Achieving Fairness Without Harm via Selective Demographic Experts

Xuwei Tan[1], **Yuanlong Wang**[1,2], **Thai-Hoang Pham**[1,2], **Ping Zhang**[1,2], **Xueru Zhang**[1]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Biomedical Informatics, The Ohio State University, USA
{tan.1206, wang.16050, pham.375, zhang.10631, zhang.12807}@osu.edu

Abstract

As machine learning systems become increasingly integrated into human-centered domains such as healthcare, ensuring fairness while maintaining high predictive performance is critical. Existing bias mitigation techniques often impose a trade-off between fairness and accuracy, inadvertently degrading performance for certain demographic groups. In high-stakes domains like clinical diagnosis, such trade-offs are ethically and practically unacceptable. In this study, we propose a fairness-without-harm approach by learning distinct representations for different demographic groups and selectively applying demographic experts consisting of group-specific representations and personalized classifiers through a no-harm constrained selection. We evaluate our approach on three real-world medical datasets—covering eye disease, skin cancer, and X-ray diagnosis—as well as two face datasets. Extensive empirical results demonstrate the effectiveness of our approach in achieving fairness without harm.

Code — https://github.com/osu-srml/FairSDE
Extended version — https://arxiv.org/abs/2511.06293

1 Introduction

As machine learning (ML) is increasingly used in high-stakes domains like healthcare, finance, and criminal justice, concerns about algorithmic bias and its impact on marginalized groups have grown. In medical domain, fairness is particularly critical: diagnostic tools that underperform on certain age, gender, or racial groups can lead to misdiagnoses, delayed treatments, and ultimately, harm to patients. For instance, a model developed to diagnose skin diseases (Maron et al. 2019) was shown to exhibit age bias, with up to a 13% gap in area under the curve (AUC) between younger and older patients (Figure 1). Such performance disparities highlight the urgent need for effective bias mitigation strategies to promote equitable outcomes across diverse social groups.

To quantify model unfairness, existing studies have proposed many notions of **group fairness**, which require statistical measures (e.g., accuracy, true/false positive rate) to be equal across different groups. Commonly used notions include *demographic parity* (Dwork et al. 2012), *equal opportunity* and *equalized odds* (Hardt, Price, and Srebro 2016).

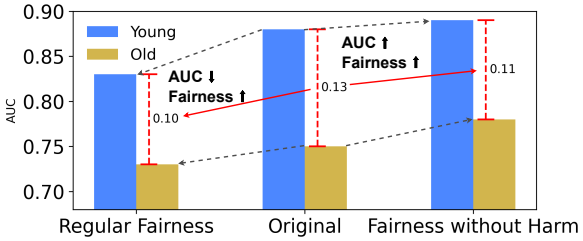


Figure 1: AUC of ResNet-18 on `Ham10000` for malignant prediction using age as the sensitive group. The 13% gap in AUC between age groups indicates bias. Regular approaches that enforce group constraints (e.g., left *Young* and *Old*) reduce this gap but degrade performance for both groups. In contrast, our method improves fairness without harming either group's performance, as seen in right *Young* and *Old*.

Based on them, various approaches have been developed and can be broadly categorized into three types: (i) ***pre-processing***, by modifying the original dataset such as removing certain features or reweighing samples (Kamiran and Calders 2012; Zemel et al. 2013; Gordaliza et al. 2019); (ii) ***in-processing***, by modifying the learning algorithms such as imposing fairness constraints or changing objective functions (Zafar et al. 2019, 2017a; Agarwal et al. 2018); (iii) ***post-processing***, by adjusting model outputs based on sensitive attributes (Hardt, Price, and Srebro 2016; Khalili, Zhang, and Abroshan 2021). However, these methods often enhance fairness at the cost of reduced accuracy. Notably, performance deterioration may happen to all groups, including those disadvantaged (see Figure 1 for an illustration).

In high-stakes domains like healthcare, sacrificing any group's model performance for fairness is unacceptable, as it conflicts with the ethical principles of *beneficence* (doing good) and *non-maleficence* (avoiding harm) (Beauchamp and Childress 1994). A more desirable goal is to improve outcomes for disadvantaged groups without compromising performance for others, as illustrated in Figure 1.

This paper aims to achieve fairness in ML without compromising the model performance of any group. The most closely related works are (Martinez and Bertran 2019; Ustun, Liu, and Parkes 2019; Yin et al. 2024; Cai, Khalili, and Zhang 2025). Specifically, Martinez and Bertran (2019)

focused on finding a *Pareto-optimal* fair model that minimizes the performance gap among different groups without *unnecessary* harm (i.e., minimal accuracy reduction for any group); this differs from this paper where we aim to avoid performance degradation for any group. To prevent harm, Ustun, Liu, and Parkes (2019) suggested utilizing individuals' demographic information (i.e., sensitive attributes) to train a set of *decoupled* models. Rather than equalizing outcomes across different groups, their goal is to ensure that each group achieves the best performance with its assigned decoupled model, compared to the pooled model (trained on all groups) and the decoupled models of other groups. More recently, (Cai, Khalili, and Zhang 2025) considered a similar fairness notion, but introduced a demographic-agnostic method for learning decoupled models. Another study (Yin et al. 2024) proposed a post-processing method that utilizes *abstention* to achieve group fairness without harming accuracy, i.e., adjusting the output of a pre-trained ML model that selectively abstains from making predictions on certain samples, deferring decisions to humans. However, this approach requires human involvement and has limited scalability.

