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

Milad Masroor(✉), Cuong Nguyen, Kevin Wells, and Gustavo Carneiro

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
{m.masroor,c.nguyen,k.wells,g.carneiro}@surrey.ac.uk

Abstract. Medical image analysis models can exhibit performance disparities across patient subgroups, threatening clinical safety and fairness. Existing methods typically address this issue by optimizing accuracy and fairness metrics for visible demographic attributes (e.g., sex or age) considered in isolation. This strategy not only overlooks potentially more informative latent stratifications, which may reveal deeper sources of model error and inequity, but also fails to scale when multiple demographic attributes are considered simultaneously due to the resulting sparsity of training data within each subgroup. We deal with these issues by introducing the label-free hidden-cohort fairness (LHCF) training paradigm that instead of maximizing fairness over visible demographic attributes, it optimizes fairness across latent subpopulations discovered from image appearance. By clustering images into K appearance-based cohorts and applying fairness optimization over them, LHCF uncovers underlying sources of model error and avoids the combinatorial sparsity of multi-demographic attributes, reducing disparities across both single and multiple demographic attributes. We demonstrate on our proposed fairness benchmark, HIDFairBench, that LHCF provides state-of-the-art fairness results on single and multiple demographic attributes without requiring demographic labels for training, improving practicality in real-world clinical settings where metadata may be incomplete or unavailable. Our results position hidden-cohort fairness as a practical, scalable, and robust alternative to demographic-based fairness optimization for trustworthy medical image analysis. Project page is available here.[1]

Keywords: Hidden Cohort Fairness · Demographic-Label-Free Fairness · Trustworthy AI in Healthcare.

1 Introduction

Medical image analysis (MIA) systems often exhibit performance disparities across visible demographic attributes (e.g., sex, age, ethnicity) [1, 6, 12, 18, 24, 32], raising concerns about clinical safety and fairness [6, 32]. Although these

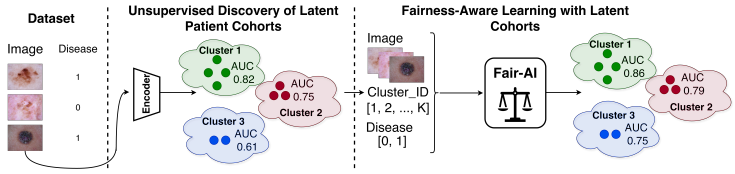
[1] Supported by the Engineering and Physical Sciences Research Council (EPSRC) through grant EP/Y018036/1.

attributes are assumed to be the main drivers of such gaps, recent studies have shown that they may align only weakly with latent factors that may better explain model behavior [1, 24, 31]. Indeed, performance disparities can be much larger between hidden patient cohorts (formed by visually coherent but unlabeled subsets of the data) than across demographic groups, creating low-performing subpopulations often referred to as underperforming slices or blind spots [10, 25, 27, 39]. This challenge becomes more severe when multiple attributes are considered jointly, as the number of intersectional subgroups expands while available samples per subgroup shrink [2]. These issues reflect the hidden stratification phenomenon, in which coarse labels fail to capture meaningful clinical variation [24, 40], causing models to perform poorly on underrepresented cohorts [8, 28, 29, 33] and ultimately increasing the risk of disproportionate patient harm [15].

