www.miccai2022.org

MICCAI2022 Singapore
25th International Conference on Medical Image Computing and Computer Assisted Intervention
September 18-22, 2022
Resorts World Convention Centre Singapore
MICCAI

FairPrune: Achieving Fairness Through Pruning for Dermatological Disease Diagnosis

Yawen Wu*[1], Dewen Zeng*[2], Xiaowei Xu[3], Yiyu Shi[2], Jingtong Hu[1]

[1] University of Pittsburgh, [2] University of Notre Dame,

[3] Guangdong Provincial People's Hospital, China

* Equal contributions

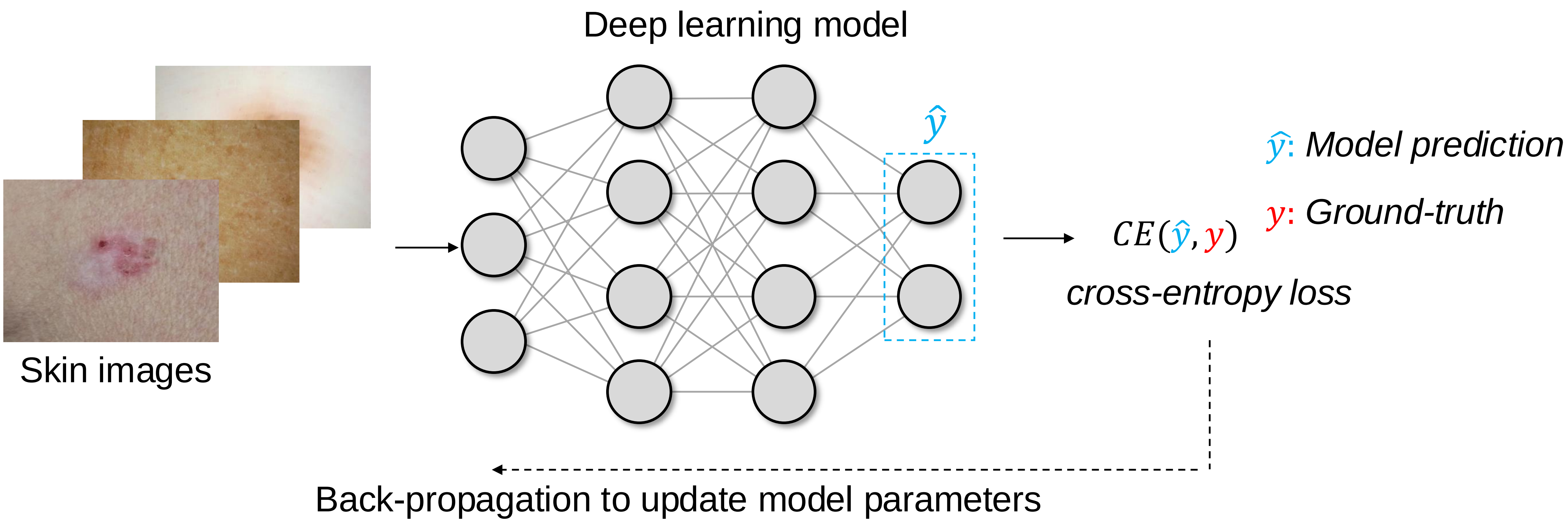
University of Pittsburgh

UNIVERSITY OF NOTRE DAME

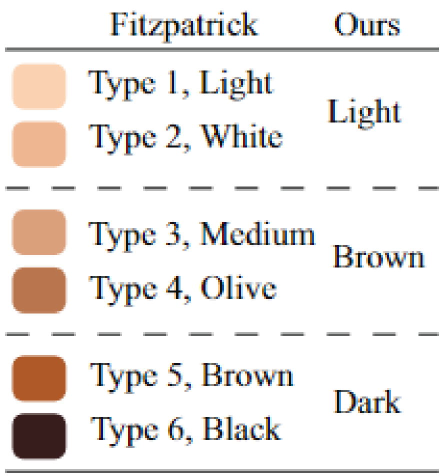
Outline

- Background and Motivation
- FairPrune
- Experimental Results

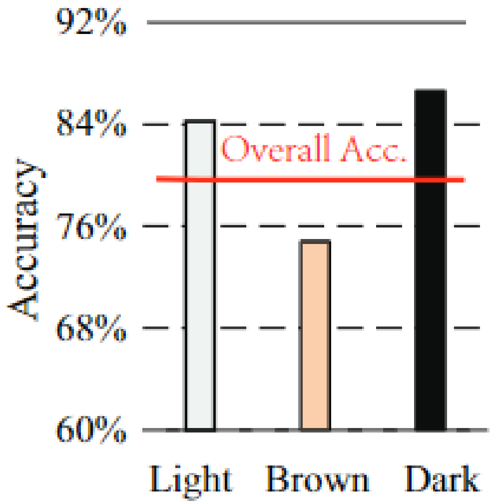
Background of Dermatological Disease Diagnosis by Machine Learning



Motivation



Skin tones



Different accuracy of different skin tones

Machine learning-based dermatological disease diagnosis methods usually targets a high accuracy.

