Medical Image Analysis 109 (2026) 103950

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Rethinking fairness in medical imaging: Maximizing group-specific performance with application to skin disease diagnosis

Gelei Xu [a,*], Yuying Duan [a], Jun Xia [a], Ching-Hao Chiu [a], Michael Lemmon [a], Wei Jin [b], Yiyu Shi [a,*]

[a] University of Notre Dame, Notre Dame, IN, 46556, USA
[b] Emory University, Atlanta, GA, 30322, USA

ARTICLE INFO

Keywords:
Fairness
Skin disease
Group-specific model

ABSTRACT

Recent efforts in medical image computing have focused on improving fairness by balancing it with accuracy within a single, unified model. However, this often creates a trade-off: gains for underrepresented groups can come at the expense of reduced accuracy for groups that were previously well-served. In high-stakes clinical contexts, even minor drops in accuracy can lead to serious consequences, making such trade-offs highly contentious. Rather than accepting this compromise, we reframe the fairness objective in this paper as maximizing diagnostic accuracy for each patient group by leveraging additional computational resources to train group-specific models. To achieve this goal, we introduce SPARE, a novel data reweighting algorithm designed to optimize performance for a given group. SPARE evaluates the value of each training sample using two key factors: utility, which reflects the sample's contribution to refining the model's decision boundary, and group similarity, which captures its relevance to the target group. By assigning greater weight to samples that score highly on both metrics, SPARE rebalances the training process-particularly leveraging the value of out-of-group data-to improve group-specific accuracy while avoiding the traditional fairness-accuracy trade-off. Experiments on two skin disease datasets demonstrate that SPARE significantly improves group-specific performance while maintaining comparable fairness metrics, highlighting its promise as a more practical fairness paradigm for improving clinical reliability.

1. Introduction

Machine learning-based medical diagnosis systems have become increasingly prevalent. During the diagnosing process, these systems typically employ a "one-size-fits-all" approach, using models trained on data from diverse populations to maximize overall accuracy. However, due to substantial differences in disease prevalence and manifestations across patient groups, such generalized models typically achieve satisfactory performance only for well-represented populations, while underperforming for others. For instance, in dermatology, individuals with lighter skin have lower melanin levels (Caini et al., 2009), making them more susceptible to melanoma. Consequently, they are overrepresented in training datasets and are prioritized during the model's learning process, leading to better performance for lighter skin types compared to dark skin types. This discriminatory behavior can have serious societal consequences, including misdiagnoses or delayed treatments for certain groups, ultimately exacerbating existing healthcare disparities. To mitigate such biases, traditional fairness-aware algorithms (Aayushman et al., 2024; Chiu et al., 2024; Zhang et al., 2022; Zong et al., 2022; Mehrabi et al., 2021; Puyol-Antón et al., 2021) aim to reduce the accuracy gap between groups. This conventional approach, however, often improves accuracy for underrepresented groups at the expense of reducing performance for those originally well-served, embodying the well-known fairness-accuracy trade-off (Rodolfa et al., 2021).

In clinical settings, where even minor losses in accuracy can lead to severe consequences (Chen et al., 2018), sacrificing precision for the sake of fairness is simply not an option. Unlike non-critical domains-such as online search engines (Alfiana et al., 2023) or content recommendation systems (Jesse and Jannach, 2021)-where slight degradations in performance may be acceptable, in healthcare they translate to missed or incorrect diagnoses that endanger patient safety. More critically, these reductions often occur near the decision boundary-precisely where clinical cases are most ambiguous and clinicians are most likely to make errors (Yuan et al., 2021). In such high-uncertainty scenarios, even a slight drop in accuracy can disproportionately increase the risk of misdiagnosis, amplifying clinical harm. At the same time, it is crucial to recognize that improving outcomes for underrepresented groups does not inherently require sacrificing performance for well-served ones.

* Corresponding authors.
E-mail addresses: gxu4@nd.edu (G. Xu), yshi4@nd.edu (Y. Shi).

https://doi.org/10.1016/j.media.2026.103950
Received 19 May 2025; Received in revised form 8 November 2025; Accepted 14 January 2026
Available online 16 January 2026

Fairness algorithms are often framed as redistributing performance, but in practice, drops in accuracy for well-served groups are neither necessary nor causally tied to gains for others. In fact, once improvements are achieved for disadvantaged groups, it is both feasible and desirable to retain the original model for groups already performing well. Therefore, we propose rethinking fairness in healthcare by shifting its definition: rather than merely closing the gap between groups while keeping average accuracy, it should be achieved by enabling each group to reach its maximum possible accuracy.

Central to our approach is a novel trade-off triangle involving fairness, group-specific accuracy, and computing resources. While every model inherently grapples with these three factors, previous methods have largely confined their efforts to balancing group accuracy and fairness within a single model, implicitly assuming fixed computational resources and tolerable performance drops. In contrast, in this paper, we explicitly leverage additional computing power to train group-specific models, bypassing the traditional fairness-accuracy trade-off. This computational investment aims to maximize diagnostic precision for each group, a requirement indispensable for clinical reliability in high-stakes applications. While many domains accept modest accuracy losses, in healthcare the imperative to optimize performance across all groups justifies the investment in additional computational capacity.

This paper focuses on developing methods to maximize performance for a target group and train dedicated models for each group. This group-specific approach enables more precise modeling and leads to improved performance across all subpopulations. While tailoring models to individual patient groups holds great promise, a central challenge remains: how to identify the most effective training data for each target group. The most straightforward approach is to train exclusively on data from the target group, but this often leads to suboptimal performance due to data scarcity, limiting the model's ability to capture robust patterns. Conversely, leveraging the entire dataset may introduce distributional shifts that obscure the unique characteristics of the target group. This situation presents an inherent trade-off: while out-of-group data can enhance generalizability, it may simultaneously compromise the model's ability to learn group-specific features. This trade-off is empirically shown in Section 2, and theoretically shown in Section 4.2. However, selectively incorporating a subset of out-of-group samples remains challenging without a reliable metric to assess each sample's contribution. Therefore, a data-driven approach is crucial for effectively integrating out-of-group data, optimizing model performance for each target group, and ensuring the highest possible diagnostic precision across diverse populations.

To address these challenges, we propose SPARE (**S**ubgroup **P**erformance-**A**ware **R**eweighting m**E**thod)-a unified sample reweighting framework that aims to maximize subgroup performance by intelligently selecting and weighting out-of-group data. A core difficulty in this task lies in balancing two conflicting needs: identifying samples that contribute to improving the target group's model while avoiding those that introduce harmful distributional shifts. Instead of handling these two objectives independently, SPARE provides an elegant, unified solution by framing both utility and distribution similarity through a shared perspective: the distance between a sample and its relevant decision boundaries. Specifically, samples closer to the boundary of the diagnostic classifier are more informative for model refinement, while those near the boundary of the group label predictor better align with the target group's distribution. By integrating these dimensions into a unified scoring function, SPARE prioritizes samples that are both valuable and distributionally compatible, selectively incorporating beneficial out-of-group data while mitigating the risks of distributional shift. This scoring-based approach enables principled integration of out-of-group data to optimize subgroup performance for group-specific fmodels.

