Achieving Fairness Through Channel Pruning for Dermatological Disease Diagnosis

Qingpeng Kong[1], Ching-Hao Chiu[2], Dewen Zeng[1], Yu-Jen Chen[2], Tsung-Yi Ho[3], Jingtong hu[4], and Yiyu Shi[1(✉)]

[1] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
{qkong2, dzeng2, yshi4}@nd.edu
[2] Department of Computer Science, National Tsing Hua University, Taiwan
gwjh101708@gapp.nthu.edu.tw, chenjuzen@gmail.com
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong tyho@cse.cuhk.edu.hk
[4] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA jthu@pitt.edu

Abstract. Numerous studies have revealed that deep learning-based medical image classification models may exhibit bias towards specific demographic attributes, such as race, gender, and age. Existing bias mitigation methods often achieve a high level of fairness at the cost of significant accuracy degradation. In response to this challenge, we propose an innovative and adaptable Soft Nearest Neighbor Loss-based channel pruning framework, which achieves fairness through channel pruning. Traditionally, channel pruning is utilized to accelerate neural network inference. However, our work demonstrates that pruning can also be a potent tool for achieving fairness. Our key insight is that different channels in a layer contribute differently to the accuracy of different groups. By selectively pruning critical channels that lead to the accuracy difference between the privileged and unprivileged groups, we can effectively improve fairness without sacrificing accuracy significantly. Experiments conducted on two skin lesion diagnosis datasets across multiple sensitive attributes validate the effectiveness of our method in achieving a state-of-the-art trade-off between accuracy and fairness. Our code is available at https://github.com/Kqp1227/Sensitive-Channel-Pruning.

Keywords: Dermatological Disease Diagnosis · AI Fairness · Medical Image Analysis · Channel Pruning

1 Introduction

In AI-powered medical image analysis, deep neural networks (DNNs) are adept at extracting vital statistical details like colors and textures from the provided training data. This data-centric approach enables the network to grasp task-specific characteristics, thereby enhancing accuracy in the intended task. Nonetheless,

in the pursuit of optimal accuracy, the network might rely on certain data attributes present in some instances but not in others, potentially leading to biases against specific demographics, such as skin tone or gender. For example, in dermatological disease diagnosis, DNNs tend to exhibit different prediction accuracy across patients of different skin colors or genders [9,14,18,20].

Various methods have been proposed to mitigate bias in machine learning models. These studies can be classified based on the stage when the debiasing techniques are applied: pre-processing, in-processing, and post-processing. Pre-processing methods [13,25,27,28] are deployed to mitigate biases within the dataset, thus fostering fairness. In-processing methods [1,5,6,17,22,23,29] typically involve the development of distinct network architectures to minimize fairness discrepancies. As for post-processing methods [7,10,28], calibration is done at inference time by considering both the model's prediction and the sensitive attribute as input. While these methods can enhance fairness and mitigate bias, they often struggle to uphold high accuracy levels while achieving great fairness improvement.

To attain better accuracy-fairness trade-off, we introduce a novel framework that leverages channel pruning, which is orthogonal to and can be applied in conjunction with any existing bias mitigation method. Our approach is based on the observation that the output channels of convolutional layers within a convolutional neural network exhibit varying sensitivities to different sensitive attributes, such as skin color or gender. Those channels that are sensitive to these attributes contribute greatly to the accuracy gap between privileged and unprivileged groups (defined as "sensitive channels" in this paper), while those that are not would have little impact (defined as "insensitive channels"). Conventionally, the channel pruning technique selects channels that are not important to accuracy and prunes them to accelerate DNN inference. We can make "off-label" use of this technique and instead prune the sensitive channels to promote feature diversity and diminish the impact of discriminatory features, thus mitigating bias. Specifically, we utilize the Soft Nearest Neighbor Loss (SNNL) [8] to measure the entanglement level between different sensitive attributes in the representation space, and identity the sensitive channels as those with feature maps exhibiting low SNNL values.

