MEDFAIR: Benchmarking Fairness for Medical Imaging

ICLR
THE UNIVERSITY of EDINBURGH
@yongshuozong
yongshuo.zong@ed.ac.uk

Yongshuo Zong[1], Yongxin Yang[1] and Timothy Hospedales[1,2]

[1] University of Edinburgh [2] Samsung AI Centre, Cambridge

Paper & Code
Samsung Research
Biomedical AI UKRI CENTRE FOR DOCTORAL TRAINING

Motivation

- **Harmful:** Machine learning models are found to be biased against specific subgroups.
- **Unclear**: No unified fairness notions for medical imaging.
- **Inconsistent**: Previous studies use different experimental settings.

A fairness benchmark for medical imaging is needed!

Fairness Definitions in Medicine

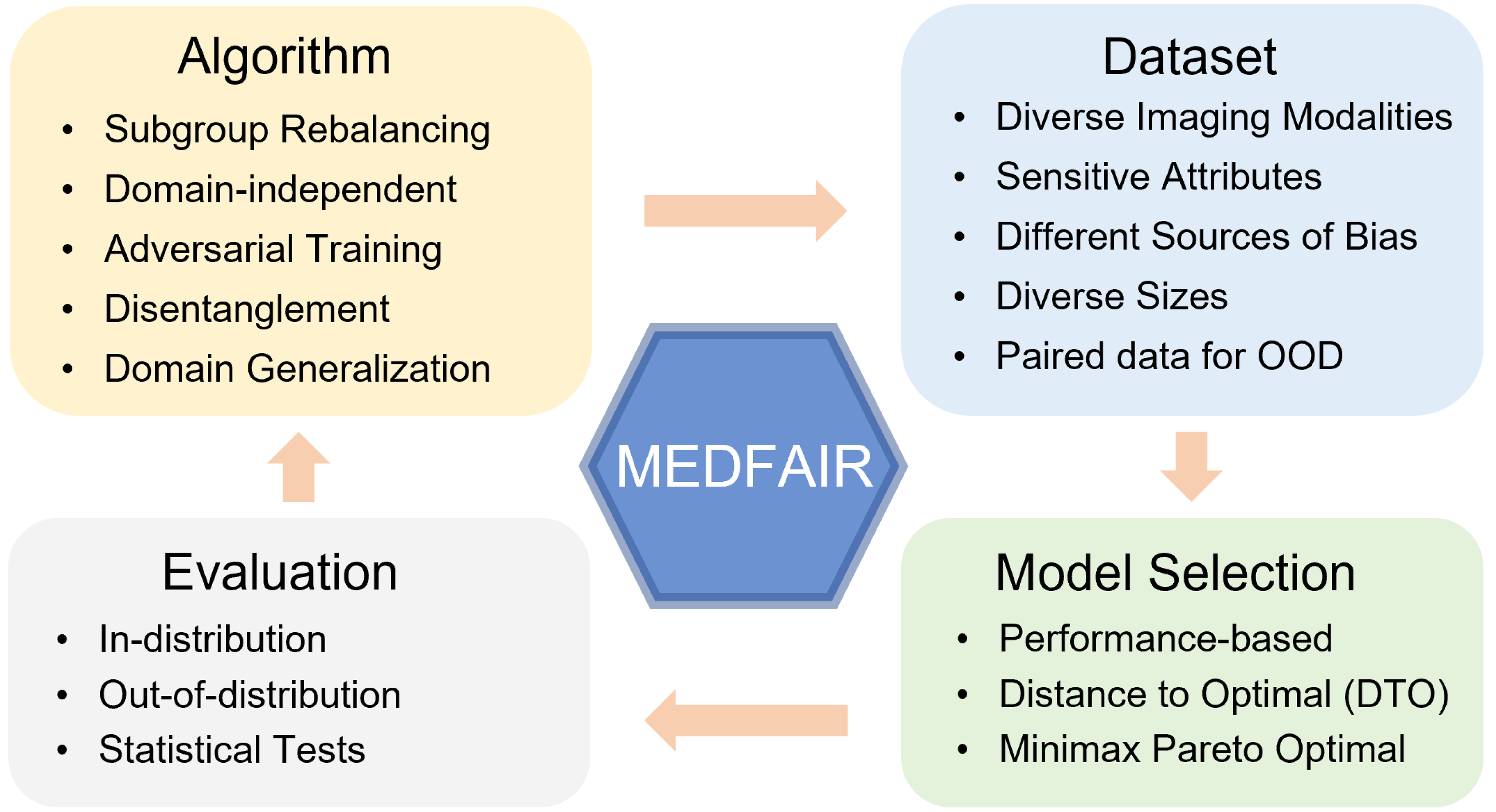
- **Group Fairness**: parity of predictive performance across subgroups [1].
  - Suitable for resource allocation (zero-sum game)
  - Metric: Gap of AUC (separation, $\hat{Y} \perp S \mid Y$)
- **Max-Min Fairness**: the worst-off group should be improved [2].
  - Suitable for diagnosis (non-zero-sum game)
  - Metric: Worst-case AUC