To validate the effectiveness of our approach, we focus on one of the most extensively studied tasks in medical fairness: skin disease diagnosis (Ansari et al., 2024; Chiu et al., 2023). Experiments were conducted on two widely used dermatology datasets, where SPARE was evaluated against the current state-of-the-art fairness methods. These baseline methods fall into two primary categories: those aiming to minimize performance disparities between groups, and those employing group-specific models or modules. Experimental results show that, rather than compromising overall model performance to reduce disparities, SPARE substantially improves performance for both groups. For instance, on the Fitzpatrick-17k dataset, SPARE achieves a 3.7% improvement for dark skin types and a 4.0% improvement for light skin types over state-of-the-art methods. Furthermore, although SPARE does not explicitly optimize for gap reduction, it nevertheless reduces performance disparities across groups. On classical fairness metrics used to measure such disparities, SPARE exceeds or performs comparably to the best existing methods. This outcome indicates that the process of training group-specific models under SPARE confers greater performance gains to underrepresented groups. A plausible explanation is that conventional training tends to allocate model capacity disproportionately toward privileged groups, whereas SPARE rebalances this allocation during group-specific training. Overall, these results demonstrate that, relative to conventional fairness algorithms, SPARE provides a more practical and effective solution for improving fairness in clinical applications. By enhancing performance across all groups without compromising diagnostic accuracy, SPARE contributes to a promising direction for advancing equitable medical AI systems.

The main contributions of this paper are as follows:

- We present a practice-oriented view that frames fairness as maximizing performance within each subgroup. Guided by this view, we investigate the subgroup-specific data selection problem, where using out-of-group data can improve generalizability but may dilute group-specific features.
- We develop SPARE, a sample-wise reweighting method that quantifies each sample's value through two factors-utility and similarity-to balance generalizability and group-specific representational fidelity during the training process.
- Extensive experimental results demonstrate that our approach substantially improves performance across all subgroups, while matching or exceeding state-of-the-art methods on fairness metrics across multiple skin disease diagnosis datasets.

2. Empirical analysis of group-specific training

In this section, we empirically demonstrate the importance of selecting appropriate training data to achieve optimal performance for specific demographic groups. We adopt the Fitzpatrick-17k dataset for a skin disease classification task (Groh et al., 2021), where skin types T1-T3 are grouped as light skin and T4-T6 as dark skin. The dataset contains 16,577 images representing 114 different skin conditions, and is split into train/validation/test sets with a ratio of 6:2:2. We use the same preprocessing and training settings as in Section 5, where full experimental details are provided. In our experiments, we divide the dataset into several subsets based on Fitzpatrick skin types and evaluate the performance of different combinations of training subsets on both light and dark skin test groups. The results are presented in Fig. 1, where the x-axis denotes the specific combinations of training subsets, and the y-axis shows the testing accuracy for the light skin group (lighter line) and the dark skin group (darker line). From this figure, we observe the following key findings:

For the light-skin subgroup, performance does not peak when trained on the full dataset. Instead, using only skin types T1-T4 yields the best results, achieving a 2.1% improvement over training on the full dataset, as confirmed by McNemar's test on paired predictions ($p < .05$). This result underscores the benefits of group-specific training, particularly for lower-performing groups in a single-model setting. Second, adding dark skin samples initially enhances light skin accuracy; however, as the proportion of dark skin samples increases further, the performance on light skin declines. While out-of-group data can enhance generalization,

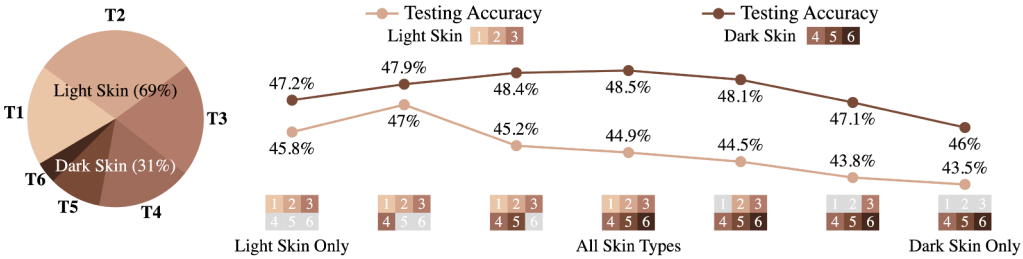


Fig. 1. Left: Distribution of data volume across skin types in the Fitzpatrick-17k dataset. Right: Model performance trained with different skin type subsets and evaluated separately on light and dark skin. Adding dark skin samples initially enhances light skin accuracy; however, as the proportion of dark skin samples increases further, the performance on light skin declines.

it can also introduce distributional shifts that negatively impact performance. This highlights the importance of quantifying both the benefits and potential risks of incorporating out-of-group data to optimize group-specific performance. Third, for the originally well-performed dark-skin type, training exclusively on dark-skin data yields lower accuracy than training exclusively on light-skin data. A likely explanation is that the light-skin subset is substantially larger in size, so the benefit of its greater data volume outweighs the distribution shift it introduces, ultimately providing a more reliable signal for the dark-skin model.

These empirical findings highlight the need for data-driven approaches that optimize performance for individual subgroups. In particular, the results suggest that tailoring models to specific demographic groups can be especially beneficial for those that are originally underrepresented in the training data. This may be attributed to the limited influence of underrepresented subgroups on the shared feature space in generalized models, which often fail to capture their specific characteristics. In contrast, subgroup-specific models can dedicate their full representational capacity to the target group, resulting in improved performance, particularly for minority populations (Afrose et al., 2022).

3. Related works

3.1. Fairness falgorithms in healthcare

Most fairness methods in healthcare focus on reducing disparities across demographic groups by minimizing differences in performance metrics such as true positive rates and false positive rates. This is commonly formalized through criteria like *Equalized Opportunity* and *Equalized Odds* (Hardt et al., 2016), which aim to ensure similar outcomes across sensitive attributes for patients with the same ground truth label.

Existing fairness approaches are commonly categorized into three groups: pre-processing, in-processing, and post-processing methods. Pre-processing techniques aim to achieve fairness by modifying the training data before model development. For example, Xu et al. (2018), Ngxande et al. (2020), Lu et al. (2020) applies specific data transformations to remove discriminatory patterns, while Kamiran and Calders (2012) assigns varying weights to individual samples to suppress the influence of sensitive attributes. In-processing methods intervene during model training to balance multiple objectives, typically aiming to jointly optimize for accuracy and fairness. A widely adopted strategy in this category is adversarial training, where an auxiliary adversary network attempts to predict sensitive attributes from learned representations, while the main model is trained to minimize the adversary's success, thereby reducing the encoding of sensitive information (Alvi et al., 2018; Zhang et al., 2018; Wang et al., 2022). Another line of work focuses on regularization-based methods, which penalize correlations between sensitive attributes and the model's output to encourage fairness (Jung et al., 2021; Quadrianto et al., 2019). For instance, Gretton et al. (2012) learns fair representations by distilling the fair information from a teacher model into a student model using the Maximum Mean Discrepancy loss. More recently, techniques such as pruning and quantization have been explored to reduce bias by removing model components that disproportionately contribute to disparities across sensitive groups (Chiu et al., 2023; Guo et al., 2024). Post-processing methods operate after the model has been trained, adjusting its outputs to enhance fairness. A typical approach involves threshold adjustment, where different prediction thresholds are applied to different sensitive groups to satisfy fairness criteria (Hardt et al., 2016; Valera et al., 2018). In addition, Du et al. (2020) improves fairness by calibrating the model's output distribution to align with a specified fairness metric, using both the model's raw predictions and sensitive attribute information as inputs.