Through extensive experimentation on different skin disease datasets and backbone models, we demonstrated that our framework can achieve the state-of-the-art trade-off between accuracy and fairness. We also demonstrated that it can be effectively applied to various existing bias mitigation frameworks to boost their trade-offs.

2 Method

2.1 Problem Formulation

Consider a dataset $D = \{x_i, y_i, c_i\}, i = 1, ..., N$, where $x_i \in \mathbf{x}$ represents an input image ($\mathbf{x}$ represents all the input images), y_i is the class label, and c_i represents a

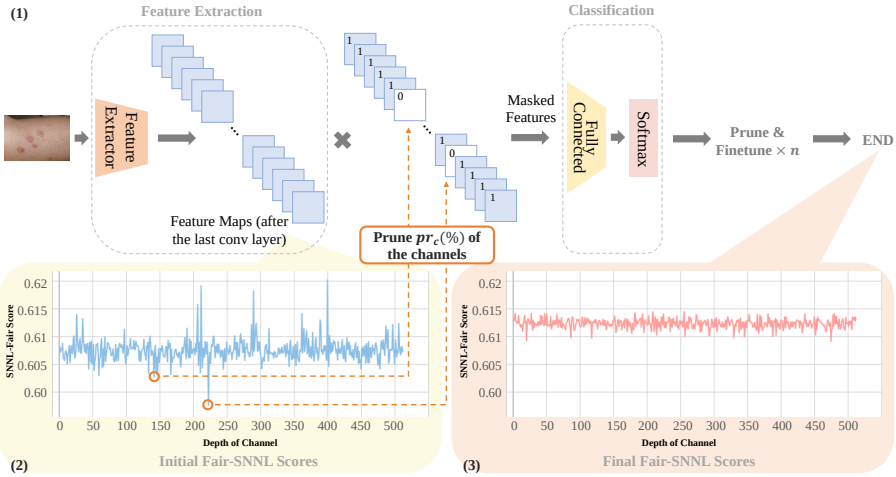

Fig. 1. Illustration of (1) proposed channel pruning framework and (2) distribution of initial SNNL-Fair scores, generated using the Fitzpatrick-17k dataset and VGG-11 backbone and (3) distribution of the final SNNL-Fair scores after n iterations of channel pruning. pr_c represents channel pruning ratio in one iteration, and n is the total number of pruning and fine-tuning iteration(s) needed under the stopping criteria.

sensitive attribute. Additionally, consider a pre-trained classification model $f_\theta(\cdot)$ with parameters θ that maps input x_i to a predicted label $\hat{y}_i = f_\theta(x_i)$.

Our goal is to selectively prune channels from a specific layer of the model $f_\theta(\cdot)$ that are sensitive to attribute c_i, thereby mitigating their influence on the overall fairness. In this study, we only consider binary sensitive attributes, where $c_i \in \{0, 1\}$. Here, $c_i = 0$ corresponds to unprivileged samples, indicating discrimination by the model, while $c_i = 1$ corresponds to privileged samples.

2.2 Channel Pruning for Fairness

Using SNNL to reflect entanglement between different sensitive groups. Conventionally, as mentioned in [8], the Soft Nearest Neighbor Loss (SNNL) measures the entanglement state of features based on their class representation. A higher SNNL value indicates that the features belonging to different classes are intertwined, whereas a lower value suggests that the features are more separated by class. We can extend the concept and use it to evaluate the entanglement state of features based on their sensitive attributes. We define $\mathbf{m}$ for the feature maps extracted after the specific convolutional layer using the input images $\mathbf{x}$. For the t-th batch, let b denote the batch size. $\mathbf{m}^{(t),k} = \{m_1^{(t),k}, m_2^{(t),k}, ..., m_b^{(t),k}\}$ denotes the feature maps extracted from the specific channel k for the batch, and $\mathbf{c}^{(t)} = \{c_1^{(t)}, c_2^{(t)}, ..., c_b^{(t)}\}$ represents the corresponding sensitive labels. The modified SNNL $l_{sn}^{(t),k}$ to reflect sensitive attribute entanglement for channel k of this batch t, with temperature T, can be calculated as follows:

