Do deep learning chest X-ray classifiers discriminate?

Laleh Seyyed-Kalantari[1], Guanxiong Liu[1], Matthew McDermott[2], Irene Chen[2], Marzyeh Ghassemi[1]

[1]University of Toronto, Vector Institute [2] Massachusetts Institute of Technology

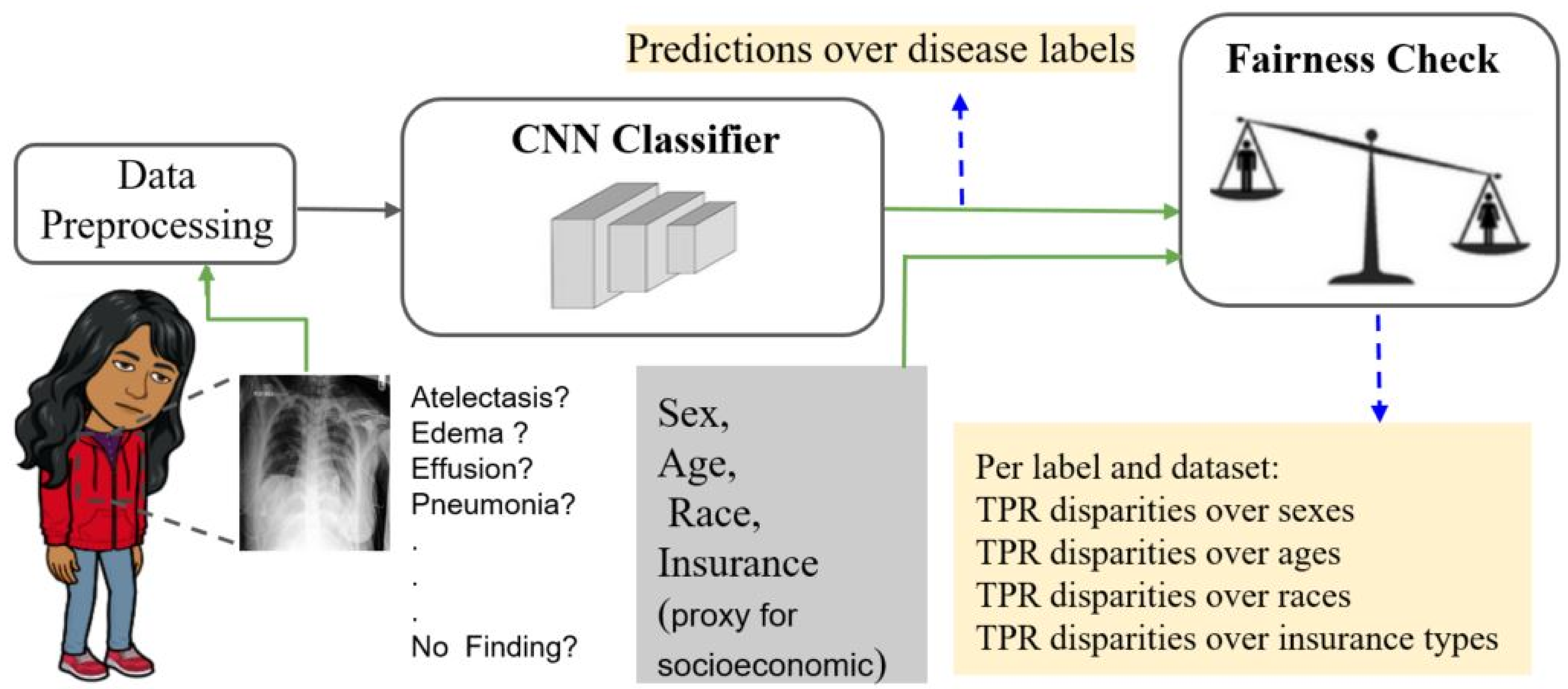
Contributions

- We show the state-of-the-art (SOTA) deep learning classifiers trained to yield diagnostic labels from X-ray images display systematic bias over patient's **sex**, **age**, **race** and **insurance type** (as a proxy of socialeconomic status).
- We quantify biases by evaluating the *TPR disparity* – differences in true positive rates (TPR) – among different protected attributes.
- As clinical models move from papers to products, we encourage clinical decision makers to carefully audit for algorithmic disparities prior to deployment.