These fairness-aware methods inherently face limitations at the Pareto frontier, where it becomes infeasible to simultaneously improve the performance of all groups (Dehdashtian et al., 2024). Consequently, such methods often leave at least one group in a suboptimal state. More concerningly, fairness constraints can lead to performance degradation across all groups in some cases, undermining the overall utility of the model (Wu et al., 2022; Duan et al., 2025). A notable example lies in the common practice of suppressing sensitive attributes to mitigate bias. While this strategy may appear effective in improving fairness metrics, it neglects the critical role these attributes play in clinical decision-making. Attributes like skin type, race, and gender are not merely confounders but often inform diagnoses. For instance, skin type provides crucial information for assessing UV susceptibility (Caini et al., 2009; Narayanan et al., 2010), and disease prevalence varies by race and gender (Narayanan et al., 2010; Gordon, 2013). In high-stakes clinical environments, compromising diagnostic accuracy for fairness can often be impractical and potentially hazardous. Therefore, fairness and accuracy should be treated not as trade-offs but as complementary goals to ensure reliable and effective medical care.

3.2. Fairness through subgroup performance maximization

Beyond gap-reduction approaches, several studies have pursued fairness by directly maximizing performance within each subgroup. In medical imaging, Puyol-Antón et al. (2021) ftrained independent models for each demographic group using only in-group data. More recently, Zhang et al. (2022) introduced *Stratified ERM* for chest X-rays, which partitions data by subgroup and learns distinct empirical risk minimizers, while the MEDFAIR benchmark (Zong et al., 2022) formalized a similar perspective under the notion of *domain independence* across multiple imaging modalities. These approaches share a common goal of enhancing subgroup-specific accuracy rather than minimizing disparities across groups.

Similar ideas appear outside the medical imaging domain, where fairness has been more broadly conceptualized in terms of subgroup-specific performance. For instance, Wang et al. (2020) propose group-specific classifiers with shared parameters that are optimized separately for each subgroup, while Dwork et al. (2018) design a decoupled classification framework where distinct classifiers are trained for different groups. In addition, Mehrabi et al. (2021) provide a comprehensive review of fairness methods, explicitly highlighting strategies that

emphasize improving subgroup-level performance. Collectively, these works reinforce the perspective that fairness need not always be equated with reducing inter-group disparities, but can instead be framed as ensuring that each subgroup reaches its maximum achievable accuracy.

Beyond fairness-specific literature, adjacent paradigms also resonate with this perspective. Multi-task learning (Evgeniou and Pontil, 2004; Agiza et al., 2024) and personalized federated learning (Luo and Wu, 2022; Tan et al., 2022) both aim to balance knowledge transfer across groups or domains with the need for subgroup-specific adaptation. While these methods are not explicitly designed for fairness, their principle of combining global generalization with local specialization aligns closely with our objective of maximizing subgroup performance.

While prior approaches also aim to achieve fairness without compromising accuracy by improving subgroup-specific performance, they typically regard each group as an isolated entity and train models using only in-group data. In contrast, our method extends this line of work by exploring the utility of out-of-group samples, showing that when incorporated strategically, such data can further enhance subgroup performance. We validate this insight empirically and include representative subgroup-specific methods as baselines for comparison. It is also worth noting that, although optimizing subgroup performance does not inherently guarantee improvements in traditional gap-based fairness metrics discussed in Section 3.1, in practice we often observe such metrics improving as a byproduct, likely because lower-performing groups tend to achieve larger relative gains when subgroup-specific performance is maximized.

3.3. Bridging fairness, domain shift and out-of-distribution generalization

A central cause of fairness issues in machine learning is the uneven distribution of data across demographic groups. When certain groups are underrepresented or exhibit unique feature-label relationships, models trained on aggregate data often generalize poorly to these groups. This performance gap is closely related to challenges studied in out-of-distribution (OOD) generalization and domain adaptation, where models fail under distribution shifts between training and deployment conditions (Quiñonero-Candela et al., 2022; Sun and Saenko, 2016).

In OOD generalization, the goal is to learn representations that remain robust across diverse domains, minimizing reliance on spurious correlations often introduced by majority-group patterns (Sagawa et al., 2019). To address this, researchers have proposed techniques such as minimizing worst-case loss (Sagawa et al., 2019), pruning biased samples (Jain et al., 2024), and learning invariant representations (Arjovsky et al., 2019). These strategies are conceptually aligned with fairness methods that aim to reduce performance disparities by enforcing group-invariant features. However, such approaches often discard group-specific characteristics, limiting their ability to achieve optimal performance within each individual distribution.

In contrast, domain adaptation focuses not on learning shared features across all domains, but on improving performance for a particular domain by transferring knowledge from related distributions (Ben-David et al., 2010). Adaptation methods include feature alignment (Ganin et al., 2016), domain-invariant representation learning (Cortes et al., 2019), and weighted empirical risk minimization (Zhang et al., 2012; Bu et al., 2022). These methods are more aligned with our goal of group-specific optimization, where the objective is to maximize performance for a specific subgroup. However, our approach departs from traditional domain adaptation in two key ways. First, instead of treating each group as a monolithic domain, we evaluate cross-group samples at a finer, sample-wise granularity to assess their utility for improving target group performance. Second, we introduce a principled mechanism that jointly considers both utility and distributional similarity when selecting which out-of-group samples to incorporate. This allows us to move beyond rigid domain boundaries and adaptively leverage the most beneficial examples, regardless of origin.

Recent studies have begun exploring fairness generalization across domains and distributions, aiming to preserve fairness established in a source environment when deploying a model under distribution shifts in a new target environment. For example, Pham et al. (2023) seek to maintain both fairness and accuracy in domain generalization settings, while Liang et al. (2023) and Stan and Rostami (2024) investigate domain adaptation techniques to safeguard fairness in the face of distributional change. While such work on fairness robustness is valuable, our study focuses on a more fundamental challenge at the source. Rather than seeking to preserve fairness only after an external domain shift, we use the theoretical tools of domain adaptation to address the initial fairness disparities arising from subgroup differences. In our view, a "domain" need not be limited to environmental changes encountered after deployment; instead, each demographic group can be framed as its own domain. This reframing allows fairness to be pursued through a domain-aware perspective. To the best of our knowledge, we present the first systematic formulation of this objective, introducing a framework that links fairness considerations with data-driven strategies for optimizing performance across subgroups.