Moreover, methods introduced in (Martinez and Bertran 2019; Ustun, Liu, and Parkes 2019; Yin et al. 2024; Cai, Khalili, and Zhang 2025) were primarily evaluated on simple tabular data, such as `Adult` (Asuncion, Newman et al. 2007), `COMPAS` (Bellamy et al. 2018), and `Law` (Bellamy et al. 2018), while their performance on high-dimensional image data remains less explored. As we will demonstrate in this paper, high-dimensional features, such as medical images, are often not easily separable by sensitive attributes in the latent space, making it highly challenging to achieve fairness without harm using decoupled or abstained classifiers.

In this paper, we propose methods that **improve model fairness without reducing performance for any group**, relative to an unconstrained pooled empirical risk minimization (ERM) baseline trained on the full population. Our approach is well-suited for **high-dimensional** feature spaces. Inspired by Ustun, Liu, and Parkes (2019), we adopt personalized classifiers for each demographic group to achieve fairness without harm. The use of demographic information is motivated by real-world medical practice, where attributes such as age, sex, or ethnicity routinely inform diagnosis and treatment for specific subpopulations. A key challenge is that naively personalized classifiers may underperform pooled ERM when group distributions are similar, as the pooled model benefits from more data to learn shared patterns. To address this, we learn *demographic experts*—group-specific representations paired with personalized classifiers—that better capture group distributional differences. Building on experts, we develop a post-processing method that dynamically selects between expert models and the pooled model via a *fairness without harm* guided combinatorial optimization. Our main contributions are:

- We explore the feasibility of achieving fairness without harm on complex image datasets, addressing a critical gap in current research, which has primarily focused on simpler tabular datasets.
- We propose **FairSDE**, an extension of decoupled classifiers to **Fair S**elective **D**emographic **E**xperts, which decouples both representations and classifiers to enable group-specific models to effectively capture group-specific patterns. Then, we selectively adjust predictions by dynamically choosing between the expert and pooled models via a combinatorial optimization framework.
- We conduct extensive experiments on real-world medical imaging data, demonstrating that FairSDE consistently achieves fairness without harm. This highlights its potential for practical deployment in performance-critical applications. In contrast, existing methods often achieve fairness by sacrificing performance for certain groups.

2 Related Work

Fairness notions. Many notions have been proposed to measure the algorithmic unfairness, which can be broadly classified into the following. *Unawareness* prohibits the use of sensitive attributes in the training and decision-making process. *Parity-based fairness* requires certain statistical measures to be equalized across different groups, including Demographic Parity (Dwork et al. 2012; Zhang et al. 2020), Equal Opportunity (Hardt, Price, and Srebro 2016; Khalili et al. 2021), Equalized Odds (Hardt, Price, and Srebro 2016; Pham, Zhang, and Zhang 2023), Predictive Parity (Chouldechova 2017), Accuracy Parity (Khalili, Zhang, and Abroshan 2023; Zhang et al. 2019), etc. *Preference-based fairness*, inspired by fair division and envy-freeness in economics, ensures that each group prefers its own treatment over others, regardless of inter-group disparities (Zafar et al. 2017b; Ustun, Liu, and Parkes 2019). *Counterfactual fairness* holds when an individual's outcome remains unchanged in a hypothetical world where their sensitive attribute is different (Kusner et al. 2017; Zuo et al. 2024; Zuo, Khalili, and Zhang 2023). Finally, *individual fairness* treats similar individuals similarly for individual level fairness (Dwork et al. 2012).

Approaches to mitigating unfairness. Many fairness algorithms have been proposed to address bias, which can be categorized into pre-processing, in-processing, and post-processing. Pre-processing methods manipulate datasets to mitigate bias in the data, such as reweighing (Kamiran and Calders 2012), resampling (Kamiran and Calders 2012), and data preprocessing (Celis, Keswani, and Vishnoi 2020). In-processing mitigation methods refer to regularizing the objective to guide the model in learning a fair classification. For example, adversarial training (Zhang, Lemoine, and Mitchell 2018; Han, Baldwin, and Cohn 2021b,a) learn a sensitive discriminator and reverse gradient to learn the group-invariant classification. Shen et al. (2021); Park et al. (2022) utilize contrastive learning to align the data from the same group and Creager et al. (2019); Park et al. (2021); Lee et al. (2021) disentangle features to de-bias the model. Post-processing methods alter the prediction to improve the fairness. Hardt, Price, and Srebro (2016) creates separate thresholds for each sensitive group and alters the results to satisfy the specified fairness criteria, which results in lower accuracy in many cases. Yin et al. (2024) train surrogate models based on the results of a baseline model to adjust predictions.

3 Problem Statement

Let $\mathcal{S}$ be a dataset of n individuals, where each individual i is represented as (x_i, y_i, a_i) drawn from the joint distribution $P(X, Y, A)$. Here, $x_i = [x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^d$ is a d-dimensional feature vector, $y_i \in \mathcal{Y}$ is the label, and $a_i \in \mathcal{A}$ is a sensitive attribute (e.g., gender, race, age). We aim to learn a *representation function* $f : \mathbb{R}^d \to \mathbb{R}^m$ that maps inputs to representations $z_i = f(x_i)$, and a set of group-specific *classifiers* $h_a : \mathbb{R}^m \to \mathcal{Y}$ such that $\hat{y}_i = h_a(z_i)$ for individuals in group $a \in \mathcal{A}$. We use capital letters for random variables and lowercase for their realizations.