***One should focus on different fairness definitions depending on specific clinical application.**

Introducing MEDFAIR

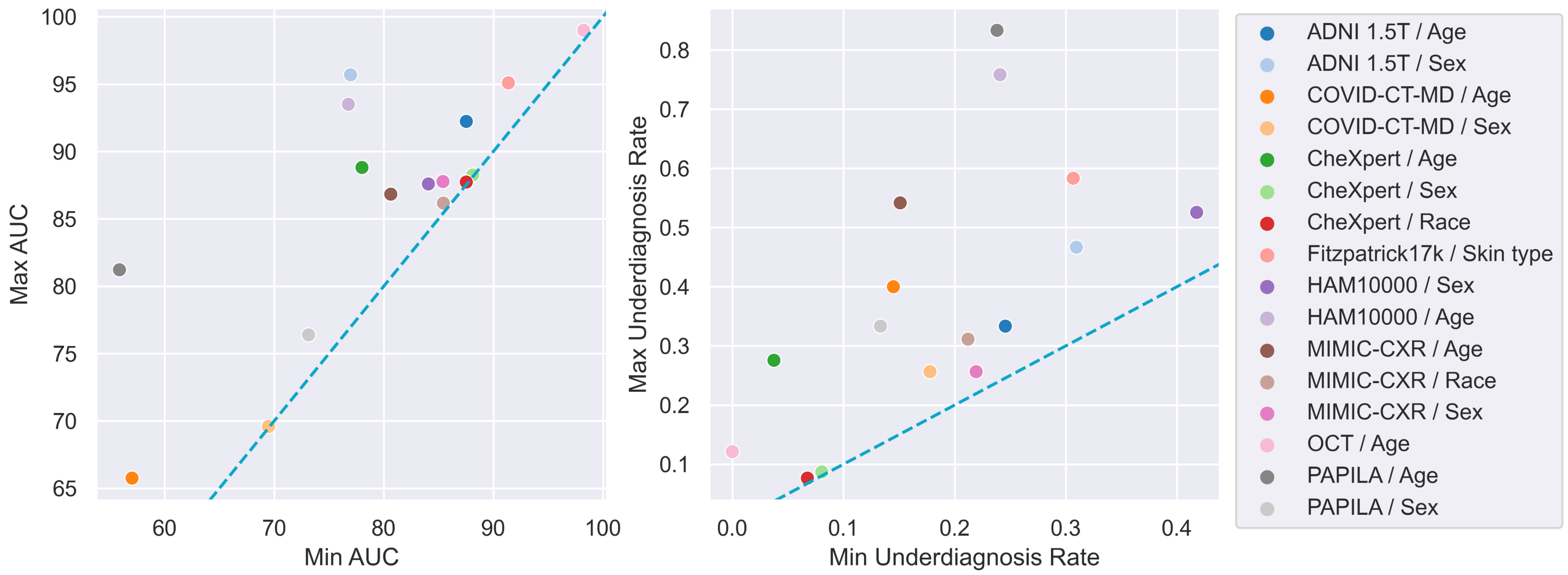
What MEDFAIR has:

- *11* state-of-the-art bias mitigation algorithms
- *10* datasets from different imaging modalities
- *3* popular model selection strategies
- Extensive evaluation with rigorous statistical tests
- Over *7,000* models trained – *~0.7x* A100-80GB GPU years
- Easy to extend new algorithms and datasets