4. Methodology

4.1. Problem formulation

Consider a dataset $D = \{(x_i, (y_i, c_i))\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ represents an input sample, and each sample is associated with a pair of labels (y_i, c_i). Here, $y_i \in \mathcal{Y} = \{1, 2, \ldots, k\}$ denotes the class label, while $c_i \in C = \{0, 1\}$ represents a binary group label (e.g., gender, race). In this paper we focus on binary group labels for clarity, though our approach extends easily to multiple groups with minor modifications. We partition the dataset into two subsets, D_0 and D_1, corresponding to groups 0 and 1, respectively. Our approach builds upon a basic ensemble model that dynamically selects the group-specific classifier based on the input. This ensemble consists of two group-specific models, $f_0 : \mathcal{X} \to \mathcal{Y}$ and $f_1 : \mathcal{X} \to \mathcal{Y}$, trained exclusively on D_0 and D_1, respectively. Additionally, a group label predictor $f_g : \mathcal{X} \to C$ determines which group-specific model to apply for a given input.

Our goal is to optimize the group-specific models f_0 and f_1, such that f_0 maximizes accuracy on D_0 and f_1 maximizes accuracy on D_1. To achieve this, we incorporate out-of-group data to improve each model. Without loss of generality, the following method description focuses on maximizing the performance of Group 0. That is, we treat Group0's dataset (D_0) as the primary set and Group1's dataset (D_1) as the auxiliary set, selecting relevant samples from D_1 to enhance f_0. The process is symmetric when optimizing f_1. In the following sections, we denote f_0 as f_c to emphasize its role as the classifier for Group 0. A direct approach to improving ff_c is to identify samples from D_1 that improve its performance on D_0 when incorporated. However, this selection process is NP-hard and inherently imposes a binary classification of sample importance. To address this problem, instead of selecting samples, this paper assigns each sample a weight that reflects its importance and can be anywhere between 0 and 1, i.e., $w \in [0, 1]$.

4.2. The value of additional data: Utility vs. similarity

To improve the group-specific model f_c trained on Group 0 data D_0, we examine how incorporating additional samples from the auxiliary group D_1 impacts a model trained for Group 0. Intuitively, more training data can reduce generalization error by decreasing variance from data fluctuations, generally improving performance (Hastie et al., 2009; Duda et al., 1973). However, the benefit of using samples from D_1 depends critically on its alignment with the distribution of D_0. While additional data can stabilize learning, incorporating samples from a mismatched distribution can introduce bias and degrade performance.

To guide the design of our sample weighting method, we formalize this trade-off and derive an upper bound on the generalization error

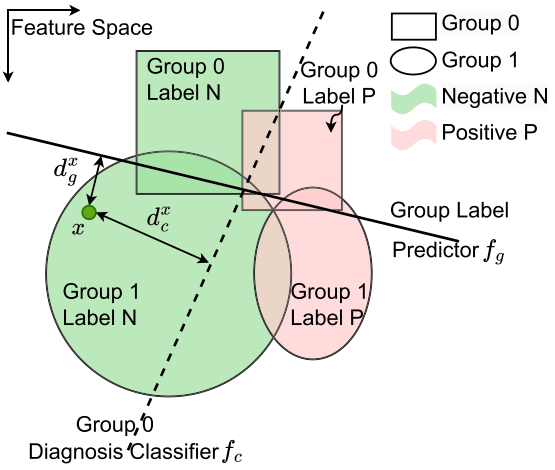


Fig. 2. The illustration of the distances from a sample to the decision boundaries of the diagnosis classifier and the group predictor. These distances can serve as proxies for utility and similarity, respectively.

of a model f_w trained on a weighted dataset $\mathcal{D}_w$, where each sample $x_i \in \mathcal{D}_0 \cup \mathcal{D}_1$ is assigned a weight $w_i \in [0, 1]$. Let $\Pr_{\mathcal{D}_w}(x_i) \propto w_i$ represent the empirical distribution induced by these weights. The bound on the excess error when evaluating on Group 0 is given by:

$$e_0(f_w) - e_0(f_0^*) \le \sum_x \underbrace{G_i}_{\text{Utility}} \cdot \underbrace{\left|\Pr_{\mathcal{D}_w}(x_i) - \Pr_{\mathcal{D}_0}(x_i)\right|}_{\text{Similarity}} + \underbrace{\sqrt{\frac{\log(4/\delta)}{2|\mathcal{D}_w|}}}_{\text{Empirical Sample Size}} + c, \tag{1}$$

where f_0^* is the ideal optimal model for Group 0, $G_i = \ell(f_w(x_i), y_i)$ is the empirical loss incurred on sample x_i, $\left|\Pr_{\mathcal{D}_w}(x_i) - \Pr_{\mathcal{D}_0}(x_i)\right|$ captures the degree of distributional mismatch between the sample-wise weighted distribution $\mathcal{D}_w$ and the Group 0 distribution $\mathcal{D}_0$, $|\mathcal{D}_w|$ is the effective sample size of the weighted dataset, and c is a constant. The detailed proof of Eq. (1) is in the Appendix A.1.

This bound exposes three competing factors that determine the benefit of the auxiliary data from $\mathcal{D}_1$: (i) utility G_i of each sample for training the model, (ii) similarity between the sample's distribution under $\mathcal{D}_w$ and that of $\mathcal{D}_0$, and (iii) the effective dataset size induced by the weighting. Accordingly, our goal is to learn sample weights that jointly optimize these dimensions-prioritizing samples that are both informative and distributionally aligned, while naturally benefiting from increased data volume to enhance the model performance on $\mathcal{D}_0$.

Guided by this theoretical insight, we develop a weighting mechanism that quantifies both utility and similarity, as described in the following section. We do not explicitly regularize for sample size, as our weighting function inherently balances informativeness and distributional alignment in a continuous manner. As a result, the effective size of the training set arises implicitly from the learned weight distribution. Introducing an additional constraint on sample count would be redundant and unlikely to offer further benefit over the selection process.

4.3. Quantifying utility and similarity

4.3.1. Distance to decision boundaries as a proxy

To effectively incorporate out-of-group data, it is essential to quantify both a sample's utility in improving classification and its similarity to Group 0. These two factors are illustrated in Fig. 2, where the feature space contains samples from both Group 0 and Group 1, each labeled as positive (P) or negative (N). The decision boundary of Group 0's diagnosis classifier f_c (dashed line) separates positive and negative cases, while the group label predictor f_g (solid line) distinguishes between Group 0 and Group 1 samples. A sample x's utility is determined by its distance to the decision boundary of f_c-the closer it is, the more likely it is to be misclassified, making it more valuable for refining Group0/s model. Similarly, the similarity is measured by f_g's classification. If f_g predicts x belongs to Group 0, it introduces no distribution shift. If it is instead classified as Group 1, its similarity increases as it gets closer to Group 0, indicating a smaller distribution gap.

Motivated by this observation, we define two key distances for a sample x: fd_c^x, the distance to the decision boundary of f_c, which quantifies utility as samples near the boundary provide greater value for model refinement, and d_g^x, the distance to the decision boundary of f_g, which represents distribution similarity—a smaller d_g^x indicates greater similarity to Group 0. If f_g classifies x as part of Group 0, we assume no distribution shift and set $d_g^x = 0$.

To balance these two distances, we define the combined distance as:

$$d(x) = \alpha d_c^x + (1 - \alpha) d_g^x. \tag{2}$$

Here, α controls the trade-off between utility and similarity. A smaller $d(x)$ indicates a higher weight to x, ensuring the framework prioritizes samples that are both informative and distributionally relevant.

