The Role of Subgroup Separability in Group-Fair Medical Image Classification

Charles Jones[1†], Mélanie Roschewitz[1], Ben Glocker[1]

Fairness methods and metrics assume that individuals belong to clearly defined subgroups, but what if this information is only partially encoded in medical images?

Takeaway 1: The amount of sensitive information encoded in medical images (*subgroup separability*) varies substantially across modalities and sensitive attributes.

Table 1. Separability of protected subgroups in real-world datasets, measured by test-set AUC of classifiers trained to predict the groups. Mean and standard deviation are reported over ten random seeds, with results sorted by ascending mean AUC.

Dataset-Attribute	Modality	Subgroups		AUC	
		Group 0	Group 1	μ	σ
PAPILA-Sex	Fundus Image	Male	Female	0.642	0.057
HAM10000-Sex	Skin Dermatology	Male	Female	0.723	0.015
HAM10000-Age	Skin Dermatology	< 60	≥ 60	0.803	0.020
PAPILA-Age	Fundus Image	< 60	≥ 60	0.812	0.046
Fitzpatrick17k-Skin	Skin Dermatology	I-III	IV-VI	0.891	0.010
CheXpert-Age	Chest X-ray	< 60	≥ 60	0.920	0.003
MIMIC-Age	Chest X-ray	< 60	≥ 60	0.930	0.002
CheXpert-Race	Chest X-ray	White	Non-White	0.936	0.005
MIMIC-Race	Chest X-ray	White	Non-White	0.951	0.004
CheXpert-Sex	Chest X-ray	Male	Female	0.980	0.020
MIMIC-Sex	Chest X-ray	Male	Female	0.986	0.008

Takeaway 2: Disease classifiers implicitly learn to rely on sensitive information when trained on data with underdiagnosis bias.

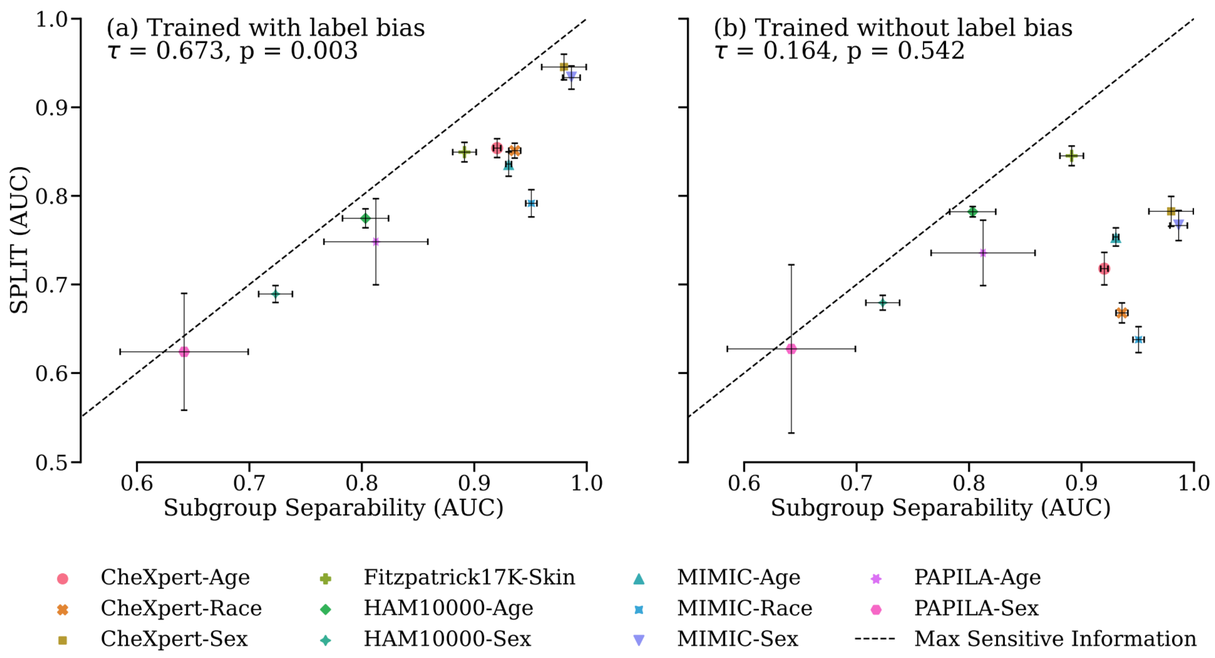

Fig. 2. AUC of the SPLIT test for sensitive information encoded in learned representations, plotted against subgroup separability. Along the maximum sensitive information line, models trained for predicting the disease encode as much sensitive information in their representations as the images do themselves.