$$l_{sn}^{(t),k}(\mathbf{m}^{(t),k}, \mathbf{c}^{(t)}, T) = -\frac{1}{b} \sum_{i \in 1..b} \log \left(\frac{\sum\limits_{\substack{j \in 1,\ldots,b \\ j \neq i \\ c_i^{(t)} = c_j^{(t)}}} e^{-\frac{\| m_i^{(t),k} - m_j^{(t),k} \|^2}{T}}}{\sum\limits_{\substack{p \in 1,\ldots,b \\ p \neq i}} e^{-\frac{\| m_i^{(t),k} - m_p^{(t),k} \|^2}{T}}} \right) \tag{1}$$

Assume that the entire dataset forms a total of n_b batches, then we will obtain a total of n_b SNNL values $l_{sn}^{(1),k}, l_{sn}^{(2),k}, ..., l_{sn}^{(n_b),k}$. We then average these values to obtain a single value, defined as SNNL-Fair score S^k for channel k, i.e., $S^k = \frac{1}{n_b} \sum_{t=1}^{n_b} l_{sn}^{(t),k}$. For interested readers, a visual illustration of the SNNL-Fair calculation is included our supplementary materials.

Using channel pruning to achieve fairness. In our methodology, we adopt an in-processing approach that entails pruning the sensitive channels. Even though our method can be applied to any layers, we choose to only prune the output channels from the last convolutional layer. This is because feature maps from deeper layers contain more detailed information about the input and thus providing better estimation of SNNL and more effective pruning. In the experiments we will present ablation studies to show that pruning the output channels of the last convolutional layer indeed presents the best opportunity for bias mitigation and accuracy preservation.

As discussed, the SNNL-Fair scores reflect the entanglement state of different sensitive attributes in the feature maps of each channel. A smaller SNNL-Fair score for a channel indicates a stronger ability to differentiate between different sensitive attributes in its feature maps, potentially creating bias. As we can see from Fig. 1(2), using VGG-11 as backbone and Fitzpatrick-17k dataset as an example, there indeed exist certain channels with significantly lower SNNL-Fair scores than others (circled in orange), which means that different sensitive attributes are less entangled in these channels. To enhance fairness, we propose to prune them. Specifically, we sort the channels based on their SNNL-Fair scores, and prune the bottom pr_c percent, where pr_c is the channel pruning ratio. In the ablation study, we will study its impact on accuracy and fairness.

Note that similar to existing channel pruning methods for inference speedup, we can adopt the iterative pruning for fairness enhancement. After each iteration of pruning, we fine-tune the pruned model on the training dataset to update the weights. The iterations stop either when the accuracy drop compared with the original unpruned model is above a threshold th_{acc}, or when the fairness improvement compared with the previous iteration is smaller than a threshold th_{fair}.

We illustrate our framework in Fig. 1. Upon implementing our framework, we observed improvements in the overall SNNL-Fair scores, as illustrated in Fig. 1(3). The blue and red curves represent the SNNL-Fair score of each output

channel in the last convolutional layer of VGG-11 before and after applying our framework, respectively. It is interesting to see that even though we only choose to prune the channels with lowest SNNL-Fair socre, after several iterations of pruning and fine-tuning, the SNNL-Fair scores across all the channels have increased. As previously mentioned, higher SNNL-Fair scores suggest improved fairness, implying that the model's fairness has been enhanced.

