MICCAI 2026

UNIVERSITY OF SURREY

Fairness Beyond Demographics: Optimizing Performance Across Appearance-Based Hidden Cohorts in Medical Imaging

Milad Masroor   Cuong Nguyen   Kevin Wells   Gustavo Carneiro
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

Motivation

? Why go beyond demographic attributes for fairness in medical imaging?

1. Demographics may not capture model failures

Visible demographic attributes may only weakly align with latent factors driving model behaviour.

Gender • Age • Ethnicity ≠ all sources of disparity

Visible groups may overlook important sources of disparity.

⇓

2. Hidden stratification

Meaningful vulnerable cohorts may be hidden within image appearance.

Appearance-based hidden cohorts

Unlabeled cohorts can expose model blind spots and performance disparities.

⇓

3. Demographic-dependent fairness is difficult to deploy and scale

(a) Incomplete or unavailable metadata

Sensitive attributes may be **missing, noisy, incomplete, or unavailable** due to privacy constraints.

Demographic metadata	Potential limitation
Gender	Missing / noisy
Age	Incomplete
Ethnicity	Unavailable / privacy-constrained

(b) Intersectional groups lead to data sparsity

Attributes	# Groups
Gender	2
Gender × Age	4
Gender × Age × Ethnicity	8
Gender × Age × Ethnicity × ...	16+

More groups → fewer samples per subgroup

→ Can we optimize fairness across appearance-based hidden cohorts without requiring demographic labels?

Main Contributions

1 **LHCF**
Label-Free Hidden-Cohort Fairness
A demographic-label-free training paradigm that **discovers appearance-based hidden cohorts** and **optimizes fairness across them.**

2 **HIDFairBench**
Hidden-Demographic Fairness Benchmark
A benchmark for evaluating **hidden-cohort quality** and stress-testing fairness under **increasingly complex demographic partitions.**