The paper and link to the code are available in: https://arxiv.org/abs/2003.00827

Corresponding email: laleh@cs.toronto.edu

Methods



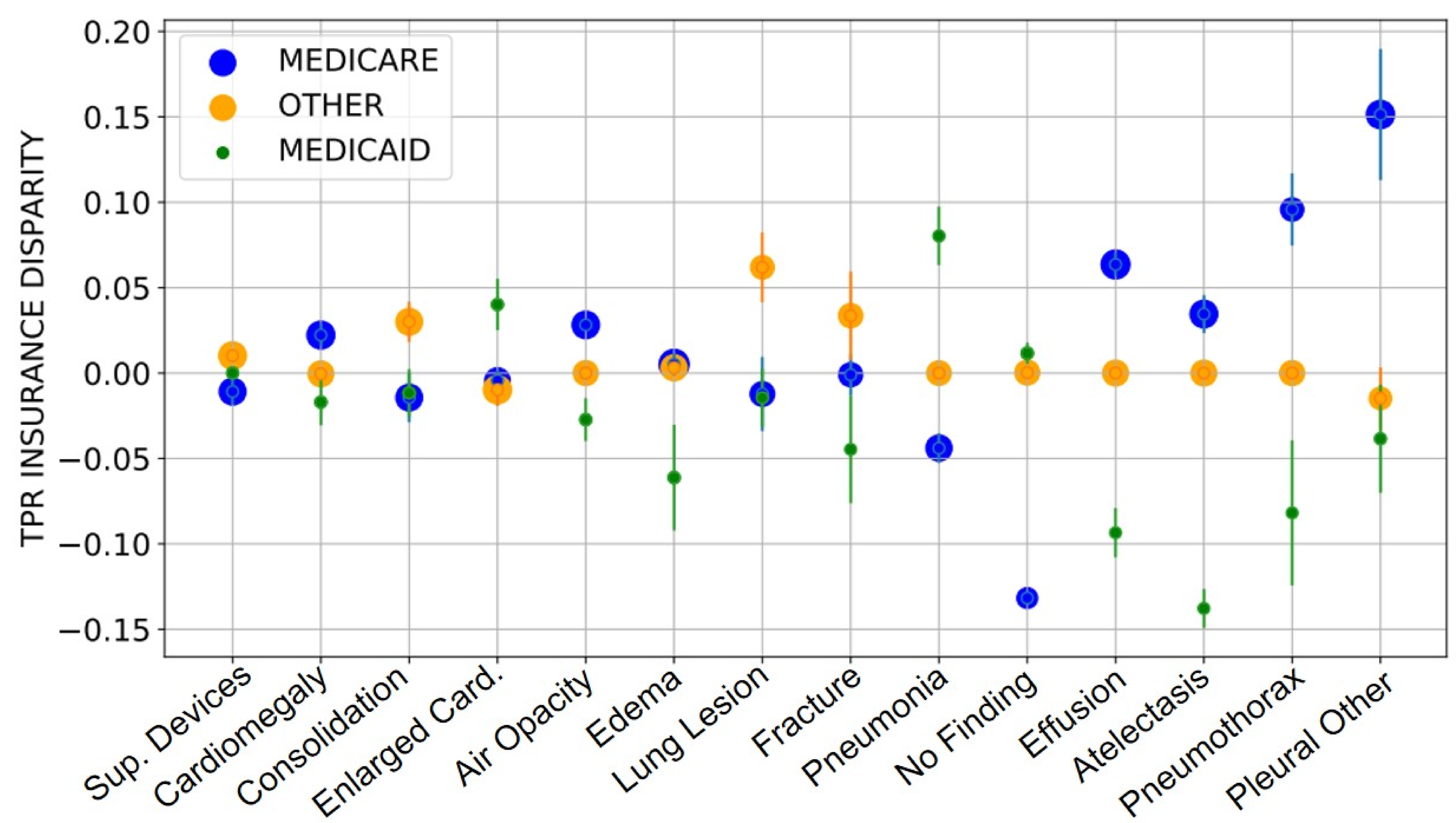
Classification performance on X-ray diagnoses

Dataset: MIMIC-CXR (CXR), CheXpert (CXP), Chest-Xray8 (NIH) and all of those multi-site aggregation (ALL).
The AUC for chest X-ray classifiers trained on CXP, CXR, NIH, and ALL averaged over 5 runs (different seeds) ±95%CI.

