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





Ramon Correa¹, Jiwoong Jason Jeong¹, Bhavik Patel^{2,1}, Hari Trivedi³, Judy W. Gichoya³, Imon Banerjee^{2,1} ¹SCAI, Arizona State University, USA, ²Department of Radiology, Mayo Clinic, Arizona, USA, ³Department 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. \geq 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
- \geq 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



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 twostep training process:
- 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 debiasing maintained good performance while reducing TPR disparity (0.8 - 1.2 no disparity range).

partial

Metric Model Comparison Baseline Partial Full AUC 0.865 0.870 0.873 Atelectasis Precision 0.889 0.891 0.893 Recall 0.925 0.924 0.926 TPR Black 1.1 1.12 1.06 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 AUC 0.829 0.837 0.857 Precision 0.558 0.586 0.591 Precision 0.558 0.586 0.591 Recall 0.460 0.505 0.512 TPR Black 0.74 0.91 0.89 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 Recall 0.188 0.313 0.313 TPR Black	Diagnosis from Chest X-ray					
Disease Metric Baseline Partial Full AUC 0.865 0.870 0.873 Atelectasis Precision 0.889 0.891 0.893 Recall 0.925 0.924 0.926 TPR Black 1.1 1.12 1.06 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 AUC 0.829 0.837 0.857 Precision 0.558 0.586 0.591 Precision 0.558 0.586 0.591 Recall 0.460 0.505 0.512 TPR Black 0.74 0.91 0.89 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 Recall 0.188 0.313 0.313 TPR Black 0.855 0.73			Model Comparison			
AtelectasisAUC0.8650.8700.873AtelectasisPrecision0.8890.8910.893Recall0.9250.9240.926TPR Black1.11.121.06AUC0.8980.8830.884Precision0.5030.4570.405Recall0.5110.5250.484TPR Black1.331.111.15Precision0.5580.5860.591Precision0.5580.5860.591Precision0.5580.5860.591Recall0.4600.5050.512TPR Black0.740.910.89No FindingAUC0.8660.8890.846Precision0.3460.3360.349Recall0.1880.3130.313TPR Black0.8550.730.800MetricBaselinePartialFullBreast DensityMetricBaselinePartialAUC0.9650.960.942Precision0.6370.7090.6371Recall0.8580.6770.6821Recall0.8580.6770.6821Recall0.8580.6770.6821Recall0.8980.6770.6821Recall0.8980.6770.6821Recall0.8980.6770.6821Recall0.8990.8960.8761Recall0	Disease	IVIETIC	Baseline	Partial	Full	
Atelectasis Precision 0.889 0.891 0.893 Recall 0.925 0.924 0.926 TPR Black 1.1 1.12 1.06 TPR Black 1.1 1.12 1.06 Bedema 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 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 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 Breast Density Score Metric Baseline Partial Full AUC 0.	Atelectasis	AUC	0.865	0.870	0.873	
Recall0.9250.9240.926TPR Black1.11.121.06TPR Black1.11.121.06AUC0.8980.8830.884Precision0.5030.4570.405Recall0.5110.5250.484TPR Black1.331.111.15PneumothoraxAUC0.8290.8370.857Precision0.5580.5860.591Recall0.4600.5050.512TPR Black0.740.910.89No FindingAUC0.8660.8890.846Precision0.3460.3360.349Recall0.1880.3130.313TPR Black0.850.730.80Precision0.8650.960.942Recall0.1880.3130.313TPR Black0.6370.6370.637AUC0.9650.960.942Precision0.6370.7090.637AUC0.8680.6770.682TPR Asian3.3751.8951.674TPR Black1.3761.3071.111AUC0.8990.8900.879		Precision	0.889	0.891	0.893	
TPR Black 1.1 1.12 1.06 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 AUC 0.829 0.837 0.857 Precision 0.558 0.586 0.591 Precision 0.558 0.586 0.591 Recall 0.460 0.505 0.512 Precision 0.346 0.336 0.349 No Finding AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 No Finding Recall 0.188 0.313 0.313 TPR Black 0.855 0.73 0.80 Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.637 0.682 TPR Asian 3.37		Recall	0.925	0.924	0.926	
AUC0.8980.8830.884Precision0.5030.4570.405Recall0.5110.5250.484TPR Black1.331.111.15AUC0.8290.8370.857Precision0.5580.5860.591Recall0.4600.5050.512Recall0.4600.5050.512TPR Black0.740.910.89AUC0.8660.8890.846Precision0.3460.3360.349AUC0.8660.3360.349Recall0.1880.3130.313TPR Black0.8550.730.80Precision0.8560.730.80Precision0.6350.730.80Precision0.6370.700.637Precision0.6370.9650.96AUC0.9650.960.942Precision0.6370.7090.637AUC0.8580.6770.682TPR Asian3.3751.8951.674TPR Black1.3761.3071.111AUC0.8990.8390.879		TPR Black	1.1	1.12	1.06	
Edema Precision 0.503 0.457 0.405 Recall 0.511 0.525 0.484 TPR Black 1.33 1.11 1.15 AUC 0.829 0.837 0.857 Precision 0.558 0.586 0.591 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 Recall 0.188 0.313 0.313 TPR Black 0.855 0.73 0.80 Score Metric Baseline Partial Full 1 AUC 0.965 0.96 0.942 Precision 0.637 0.637 0.637 Recall		AUC	0.898	0.883	0.884	
Edema Recall 0.511 0.525 0.484 TPR Black 1.33 1.11 1.15 AUC 0.829 0.837 0.857 Precision 0.558 0.586 0.591 Recall 0.460 0.505 0.512 Precision 0.366 0.889 0.846 Precision 0.346 0.336 0.349 No Finding AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 No Finding Recall 0.188 0.313 0.313 Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Breast Density Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 1 Recall 0.858 0.677 0.6	Edomo	Precision	0.503	0.457	0.405	