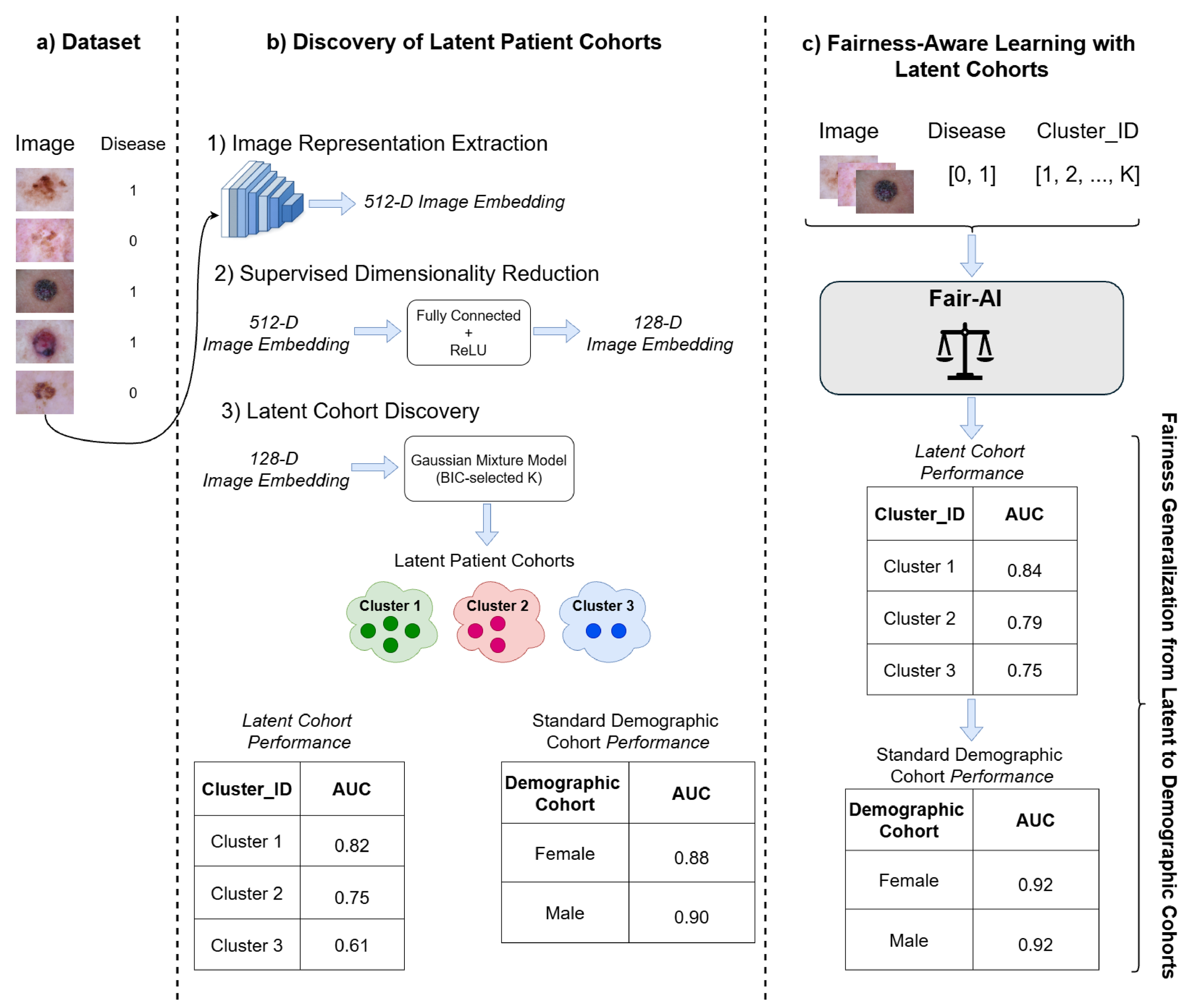
Future Directions

→ **Expert-Guided Validation of Hidden-Cohort Quality**
Incorporate expert assessment to evaluate the **clinical relevance** of discovered hidden cohorts.

→ **Reproducible Hidden-Cohort Discovery**
Stabilize the unsupervised clustering stage to improve **LHCF** reproducibility.

Methodology

- **Latent cohort discovery:** Discover K latent patient cohorts from image embeddings using GMM, with K selected by BIC.
- **Label-free fairness optimization:** Use **Cluster_ID** as the sensitive attribute to optimize fairness without demographic labels.
- **FairAI-compatible:** **LHCF** is **model- and fairness-loss agnostic** and integrates with existing fairness methods.



Experiments

Hidden-Cohort Quality

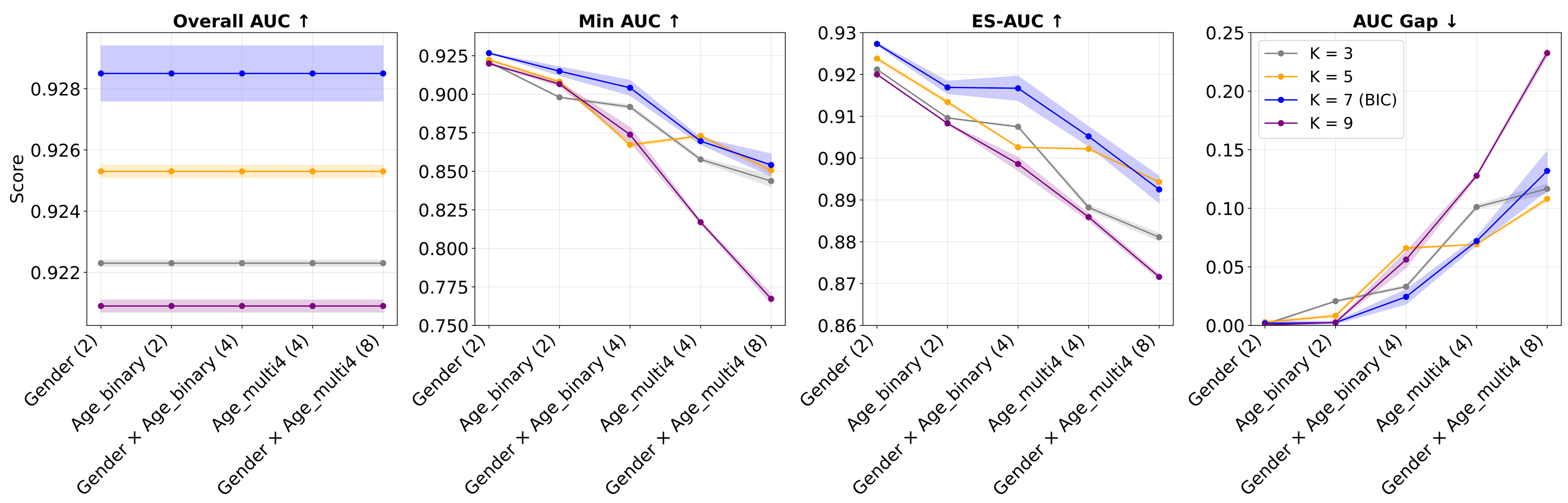
Clt.ID	AUC	BS	(Mal)(Ben)	(Male)(Female)	($\leq$60)($>$60)
0	–	5.3E-5	(0.00)(100.00)	(48.66)(51.34)	(78.61)(21.39)
1	0.6179	0.1506	(18.55)(81.45)	(39.18)(60.82)	(70.10)(29.90)
2	0.8247	0.0241	(2.63)(97.37)	(46.24)(53.76)	(74.44)(25.56)
3	0.7507	0.1759	(27.27)(72.73)	(45.45)(54.55)	(54.54)(45.46)
4	0.6113	0.2255	(36.36)(63.64)	(63.64)(36.36)	(35.06)(64.94)
5	0.7557	0.1090	(13.65)(86.35)	(59.03)(40.97)	(56.38)(43.62)
6	0.7299	0.1956	(66.39)(33.61)	(69.75)(30.25)	(23.53)(76.47)