Label (Abbr.)	**CXR**	**CXP**	**NIH**	**ALL**
Airspace Opacity (AO)	0.782 ± 0.002	0.747 ± 0.001	—	—
Atelectasis (A)	0.837 ± 0.001	0.717 ± 0.002	0.814 ± 0.004	0.808 ± 0.001
Cardiomegaly (Cd)	0.828 ± 0.002	0.855 ± 0.003	0.915 ± 0.002	0.856 ± 0.001
Consolidation (Co)	0.844 ± 0.001	0.734 ± 0.004	0.801 ± 0.005	0.805 ± 0.001
Edema (Ed)	0.904 ± 0.002	0.849 ± 0.001	0.915 ± 0.003	0.898 ± 0.001
Effusion (Ef)	0.933 ± 0.001	0.885 ± 0.001	0.875 ± 0.002	0.922 ± 0.001
Emphysema (Em)	—	—	0.897 ± 0.002	—
Enlarged Card (EC)	0.757 ± 0.003	0.668 ± 0.005	—	—
Fibrosis	—	—	0.788 ± 0.007	—
Fracture (Fr)	0.718 ± 0.007	0.790 ± 0.006	—	—
Hernia (H)	—	—	0.978 ± 0.004	—
Infiltration (In)	—	—	0.717 ± 0.004	—
Lung Lesion (LL)	0.772 ± 0.006	0.780 ± 0.005	—	—
Mas (M)	—	—	0.829 ± 0.006	—
Nodule (N)	—	—	0.779 ± 0.006	—
No Finding (NF)	0.868 ± 0.001	0.885 ± 0.001	—	0.890 ± 0.000
Pleural Thickening (PT)	—	—	0.813 ± 0.006	—
Pleural Other (PO)	0.848 ± 0.003	0.795 ± 0.004	—	—
Pneumonia (Pa)	0.748 ± 0.005	0.777 ± 0.003	0.759 ± 0.012	0.784 ± 0.001
Pneumothorax (Px)	0.903 ± 0.002	0.893 ± 0.002	0.879 ± 0.005	0.904 ± 0.002
Support Devices (SD)	0.927 ± 0.001	0.898 ± 0.001	—	—
Average	**0.834 ± 0.001**	**0.805±0.001**	**0.840 ± 0.001**	**0.859 ± 0.001**

Disparities exist

We demonstrate that **TPR disparities exist in the SOTA classifiers in all datasets, for all clinical tasks, and all protected attributes, sex, age, race and insurance type**.
As an illustrative example we show the insurance type sorted TPR disparities distribution in MIMIC-CXR dataset. The scatter plot's circle area is proportional to the membership. In a fair setting disparities are zero. Negative and positive disparities denote bias against and in favor of a subgroup.



Disparities overview over attributes and datasets, in the following table shows the average cross-label gap between the least and most favorable subgroup's TPR disparities. The most frequent "Unfavorable" and "Favorable" subgroups are the ones that experience TPRs disparities below or above the zero gap line frequently.

Attribute	Dataset	Average Cross-Label Gap			Unfavorable	Favorable
		Gap	Lowest	Greatest		
Sex	ALL	0.045	Ef:0.001	Pa:0.105	Female (4/7)	Male (4/7)
	CXP	0.062	Ed:0.000	Co:0.139	Female (7/13)	Male (7/13)
	CXR	0.072	Ed:0.011	EC.:0.151	Female (10/13)	Male (10/13)
	NIH	0.190	M:0.001	Cd:0.393	Female (8/14)	Male (8/14)
Age	ALL	0.215	Ef:0.115	NF:0.444	0-20 (5/7)	40-60,60-80(5/7)
	CXR	0.245	SD:0.091	Cd:0.440	0-20, 20-40 (7/13)	60-80 (10/13)
	CXP	0.270	SD:0.084	NF:0.604	0-20, 20-40, 80- (7/13)	40-60 (8/13)
	NIH	0.413	In:0.188	Em:1.00	60-80 (7/14)	20-40 (9/14)
Race	CXR	0.226	NF:0.119	Pa:0.440	Hispanic (9/13)	White (9/13)
Insurance	CXR	0.100	SD:0.021	PO:0.190	Medicaid (10/13)	Other (10/13)

The multi-source dataset corresponds to the smallest disparities, suggesting one way to reduce bias.

Disparity in proportion to membership

The Pearson correlation coefficient between the TPR disparities and patients proportion per label across all subgroups/datasets shows that **TPR disparities are not often significantly correlated with a subgroup's proportional disease burden.**

Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC, funding number PDF-516984), Microsoft Research, CIFAR, NSERC Discovery Grant, and high performance computing platforms of Vector Institute. We also thank Dr. Alistair Johnson, Dr. Errol Colak and Grey Kuling for productive discussions.