Let $R : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be the risk/loss function measuring the discrepancy between prediction and ground truth, and n_a be the number of samples in group $a \in \mathcal{A}$. Our goal is to learn a representation function f and group-specific classifiers $\{h_a\}_{a\in\mathcal{A}}$ that minimize the risk of the entire population while ensuring fairness without harm:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_A \left[\mathbb{E}_{X,Y|A} \left[R\left(h_A\left(f(X)\right), Y\right)\right]\right] \\ \text{s.t.} \quad & \text{No-harm constraint} \\ & \text{Fairness constraint} \end{aligned} \tag{1}$$

No-harm constraint. Denote $h_{\text{erm}} : \mathbb{R}^d \to \mathcal{Y}$ as the

$$h_{\text{erm}} = \arg\min_{h\in\mathcal{H}} \sum\nolimits_{i=1}^{n} R(h(x_i), y_i) \tag{2}$$

where $\mathcal{H}$ is the hypothesis class. It defines the baseline for the "no harm" criterion: predictions using the representation function f and group-specific classifiers $\{h_a\}_{a\in\mathcal{A}}$ cause no harm if every group's loss is no greater than that under h_{erm}, i.e., $\forall a \in \mathcal{A}$, the following should hold:

$$\mathbb{E}_{X,Y|A=a} \left[R\left(h_a\left(f(X)\right), Y\right)\right] \leq \mathbb{E}_{X,Y|A=a} \left[R\left(h_{\text{erm}}(X), Y\right)\right] \tag{3}$$

Fairness constraint. We adopt two widely used metrics:

- *Overall accuracy parity* (Berk et al. 2021) requires similar overall accuracy across all groups.:

$$\mathbb{P}[\hat{Y} = Y \mid A = a] = \mathbb{P}[\hat{Y} = Y \mid A = a'], \quad \forall a, a' \in \mathcal{A} \tag{4}$$

- *Max-min fairness* (Lahoti et al. 2020) maximizes the worst-group performance to reduce disparities across groups. It is formally defined as:

$$\max_{a\in\mathcal{A}} \mathbb{E}_{X,Y|A=a} \left[R\left(h_a\left(f(X)\right), Y\right)\right] \tag{5}$$

4 Methodology

Next, we present our approach to achieving fairness without harm. We first describe how to decouple representations to learn *demographic experts*, followed by a dynamic expert selection method that adjusts predictions to enforce fairness. An overview is shown in Figure 2.

4.1 Decoupling Representations for Experts

To advance group fairness while preserving task-specific discriminability, we propose a structured framework that learns demographic experts—specialized representations and classifiers for each subgroup defined by the sensitive attribute A and the class Y. Unlike adversarial methods, which aim to eliminate the dependence between the sensitive attribute and certain variables (e.g., achieving accuracy parity by learning a predictor independent of the sensitive attribute), our approach explicitly models group-class distributions through expert representations. This ensures separability in latent space while maintaining intra-group cohesion. We optimize this objective at three levels: group-wise, class-wise, and sample-wise, achieving a structured and hierarchical disentanglement of representations.

Specifically, we explicitly enforce a *group-wise* dependence between the representation $f(X)$ and the sensitive attribute A, i.e., by maximizing $P(A|f(X))$. This is achieved by introducing a discriminator $D : \mathbb{R}^m \to [0, 1]^{|\mathcal{A}|}$ that predicts A given the representation $f(X)$. We minimize $\mathcal{L}_{\text{disc}}$ to **link the representation $f(X)$ with its sensitive attribute** A, as defined below.

$$\mathcal{L}_{\text{disc}} = -\sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \mathbb{I}(a_i = a) \log D(f(x_i))_a \tag{6}$$

With the enforced group-wise dependence between representations and sensitive attributes, we then learn separable demographic representations that effectively diversify the representations across different groups. However, directly computing the representation distributions in the latent space is challenging. To address this, we introduce *virtual centers* $V_{a,y}$ at class-wise level as proxies for the representation distributions. For each class y and each sensitive group a, a virtual center is learned to represent the mean of the representation distribution. The similarity between each $V_{a,y}$ and the representation $f(x_i)$ of a sample from its respective class and sensitive group is measured by:

$$d\left(V_{a,y}, f(x_i)\right) = \frac{V_{a,y} \cdot f(x_i)}{\|V_{a,y}\| \|f(x_i)\|} \tag{7}$$

By mutually aligning these similarities, we ensure that the virtual centers accurately reflect the central tendencies of their respective representation distributions and draw the related samples closer to these centers in the latent space. This alignment is achieved by maximizing bi-directional similarity, which is equal to minimizing $\mathcal{L}_{\text{virt}}$:

$$\mathcal{L}_{\text{virt}} = -\sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \log \left(\frac{\exp(d(V_{a,y}, f(x_i)))}{\sum_{y'\in\mathcal{Y}} \exp(d(V_{a,y'}, f(x_i)))} \right) \tag{8}$$

where $\mathcal{Y}$ denotes all possible classes, respectively. By incorporating these virtual centers and aligning the similarities, we achieve a dual objective: (i) the virtual centers accurately represent the distribution centers, and (ii) the samples are naturally drawn toward these centers. As we associate representations with sensitive attributes, we can effectively learn the distribution for each class-group pair.

