. 151, no. 8, pp. 902–903, Aug. 2015.

[20] Eman Rezk, Mohamed Eltorki, Wael El-Dakhakhni, et al., "Improving skin color diversity in cancer detection: deep learning approach," *JMIR Dermatology*, vol. 5, no. 3, pp. e39143, 2022.

[21] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *arXiv preprint arXiv:2207.02942*, 2022.

[22] Peter J. Bevan and Amir Atapour-Abarghouei, "Skin Deep Unlearning: Artefact and Instrument Debiasing in the Context of Melanoma Classification," 2021.

[23] Neda Alipour, Ted Burke, and Jane Courtney, "Skin Type Diversity: a Case Study in Skin Lesion Datasets," July 2023.

[24] Stephan Dreiseitl, Melanie Osl, Christian Scheibböck, and Michael Binder, "Outlier detection with one-class svms: an application to melanoma prognosis," in *AMIA annual symposium proceedings*. American Medical Informatics Association, 2010, vol. 2010, p. 172.

[25] Christian Desrosiers Xuan Li, Yuchen Lu and Xue Liu, "Out-of-distribution detection for skin lesion images with deep isolation forest," in *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*. 2020, pp. 91–100, Springer International Publishing.

[26] Yuchen Lu and Peng Xu, "Anomaly detection for skin disease images using variational autoencoder," *arXiv preprint arXiv:1807.01349*, 2018.

[27] Muhammad Zaida, Shafaqat Ali, Mohsen Ali, Sarfaraz Hussein, Asma Saadia, and Waqas Sultani, "Out of distribution detection for skin and malaria images," *arXiv preprint arXiv:2111.01505*, 2021.

[28] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*. Springer, 2011, pp. 52–59.

[29] Thomas B Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.