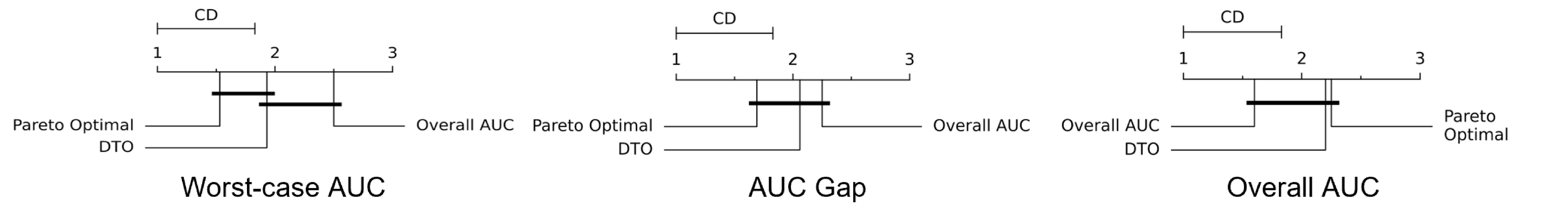
Results

1. Bias widely exists in ML models



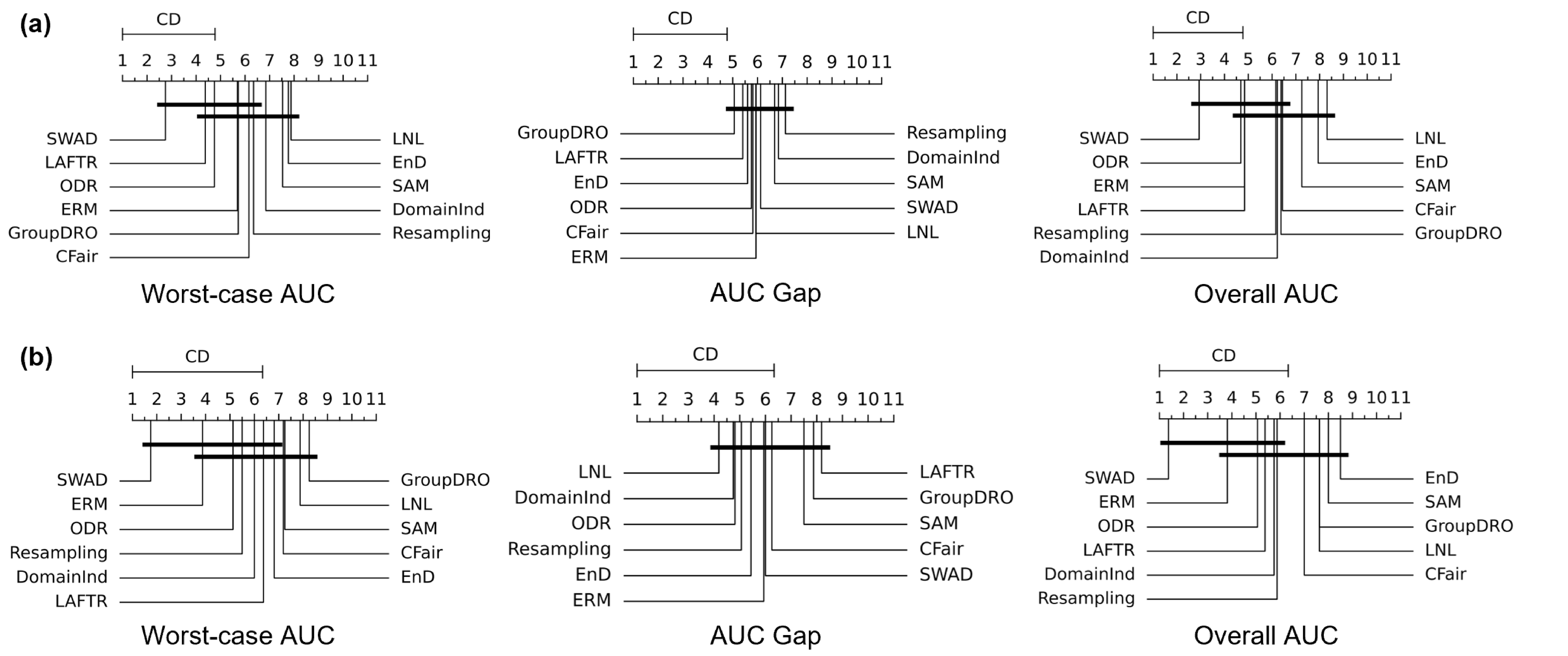
Most points are off the blue equality line for both AUC and underdiagnosis rate, when training with ERM.