To ensure distinct representations across different groups, we aim for these distributions to exhibit distinct means. Beyond mean divergence, variations in variance also influence the separability of the distributions. Specifically, if the distributions have low variance and distinct means, they will be more easily distinguishable in latent space. Formally, given

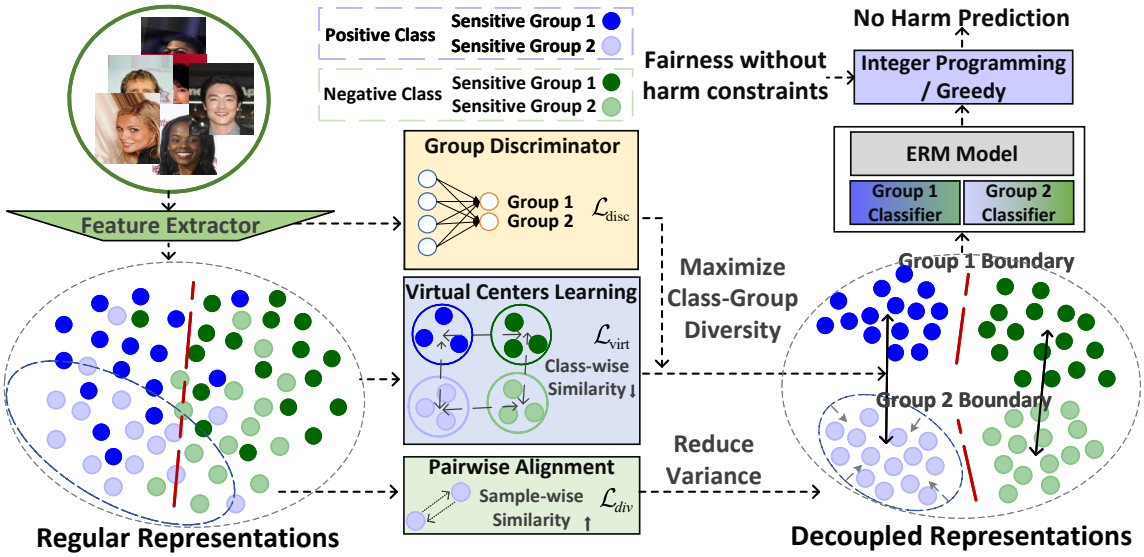


Figure 2: The illustration of FairSDE. Decoupled Representations aims to create distinct representations for diverse sensitive groups by ensuring inner-group similarity and inter-group and inter-class diversity. In particular, we estimate the mean of the representation distribution using virtual centers and mutually align them to the samples from a specific class and group to ensure the representations of different groups are well-separated in the latent space. Additionally, by minimizing the pairwise distances within each group, we achieve a more compact group representation, leading to reduced variance.

the estimated center $V_{a,y}$ for each group and class, our goal is to maximize the distance between these centers while minimizing the number of samples located near the boundaries of the distributions. For each sample x_i, we penalize its similarity to centers from other classes or groups, thereby enhancing the diversity among the estimated centers. Furthermore, we introduce pairwise representation alignment to reduce the variance of the representation distributions. Specifically, we randomly select a sample x_i' from the same class y_i and group a_i, and a sample x_i^- from a different class or group, optimizing both pairwise representation alignment and center diversity simultaneously:

$$\mathcal{L}_{\text{div}} = -\sum_{i=1}^{n} \log \frac{\exp\left(f(x_i)\cdot f(x_i')\right) + \exp\left(d\left(V_{a_i,y_i}, f(x_i)\right)\right)}{\exp\left(f(x_i)\cdot f(x_i^-)\right) + \sum_{a\neq a_i, y\neq y_i} \exp\left(d\left(V_{a,y}, f(x_i)\right)\right)} \tag{9}$$

Through random pairwise alignment and similarity penalization, we can better **diversify and compact the representation distribution**, particularly by pushing samples with high variance toward the center, which effectively reduces the variance of the representation distribution. This process results in separable representations that are more distinct and well-defined for each group and class.

4.2 Experts Dynamic Selection

With separable representations, we then train decoupled group-specific classifiers $\{h_a\}_{a\in\mathcal{A}}$ using samples from their respective groups to build the demographic experts. Each classifier is trained using the standard cross-entropy loss. For each group, we form the expert by coupling the group-specific representation with the corresponding classifier h_a. However, using each expert independently can be problematic, particularly for groups with limited data, as these classifiers may overfit or lack sufficient information to generalize effectively, which could result in worse performance or amplified bias. To address these challenges and ensure fairness without harm, we introduce a heuristic *No-Harm Selection* to dynamically select either the experts or the pooled model. Following Dutt et al. (2024), we assume that the validation and test sets share a similar distribution. We evaluate our models on the validation set and adopt two strategies to satisfy max-min fairness or overall accuracy parity fairness.

Max-min fairness selection. As the goal is to maximize the worst performance across all groups, this can be achieved using a *greedy strategy*. Specifically, for each group $a \in \mathcal{A}$, we compare our model $h_a \circ f$ with the ERM baseline h_{erm} and adopt the better-performing model.

Overall accuracy parity fairness selection. To maximize overall performance while minimizing the performance gap among all groups without imposing harm, we treat this as a constrained combinatorial optimization problem that selects the optimal combination of models to achieve a smaller fairness gap without compromising overall performance. Given that the number of sensitive attributes is relatively small in fairness studies, it is feasible to find the exact solution using integer programming. Specifically, for each group $a \in \mathcal{A}$, define a binary decision variable $v_a \in \{0,1\}$, where $v_a = 1$ if we choose our model $h_a \circ f$ for group a, and $v_a = 0$ if we choose h_{erm}. Let $\alpha_{\text{expert},a}$ and $\alpha_{\text{erm},a}$ be the model accuracy of group a attained under expert $h_a \circ f$ and h_{erm}, respectively, then the model accuracy of group a, denoted as α_a, can be equivalently written as:

$$\alpha_a = v_a \cdot \alpha_{\text{expert},a} + (1 - v_a) \cdot \alpha_{\text{erm},a} \tag{10}$$

Let p_a be the proportion of group a in the entire population. We have the following constrained optimization:

$$\begin{aligned} \min \quad & \Delta - \lambda \sum_{a \in \mathcal{A}} p_a \cdot \alpha_a \\ \text{subject to:} \quad & \alpha_a \geq \alpha_{\text{erm},a} \quad \forall a \in \mathcal{A} \quad \text{(No-harm)} \\ & \Delta \geq \alpha_i - \alpha_j \quad \forall i, j \in \mathcal{A} \quad \text{(Fairness)} \\ & v_a \in \{0, 1\} \quad \forall a \in \mathcal{A} \end{aligned} \tag{11}$$

Here Δ represents the maximum difference between accuracies across groups, and $-\lambda \sum_a p_a \cdot \alpha_a$ directs the optimization to find a solution with higher accuracy.