- The learned models show **discrimination** towards certain demographic groups.
  - Models show a high accuracy on some demographic groups, but low on others.
- Caused by that the models use the information present in some data but not in other data.
  - Information such as skin tones, genders
- It is necessary to effectively remove this information for a fair model.

Challenges of Achieving Fair Dermatological Diagnosis Models

1. Completely removing the model's arability to predict a protected attribute is challenging since this attribute can also be predicted from the combination of other attributes.
2. Aggressive suppression of sensitive information will greatly degrade the model's accuracy.

Contributions

1. We propose FairPrune, a technique to achieve fairness via pruning.
    - Conventionally, pruning is used to reduce the model size for efficient inference.
    - We show that pruning can be a powerful tool for fairness.
2. By controlling the parameters to prune, we can reduce the accuracy difference between the privileged group and the unprivileged group.
    - Improving fairness while keeping their overall accuracy as high as possible.
3. We measure the importance of each parameter to different groups by its saliency.

Revisiting Parameter Saliency

Saliency reflects the increase of prediction error after pruning some parameters.

The saliency of parameter θ_i: $\boxed{\Delta E = h_{ii} \frac{\partial^2 E}{\partial^2 \theta_i}}$

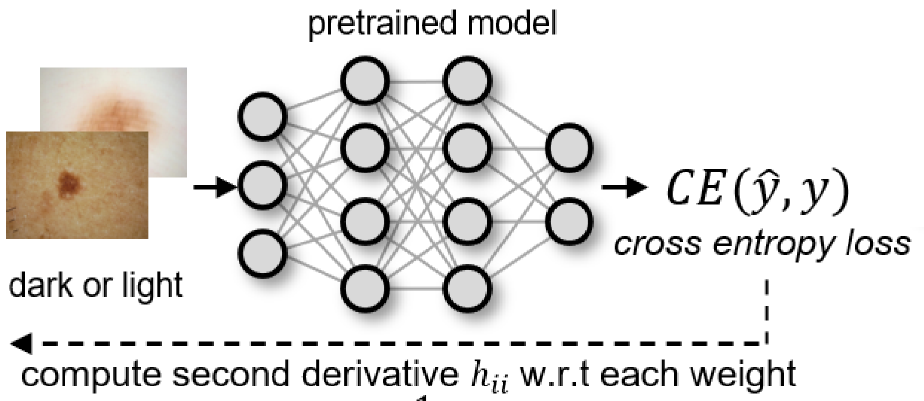
$$
\begin{aligned}
\Delta E &= E(D|\Theta = 0) - E(D) \\
&= -\sum_i g_i \theta_i + \frac{1}{2}\sum_i h_{ii}\theta_i^2 + \frac{1}{2}\sum_{i \neq j} h_{ij}\theta_i\theta_j + O(||\Theta||^3) = \boxed{\frac{1}{2}\sum_i h_{ii}\theta_i^2}.
\end{aligned}
$$

1. $g_i = \frac{\partial E}{\partial \theta_i}$ the gradient of E with respect to θ_i. Close to 0 for pre-trained models.

2. $h_{ii} = \frac{\partial^2 E}{\partial^2 \theta_i}$ The diagonal element in row i and column i of the second derivate Hessian matrix H.