2. Model selection matters



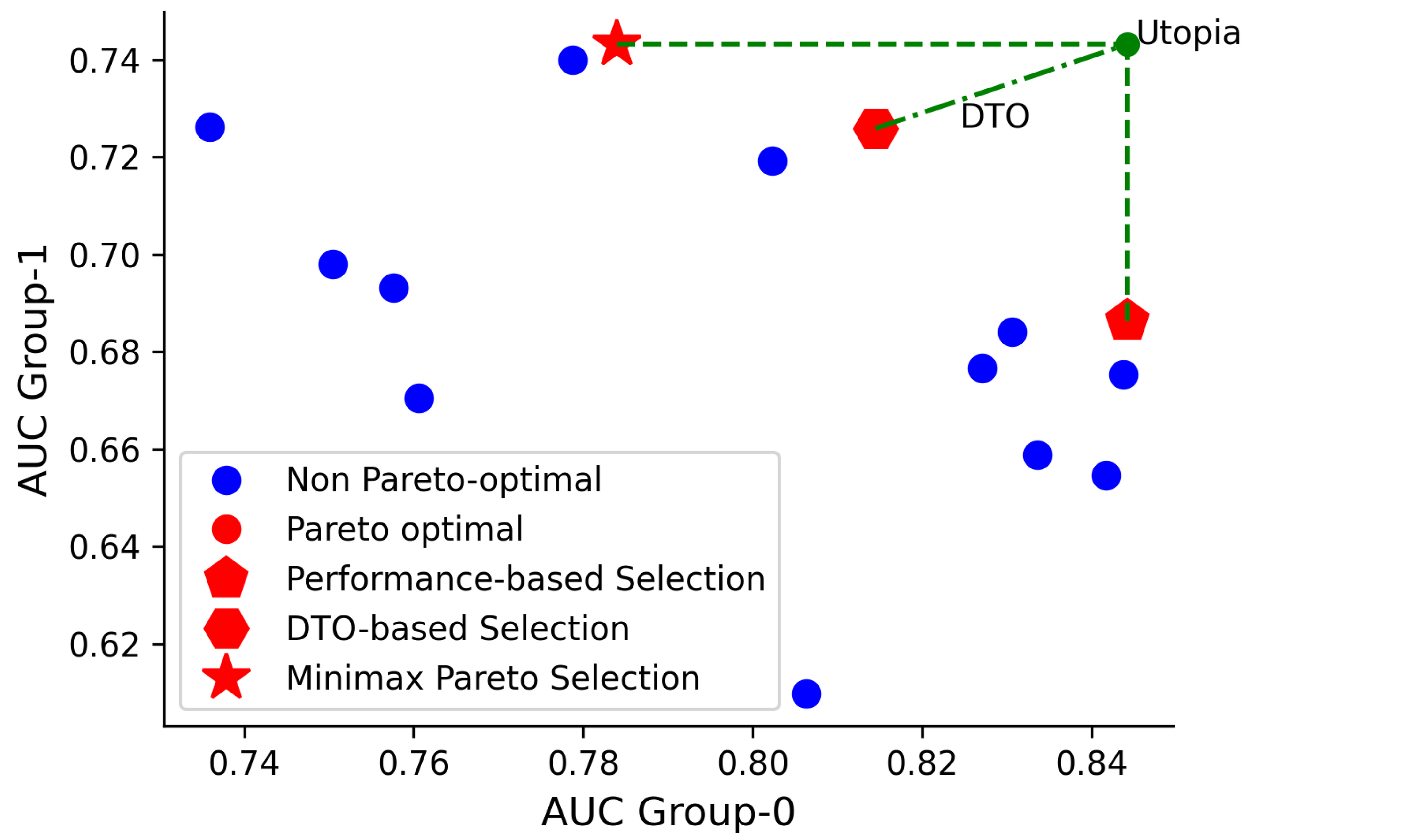
Influence of model selection strategies on ERM: *even without any explicit bias mitigation algorithm, Max-Min fairness can be significantly improved by adopting the Pareto optimal strategy in place of the overall strategy.*

3. No method performs significantly better than ERM



Performance of bias mitigation algorithms summarized across all datasets as average rank CD diagrams. (a) in-distribution, (b) out-of-distribution. No method performs significantly better than ERM.

Model Selection



Three model selection strategies: 1) overall performance, 2) distance to optimal (DTO), and 3) Minimax Pareto optimal. Each data point represents a different hyperparameter combination for one algorithm, where the red points are the models lying on the Pareto front.

Discussion

- **Failure of the bias mitigation algorithms:**
  - Multiple confounding effects can lead to bias, while most of the algorithms are designed for specific factors.
  - Understandable, as some are not originally designed for medical imaging.
- **Relation of domain generalization and fairness**
  - Share the eventual goal: being robust to changes in distribution across different sub-populations.
  - Some DG methods consistently improve the performance of all subgroups (e.g., SWAD).
- **Is the current evaluation enough?**

  – MEDFAIR will be a living codebase for more algorithms, datasets, and tasks.

References

[1] Dwork, et al. Fairness through awareness. ITCS'12.
[2] Lahoti, et al. Fairness without Demographics through Adversarially Reweighted Learning. NeurIPS'20.
[3] Han, et al. Balancing out Bias: Achieving Fairness Through Balanced Training. EMNLP'22.
[4] Martinez, et al. *Minimax Pareto Fairness: A Multi Objective Perspective*. ICML'20.