4.3.2. Computing decision boundary distance via minimal perturbation

While the simplified illustration in Fig. 2 suggests that a perpendicular distance could be computed analytically, in practice this is difficult to obtain directly. In high-dimensional, non-linear classifiers such as deep neural networks, the decision boundary forms a highly complex, non-convex surface for which no closed-form representation exists (Goodfellow et al., 2014). To estimate how close a sample x is to a model's decision boundary, we draw inspiration from adversarial attack techniques (Carlini and Wagner, 2017), which are designed to find small perturbations that change a model's prediction. These techniques offer a principled way to quantify the local robustness of a model's prediction for a given input. We adopt this approach for two reasons. First, directly measuring the minimal perturbation required to flip a classification provides a concrete, model-sensitive estimate of how close x lies to the decision boundary. Compared to proxy measures such as confidence scores or margin values, this perturbation-based metric more accurately reflects the local geometry of the classifier's decision surface. Second, because our method ultimately involves weighting images based on their relevance and informativeness, it is crucial to ground these metrics in the model's actual behavior under input variations, rather than heuristic approximations.

As illustrated in Fig. 3, for a given input x, we compute its distance to a decision boundary by determining the smallest perturbation δ such that the model's prediction changes. Formally, this is defined as:

$$\min_{\delta} D(x, x + \delta) \quad \text{s.t.} \quad C(x + \delta) \neq C(x), \tag{3}$$

where $D(\cdot)$ is a distance function-specifically the L_2 norm-and $C(\cdot)$ is the model's hard classification function (either f_c for utility or f_g for similarity). A smaller perturbation norm $\|\delta\|_2$ indicates that the sample x is closer to the decision boundary, and thus either more informative for refining decision boundaries (in the case of f_c) or more similar to the target $\mathcal{D}_0$'s distribution (in the case of f_g).

However, solving this problem directly is challenging due to the non-linearity and discontinuity of the classification constraint. To make it tractable, we adopt a relaxed formulation based on soft labels. Specifically, we maximize the cross-entropy loss between the predicted softmax outputs before and after perturbation, encouraging a change in prediction without relying on a hard decision threshold. The relaxed optimization becomes:

$$\min_{\delta} \|\delta\|_2 - c \cdot \mathcal{L}(\hat{C}(x), \hat{C}(x + \delta)), \tag{4}$$

where $\hat{C}(\cdot)$ denotes the softmax output of the model and $\mathcal{L}$ is the cross-entropy loss between original and perturbed outputs. The constant c

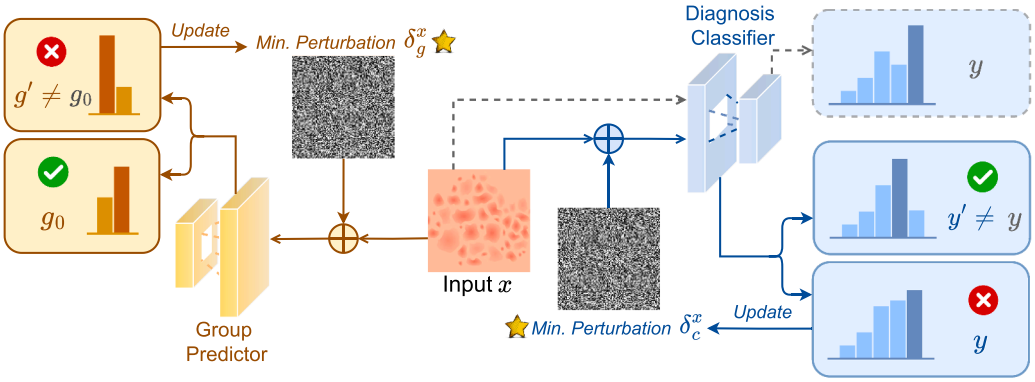


Fig. 3. The distance between a sample and a model's decision boundary is estimated using adversarial perturbations. For the group predictor, the minimal perturbation required to alter the prediction to group 0 is computed. For the diagnosis classifier, the predicted class y is first identified, and the smallest perturbation needed to change the output to any $y' \neq y$ is then computed.

Algorithm 1 Training algorithm for SPARE.

Input Full dataset $D = D_0 \cup D_1$; diagnosis model f_c; group model f_g
Output Optimized group-specific model f_c^*

Phase 1: Compute sample weights
for each sample $x \in D$ **do**
  Compute d_c^x using f_c and d_g^x using f_g (Section 4.3)
  Compute $d(x) = \alpha \cdot d_c^x + (1 - \alpha) \cdot d_g^x$
  Set $w_x = e^{-d(x)}$
end for
Phase 2: Train group-specific model
Train f_c^* on D using fixed weights $\{w_x\}$
return f_c^*

controls the balance between minimizing perturbation size and maximizing prediction change. We solve this optimization using an iterative gradient-based approach that refines δ, progressively minimizing the objective until a classification flip is achieved.

4.3.3. Final weighting function

The resulting perturbation norm $\|\delta\|_2$ serves as the distance metric for d_c^x and d_g^x. These distances are then combined (by Eq. (2)) to determine the weight assigned to x, ensuring that SPARE prioritizes samples that are both close to decision boundaries and distributionally aligned with D_0. Since the distribution of the calculated $d(x)$ is highly skewed, we ultimately use $w_x = e^{-d(x)}$ to map $d(x)$ inversely to a range between 0 and 1 to obtain the final weight of an image x. Algorithm 1 shows the training procedure for obtaining the group-specific model for Group 0.

5. Experiments and results

To evaluate the effectiveness of the proposed method, we conduct comprehensive experiments designed to answer the following research questions (RQs):

- **RQ1: General performance comparison.** How does the proposed method perform compared to state-of-the-art bias mitigation approaches across different backbone architectures?
- **RQ2: Weight distribution analysis.** How do the weight distributions differ across demographic groups in group-specific models trained using our method?
- **RQ3: Impact of weighting strategies.** Our method assigns individual sample weights, subsequently normalized to the range [0, 1] via an exponential mapping. To what extent do alternative weighting strategies influence model performance?
- **RQ4: Utility vs. similarity comparison.** How does the hyperparameter α, which balances utility and similarity in the weighting function, affect overall performance?
- **RQ5: Ablation study of the combined distance.** How do the utility and similarity components, individually and in alternative formulations, contribute to the overall effectiveness of the proposed combined distance?
- **RQ6: Resource-performance trade-off.** In scenarios with limited computational resources, where training separate models per group is infeasible, how does the use of partially shared models impact performance?