2.3 Pruning Recipe

Our framework prunes a pre-trained model by the following steps:

1. For each channel k in the last convolutional layer, compute the SNNL-Fair Scores S^k using the given training dataset D.
2. Sort the channels in descending order of S^k.
3. Prune pr_c percent of the channels with smallest SNNL-Fair scores, resulting in a new model $f'_\theta(\cdot)$.
4. Fine-tune the new model $f'_\theta(\cdot)$ on D.
5. Repeat steps 1-4 until the stopping criteria are met.

3 Experiments and Results

3.1 Experiments

Dataset and Models The proposed methods are evaluated on two dermatology datasets for disease classification: the Fitzpatrick-17k dataset[9] and the ISIC 2019 challenge dataset[3]. The Fitzpatrick-17k dataset comprises 16,577 images representing 114 different skin conditions. Skin tones are categorized from 1 to 6, with 1 being the lightest and 6 the darkest. We group types 1-3 as light skin and types 4-6 as dark skin. For our analysis, we consider skin tones as the sensitive attribute, representing the situation of "explicit" attribute (observable from the images). Containing a diverse collection of 25,331 dermoscopic images, the ISIC 2019 dataset offers resource for studying skin conditions across various 8 diagnostic categories. We consider gender as the sensitive attribute, representing the situation of "implicit" attribute (unobservable from the images). We follow existing works and use VGG-11 [19] as the backbone for Fitzpatrick-17k and ResNet18 [11] for ISIC2019. This allows us to assess the adaptability of our framework across different backbones and various datasets. All the training settings were the same with those described in [2,26].

A standard preprocessing step for both datasets involves resizing all the images to a uniform size of 128×128 pixels. To augment the data and improve generalization, we apply various techniques such as random horizontal flipping, vertical flipping, rotation, scaling, and autoaugment, consistent with [2,26].

Fairness Metrics We employ the multi-class equalized opportunity (*Eopp*) and equalized odds (*Eodd*) metrics [10] to evaluate the fairness of our model. We follow the approach of [26] for calculating these metrics. As there is always

trade-off between accuracy and fairness, to better compare different methods, we employ the *FATE* metric proposed by [27] to evaluate the overall performance of bias mitigation methods. A higher *FATE* score indicates that the model reaches a better trade-off between fairness and accuracy. [1] It can be calculated as $FATE_{FC} = \frac{ACC_m - ACC_b}{ACC_b} - \lambda \frac{FC_m - FC_b}{FC_b}$, where FC can use one of the values from *Eopp0*, *Eopp1*, or *Eodd*. ACC represents accuracy, which can be chosen from accuracy scores, including F1-score, Recall, and Precision. Here, we choose the F1-score for comparison. The subscripts m and b represent the bias mitigation and baseline models, respectively. The parameter λ is used to modify the significance of fairness in the ultimate evaluation, and we set $\lambda = 1.0$ following the same setting used in [27].

Table 1. Results of accuracy and fairness of various methods on the Fitzpatrick-17k and VGG-11 backbone, using skin tone as the sensitive attribute. The dark skin is the privileged group. *FATE* metrics are evaluated using the vanilla VGG-11 as the baseline.