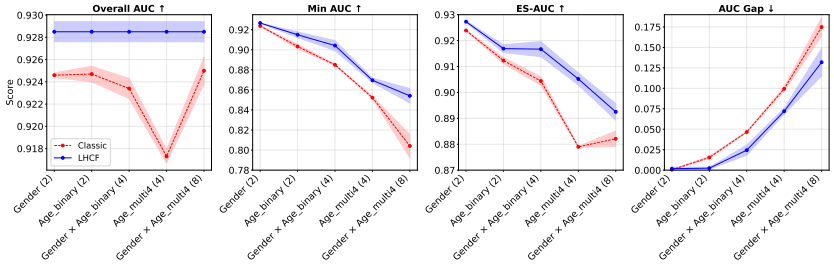
Despite growing evidence that hidden patient cohorts represent a major driver of model failures, fairness methods in medical imaging still focus primarily on visible demographic attributes [19, 20, 22, 34, 35], while work on hidden cohorts has mostly diagnosed, instead of mitigating, the reported disparities [1, 5, 7, 25]. Demographic-focused fairness approaches also scale poorly as intersectional groups increase and per-group sample sizes collapse [2]. Taken together, these limitations point to latent cohorts, with sufficient data and richer structure, as a more reliable basis for fairness optimization than visible demographic groups. From a real-world perspective, latent-cohort fairness optimization addresses a key limitation of demographic-based fairness methods, which rely on predefined sensitive attributes that may be missing, noisy, incomplete, or unavailable due to privacy constraints. By discovering appearance-driven hidden cohorts directly from learned representations, it enables more scalable and robust fairness optimization without requiring explicit demographic information. To illustrate this paradigm shift, we compare two approaches: (i) classic fairness training that optimizes across visible demographic attributes; and (ii) label-free hidden-cohort fairness (LHCF), which optimizes fairness across clusters formed with appearance-based embeddings. Fig. 1b shows that LHCF yields the best classification and fairness results not only for isolated demographic groups but also for multi-attribute intersections.

In this paper, we propose LHCF (Fig. 1a), which optimizes fairness over appearance-based latent cohorts without demographic labels and is compatible with existing fairness approaches. We also propose the HIdden-Demographic Fairness Benchmark (HIDFairBench), a new benchmark to stress-test fairness methods. This paper has the following contributions:

- The new LHCF training that clusters images into K hidden clusters and applies fairness optimization to balance accuracy and fairness across these clusters. This approach avoids the combinatorial sparsity of multi-attribute demographic groups and reduces performance disparities not only for the hidden clusters but also across single and multiple demographic attributes.



(a) LHCF training.



(b) Classic vs. LHCF Results.

Fig. 1: (a) LHCF training. (b) Overall AUC (left) and fairness metrics (Min AUC, ES-AUC, AUC Gap; second-to-fourth plots) across visible cohort complexity for Classic demographic fairness training (i) and LHCF training (ii) of FairDi [22] (using ResNet18 backbone) on HAM10000 [36]. Visible-cohort complexity increases from Gender (2 groups) to Age_binary (2 groups), Age_multi4 (4 groups), and intersectional cohorts Gender×Age_binary (4 groups) and Gender×Age_multi4 (8 groups) – Sec. 2.2 explains this experimental setup.

– The new HIDFairBench, a benchmark that evaluates hidden-cohort quality and tests fairness methods under increasingly complex multi-demographic partitions, exposing the scalability limits of fairness approaches.

Our experiments show that LHCF achieves state-of-the-art accuracy and fairness across single and multiple demographic attributes from HIDFairBench, despite never using demographic labels for training, establishing LHCF as a scalable and robust alternative to demographic-based fairness optimization in MIA.

2 Method

In this section, we present the LHCF training and the HIDFairBench benchmark.

2.1 LHCF Training

The **LHCF training has two steps** (Fig. 1a): **(i)** unsupervised discovery of latent cohorts from image embeddings, and **(ii)** fairness-aware training, where the discovered cohort identities are treated as sensitive attributes.