5.1. Datasets, training protocol and metrics

Dataset and Training Details. The proposed methods are evaluated on two skin disease classification datasets: the Fitzpatrick-17k dataset (Groh et al., 2021) and the ISIC 2019 challenge dataset (Combalia et al., 2019; Tschandl et al., 2018). The Fitzpatrick-17k dataset comprises 16,577 images representing 114 different skin conditions. We group skin types 1–3 as light skin and 4–6 as dark skin, following the same settings as in Section 2. ISIC 2019 dataset contains f25,331 images across 8 categories. While gender is frequently selected as the sensitive attribute in fairness-aware learning, we instead choose age due to its quasi-continuous nature, which facilitates finer subgroup partitioning and enables more nuanced downstream visualization. Accordingly, we divided the dataset into young and old groups for analysis. A standard preprocessing step for both datasets involves resizing all the images to a uniform size of 128×128 pixels. Various techniques such as random horizontal flipping, vertical flipping, rotation, scaling, and autoaugment are used to augment the data, consistent with (Cubuk et al., 2018). The dataset is split into train, validation, and test with a ratio of 6:2:2. Unless otherwise specified, we use ResNet-18 (He et al., 2016)

as the backbone. All models are trained for 200 epochs using the SGD optimizer, with a fixed batch size of 128 to ensure consistency across methods. For the learning rate, we performed a small validation-based search over $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and selected 10^{-3}, which yielded the best performance; this setting was applied uniformly to both SPARE and all baselines. For the proposed SPARE method, the trade-off parameter α in Eq. (2) is analyzed in detail in Section 5.2.4, with $\alpha = 0.5$ used as the default unless otherwise specified. The adversarial perturbation balancing parameter c is set to 100 following (Carlini and Wagner, 2017). We repeat this process three times and report the average to ensure consistency of the results. Since SPARE is designed to train separate models for each group, we report the performance of each group using its corresponding group-specific model in the results.

Metrics. Since the fairness objective of this paper is to improve diagnostic performance for every specific group, we adopt commonly used classification metrics-*precision*, *recall*, and *F1-score*-as the main criteria and report the group-wise result to assess model effectiveness. While our proposed approach does not explicitly minimize inter-group disparities in the traditional fairness optimization sense, we report widely used fairness metrics to enable comprehensive comparisons with prior state-of-the-art methods. Specifically, we evaluate fairness using multi-class versions of *Equalized Opportunity (Eopp)* and *Equalized Odds (Eodd)*, following the definitions and implementations in Wu et al. (2022). To assess the overall trade-off between fairness and predictive performance, we also report the *Fairness-Aware Trade-off Evaluation (FATE)* metric introduced by Xu et al. (2023), where higher *FATE* scores indicate a more favorable balance between accuracy and fairness. Specifically, the *FATE* metric is computed as:

$$FATE_{FC} = \frac{ACC_m - ACC_b}{ACC_b} - \lambda \frac{FC_m - FC_b}{FC_b} \tag{5}$$

Here, ACC denotes the model's predictive performance, for which we use the F1-score, and FC refers to the fairness criterion (e.g., Eopp or Eodd). The subscripts m and b represent the bias-mitigated and baseline models, respectively. The hyperparameter λ controls the relative weight of fairness in the overall evaluation; following (Xu et al., 2023), we set $\lambda = 1$ in all experiments.

5.2. Result and discussions

5.2.1. RQ1: Performance comparison with state-of-the-art

Baselines. We compare SPARE with various bias mitigation baselines. Vanilla refers to models trained directly on ResNet-18 without any fairness intervention. FairAdaBN adapts batch normalization layers to sensitive attributes (Xu et al., 2023). For this method, we followed the grid search range reported in the original paper ($\{0.1, 1.0, 2.0\}$) for its fairness-constraint parameter α and selected $\alpha = 1.0$. SCP-FairPrune (Kong et al., 2024) and FairQuantize (Guo et al., 2024) enhance fairness through pruning and quantization, respectively. Since our dataset and backbone settings match those used in their papers, we directly adopted the reported hyperparameters: for SCP-FairPrune, $\texttt{prc} = 2\%$ and $n = 3$; for FairQuantize, we used a quantization ratio of 80% with $\beta = 0.778$ on ISIC2019, and a ratio of 20% with $\beta = 1.0$ on Fitzpatrick-17k. We also evaluate fairness methods based on group-specific training. GroupModel (Puyol-Antón et al., 2021) trains a separate model per group using only its in-group data, while DomainIndep (Wang et al., 2020) learns group-specific classifiers with shared parameters. In addition, we include two methods originally developed for broader group/domain adaptation but conceptually aligned with subgroup performance maximization. Regularized Multi-Task Learning (MTL) (Evgeniou and Pontil, 2004) encourages related groups to share information through joint parameterization while still allowing group-specific specialization, making it a natural baseline in our setting. APPLE (Luo and Wu, 2022) learns group-specific Directed Relationship (DR) weights that determine how much each subgroup borrows from others, which parallels our goal of leveraging out-of-group samples to improve subgroup models.

Results on ISIC 2019 dataset. Table 1 reports the results on the ISIC 2019 dataset, demonstrating that SPARE consistently outperforms all baselines in accuracy for both the young and old groups. For instance, compared to the ResNet-18 backbone, SPARE achieves a 3.7% improvement in F1-score for the young group and a 3.8% improvement for the old group. Such gains are particularly valuable in high-stakes medical applications, where it is crucial to avoid compromising the performance of well-served groups or failing to capture the group-specific characteristics of underrepresented populations. By training group-specific models, SPARE addresses both concerns, enabling more equitable and effective representation across demographic groups.

Meanwhile, SPARE also demonstrates strong performance on fairness metrics. It ranks first in both Eopp0 and Eodd, and second in Eopp1. Compared to the baseline, SPARE reduces Eopp0 by 16.7%, Eopp1 by 26.5%, and Eodd by 30.5%. This suggests that even without explicitly incorporating bias mitigation constraints, SPARE effectively narrows the disparity between demographic groups. In particular, the group-specific approaches (GroupModel, DomainIndep, MTL, APPLE and SPARE) yield larger performance gains for the underrepresented younger subgroup compared to non-group-specific baselines FairAdaBN, SCP-FairPrune, and FairQuantize. For example, group-specific models improve the younger group's F1-score from 0.743 (vanilla ResNet-18) to 0.780 (SPARE) and 0.752 (DomainIndep), whereas some non-group-specific methods such as FairQuantize and DomainIndep fail to improve the younger subgroup and even lower its performance. This may be because global models trained to fit all groups tend to focus disproportionately on overrepresented groups, thereby suppressing the learning of group-varying features associated with minority groups. In contrast, training group-specific models enables better representation of each group's unique features, with underrepresented groups benefiting more substantially from this tailored optimization. Given its strong performance on both accuracy and fairness metrics, SPARE achieves substantially higher *FATE* scores compared to other baseline models. For instance, its *FATE* values computed using Eopp0, Eopp1, and Eodd surpass those of the second-best baselines by 36.5%, 11.7%, and 47.9%, respectively.

Notably, although GroupModel, DomainIndep, MTL, APPLE and SPARE all adopt group-specific training strategies and achieve relatively higher accuracy compared to other bias mitigation methods, SPARE stands out by attaining the best overall performance in both accuracy and fairness. These results suggest that achieving effective group-specific training is non-trivial, underscoring the unique advantage of SPARE's weighting mechanism, which integrates both sample-level similarity and utility to guide model learning.

Results on Fitzpatrick-17k dataset. Table 2 presents the results of our method applied to the ResNet-18 backbone on the Fitzpatrick-17k dataset. SPARE outperforms all baseline methods across all accuracy metrics. In terms of fairness, it achieves the best performance on Eopp1 and ranks second on both Eopp0 and Eodd. Consistent with the results observed on the ISIC 2019 dataset, these findings further support the effectiveness of training group-specific models in both narrowing inter-group performance disparities and maximizing per-group performance-an approach that is particularly well-suited for high-stakes medical applications. Furthermore, SPARE demonstrates the highest *FATE* scores by a considerable margin. Specifically, its *FATE* values based on Eopp0, Eopp1, and Eodd exceed those of the second-best methods by 25.6%, 36.1%, and 63.6%, respectively.

