CheXclusion: Fairness gaps in deep chest X-ray classifiers

ness
gaps

Data set: three large, public chest X-ray datasets: MIMIC-CXR , CheXpert, and Chest-Xray8 .

Disease Labels

The protected attributes patients sex (Male and Female), age (0-20, 20-40, 40-60, 60-80, and 80-), race (White, Black, Other, Asian, Hispanic, and Native) and insurance type (Medicare, Medicaid, and Other).

	MIMIC-CXR	CheXpert	Chest-Xray8
Abbr.	**CXR[22]**	**CXP[21]**	**NIH[45]**
# Images	371,858	223,648	112,120
# Patients	65,079	64,740	30,805
View	Front/Lat	Front/Lat	Front
Female	47.83%	40.64%	43.51%
Male	52.17%	59.36%	56.49%
0-20	2.20%	0.87%	6.09%
20-40	19.51%	13.18%	25.96%
40-60	37.20%	31.00%	43.83%
60-80	34.12%	38.94%	23.11%
80-	6.96%	16.01%	1.01%
White	65.01%	N/A	N/A
Black	17.86%	N/A	N/A
Other	3.68%	N/A	N/A
Asian	3.12%	N/A	N/A
Hispanic	6.16%	N/A	N/A
Native	0.28%	N/A	N/A
Unknown	3.89%	N/A	N/A
Medicare	46.07%	N/A	N/A
Medicaid	8.98%	N/A	N/A
Other	44.95%	N/A	N/A

per dataset, some disease may commonly appear with larger or smaller gap between least / most favorable subgroups.

TPR disparity

unfavorable
favorable subgroups

Methods

CNN-based models
train separate models for CXR , CXP , and NIH
sperformance :sex and age.
CXR: patient race and insurance type.

Models

tune the learning rate, tune the degree of random rotation data augmentation. Finally we tune all the hyperparameters of the best model and train four extra models with the same hyperparameters

		CXP	CXR	NIH	NIH		
Label	**Abbr.**	**AUC**	**AUC**	**Label**	**Abbr.**	**AUC**	**[34] AUC**
Airspace Opacity	AO	0.747±0.001	0.782±0.001	Atelectasis	A	0.814±0.004	0.862
Atelectasis	A	0.717±0.002	0.837±0.001	Cardiomegaly	Cd	0.915±0.002	0.831
Cardiomegaly	Cd	0.855±0.003	0.828±0.001	Consolidation	Co	0.801±0.005	0.893
Consolidation	Co	0.734±0.004	0.844±0.001	Edema	Ed	0.915±0.003	0.924
Edema	Ed	0.849±0.001	0.905±0.001	Effusion	Ef	0.875±0.002	0.901
Enlarged Card	EC	0.668±0.005	0.758±0.004	Emphysema	Em	0.897±0.002	0.704
Fracture	Fr	0.790±0.006	0.717±0.007	Fibrosis	Fb	0.788±0.007	0.806
Lung Lesion	LL	0.780±0.005	0.773±0.005	Hernia	H	0.978±0.004	0.851
No Finding	NF	0.885±0.001	0.869±0.001	Infiltration	In	0.717±0.004	0.721
Pleural Effusion	PE	0.885±0.001	0.933±0.001	Mass	M	0.829±0.006	0.909
Pleural Other	PO	0.795±0.004	0.846±0.003	Nodule	N	0.779±0.006	0.894
Pneumonia	Pa	0.777±0.003	0.748±0.005	Pleural_Thickening	PT	0.813±0.006	0.798
Pneumothorax	Px	0.893±0.002	0.903±0.002	Pneumonia	Pa	0.759±0.012	0.851
Support Devices	SD	0.898±0.001	0.927±0.001	Pneumothorax	Px	0.879±0.005	0.944
Average		**0.805±0.001**	**0.834±0.001**	**Average**		**0.840±0.001**	**0.849**

2: The AUC for chest X-ray classifiers trained on CXP, CXR, and NIH, averaged over 5 runs± 95%CI, where all runs have hyperparameters but different random seed. ('Airspace Opacity' in [22] and 'Lung Opacity' in [21] denote a same label.)

Classier Disparity Evaluation

TPR disparities for binary a ributes
TPR disparities for non-binary a ributes
Underdiagnosis rate

In this paper, we illustrate the TPR disparity of SOTA chest X-ray pathology classi ers trained on three di er- ent datasets, (MIMIC-CXR, ChestX-ray8, and CheXpert) across 14 disease labels. We quantify the TPR disparity across experimental studies along sex, age, race and insurance type. We also spot some subsection of the population are chronically underdiagnosed.

Results indicate that high-capacity models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modication