Step (i) assumes the availability of a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, containing medical images $x \in \mathcal{X}$ and diagnostic labels $y \in \mathcal{Y} \subseteq \Delta^{|\mathcal{Y}|-1}$, but without demographic annotations. Our goal is to learn the model $f_\theta = d_{\theta_d} \circ e_{\theta_e}$, where $\theta = (\theta_e, \theta_d)$, $e_{\theta_e} : \mathcal{X} \to \mathcal{Z}$ represents the encoder, and $d_{\theta_d} : \mathcal{Z} \to \mathcal{Y}$ denotes the decoder. We assume that the model $f_\theta(.)$ has been pre-trained either with self-supervised learning [17] or with supervised learning using $\mathcal{D}$. In this Step (i), the method first finds the hidden cohorts via Gaussian Mixture Model (GMM) [9], where the number of clusters is estimated with the Bayesian Information Criterion (BIC) [30]. This clustering happens with the image embeddings in $\mathcal{Z} = \{z_i | z_i = e_{\theta_e}(x_i), (x_i, y_i) \in \mathcal{D}\}$. For $K \in \mathbb{N}$, a Gaussian mixture is defined by $p(z \mid \Phi_K) = \sum_{k=1}^K \pi_k \mathcal{N}(z \mid \mu_k, \Sigma_k), \quad \Phi_K = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, with log-likelihood $\ell(\Phi_K; \mathcal{Z}) = \sum_{i=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(z_i \mid \mu_k, \Sigma_k)$. We estimate Φ_K^* by Expectation Maximization (EM) [9], yielding the responsibility $\gamma_{ik} = p(k \mid z_i, \Phi_K^*) = \frac{\pi_k \mathcal{N}(z_i|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(z_i|\mu_j, \Sigma_j)}$. The number of cohorts can be estimated with

$$K^* \in \operatorname{argmin}_{K \in \{1,\dots,N\}} \mathrm{BIC}(K) = \operatorname{argmin}_{K \in \{1,\dots,N\}} p_K \log N - 2\,\ell(\Phi_K^*; Z), \quad (1)$$

where $p_K = (K-1) + Kd + K\,\frac{d(d+1)}{2}$ for full covariances. The hidden cohort dataset to be used in subsequent fairness optimization is then defined by $\widehat{\mathcal{D}} = \{(x_i, y_i, c_i) | (x_i, y_i) \in \mathcal{D}, c_i = \arg\max_{k \in \{1,\dots,K^*\}} \gamma_{ik}\}$, where c_i represents the hidden cohort label of the sample (x_i, y_i) computed with the responsibility γ_{ik}.

Step (ii) uses the hidden cohort dataset $\widehat{\mathcal{D}}$ to optimize

$$\theta^* = \operatorname{argmin}_\theta \ \tfrac{1}{N} \textstyle\sum_{i=1}^N \ell_{\text{clss}}\big(y_i, f_\theta(x_i)\big) + \lambda\, \ell_{\text{fair}}\big(\widehat{\mathcal{D}}, \theta\big), \quad (2)$$

where $\lambda \geq 0$ is a hyper-parameter, $\ell_{\text{clss}}(.)$ is a classification loss and $\ell_{\text{fair}}(.)$ denotes a fairness loss that can be defined in different ways, such as:

$$\ell_{\text{fair}}^{\text{wst}}\big(\widehat{\mathcal{D}}, \theta\big) = \max_k R_k(\theta) \quad \text{or} \quad \ell_{\text{fair}}^{\text{gap}}\big(\widehat{\mathcal{D}}, \theta\big) = \max_k R_k(\theta) - \min_k R_k(\theta), \quad (3)$$

with $R_k(\theta) = \frac{1}{|C_k|} \sum_{(x_i,y_i) \in C_k} \ell_{\text{clss}}\big(y_i, f_\theta(x_i)\big)$ denoting classification loss of samples in the cluster C_k, $\ell_{\text{fair}}^{\text{wst}}(\cdot)$ minimizing the worst-cohort classification loss, and $\ell_{\text{fair}}^{\text{gap}}(\cdot)$ minimizing the gap between the best- and worst-cohort losses. LHCF is agnostic to the model and fairness loss, which together define a FairAI method (e.g., SWAD [4], FIS[20], FEBS [35], FairCLIP [19], FaMI [34], FairDi [22]).