Comparison with state-of-the-art in different backbone. To further evaluate the generalizability of our approach, we replaced the backbone model with VGG-11 (Simonyan and Zisserman, 2014) and replicated the experiments described in Section 5.2.1. The results are presented in Table 3, which report performance on the ISIC 2019 and Fitzpatrick-17k datasets. Our method achieves the highest performance across all accuracy metrics on both datasets, while also maintaining competitive results

Table 1
Results of accuracy and fairness on ISIC 2019 dataset using ResNet-18 backbone.

		Accuracy			Fairness		
Method	Age	Precision	Recall	F1-score	Eopp0 ↓ / FATE ↑	Eopp1 ↓ / FATE ↑	Eodd ↓ / FATE ↑
ResNet-18	Young Old	0.718 0.764	0.786 0.765	0.743 0.758	0.018 / 0.000	0.102 / 0.000	0.558 / 0.000
GroupModel	Young Old	0.724 0.777	0.782 0.758	0.749 0.762	0.021 / −0.160	0.116 / −0.131	0.560 / 0.003
DomainIndep	Young Old	0.723 0.740	0.777 0.769	0.752 0.758	0.016 / 0.117	0.075 / 0.271	0.492 / 0.124
MTL	Young Old	0.745 0.767	0.756 0.766	0.747 0.766	0.018 / −0.013	0.082 / 0.207	0.555 / 0.014
APPLE	Young Old	0.743 0.773	0.769 0.764	0.749 0.764	0.016 / 0.118	0.075 / 0.277	0.539 / 0.044
FairAdaBN	Young Old	0.712 0.739	0.772 0.755	0.739 0.752	0.016 / 0.104	**0.073** / 0.278	0.458 / 0.173
SCP-FairPrune	Young Old	0.722 0.759	0.780 0.764	0.746 0.760	0.016 / 0.114	0.089 / 0.131	0.521 / 0.070
FairQuantize	Young Old	0.707 0.762	0.781 0.765	0.738 0.752	**0.015** / 0.159	0.088 / 0.130	0.420 / 0.240
SPARE	Young Old	**0.768** **0.809**	**0.803** **0.785**	**0.780** **0.796**	**0.015** / **0.217**	0.075 / **0.315**	**0.388** / **0.355**

Table 2
Results of accuracy and fairness on Fitzpatrick-17k dataset using ResNet-18 backbone.

		Accuracy			Fairness		
Method	Skin Tone	Precision	Recall	F1-score	Eopp0 ↓ / FATE ↑	Eopp1 ↓ / FATE ↑	Eodd ↓ / FATE ↑
ResNet-18	Dark Light	0.512 0.467	0.511 0.468	0.490 0.449	0.0031 / 0.000	0.332 / 0.000	0.180 / 0.000
GroupModel	Dark Light	0.511 0.475	0.512 0.472	0.492 0.453	0.0030 / 0.045	0.320 / 0.048	0.164 / 0.107
DomainIndep	Dark Light	0.501 0.465	0.522 0.479	0.492 0.459	0.0030 / 0.041	0.320 / 0.045	0.163 / 0.103
MTL	Dark Light	0.514 0.458	0.529 0.488	0.500 0.454	0.0030 / 0.051	0.302 / 0.110	0.164 / 0.108
APPLE	Dark Light	0.521 0.477	0.517 0.483	0.491 0.466	0.0030 / 0.047	0.312 / 0.075	0.171 / 0.064
FairAdaBN	Dark Light	0.493 0.450	0.495 0.443	0.469 0.435	0.0030 / 0.007	0.302 / 0.065	0.171 / 0.024
SCP-FairPrune	Dark Light	0.512 0.490	0.512 0.472	0.490 0.469	0.0030 / 0.049	**0.289** / 0.147	0.164 / 0.106
FairQuantize	Dark Light	0.498 0.459	0.513 0.470	0.480 0.448	**0.0028** / 0.082	0.291 / 0.109	**0.156** / 0.118
SPARE	Dark Light	**0.534** **0.508**	**0.542** **0.499**	**0.517** **0.488**	0.0030 / **0.103**	**0.289** / **0.200**	0.158 / **0.193**

in fairness metrics. Compared to the baseline VGG-11 model, SPARE yields average improvements across two datasets of 21.1%, 36.4%, and 29.5% in Eopp0, Eopp1, and Eodd, respectively. Additionally, it attains the highest *FATE* scores among all methods on both datasets. These findings underscore the robustness of our approach across different neural network architectures.

5.2.2. RQ2: Weight distribution analysis on group-specific models

Fig. 4a presents the weight distribution boxplots for data with different Fitzpatrick skin types from the Fitzpatrick-17k dataset, evaluated using two group-specific models: the light-skin model (left) and the dark-skin model (right). The results show that for in-group data, most weights remain high, typically exceeding 0.8. In contrast, for out-of-group data, weights tend to decrease as the distance from the group increases in terms of Fitzpatrick skin type. Furthermore, we observe that the light-skin model assigns relatively lower weights to dark-skin samples, whereas the dark-skin model tends to place slightly higher weights on light-skin data. A Mann–Whitney U test confirmed that this difference is statistically significant ($p < .001$), although the effect size is small (Cliff's $\delta \approx -0.04$). These nuanced but consistent asymmetries suggest that the light-skin model performance may be more dependent on in-group data, potentially due to a greater distributional mismatch between dark-skin samples and the light-skin subgroup. This interpretation aligns with our empirical findings in Section 2: despite having a larger training set, the light-skin group underperforms in the global model. This may be due to dark-skin samples being farther from the

Table 3
Results of accuracy and fairness on Fitzpatrick-17k dataset and ISIC 2019 dataset using VGG-11 backbone.