Visible Cohort Evaluation

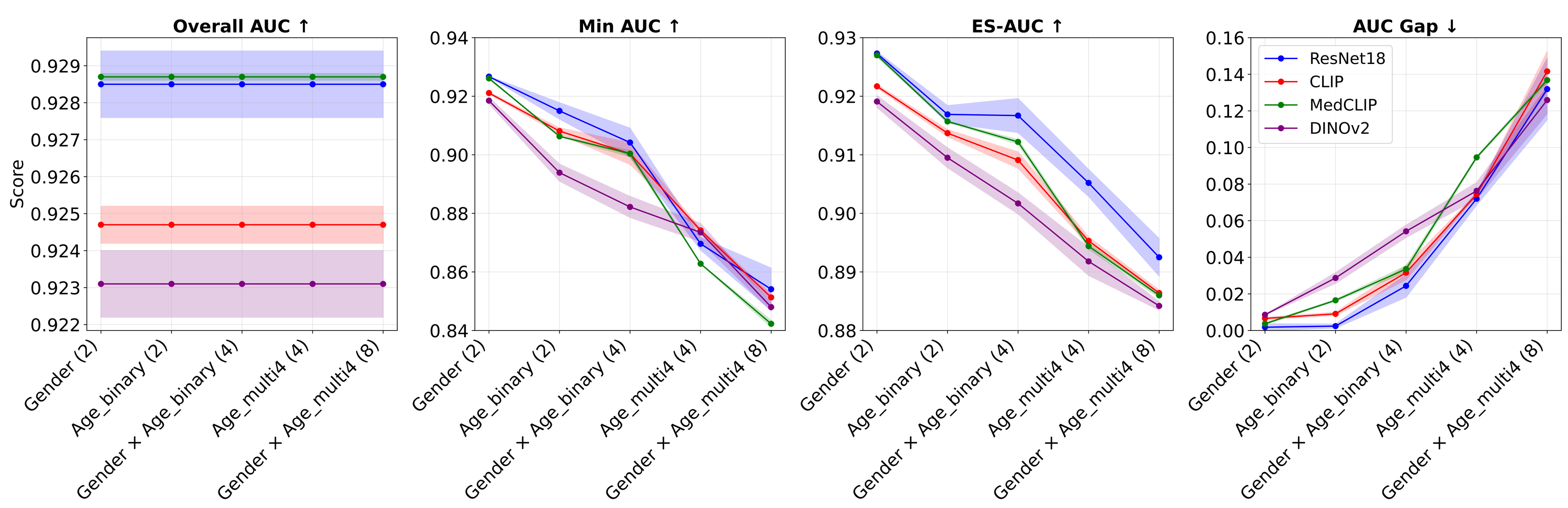
	Measures and Ranks	ERM	SWAD	FIS	FaMI	FEBS	FairCLIP-OT	FairDi
Classic	Avg. Overall AUC Score↑	0.8775	0.8867	0.8806	0.8582	0.8731	0.8692	0.9014
	Avg. Min. AUC Score↑	0.8213	0.8340	0.8212	0.8113	0.8246	0.8083	0.8511
	Avg. ES-AUC Score↑	0.8447	0.8578	0.8477	0.8304	0.8444	0.8374	0.8744
	Avg. Gap AUC Score↓	0.1053	0.0905	0.1019	0.0953	0.0940	0.1012	0.0824
	Avg. Mean PSD Score↓	0.0478	0.0406	0.0487	0.0448	0.0444	0.0473	0.0354
	Avg. Max PSD Score↓	0.1220	0.1042	0.1186	0.1134	0.1104	0.1192	0.0915
LHCF	Avg. Overall AUC Score↑	0.8775	0.8889	0.8804	0.8668	0.8827	0.8703	**0.9050**
	Avg. Min. AUC Score↑	0.8213	0.8418	0.8266	0.8106	0.8369	0.8127	**0.8686**
	Avg. ES-AUC Score↑	0.8447	0.8617	0.8486	0.8339	0.8519	0.8399	**0.8817**
	Avg. Gap AUC Score↓	0.1053	0.0843	0.0967	0.1061	0.0929	0.0963	**0.0666**
	Avg. Mean PSD Score↓	0.0478	0.0386	0.0461	0.0498	0.0450	0.0451	**0.0312**
	Avg. Max PSD Score↓	0.1220	0.0976	0.1124	0.1245	0.1080	0.1132	**0.0762**

Ablation Studies

Number of Hidden Cohorts (K) – BIC-selected $K = 7$ achieves the best overall AUC–fairness trade-off.



Embedding Backbone – ResNet18 and MedCLIP provide the best overall accuracy–fairness performance.



Demographic-Aware Clustering (DAC) vs. LHCF – DAC clusters image + demographic features.