Theoretical analysis We show that under mild clustering assumptions, the fairness loss $\ell_{\text{fair}}^{\text{wst}}(\widehat{\mathcal{D}}, \theta)$ upper-bounds the loss of any visible cohort; thus, minimizing it yields fairness across all visible cohorts [29]. Specifically, we assume that the GMM with hard assignments partitions the data into K disjoint clusters $\{C_1, \dots, C_K\}$ with $C_k \cap C_{k'} = \varnothing$ for $k \neq k'$, and that each visible cohort G_t is a union of these clusters, i.e., $G_t = \bigcup_{k \in \mathcal{K}} C_k$, where $\mathcal{K} \subseteq \{1, \dots, K\}$.

Lemma 1. *The empirical risk of any visible cohort is upper-bounded by* $\ell_{\text{fair}}^{\text{wst}}(\widehat{\mathcal{D}}, \theta)$, *meaning:* $\mathsf{L}(G_t) = \frac{1}{|G_t|} \sum_{(x_i,y_i) \in G_t} \ell_{clss}(y_i, f_\theta(x_i)) \leq \max_k R_k(\theta)$.

Proof. Assuming $G_t = \cup_{k\in\mathcal{K}} C_k$, $\mathbb{I}((x_i, y_i) \in G_t) = \sum_{k\in\mathcal{K}} \mathbb{I}((x_i, y_i) \in C_k)$, where $\mathbb{I}(.)$ is the indicator function, we re-write the risk on G_t as:

$$\begin{aligned}
\mathsf{L}(G_t) &= \tfrac{1}{|G_t|} \textstyle\sum_{(x_i,y_i)} \ell_{\text{clss}}(y_i, f_\theta(x_i)) \mathbb{I}((x_i, y_i) \in G_t) \\
&= \tfrac{1}{|G_t|} \textstyle\sum_{(x_i,y_i)} \sum_{k\in\mathcal{K}} \ell_{\text{clss}}(y_i, f_\theta(x_i)) \mathbb{I}((x_i, y_i) \in C_k) \\
&= \tfrac{1}{|G_t|} \textstyle\sum_{k\in\mathcal{K}} \sum_{(x_i,y_i)\in C_k} \ell_{\text{clss}}(y_i, f_\theta(x_i)) \\
&= \tfrac{1}{|G_t|} \textstyle\sum_{k\in\mathcal{K}} |C_k| R_k(\theta) \leq \sum_{j\in\mathcal{K}} {}^{|C_j|}\!/_{|G_t|} \max_k R_k(\theta) \leq \max_k R_k(\theta). \quad (4)
\end{aligned}$$

□

2.2 HIDFairBench Benchmark

Datasets. The benchmark relies on three public medical imaging datasets: *Fitzpatrick17K* [14], *HAM10000* [36], and *CMMD* [3]. For HAM10000 (9,948 dermoscopic images), we follow prior work [21] and binarize diagnoses into benign vs. malignant. Visible cohort partitions include *Gender*, *Age_ binary* ($\leq$60, >60), *Age_ multi4* ($\leq$39, 40–59, 60–79, $\geq$80), and intersectional *Gender* $\times$ *Age_ binary* and *Gender* $\times$ *Age_ multi4* groups. For Fitzpatrick17K, we group lesions into benign (non-neoplastic + benign) vs. malignant, and use Fitzpatrick *skin types* (0–5) as visible demographic attributes. For both HAM10000 and Fitzpatrick17K, we use an 80%/10%/10% train/validation/test split. For CMMD (5,152 mammography images), we frame classification as non-cancer vs. cancer, use binarized *patient age* ($\leq$55, >55) as the sensitive attribute, and rely on a 70%/10%/20% train/validation/test split.

Evaluation & Statistical Tests. The hidden-cohort quality is measured using the *Brier Score (BS)* [25], which captures a combination of calibration and discrimination, and *Average Purity (AP)* [1], which quantifies alignment between discovered cohorts and visible demographic attributes. For visible cohort evaluation, we report *AUC* and four fairness metrics: *AUC Gap*, *Worst-Case AUC* [41], *ES-AUC* [20], and *Performance-Scaled Disparity (PSD)*. Statistical significance uses the *Friedman test* [13, 41] with *Nemenyi* post-hoc comparisons [23] ($p < 0.05$), visualized via Critical Difference (CD) diagrams.