Feasibility Analysis. In the worst case, considering the trivial solution where we choose the ERM model for all groups $\alpha_a = \alpha_{\text{erm},a}$, the accuracy constraint $\alpha_a \geq \alpha_{\text{erm},a}$ is trivially satisfied for all groups. The fairness constraint reduces to: $\Delta \geq \alpha_{\text{erm},i} - \alpha_{\text{erm},j}, \forall i, j \in \mathcal{A}$. Let Δ_{erm} represent the maximum accuracy difference between any two groups under the ERM model, setting $\Delta = \Delta_{\text{erm}}$ satisfies this constraint. Therefore, the trivial solution is feasible.

Consider a more general case where we select our models $h_a \circ f$ for some groups and the baseline h_{erm} for others. The accuracy constraint guarantees that the accuracy of our model is at least as good as h_{erm} for each group, while the fairness constraint minimizes the maximum accuracy difference across all groups. Although the trivial solution, where all groups use the ERM model, is feasible, we aim to explore better alternatives that satisfy both constraints and optimize the overall objective. This allows us to potentially achieve better fairness and performance across all groups.

5 Experiments

Experiments are run on a server with NVIDIA A5000 GPUs and AMD EPYC 7313 CPUs. ResNet-18 (He et al. 2016) serves as the backbone f, with separate linear classifiers h for each group. Implementation details are in Appendix B.

5.1 Datasets and Baselines

We focus on medical datasets, where access to sensitive attributes is typically feasible and required by our method. Specifically, we evaluate FairSDE on three diagnostic tasks: skin disease classification with **Ham10000** (Maron et al. 2019), chest X-ray interpretation with **MIMIC-CXR** (Johnson et al. 2019), and glaucoma detection using **Harvard-GF** (Luo et al. 2024). To provide a broader comparison, we also include two facial image datasets: **CelebA** (Liu et al. 2015) and **UTKFace** (Zhang, Song, and Qi 2017). Dataset details are provided in Appendix A. As medical datasets often exhibit class imbalance, we report the AUC for disease prediction tasks and accuracy for facial recognition tasks.

We compare against the following baselines: 1) **ERM**: Serves as a no-harm baseline; methods must outperform ERM on each group to claim fairness without harm. 2) **Decoupled Classifiers**: Assigns each group a classifier $h(X, A)$ trained with a shared feature extractor, adapted from Ustun, Liu, and Parkes (2019). 3) **Adversarial Training** (Zhang, Lemoine, and Mitchell 2018): Employs an adversary to predict the sensitive attribute, which the model tries to obscure. 4) **CFL** (Shen et al. 2021): Uses contrastive learning to align representations of samples with similar characteristics across groups. 5) **FSCL** (Park et al. 2022): Encourages class-wise compactness while removing sensitive information from representations. 6) **FIN** (Luo et al. 2024): Uses group-specific feature normalization with learnable statistics to balance feature importance. 7) **GroupDRO** (Sagawa et al. 2020): Applies group distributional robustness to mitigate spurious correlations. 8) **FIS** (Pang et al. 2025): Selects training samples based on utility and fairness influence to improve fairness without harm.

5.2 Results

The results for the medical and facial datasets are shown in Tables 1 and 2. Experiments are repeated three times, and we report the average values. Additional results are provided in the Appendix, including: **equalized odds**, **standard deviations**, and **ablation studies on each module**. We find that **FairSDE is the only method that consistently achieves fairness without harm**. Based on the results, we address the following research questions:

Why Do We Need Fairness Without Harm? As shown in Table 1, methods like *Adversarial* and *FSCL+* significantly reduce AUC gap between male and female groups, suggesting improved fairness over baselines like *ERM*. However, this fairness gain comes at a cost: *FSCL+* reduces overall AUC by 1.89% and lowers performance for the disadvantaged group by 1.02%. A similar pattern emerges in glaucoma assessment on Harvard-GF data, where fairness methods compromise diagnostic accuracy. In healthcare, sacrificing accuracy for fairness is problematic and may have harmful consequences such as misdiagnosis or delayed treatment. Therefore, the concept of *fairness without harm* is crucial in real world. It emphasizes the need to develop models that not only promote fairness across groups but also maintain high levels of accuracy and reliability for patients.

Comparison and Discussion. We conducted a comparative analysis of model performance across medical and face datasets, using a down arrow symbol (↓) to indicate performance degradation relative to ERM. Note that **FairSDE is not designed to significantly outperform other methods in accuracy or fairness alone, but rather to ensure that fairness is consistently achieved without sacrificing performance**. Its expert selection mechanism may deliberately choose a model with slightly lower accuracy if it better satisfies the fairness without harm criterion. Among the baselines, *FSCL+* performed well on face datasets, improving both fairness and accuracy over ERM. However, this improvement did not generalize to medical datasets, where *FSCL+* yielded lower AUC than ERM. Other fairness methods showed similar performance degradation. **In contrast, FairSDE is the only method that consistently achieves fairness without harm, outperforming ERM across both AUC and ACC metrics without reducing performance.** This consistency highlights the strength of FairSDE, particularly in tasks like medical diagnosis, where it achieves fairness while simultaneously improving accuracy through the learned distinct representations and personalized classifiers.