TPR Black 1.33 1.11 1.15 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 No Finding Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Score Metric Baseline Partial Full Metric Baseline Partial Full 1 AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111	Edema	Recall	0.511	0.525	0.484	
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 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 No Finding Recall 0.188 0.313 0.313 Recall 0.188 0.313 0.313 0.313 TPR Black 0.85 0.73 0.80 Breast Density Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 AUC 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		TPR Black	1.33	1.11	1.15	
Pneumothorax Precision 0.558 0.586 0.591 Recall 0.460 0.505 0.512 TPR Black 0.74 0.91 0.89 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 Recall 0.188 0.313 0.313 Precision 0.85 0.73 0.80 Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Metric Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.637 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 AUC 0.899 0.896 0.879		AUC	0.829	0.837	0.857	
Pneumotnorax Recall 0.460 0.505 0.512 TPR Black 0.74 0.91 0.89 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 Recall 0.188 0.313 0.313 Precision 0.855 0.73 0.80 Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Metric Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		Precision	0.558	0.586	0.591	
TPR Black 0.74 0.91 0.89 AUC 0.866 0.889 0.846 Precision 0.346 0.336 0.349 Recall 0.188 0.313 0.313 Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Mammogram Breast density Breast Density Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.637 1 AUC 0.965 0.96 0.942 Precision 0.637 0.637 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879	Pneumothorax	Recall	0.460	0.505	0.512	
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 Metric Model Comparison Breast Density Score Metric Baseline Partial Full 1 AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		TPR Black	0.74	0.91	0.89	
No Finding 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 Metric Model Comparison Breast Density Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		AUC	0.866	0.889	0.846	
No Finding Recall 0.188 0.313 0.313 TPR Black 0.85 0.73 0.80 Mammogram Breast density Metric Model Comparison Breast Density Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		Precision	0.346	0.336	0.349	
TPR Black0.850.730.80Mammogram Breast densityMatricModel ComparisonBreast Density ScoreMetricBaselinePartialFullAUC0.9650.960.942Precision0.6370.7090.6371Recall0.8580.6770.682TPR Asian3.3751.8951.674TPR Black1.3761.3071.111AUC0.8990.8960.879	No Finding	Recall	0.188	0.313	0.313	
Mammogram Breast densityModel ComparisonBreast Density ScoreMetricBaselinePartialFullAUC0.9650.960.942Precision0.6370.7090.6371Recall0.8580.6770.682TPR Asian3.3751.8951.674TPR Black1.3761.3071.111AUC0.8990.8960.879		TPR Black	0.85	0.73	0.80	
Breast Density Score Metric Model Comparison AUC Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879	Mammogram Breast density					
Breast Density Score Metric Baseline Partial Full AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879			Mode	el Compa	rison	
AUC 0.965 0.96 0.942 Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879	Score	Metric	Baseline	Partial	Full	
Precision 0.637 0.709 0.637 1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879		AUC	0.965	0.96	0.942	
1 Recall 0.858 0.677 0.682 TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879 Dracision 0.794 0.760 0.762		Precision	0.637	0.709	0.637	
TPR Asian 3.375 1.895 1.674 TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879 Dracician 0.794 0.760 0.762	1	Recall	0.858	0.677	0.682	
TPR Black 1.376 1.307 1.111 AUC 0.899 0.896 0.879 Dracicion 0.794 0.760 0.762		TPR Asian	3.375	1.895	1.674	
AUC 0.899 0.896 0.879 Dracicion 0.794 0.760 0.762		TPR Black	1.376	1.307	1.111	
Dragician 0701 0700 0700	2	AUC	0.899	0.896	0.879	
TIECISIUII U./03 U./03		Precision	0.781	0.769	0.763	
2 Recall 0.765 0.758 0.736		Recall	0.765	0.758	0.736	
TPR Asian 0.635 0.794 0.809		TPR Asian	0.635	0.794	0.809	
TPR Black 0.689 0.721 0.855		TPR Black	0.689	0.721	0.855	
AUC 0.923 0.895 0.917	3	AUC	0.923	0.895	0.917	
Precision 0.879 0.825 0.831		Precision	0.879	0.825	0.831	
3 Recall 0.739 0.727 0.821		Recall	0.739	0.727	0.821	
TPR Asian 1.976 2.84 2.148		TPR Asian	1.976	2.84	2.148	
TPR Black 1.046 1.664 0.839		TPR Black	1.046	1.664	0.839	
AUC 0.979 0.972 0.957	3	AUC	0.979	0.972	0.957	
Precision 0.413 0.324 0.482		Precision	0.413	0.324	0.482	
3 Recall 0.867 0.883 0.625		Recall	0.867	0.883	0.625	
TPR Asian n/a n/a n/a		TPR Asian	n/a	n/a	n/a	
TPR Rlack 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.



Effectiveness of debiasing techniques correlation depending between on demographic attributes.

is variable task and