FairAI Methods & Backbones. The benchmark tests many *FairAI methods*, starting from the *ERM* [37] baseline, which optimizes average classification error without addressing fairness. *SWAD* [4] improves generalization by promoting convergence to flat minima. Among fairness approaches, *FIS* [20] reweights group losses to equalize performance, while *FEBS* [35] emphasizes hard examples via error-bound scaling. We also include *FairCLIP-OT*, an optimal-transport fairness regularizer adapted from FairCLIP [19], and *FaMI* [34], which reduces dependence between representations and sensitive attributes via mutual information. Furthermore, *FairDi* [22] learns a shared fair backbone and distills knowledge from cohort-specific experts into a unified student model. For the *backbone models*, we consider *ResNet18* [16], *CLIP* [11], *MedCLIP* [38], and *DINOv2* [26].

Table 1: AUC and BS across hidden clusters (Clt. ID) on HAM10000 (ERM, ResNet18), with cohort proportions (%) of malignant/benign, male/female, and age ≤ 60 or > 60 years old.

Clt.ID	AUC	BS	(Mal)(Ben)	(Male)(Female)	(≤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)

3 Experiments

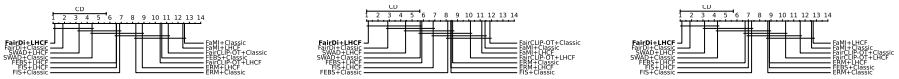
3.1 HIDFairBench Results

Hidden-Cohort Quality. Using an ERM-trained ResNet18, BS on HAM10000 reveals a clear risk stratification across hidden cohorts (Table 1): low malignancy-rate cohorts (Clusters 0,2) show low BS, intermediate malignancy-rate cohorts (1,3,5) exhibit moderate BS, and high malignancy-rate cohorts (4,6) attain the highest BS, indicating reduced reliability on clinically severe cases. These trends mirror underlying population structure, with cohorts dominated by older patients showing higher malignancy prevalence, consistent with known age-related risk patterns in dermatology. Across CMMD and Fitzpatrick17K, BS similarly tracks malignancy severity, with low malignancy-rate cohorts obtaining low calibration error and high malignancy-rate clusters showing higher BS. Despite this clear risk stratification, hidden cohorts show weak alignment with available demographic metadata: on HAM10000, AP is 0.4919±0.0183 (gender) and 0.2234±0.0065 (age); on Fitzpatrick17K, AP for skin type is 0.0998±0.0161; and on CMMD, AP for age is 0.3517±0.0067. Such APs < 0.5 confirm that hidden cohorts primarily reflect visual structure rather than demographic attributes.

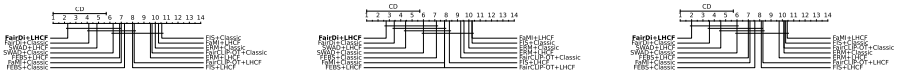
Visible Cohort Evaluation. Table 2 compares visible-demographic supervised (*Classic*) and *LHCF* training across all datasets and demographic attributes. LHCF improves AUC and fairness, especially for intersectional cohorts, without using any demographic metadata (see Fig. 1b). Despite weak alignment between hidden and demographic groups (AP$<$ 0.5), the fairness learned from appearance-based cohorts generalizes well to visible attributes, with FairDi+LHCF delivering the strongest overall results. The CD diagrams in Fig. 2 confirm the superiority of LHCF, and in particular the FairDi+LHCF as the top-ranked approach across nearly all evaluation metrics. An important question is why LHCF, trained on hidden cohorts, outperforms classical methods trained on visible cohorts. The intuition, based on Lemma 1, is that instead of minimizing the average loss over overlapping visible cohorts, which can overemphasize intersectional groups, LHCF minimizes the loss over hidden cohorts, thereby reducing extreme disparities and producing more balanced performance.