Dataset	Sensitive	Metric	ERM	Adversarial	CFL	FSCL+	FIN	GroupDRO	FIS	Decoupled	FairSDE
Ham10000	Gender	AUC	84.07	83.70 ↓	83.67 ↓	82.18 ↓	83.33 ↓	84.45	84.55	83.60 ↓	**84.75**
		MF	82.78	82.93	82.70 ↓	81.76 ↓	82.06 ↓	83.42	82.99	82.20 ↓	**83.48**
		Gap	2.76	1.67	2.52	**0.75**	2.93	2.17	3.34	3.06	2.65
	Age	AUC	84.10	82.85 ↓	84.30	81.18 ↓	83.25 ↓	82.92 ↓	84.30	83.91 ↓	**84.62**
		MF	75.13	72.88 ↓	74.56 ↓	70.67 ↓	73.78 ↓	74.48 ↓	75.17	73.73 ↓	**76.03**
		Gap	13.05	15.26	14.37	15.82	13.30	12.23	13.34	14.68	**12.17**
Mimic-CXR	Gender	AUC	82.28	79.98 ↓	82.62	**82.67**	82.22 ↓	82.09 ↓	82.07 ↓	82.27 ↓	82.60 / 82.39
		MF	81.53	79.42 ↓	81.81	81.71	81.54	81.43 ↓	81.33 ↓	81.62	**81.92**
		Gap	1.54	1.34	1.62	1.94	1.36	1.33	1.52	1.41	**1.33**
	Race	AUC	82.28	81.56 ↓	82.50	**82.66**	82.23 ↓	81.81 ↓	82.23 ↓	82.32	82.44
		MF	81.70	81.27 ↓	82.06	82.16	81.58 ↓	81.39 ↓	81.71	81.89	**82.17**
		Gap	0.44	0.20	0.30	0.45	0.47	0.25	0.31	**0.14**	0.26
Harvard-GF	Gender	AUC	81.60	82.36	82.48	79.95 ↓	81.67	81.58 ↓	83.56	**84.44**	82.31 / 81.98
		MF	80.54	81.77	81.63	79.50 ↓	80.77	80.66	82.68	**84.21**	81.20
		Gap	2.36	1.42	2.03	1.08	2.05	2.15	2.01	**0.42**	1.81
	Race	AUC	82.59	82.54 ↓	82.61	79.34 ↓	82.23 ↓	82.63	83.42	**84.81**	82.83
		MF	78.13	78.25	77.99 ↓	74.46 ↓	76.52 ↓	78.83	78.80	**80.44**	78.61
		Gap	6.58	6.66	6.99	7.61	8.01	6.28	6.97	7.17	**6.12**
Fairness w/o Harm	**Utility**	Win | Loss	-	1|5	4|2	2|4	1|5	2|4	2|4	3|3	**6|0**
	Fairness	Win | Loss	-	9|3	9|3	8|4	8|4	9|3	11|1	10/2	**12|0**

Table 1: Classification results on three medical datasets. The AUC metric is used to quantify classification performance. MF denotes max-min fairness, representing the lowest AUC across all groups. Overall accuracy parity is evaluated using the AUC metric, with the AUC gap between the most advantaged and disadvantaged groups reported as "Gap" in the table. When the *greedy strategy* and *Integer Programming* yield different solutions, both overall AUC values are reported for comparison.

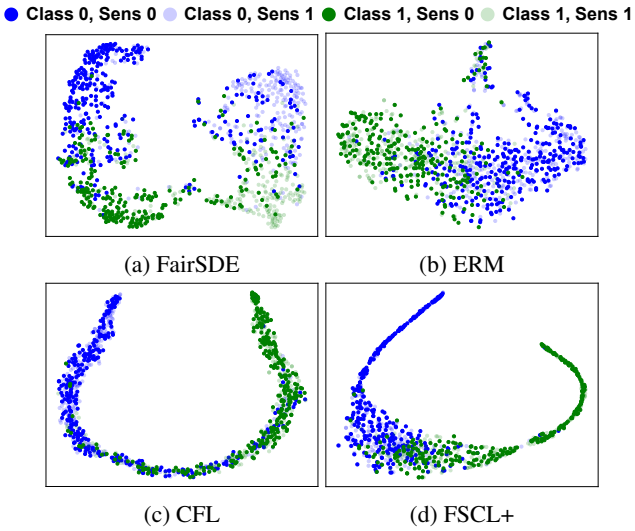
Dataset	Sensitive	Metric	ERM	Adversarial	CFL	FSCL+	FIN	GroupDRO	FIS	Decoupled	FairSDE
CelebA-Hair	Gender	ACC	81.28	79.67 ↓	81.49	81.69	81.13 ↓	79.63 ↓	80.14 ↓	81.40	**82.54**
		MF	77.60	75.71 ↓	77.72	78.35	77.28 ↓	78.35	76.64 ↓	77.70	**79.16**
		Gap	6.32	6.80	6.48	5.73	6.60	**2.20**↓	6.01	6.35	5.80
CelebA-Smiling	Gender	ACC	91.87	90.32 ↓	91.88	91.76 ↓	91.73 ↓	91.50 ↓	91.14 ↓	91.82 ↓	**92.22**
		MF	90.68	89.19 ↓	90.64 ↓	90.72	90.53 ↓	90.98	89.76 ↓	90.58 ↓	**91.21**
		Gap	2.04	1.92	2.12	1.77	2.06	**0.88**	2.36	2.12	1.71
UTK	Gender	ACC	78.06	73.40 ↓	77.89 ↓	79.97	73.93 ↓	77.84 ↓	76.32 ↓	75.50 ↓	**80.45**
		MF	76.52	71.95 ↓	76.21 ↓	78.62	70.89 ↓	76.83	74.86 ↓	72.58 ↓	**79.52**
		Gap	3.29	3.27	3.60	2.87	6.49	2.15	3.11	6.23	**1.98**
	Race	ACC	81.65	78.39 ↓	82.84	83.62	78.86 ↓	82.83	81.61 ↓	81.39 ↓	**84.17**
		MF	80.04	77.10 ↓	81.29	82.08	77.80 ↓	81.62	80.02 ↓	80.52	**82.75**
		Gap	3.81	3.04	3.68	3.67	2.51	2.86	3.77	**2.06**	3.37
Fairness w/o Harm	**Utility**	Win | Loss	-	0|4	3|1	3|1	0|4	1|3	0|4	1|3	**4|0**
	Fairness	Win | Loss	-	4|4	6|2	8|0	4|4	7|1	4|4	2|6	**8|0**