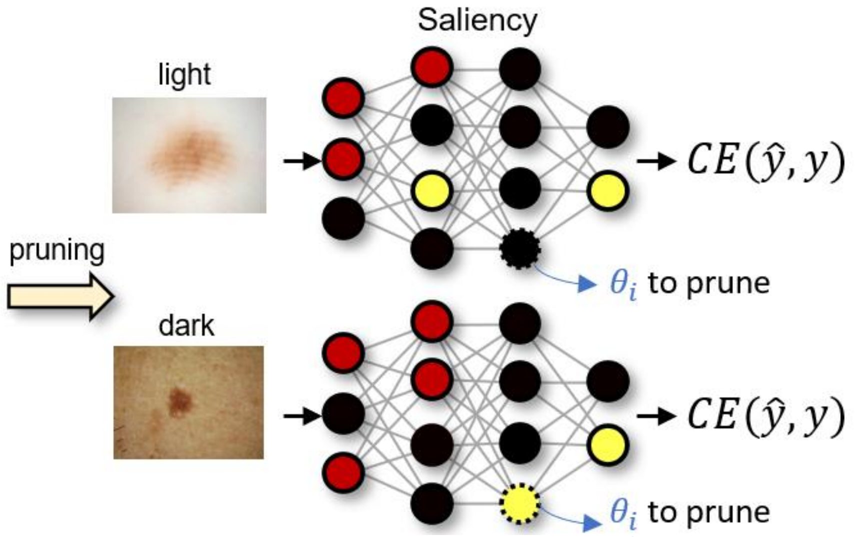
3 and 4. $\frac{1}{2}\sum_{i \neq j} h_{ii}\, \theta_i\, \theta_j + O(||\Theta||^3)$ Neglectable.

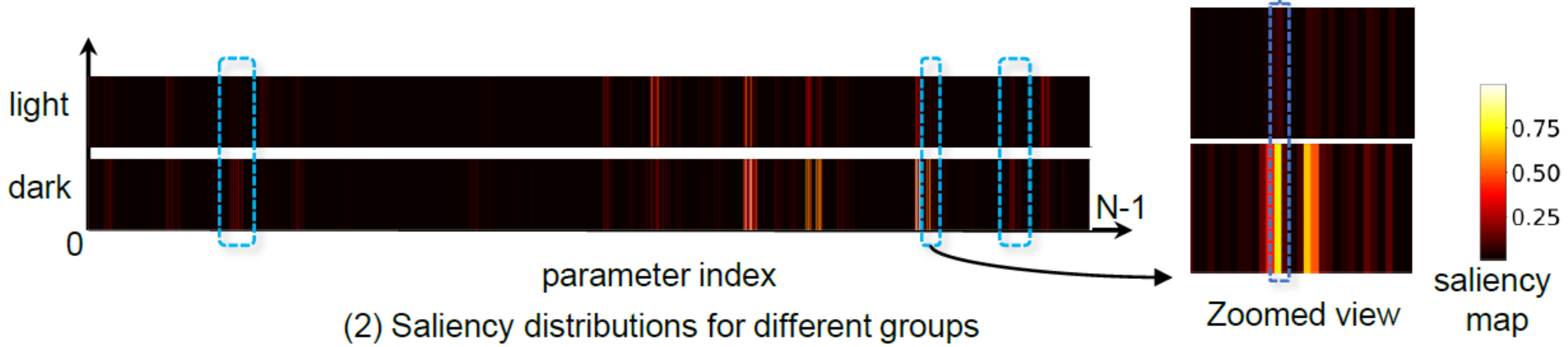
FairPrune: Achieving Fairness via Pruning



$$saliency = \frac{1}{2} h_{ii} {\theta_i}^2$$

(1) the computation of saliency





(2) Saliency distributions for different groups

$$\min \Delta E_{c=0}(\Theta), \quad \max \Delta E_{c=1}(\Theta)$$

$$\min_{\Theta} J = \Delta E_{c=0}(\Theta) - \beta \Delta E_{c=1}(\Theta) = \sum s_i,$$

$$s_i = (\frac{1}{2} h_{ii}^0 \theta_i^2) - \beta \cdot (\frac{1}{2} h_{ii}^1 \theta_i^2) = \frac{1}{2} \theta_m^2 (h_{ii}^0 - \beta \cdot h_{ii}^1).$$

Results of FairPrune

Fairness Metric: **Eopp** and **Eodd**

Equalized opportunity (**Eopp**)

- **Eopp0**: True Negative Rate difference between two groups.
- **Eopp1**: True Positive Rate difference between two groups.

$$EOpp0 = \sum_{k=1}^{K} |TNR_k^1 - TNR_k^0|, \quad EOpp1 = \sum_{k=1}^{K} |TPR_k^1 - TPR_k^0|.$$

$$TPR_k^c = \frac{TP_k^c}{TP_k^c + FN_k^c} \quad TNR_k^c = \frac{TN_k^c}{TN_k^c + FP_k^c} \quad FPR_k^c = \frac{FP_k^c}{TN_k^c + FP_k^c}$$

Equalized Odds (**Eodd**)

$$EOdd = \sum_{k=1}^{K} |TPR_k^1 - TPR_k^0 + FPR_k^1 - FPR_k^0|.$$

Results of FairPrune

Datasets

1. Fitzpatrick-17k

- 16,003 images with 114 skin conditions.
- Six levels of **skin tones**.
  - 11,057 light skins and 4,946 dark skins.
- The vanilla model has higher accuracy on dark skins.