Table 2: Classic vs. LHCF training results averaged over HAM10000 (*Gender*, *Age_ binary*, *Age_ multi4*, *Gender* × *Age_ binary*, *Gender* × *Age_ multi4*), Fitzpatrick17K (*Skin Type*), and CMMD (*Age_ binary*). Cell colors denote Better / Worse / Same between LHCF and Classic, and **Bold**=best result.

	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**



(a) Overall AUC (b) Min. AUC (c) ES-AUC



(d) Gap AUC (e) Mean PSD (f) Max PSD

Fig. 2: CD diagrams comparing Classic and LHCF learning across all datasets and demographic cohorts (Please zoom in to view the figures clearly).

3.2 Ablation Studies

Number of Hidden Cohorts (K). Fig. 3 evaluates the performance of LHCF as a function of the number of cohorts. The BIC choice (i.e., K=7) produces the best results for AUC and fairness measures for the majority of visible cohort partitions, whereas larger or smaller values of $K \in \{3, 5, 9\}$ tend to degrade performance. This supports BIC as an effective criterion for selecting K.

Embedding Backbone. Fig. 4 summarizes how representation backbones affect LHCF. ResNet18 and MedCLIP achieve the best Overall AUC. Fairness metrics separate more clearly: for Min. AUC and AUC Gap, ResNet18 performs best on simpler intersectional settings; for ES-AUC, ResNet18 is consistently strongest across all cohort complexities, with MedCLIP a solid second. CLIP and DINOv2 show larger drops in worst-case performance and higher disparity as cohort complexity increases. Overall, backbones inducing *moderate* hidden-cohort granularity, e.g., ResNet18 (K=7) and MedCLIP (K=6) selected by BIC, yield better accuracy and fairness than CLIP (K=10) and DINOv2 (K=5).

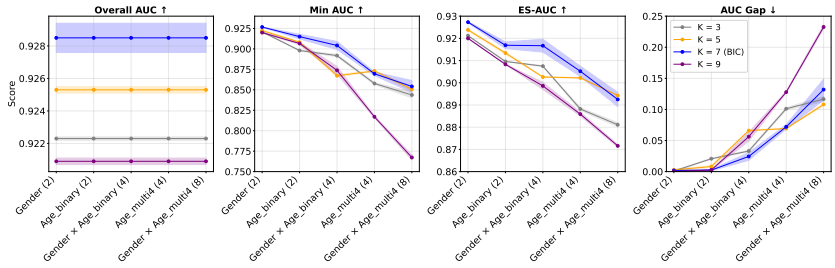


Fig. 3: Sensitivity analysis of LHCF (FairDi+ResNet18 backbone) with respect to the number of hidden cohorts (K) on the HAM10000 dataset.

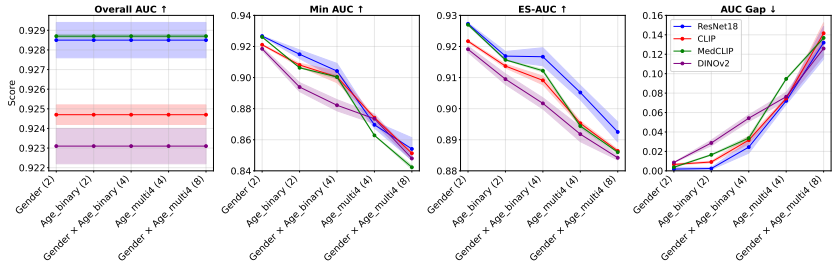


Fig. 4: Sensitivity analysis of LHCF (FairDi) with respect to the backbone (ResNet18, CLIP [11], MedCLIP [38], and DINOv2 [26]) on HAM10000.