Table 2: Classification results on face datasets. The same reporting criteria as used in Table 1 are applied. For face datasets, accuracy (ACC) is used as the performance metric instead of AUC. Standard deviations are reported in Appendix D.

However, as in many studies, we assume that distributions are identical between validation and test sets. This assumption is critical, as our no-harm selection strategy may fail to perform as intended under a significant distribution shift.

Quantitative Results of Existing Algorithms on Fairness Without Harm. Next, we examine whether existing *fairness without harm* algorithms can effectively handle more challenging datasets. For all datasets, we apply decoupled classifiers $h(X, A)$ based on sensitive attributes. Unlike the original study by Ustun, Liu, and Parkes (2019), we do not use preference guarantees to select attributes; instead, we evaluate decoupled classifiers independently for each sensitive attribute, using representations learned from the ERM model. As shown in Table 1, decoupled classifiers perform well on datasets with balanced distributions of sensitive attributes, such as Harvard-GF, which has an equal number of samples across three racial groups and nearly equal representation of male and female groups. However, on more challenging datasets with imbalanced group distributions, these classifiers struggle to capture group-specific characteristics. For instance, on Ham10000, decoupled classifiers fail to outperform ERM on both gender and age attributes, leading to a trivial solution that defaults to ERM for all sensitive groups. This degraded performance highlights that trivially decoupled classifiers $h(X, A)$ are insufficient to achieve fairness without harm. We also evaluate another fairness-without-harm method, FIS, which uses active learning to selectively sample training data based on its fairness influence on the validation set. However, instead of enforcing no-harm constraint relative to the ERM model, FIS compares fairness

improvements against previous training epochs. As a result, it may still underperform ERM in certain cases.



(a) FairSDE (b) ERM

(c) CFL (d) FSCL+

Figure 3: t-SNE visualizations of representations on the UTKFace. The two classes are represented in blue and green, while the *white* group is depicted using darker circles and the *non-white* group using lighter circles.

How Do Decoupled Representations Perform on Images? Figure 3 presents t-SNE visualizations of the representations learned by FairSDE and other methods on UTKFace data, which includes two target classes (male and female) and two sensitive groups (white and non-white). As shown in Figure 3a, FairSDE achieves a clear separation between classes and sensitive groups by learning from virtual centers and maintaining distinct group-wise mean representations. This enables the model to better capture group-specific characteristics, leading to fairer and more accurate predictions. In contrast, Figure 3b shows that ERM produces more entangled representations, with significant overlap across groups and classes. Similarly, the other two methods tend to learn representations that are less sensitive to group distinctions and rely on a pooled classifier, as opposed to FairSDE.

From the MF results in Tables 1 and 2, we also observe that using demographic experts alone consistently improves the accuracy for the worst-performing group compared to the pooled classifier (ERM). This suggests that decoupling both representations and classifiers more effectively captures the unique patterns within disadvantaged groups. In contrast, applying decoupled classifiers without adapting the representations can degrade performance for disadvantaged groups, as evidenced by the poorer MF results of *Decoupled* on the Ham10000. This highlights a key strength of FairSDE: decoupled representations are crucial for achieving fairness without compromising performance. By better modeling group-specific characteristics, FairSDE delivers fair and accurate predictions across all demographic groups.

No-Harm Selection. Using Harvard-GF as an example, Table 3 presents no-harm selection results, highlighting the distinct advantages of the *greedy (GS)* and *integer programming (IP)* strategies in improving fairness and overall AUC. The *greedy* strategy aims to maximize the minimum group performance to ensure max-min fairness, as reflected by the increased average AUC for the worst-performing group (81.2) compared to ERM baseline (80.54). *GS* often leads to better overall performance by selecting the optimal model for each group. In contrast, the *IP* strategy seeks to minimize the performance gap between advantaged and disadvantaged groups by balancing selections between ERM and FairSDE. While this may slightly reduce overall performance, it effectively reduces disparities. For example, *IP* frequently selects ERM model for Group 2 to avoid worsening performance gaps, recognizing that improving this group's performance could cause unfairness. Table 3 shows that *IP* reduces the performance gap to 1.81, with no group underperforming relative to ERM, owing to the no-harm constraint in its optimization. This confirms that both strategies successfully fulfill our goal of fairness without harm.