		Accuracy			Fairness		
Method	Skin Tone	Precision	Recall	F1-score	*Eopp0*↓ / *FATE*↑	*Eopp1*↓ / *FATE*↑	*Eodd*↓ / *FATE*↑
VGG-11 [19]	Dark	0.563	0.581	0.546	0.0013 / 0.0000	0.361 / 0.0000	0.182 / 0.0000
	Light	0.482	0.495	0.473			
	Avg.↑	0.523	0.538	0.510			
	Diff.↓	0.081	0.086	0.073			
HSIC [16]	Dark	0.548	0.522	0.513	0.0013 / -0.0196	0.331 / 0.1235	0.166 / 0.1305
	Light	0.513	0.506	0.486			
	Avg.↑	0.530	0.515	0.500			
	Diff.↓	0.040	0.018	0.029			
MFD [12]	Dark	0.514	0.545	0.503	<u>0.0011</u> / 0.0950	0.334 / 0.0160	0.166 / 0.0291
	Light	0.489	0.469	0.457			
	Avg.↑	0.502	0.507	0.480			
	Diff.↓	0.025	0.076	0.046			
FairAdaBN [27]	Dark	0.544	0.541	0.524	0.0012 / 0.0583	0.341 / 0.0368	0.171 / 0.0418
	Light	0.484	0.509	0.476			
	Avg.↑	0.514	0.525	0.500			
	Diff.↓	0.033	0.060	0.048			
FairPrune [26] ($pr = 35\%, \beta = 0.33$)	Dark	0.567	0.519	0.507	**0.0008 / 0.3317**	0.330 / 0.0329	0.165 / 0.0405
	Light	0.496	0.477	0.459			
	Avg.↑	0.531	0.498	0.483			
	Diff.↓	0.071	0.042	0.048			
ME-FairPrune [2]	Dark	0.564	0.529	0.523	0.0012 / <u>0.1005</u>	<u>0.305</u> / <u>0.1787</u>	<u>0.152</u> / <u>0.1884</u>
	Light	0.542	0.535	0.522			
	Avg.↑	0.553	0.532	0.522			
	Diff.↓	0.022	0.006	0.001			
SCP-FairPrune (Ours) ($pr_c = 2\%, n = 3$)	Dark	0.568	0.576	0.547	0.0012 / 0.0965	**0.278 / 0.2495**	**0.139 / 0.2559**
	Light	0.499	0.504	0.492			
	Avg.↑	0.533	0.540	0.520			
	Diff.↓	0.069	0.073	0.055			

3.2 Comparison with State-of-the-art

We present a comprehensive comparison of our framework against various bias mitigation baselines [2,12,15,16,19,21,24,26,27,30]. Due to space limit, we are

only able to include the most recent baselines here, and a comprehensive comparison with several other baselines are included in the supplementary material. We denote our framework as SCP (SNNL based Channel Pruning), and apply it to the model from state-of-the-art bias mitigation method FairPrune [26]. The resulting model is named SCP-FairPrune. n is the number of iterations undertaken by our framework before the stopping criteria are met, and pr_c is the channel pruning ratio in each iteration (set to 2%). **Bold** and Underline fonts denote the best and the second-best performance, respectively.

Results on Fitzpatrick-17k Dataset The results presented in Table 1 demonstrate the superior performance of our framework compared to other methods, as evidenced by the lowest *Eopp1* and *Eodd* scores. Notably, the *FATE* metric of *Eopp1* and *Eodd* demonstrate enhancements of 39.7% and 35.8% over the last state-of-the-art method ME-FairPrune respectively, which proves its effectiveness in enhancing trade-off between fairness and accuracy. Note that this improvement is achieved within only three iterations of pruning and less than 10% of the channels pruned in total.

As our framework is orthogonal to existing bias mitigation methods, we also evaluate its effectiveness when being applied to other baselines, including the vanilla VGG-11 (i.e., SCP-VGG-11) and another bias mitigation method HSIC (i.e., SCP-HSIC). Results show that the SCP-VGG-11 and SCP-HSIC both achieve positive *FATE* metrics over their respective baselines (details in the supplementary material). This demonstrates that our framework can improve the trade-off between accuracy and fairness on a wide range of methods.

Results on ISIC 2019 Dataset With ResNet-18 as backbone, the results presented in Table 2 show that our framework still achieves the highest *FATE* metric of *Eopp1* and *Eodd*, demonstrating enhancements of 22.0% and 33.9% over ME-FairPrune, respectively.

