Two-step adversarial debiasing with partial learning -- medical image case-studies

Ramon Correa[1], Jiwoong Jason Jeong[1], Bhavik Patel[2,1], Hari Trivedi[3], Judy W. Gichoya[3], Imon Banerjee[2,1]
[1]SCAI, Arizona State University, USA, [2]Department of Radiology, Mayo Clinic, Arizona, USA, [3]Department of Radiology, Emory University, Atlanta, USA

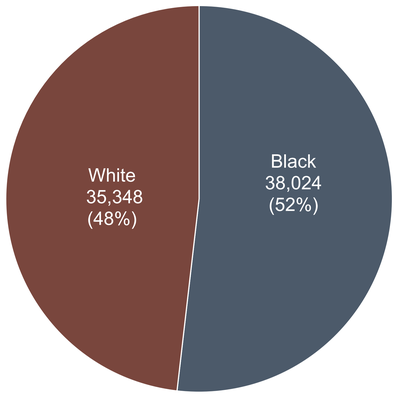
BACKGROUND

- Disparate model performance of AI models for common diagnostic imaging task has raised concerns about how demographics influence model inference.
- We propose an adversarial debiasing technique to decouple race information from task to reduce demographic disparities.

OBJECTIVES

- Characterize biases concerning the patient race for two common AI use cases chest x-ray and mammogram image interpretation
- Implementation of a novel adversarial debasing technique with partial model tuning.
- Comparison of full and partial debiasing model performance.

Chest X-ray Dataset Patient Demographics



Mammography Dataset Patient Demographics

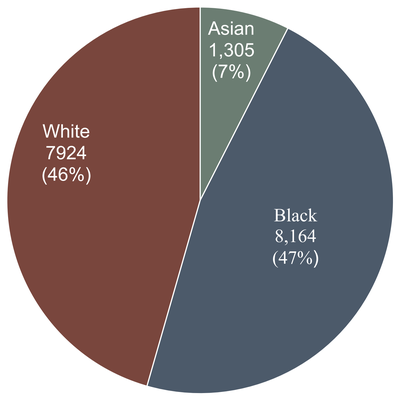


Table 1: Demographic distribution of datasets in study.

METHODS

- Two DenseNet121 models were trained to classify breast density and four common chest x-ray pathologies.
- Adversarial debiasing with finetuning was applied to layers downstream of layer with greatest impact to demographic prediction.

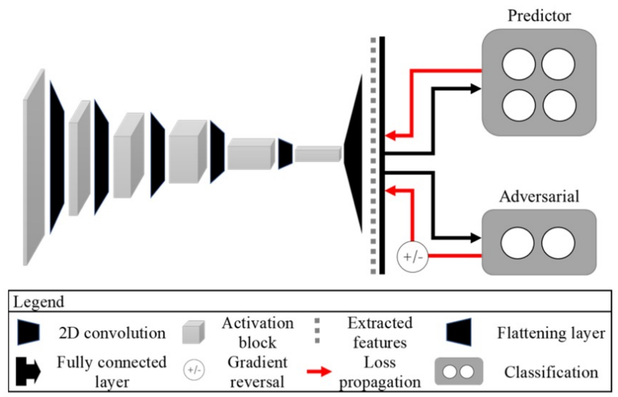


Figure 1: Model diagram demonstrating adversarial debiasing procedure

- The Adversarial debiasing procedure involves a two-step training process:

1. Both predictor and adversary branches attempt to minimize their corresponding losses.
2. The adversarial branch is penalized with a negative gradient to maximize adversarial task loss.

RESULTS

- Chest X-ray diagnosis no significant drop in performance was observed but little reduction in bias.
- For the Breast density prediction task partial debiasing maintained good performance while reducing TPR disparity (0.8 – 1.2 no disparity range).

Diagnosis from Chest X-ray				
Disease	Metric	Model Comparison		
		Baseline	Partial	Full
Atelectasis	AUC	0.865	**0.870**	**0.873**
	Precision	0.889	**0.891**	**0.893**
	Recall	**0.925**	**0.924**	**0.926**
	TPR Black	1.1	1.12	**1.06**
Edema	AUC	**0.898**	**0.883**	**0.884**
	Precision	**0.503**	0.457	0.405
	Recall	**0.511**	**0.525**	0.484
	TPR Black	1.33	**1.11**	1.15
Pneumothorax	AUC	0.829	0.837	**0.857**
	Precision	0.558	0.586	**0.591**
	Recall	0.460	0.505	**0.512**
	TPR Black	0.74	**0.91**	**0.89**
No Finding	AUC	0.866	**0.889**	0.846
	Precision	**0.346**	0.336	**0.349**
	Recall	0.188	**0.313**	**0.313**
	TPR Black	**0.85**	0.73	0.80
Mammogram Breast density				
Breast Density Score	Metric	Model Comparison		
		Baseline	Partial	Full
1	AUC	**0.965**	**0.96**	0.942
	Precision	0.637	**0.709**	0.637
	Recall	**0.858**	0.677	0.682
	TPR Asian	3.375	1.895	**1.674**
	TPR Black	1.376	1.307	**1.111**
2	AUC	**0.899**	0.896	0.879
	Precision	**0.781**	0.769	0.763
	Recall	**0.765**	0.758	0.736
	TPR Asian	0.635	0.794	**0.809**
	TPR Black	0.689	0.721	**0.855**
3	AUC	**0.923**	0.895	0.917
	Precision	**0.879**	0.825	0.831
	Recall	0.739	0.727	**0.821**
	TPR Asian	**1.976**	2.84	2.148
	TPR Black	**1.046**	1.664	0.839
3	AUC	**0.979**	**0.972**	0.957
	Precision	0.413	0.324	**0.482**
	Recall	**0.867**	**0.883**	0.625
	TPR Asian	n/a	n/a	n/a
	TPR Black	n/a	n/a	n/a

Table 2: Performance metrics comparing training techniques evaluated on our chest X-ray and mammography datasets.

CONCLUSIONS

Effectiveness of debiasing techniques is variable depending on correlation between task and demographic attributes.