2. ISIC 2019

- 25,331 images with 9 skin conditions.
- Use **gender** as the sensitive attribute.
  - 11,600 female images and 12,358 male images.
- The vanilla model has higher accuracy on female images.

Results of FairPrune on Fitzpatrick-17k

- Improved **fairness**.
- Marginally reduced diagnosis **accuracy**.

Table 1: Results of accuracy and fairness of different methods on Fitzpatrick-17k dataset, using skin tone as the sensitive attribute. The dark skin is the privileged group with higher accuracy by vanilla training. (pr is the pruning ratio).

Method	Skin Tone	Accuracy			Fairness		
		Precision	Recall	F1-score	Eopp0 ($\times 10^{-3}$) ↓	Eopp1 ↓	Eodd ↓
Vanilla	Dark	0.563	0.581	0.546	1.331	0.361	0.182
	Light	0.482	0.495	0.473			
	Avg. ↑	0.523	0.538	0.510			
	Diff. ↓	0.081	0.086	0.073			
AdvConf [29]	Dark	0.506	0.562	0.506	1.106	0.339	0.169
	Light	0.427	0.464	0.426			
	Avg. ↑	0.467	0.513	0.466			
	Diff. ↓	0.079	0.098	0.080			
AdvRev [25]	Dark	0.514	0.545	0.503	1.127	0.334	0.166
	Light	0.489	0.469	0.457			
	Avg. ↑	0.502	0.507	0.480			
	Diff. ↓	0.025	0.076	0.046			
DomainIndep [27]	Dark	0.547	0.567	0.532	1.210	0.344	0.172
	Light	0.455	0.480	0.451			
	Avg. ↑	0.501	0.523	0.492			
	Diff. ↓	0.092	0.087	0.081			
OBD [15] (pr=35%)	Dark	0.557	0.570	0.536	1.244	0.360	0.180
	Light	0.488	0.494	0.475			
	Avg. ↑	0.523	0.532	0.506			
	Diff. ↓	0.069	0.076	0.061			
FairPrune (pr=35%, β=0.33)	Dark	0.567	0.519	0.507	**0.846**	**0.330**	**0.165**
	Light	0.496	0.477	0.459			
	Avg. ↑	0.531	0.498	0.483			
	Diff. ↓	0.071	**0.042**	0.048			

Results of FairPrune on ISIC 2019

- Improved fairness
- Marginally reduced diagnosis accuracy

Table 2: Results of accuracy and fairness of different methods on ISIC 2019 dataset, using gender as the sensitive attribute. The female group is the privileged group with higher accuracy by vanilla training. (pr is the pruning ratio).

		Accuracy			Fairness		
Method	Gender	Precision	Recall	F1-score	Eopp0($\times 10^{-3}$)↓	Eopp1($\times 10^{-3}$)↓	Eodd($\times 10^{-3}$)↓
Vanilla	Female	0.758	0.733	0.744	6.1	49.7	55.8
	Male	0.766	0.684	0.716			
	Avg ↑	0.762	0.709	0.730			
	Diff ↓	0.008	0.049	0.028			
AdvConf [29]	Female	0.691	0.688	0.686	**4.0**	75.1	79.1
	Male	0.681	0.656	0.665			
	Avg ↑	0.686	0.672	0.675			
	Diff ↓	0.010	0.032	0.021			
AdvRev [25]	Female	0.638	0.714	0.670	5.0	59.2	64.2
	Male	0.642	0.666	0.650			
	Avg ↑	0.640	0.690	0.660			
	Diff ↓	0.004	0.048	0.020			
DomainIndep [27]	Female	0.782	0.693	0.729	5.0	74.7	79.7
	Male	0.783	0.653	0.697			
	Avg ↑	0.782	0.673	0.713			
	Diff ↓	**0.001**	0.040	0.032			
OBD [15] (pr=50%)	Female	0.771	0.734	0.749	6.1	55.5	61.6
	Male	0.762	0.678	0.711			
	Avg ↑	0.767	0.706	0.730			
	Diff ↓	0.009	0.056	0.038			
FairPrune (pr=50%, β=0.2)	Female	0.754	0.674	0.707	7.8	**21.0**	**28.8**
	Male	0.762	0.675	0.710			
	Avg ↑	0.758	0.675	0.709			
	Diff ↓	0.008	**0.001**	**0.003**			

Thanks!