T	Sel.	Group 1 AUC		Group 2 AUC		AUC	MF	Gap
		ERM	Ours	ERM	Ours			
1	**ERM**	81.19	-	83.80	-	82.28	81.19	2.60
	Ours	-	81.45	-	83.66	82.49	81.45	2.21
	GS	Ours: 81.45		Ours: 83.66		82.49	81.45	2.21
	IP	Ours: 81.45		ERM: 83.80		82.56	81.45	2.35
2	**ERM**	79.98	-	82.90	-	81.32	79.98	2.92
	Ours	-	80.38	-	83.83	81.99	80.38	3.45
	GS	ERM: 79.98		Ours: 83.83		81.72	79.98	3.85
	IP	ERM: 79.98		ERM: 82.90		81.32	79.98	2.92
3	**ERM**	80.45	-	82.02	-	81.21	80.45	1.57
	Ours	-	82.17	-	83.13	82.71	82.17	0.95
	GS	Ours: 82.17		Ours: 83.13		82.71	82.17	0.95
	IP	Ours: 82.17		ERM: 82.02		82.06	82.02	0.15
A	**ERM**	80.54	-	82.91	-	81.60	80.54	2.36
	GS	81.20		83.54		82.31	81.20	2.34
	IP	81.20		82.91		81.98	81.15	1.81

Table 3: Selection examples report the group AUC on Harvard-GF, along with overall AUC, max-min fairness, and AUC gap over three trials (T) and averaged values (A). We show selection results from the *greedy (GS)*, which maximizes max-min fairness, and *integer programming (IP)*, which targets overall accuracy parity. Note that selection is based on the validation set. As a result, in Trial 2, the *greedy strategy* selected ERM for Group 1, even though FairSDE achieved better performance for this group on the test set.

6 Conclusion

This paper studied the fairness without harm problem and showed that existing methods often improve fairness at the cost of group-specific performance. To address this, we proposed FairSDE, which learns distinct demographic representations and employs personalized experts to ensure fairness without degrading any group's performance. By selectively applying these experts, we ensure fairness constraints are met without compromising any group's performance. Experiments on multiple real datasets validate our approach. Despite potential concerns, using sensitive attributes is practical in medical contexts where they are routinely collected.

Acknowledgements

This work was funded in part by the National Science Foundation under award number IIS-2145625, IIS-2202699, and IIS-2416895, and by the National Institutes of Health under awards number R01AI188576.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.

Beauchamp, T. L.; and Childress, J. F. 1994. *Principles of biomedical ethics*. Edicoes Loyola.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: an extensible toolkit for detecting. *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

Cai, Z.; Khalili, M. M.; and Zhang, X. 2025. Demographic-Agnostic Fairness without Harm. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 486–497.

Celis, L. E.; Keswani, V.; and Vishnoi, N. 2020. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, 1349–1359. PMLR.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Dutt, R.; Bohdal, O.; Tsaftaris, S. A.; and Hospedales, T. 2024. Fairtune: Optimizing parameter efficient fine tuning for fairness in medical image analysis. In *International Conference on Learning Representations*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2357–2365. PMLR.

Han, X.; Baldwin, T.; and Cohn, T. 2021a. Decoupling adversarial training for fair NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 471–477.

Han, X.; Baldwin, T.; and Cohn, T. 2021b. Diverse Adversaries for Mitigating Bias in Training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2760–2765.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2021. Fair sequential selection using supervised learning models. *Advances in Neural Information Processing Systems*, 34: 28144–28155.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2023. Loss balancing for fair supervised learning. In *International Conference on Machine Learning*, 16271–16290. PMLR.

Khalili, M. M.; Zhang, X.; Abroshan, M.; and Sojoudi, S. 2021. Improving fairness and privacy in selection problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8092–8100.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740.

Lee, J.; Kim, E.; Lee, J.; Lee, J.; and Choo, J. 2021. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34: 25123–25133.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Luo, Y.; Tian, Y.; Shi, M.; Pasquale, L. R.; Shen, L. Q.; Zebardast, N.; Elze, T.; and Wang, M. 2024. Harvard Glaucoma Fairness: A Retinal Nerve Disease Dataset for Fairness Learning and Fair Identity Normalization. *IEEE Transactions on Medical Imaging*, 43(7): 2623–2633.

Maron, R. C.; Weichenthal, M.; Utikal, J. S.; Hekler, A.; Berking, C.; Hauschild, A.; Enk, A. H.; Haferkamp, S.; Klode, J.; Schadendorf, D.; et al. 2019. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer*, 119: 57–65.

Martinez, N.; and Bertran, M. 2019. Fairness With minimal harm: A Pareto-optimal approach for healthcare. *NeurIPS ML4H: Machine Learning for Health*.

Pang, J.; Wang, J.; Zhu, Z.; Yao, Y.; Qian, C.; and Liu, Y. 2025. Fairness without harm: An influence-guided active sampling approach. *Advances in Neural Information Processing Systems*, 37: 61513–61548.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2403–2411.

Park, S.; Lee, J.; Lee, P.; Hwang, S.; Kim, D.; and Byun, H. 2022. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10389–10398.

Pham, T.-H.; Zhang, X.; and Zhang, P. 2023. Fairness and Accuracy under Domain Generalization. In *The Eleventh International Conference on Learning Representations*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

Shen, A.; Han, X.; Cohn, T.; Baldwin, T.; and Frermann, L. 2021. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382. PMLR.

Yin, T.; Ton, J.-F.; Guo, R.; Yao, Y.; Liu, M.; and Liu, Y. 2024. Fair Classifiers that Abstain without Harm. In *The Twelfth International Conference on Learning Representations*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017b. From parity to preference-based notions of fairness in classification. *Advances in neural information processing systems*, 30.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, X.; Khaliligarekani, M.; Tekin, C.; and Liu, M. 2019. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in neural information processing systems*, 32.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.

Zuo, Z.; Khalili, M.; and Zhang, X. 2023. Counterfactually fair representation. *Advances in neural information processing systems*, 36: 12124–12140.

Zuo, Z.; Xie, T.; Tan, X.; Zhang, X.; and Khalili, M. M. 2024. Lookahead Counterfactual Fairness. *Transactions on Machine Learning Research*.