Demographic-Aware Clustering (DAC) vs. LHCF. As an intermediate alternative between *Classic* and *LHCF*, DAC augments image embeddings with demographic information before clustering and then optimizes fairness over the resulting metadata-aware clusters. Specifically, sex and normalized age metadata are projected into the same latent dimensionality as the image embeddings using a lightweight MLP, followed by LayerNorm normalization. The normalized image and metadata representations are then concatenated and used as input to the GMM-based clustering pipeline for metadata-aware slice discovery. Fig. 5 compares DAC, LHCF, Classic (FairDi+ResNet18 backbone), and ERM across increasing visible-cohort complexity levels. The results show that LHCF consistently outperforms DAC, Classic, and ERM on nearly all fairness-related measures, particularly under more complex and intersectional cohort definitions. These findings indicate that appearance-driven, label-free fairness optimization offers greater robustness to visible-cohort complexity than DAC while eliminating any dependence on demographic labels.

4 Conclusion

We introduced LHCF, which optimizes fairness over appearance-based latent cohorts without demographic supervision, and proposed *HIDFairBench* to evaluate

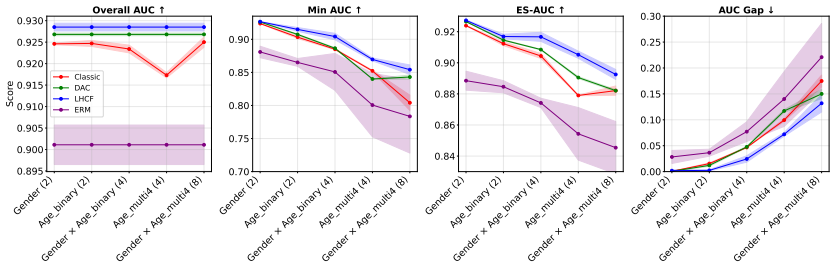


Fig. 5: Comparison of DAC, LHCF, Classic (FairDi+ResNet18 backbone), and ERM (ResNet18 backbone) under increasing visible cohort complexity on the HAM10000 dataset.

fairness under single and multiple demographic attributes. Across three datasets, hidden cohorts exhibit clinical risk stratification (BS rises with malignancy) but weak demographic alignment (AP$<$ 0.5), indicating they capture meaningful visual structure. Relative to the demographic-supervised Classic FairAI methods, LHCF improves fairness and accuracy on visible groups without using demographic labels, with solid gains for intersectional cohorts. LHCF is robust to design choices: a BIC-selected moderate number of cohorts (K=7) yields the best accuracy-fairness tradeoff, and backbones with similar cohort granularity (ResNet18, MedCLIP) provide the best performance. Compared with DAC, LHCF delivers better accuracy and fairness. In future work, we will improve the reproducibility of LHCF by stabilizing the unsupervised clustering stage and extend Lemma 1 to accommodate more fairness losses; nonetheless, the adoption of data-driven latent stratification offers a more generalizable and deployable pathway toward equitable and trustworthy medical imaging analysis.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bissoto, A. *et al.* Subgroup Performance Analysis in Hidden Stratifications. In MICCAI (2025), 594–603.
2. Buolamwini, J. *et al.* Gender shades: Intersectional accuracy disparities in commercial gender classification. In ACM FAccT (2018), 77–91.
3. Cai, H. *et al.* An online mammography database with biopsy confirmed types. Scientific Data **10,** 123 (2023).
4. Cha, J. *et al.* Swad: Domain generalization by seeking flat minima. NeurIPS **34,** 22405–22418 (2021).
5. Chakraborty, R. *et al.* Exmap: Leveraging explainability heatmaps for unsupervised group robustness to spurious correlations. In CVPR (2024), 12017–12026.
6. Christodoulou, E. *et al.* Confidence intervals uncovered: Are we ready for real-world medical imaging AI? In MICCAI (2024), 124–132.