3.3 Ablation Study

Effect of Channel Pruning Ratio pr_c To explore how altering channel pruning ratio pr_c affects accuracy and fairness, we employ our framework on a pretrained VGG-11 model sourced from the Fitzpatrick-17k dataset and experiment with various ratios. The original model without pruning is used as the baseline for *FATE* metric evaluation of *Eodd*. As depicted in Figure 2(a), the model achieves the highest *FATE* score when setting the pruning ratio to 2%. It is also interesting to note that the optimal results can be achieved by no more than three pruning iterations, demonstrating the efficiency of our proposed framework; running more-than-necessary pruning iterations will cause the performance to drop, showcasing the importance of using proper stopping criteria (Note that in this study and the study below, we allowed the pruning iterations to go beyond the stopping criteria to show the peaks).

Table 2. Results of accuracy and fairness of various methods on the ISIC 2019 and ResNet-18 backbone, using gender as the sensitive attribute. Female is the privileged group. *FATE* metrics are evaluated using the vanilla ResNet-18 as the baseline.

Method	Gender	Accuracy Precision	Recall	F1-score	Fairness *Eopp0*↓ / *FATE*↑	*Eopp1*↓ / *FATE*↑	*Eodd*↓ / *FATE*↑
ResNet-18 [11]	Female	0.793	0.721	0.746	0.006 / 0.0000	0.044 / 0.0000	0.022 / 0.0000
	Male	0.731	0.725	0.723			
	Avg.↑	0.762	0.723	0.735			
	Diff.↓	0.063	0.004	0.023			
HSIC [16]	Female	0.744	0.660	0.696	0.008 / 0.8194	0.042 / 0.0068	0.020 / 0.0559
	Male	0.718	0.697	0.705			
	Avg.↑	0.731	0.679	0.700			
	Diff.↓	0.026	0.037	0.009			
MFD [12]	Female	0.770	0.697	0.726	**0.005** / 0.9171	0.051 / -0.1482	0.024 / -0.0758
	Male	0.772	0.726	0.744			
	Avg.↑	0.771	0.712	0.735			
	Diff.↓	0.002	0.029	0.018			
FairAdaBN [27]	Female	0.776	0.724	0.746	**0.005** / **0.9273**	0.038 / 0.1548	0.019 / 0.1586
	Male	0.758	0.728	0.739			
	Avg.↑	0.767	0.726	0.743			
	Diff.↓	0.008	0.004	0.007			
FairPrune [26] ($pr = 35\%, \beta = 0.33$)	Female	0.776	0.711	0.734	0.007 / 0.8729	0.026 / 0.4039	0.014 / 0.3617
	Male	0.721	0.725	0.720			
	Avg.↑	0.748	0.718	0.727			
	Diff.↓	0.055	0.014	0.014			
ME-FairPrune [2]	Female	0.770	0.723	0.742	0.006 / 0.9018	0.020 / 0.5513	0.010 / 0.5533
	Male	0.739	0.728	0.730			
	Avg.↑	0.755	0.725	0.736			
	Diff.↓	0.032	0.005	0.012			
SCP-FairPrune (Ours) ($pr_c = 2\%, n = 3$)	Female	0.787	0.701	0.736	0.006 / 0.9018	**0.015 / 0.6724**	**0.006 / 0.7411**
	Male	0.765	0.712	0.735			
	Avg.↑	0.776	0.707	0.736			
	Diff.↓	0.022	0.012	0.001			

Effect of Pruning Channels in Different Layers The VGG-11 model features a total of 8 convolutional layers, and our framework prunes the output channels of the 8th (last) layer. We experimented with pruning the channels in other layers (index l) using a 2% pruning ratio on the Fitzpatrick-17k dataset and recorded the resulting *FATE* metrics of *Eodd*. In Fig. 2(b), though pruning the channels in other layers can also improve fairness and accuracy trade-off, it is evident that working with the last convolutional layer yields the best result.

4 Conclusion

We tackle the challenge of declining fairness in convolutional neural networks by introducing a Channel Pruning framework based on SNNL (Soft Nearest Neighbor Loss). Our framework is versatile, compatible with a range of bias mitigation techniques, and selectively prunes the sensitive channels from specific convolutional layers. Experiment results underscore the effectiveness of our framework, showcasing superior trade-offs between accuracy and fairness compared to state-of-the-art methods across two dermatological disease datasets.