Takeaway 3: Performance degradation under dataset bias is a function of subgroup separability.

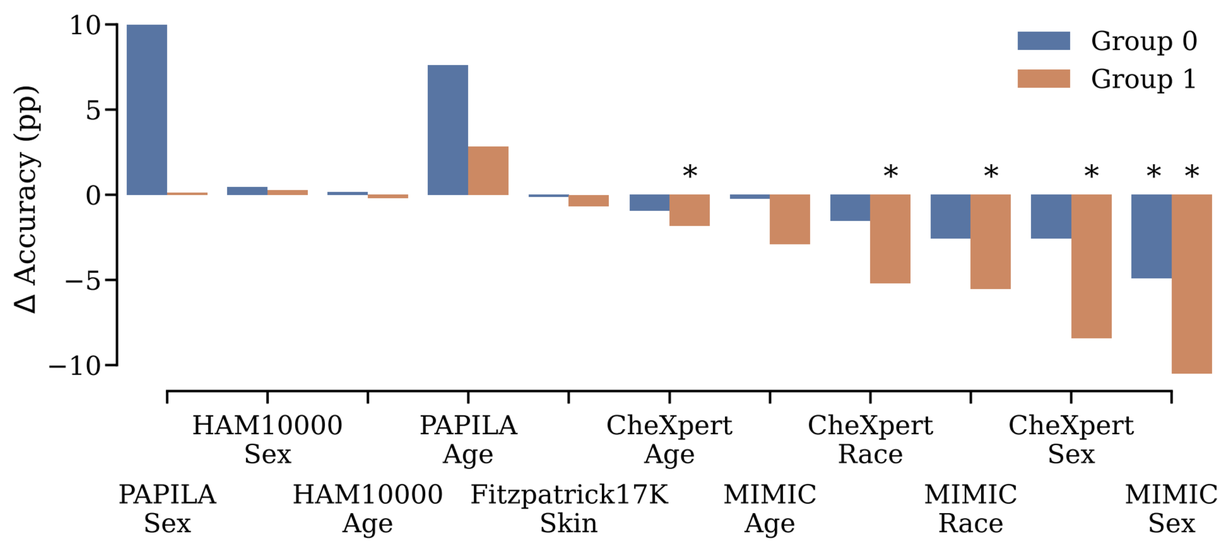

Fig. 1. Percentage-point degradation in accuracy for disease classifiers trained on biased data, compared to training on clean data. Lower values indicate worse performance for the biased model when tested on a clean dataset. Results are reported over ten random seeds, and bars marked with * represent statistically significant results. Dataset-attribute combinations are sorted by ascending subgroup separability.

Methods

We train classifiers on the 2D subset of the MEDFAIR benchmark [1]. Subgroup separability is measured by mean (over 10 random seeds) test-set AUC of classifiers trained to predict the sensitive attribute. We use the post-hoc SPLIT test [2] to measure sensitive information encoded in learned representations of disease prediction models (binary task, no-disease vs. disease). Performance degradation is measured by comparing mean test-time performance of disease prediction models trained on clean data against models trained with 25% of positive individuals in Group 1 (see Table 1) mislabelled. Test data is always uncorrupted.

References

[1] Zong, Y. et al. MEDFAIR: Benchmarking Fairness for Medical Imaging. International Conference on Learning Representations (Feb 2023).

[2] Gichoya, J.W. et al. AI recognition of patient race in medical imaging: A modelling study. Lancet Digital Health (Jun 2022).

MICCAI 2023 Vancouver CANADA

Imperial College London

[1]Department of Computing, Imperial College London, UK

[†]charles.jones17@imperial.ac.uk