7. Collevati, M. *et al.* Leveraging Neurosymbolic AI for Slice Discovery. In NESY (2024), 403–418.
8. DeGrave, A. *et al.* AI for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell. 3, 610–619. 2021.
9. Dempster, A. P. *et al.* Maximum likelihood from incomplete data via the EM algorithm. JRSSSB (methodological) **39,** 1–22 (1977).
10. Eyuboglu, S. *et al.* Domino: Discovering systematic errors with cross-modal embeddings. arXiv preprint arXiv:2203.14960 (2022).
11. Fang, Y. *et al.* Eva: Exploring the limits of masked visual representation learning at scale. In CVPR (2023), 19358–19369.
12. Finlayson, S. G. *et al.* The clinician and dataset shift in artificial intelligence. New England Journal of Medicine **385,** 283–286 (2021).
13. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association **32,** 675–701 (1937).
14. Groh, M. *et al.* Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In CVPR Workshop (2021), 1820–1828.
15. Hashimoto, T. *et al.* Fairness without demographics in repeated loss minimization. In ICML (2018), 1929–1938.
16. He, K. *et al.* Deep residual learning for image recognition. In CVPR (2016), 770–778.
17. Jing, L. *et al.* Self-supervised visual feature learning with deep neural networks: A survey. IEEE TPAMI **43,** 4037–4058 (2020).
18. Koch, L. M. *et al.* Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digital Medicine **7,** 120 (2024).
19. Luo, Y. *et al.* FairCLIP: Harnessing fairness in vision-language learning. In CVPR (2024), 12289–12301.
20. Luo, Y. *et al.* FairVision: Equitable Deep Learning for Eye Disease Screening via Fair Identity Scaling. 2024. arXiv: `2310.02492 [cs.CV]`. `https://arxiv.org/abs/2310.02492`.
21. Maron, R. C. *et al.* Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. European Journal of Cancer **119,** 57–65 (2019).
22. Masroor, M. *et al.* Fair Distillation: Teaching Fairness from Biased Teachers in Medical Imaging. arXiv preprint arXiv:2411.11939 (2024).
23. Nemenyi, P. B. Distribution-free multiple comparisons. (Princeton University, 1963).
24. Oakden-Rayner, L. *et al.* Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In CHIL (2020), 151–159.
25. Olesen, V. *et al.* Slicing through bias: Explaining performance gaps in medical image analysis using slice discovery methods. In FAIMI Workshop, MICCAI (2024).
26. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023).
27. Plumb, G. *et al.* Towards a more rigorous science of blindspot discovery in image classification models. arXiv preprint arXiv:2207.04104 (2022).
28. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine **15,** e1002686 (2018).

29. Sagawa, S. *et al.* Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In ICLR (2020).
30. Schwarz, G. Estimating the dimension of a model. The annals of statistics, 461–464 (1978).
31. Seo, S. *et al.* Unsupervised learning of debiased representations with pseudo-attributes. In CVPR (2022), 16742–16751.
32. Seyyed-Kalantari, L. *et al.* Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27,** 2176–2182 (2021).
33. Sohoni, N. *et al.* No subclass left behind: Fine-grained robustness in coarse-grained classification problems. NeurIPS **33,** 19339–19352 (2020).
34. Tian, Y. *et al.* FairDomain: Achieving Fairness in Cross-Domain Medical Image Segmentation and Classification. In ECCV (2024).
35. Tian, Y. *et al.* FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning Using Segment Anything Model with Fair Error-Bound Scaling. In ICLR (2024).
36. Tschandl, P. *et al.* The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5,** 1–9 (2018).
37. Vapnik, V. N. An overview of statistical learning theory. IEEE Trans. Neural Netw. **10,** 988–999 (1999).
38. Wang, Z. *et al.* Medclip: Contrastive learning from unpaired medical images and text. In EMNLP (2022), 3876–3887.
39. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ **369** (2020).
40. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15,** e1002683 (2018).
41. Zong, Y. *et al.* MEDFAIR: Benchmarking fairness for medical imaging. In ICLR (2023).