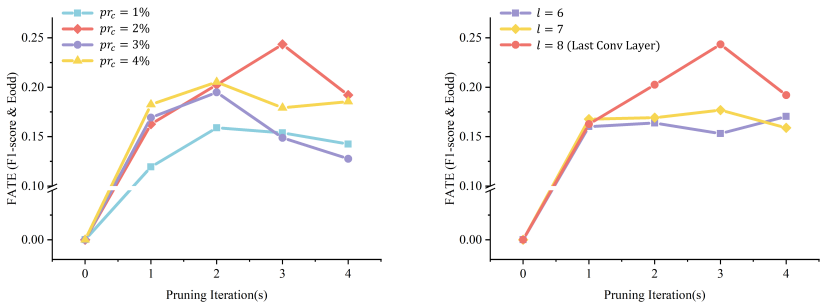

Fig. 2. *FATE* of *Eodd* v.s. pruning iterations with (a) varying channel pruning ratio pr_c and (b) varying layer from which the output channels are pruned, using VGG-11 on Fitzpatrick-17k dataset.

Declarations

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.

References

1. Chiu, C.H., Chen, Y.J., Wu, Y., Shi, Y., Ho, T.Y.: Achieve fairness without demographics for dermatological disease diagnosis. Medical Image Analysis **95**, 103188 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103188, https://www.sciencedirect.com/science/article/pii/S1361841524001130
2. Chiu, C.H., Chung, H.W., Chen, Y.J., Shi, Y., Ho, T.Y.: Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 97–107. Springer (2023)
3. 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)
4. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: Proceedings of the european conference on computer vision (eccv) workshops. pp. 0–0 (2018)
5. Deng, W., Zhong, Y., Dou, Q., Li, X.: On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations (2023)
6. Fan, D., Wu, Y., Li, X.: On the fairness of swarm learning in skin lesion classification. In: Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 2. pp. 120–129. Springer (2021)
7. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 329–338 (2019)
8. Frosst, N., Papernot, N., Hinton, G.E.: Analyzing and improving representations with the soft nearest neighbor loss. In: International Conference on Machine Learning (2019)
9. 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)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
11. 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)
12. Jung, S., Lee, D., Park, T., Moon, T.: Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12115–12124 (2021)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
14. 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–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI. pp. 320–329. Springer (2020)

15. LeCun, Y., Denker, J., Solla, S.: Optimal brain damage. Advances in neural information processing systems **2** (1989)
16. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8227–8236 (2019)
17. Roh, Y., Lee, K., Whang, S., Suh, C.: Fr-train: A mutual information-based approach to fair and robust training. In: International Conference on Machine Learning. pp. 8147–8157. PMLR (2020)
18. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., 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)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. 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)
21. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision. pp. 4068–4076 (2015)
22. Wan, M., Zha, D., Liu, N., Zou, N.: In-processing modeling techniques for machine learning fairness: A survey **17**(3) (mar 2023). https://doi.org/10.1145/3551390
23. Wang, A., Russakovsky, O.: Directional bias amplification. In: International Conference on Machine Learning. pp. 10882–10893. PMLR (2021)
24. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020)
25. Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., Ren, K.: Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10379–10388 (2022)
26. Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J.: Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 743–753. Springer (2022)
27. Xu, Z., Zhao, S., Quan, Q., Yao, Q., Zhou, S.K.: Fairadabn: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. arXiv preprint arXiv:2303.08325 (2023)
28. Yang, J., Soltan, A., Eyre, D., Yang, Y., Clifton, D.: An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ digital medicine **6**, 55 (03 2023)
29. Yao, R., Cui, Z., Li, X., Gu, L.: Improving fairness in image classification via sketching (2022)
30. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)