		Accuracy			Fairness		
Method	Group	Precision	Recall	F1-score	Eopp0 ↓ / FATE ↑	Eopp1 ↓ / FATE ↑	Eodd ↓ / FATE ↑
Fitzpatrick 17k Dataset							
VGG-11	Dark Light	0.493 0.435	0.490 0.447	0.476 0.422	0.0032 / 0.000	0.282 / 0.000	0.142 / 0.000
GroupModel	Dark Light	0.496 0.441	0.486 0.459	0.474 0.430	0.0029 / 0.101	0.273 / 0.040	0.136 / 0.049
DomainIndep	Dark Light	0.499 0.443	0.488 0.450	0.475 0.427	0.0030 / 0.068	0.276 / 0.028	0.143 / 0.003
MTL	Dark Light	0.493 0.459	0.488 0.449	0.473 0.439	0.0028 / 0.145	0.271 / 0.049	0.135 / 0.082
APPLE	Dark Light	0.497 0.457	0.482 0.452	0.473 0.440	0.0030 / 0.093	0.273 / 0.064	0.143 / 0.023
FairAdaBN	Dark Light	0.487 0.426	0.478 0.434	0.469 0.412	0.0030 / 0.044	0.277 / 0.001	0.138 / 0.012
SCP-FairPrune	Dark Light	0.494 0.451	0.496 0.454	0.478 0.438	0.0029 / 0.115	0.277 / 0.038	0.133 / 0.089
FairQuantize	Dark Light	0.482 0.422	0.474 0.427	0.461 0.402	**0.0028** / 0.086	0.268 / 0.013	**0.129** / 0.052
SPARE	Dark Light	**0.512** **0.474**	**0.517** **0.472**	**0.492** **0.469**	**0.0028** / **0.196**	**0.267** / **0.125**	0.130 / **0.160**
ISIC 2019 Dataset							
VGG-11	Young Old	0.669 0.758	0.724 0.798	0.687 0.766	0.023 / 0.000	0.150 / 0.000	0.078 / 0.000
GroupModel	Young Old	0.675 0.731	0.722 0.802	0.688 0.762	0.021 / 0.130	0.116 / 0.340	0.060 / 0.128
DomainIndep	Young Old	0.677 0.758	0.724 0.796	0.689 0.766	0.021 / 0.090	0.103 / 0.316	**0.051** / 0.345
MTL	Young Old	0.669 0.720	0.728 0.773	0.688 0.747	0.022 / 0.042	0.108 / 0.279	0.060 / 0.229
APPLE	Young Old	0.666 0.752	0.739 0.773	0.692 0.760	0.023 / 0.005	0.095 / 0.372	0.072 / 0.082
FairAdaBN	Young Old	0.643 0.727	0.717 0.770	0.663 0.748	**0.018** / 0.196	**0.084** / 0.418	0.060 / 0.209
SCP-FairPrune	Young Old	0.662 0.745	0.724 0.792	0.683 0.761	0.020 / 0.125	0.098 / 0.341	0.068 / 0.123
FairQuantize	Young Old	0.657 0.743	0.718 0.785	0.676 0.755	0.019 / 0.163	0.088 / 0.403	0.060 / 0.220
SPARE	Young Old	**0.713** **0.786**	**0.765** **0.814**	**0.724** **0.789**	**0.018** / **0.369**	0.090 / **0.565**	0.052 / **0.398**

group label predictor's decision boundary, which allows them to dominate the shared representation space and shift the model's focus toward dark-skin-specific features.

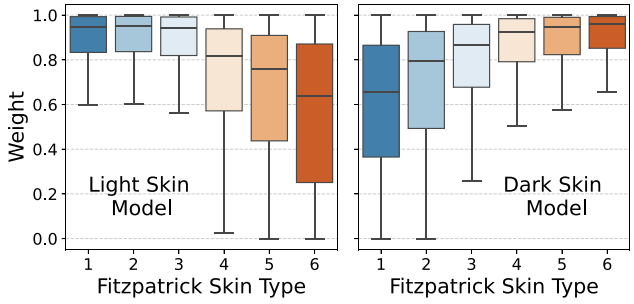
Similarly, Fig. 4b shows the weight distributions across age groups in the ISIC 2019 dataset for two group-specific models: the young model (left) and the old model (right). To enable clearer visualization, we divided the age range into six categories: 1 (0-15), 2 (16-30), 3 (31-45), 4 (46-60), 5 (61-75), and 6 (76-90). Compared to the more structured trends observed in Fitzpatrick-17k, the age-based results appear less regular. For example, in the old model, age group 2 (16-30) receives a weight similar to or slightly higher than group 3 (31-45). This may be attributed to the fact that, unlike skin tone, age-related changes are less visually distinct in skin images-particularly among individuals aged 16 to 45, where textural changes are subtle and difficult to detect visually. Overall, the observed trend-that each group-specific model assigns higher weights to in-group data and gradually decreases weights as the group difference increases-reflects the intended effect of our design, where weights are derived from distances to the corresponding group-specific decision boundary. In conjunction with the performance gains reported fin Section 5.2.1, these results suggest that the proposed group-specific weighting scheme effectively captures group distinctions and contributes to improved fairness.

5.2.3. RQ3: Impact of weighting strategies

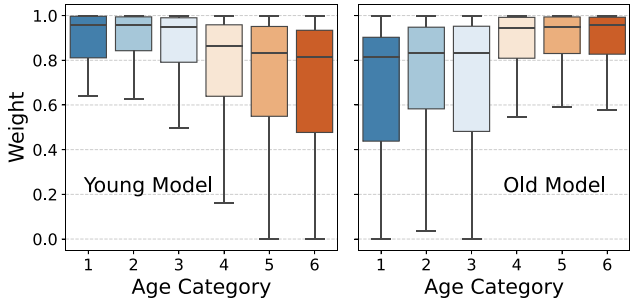
In this section, we examine alternative strategies for converting the combined distance $d(x)$ (defined in Section 4.3) into sample weights. Specifically, given a set of samples each associated with a distance score $d(x)$, the question is how to map these scores into weights. We compare our proposed continuous weighting in SPARE against several classical alternatives, as shown in Table 4. GroupWeight assigns a fixed weight to all samples within a group, without accounting for intra-group variation (Huang et al., 2016). Selection applies a binary threshold: samples with scores above the threshold receive a weight of 1, while others are assigned a weight of 0. Ranking sorts samples by their score and assigns weights based on their percentile rank (Roszkowska, 2013). Experimental results show that the weighting strategy used in SPARE

Table 4
Performance comparison across weighting strategies on Fitzpatrick-17k and ISIC 2019 datasets.

	Fitzpatrick-17k				ISIC 2019			
	Skin Tone	Precision	Recall	F1	Age	Precision	Recall	F1
GroupWeight	Light	0.483	0.477	0.457	Young	0.742	0.797	0.762
	Dark	0.514	0.515	0.495	Old	0.785	0.768	0.772
Selection	Light	0.498	0.485	0.476	Young	0.749	0.794	0.764
	Dark	0.529	0.538	0.512	Old	0.789	0.771	0.775
Ranking	Light	0.481	0.480	0.457	Young	0.731	0.792	0.754
	Dark	0.512	0.513	0.492	Old	0.779	0.756	0.764
SPARE	Light	0.508	0.499	0.488	Young	0.768	0.803	0.780
	Dark	0.534	0.542	0.517	Old	0.809	0.785	0.796



(a) Fitzpatrick-17k dataset



(b) ISIC 2019 dataset

Fig. 4. Sample weight distribution for the (a) light and dark skin models on Fitzpatrick-17k dataset and (b) young and old models on ISIC 2019 dataset.

outperforms all alternatives across both datasets. Selection ranks second, suggesting that binary filtering can yield reasonably good performance, though it lacks the granularity of continuous weighting. Both GroupWeight and Ranking perform less effectively. This may be attributed to GroupWeight's inability to capture within-group heterogeneity, and to Ranking's reliance on carefully tuned mappings between rank percentiles and assigned weights. Overall, these findings further support the effectiveness of SPARE's weighting mechanism. Its simple yet powerful design enables sample-level weighting based on both similarity and utility, providing a fine-grained way to capture the informativeness of individual data points across multiple dimensions.

5.2.4. RQ4: Utility vs. similarity comparison through different α values

Fig. 5 illustrates the performance of two group-specific models on the Fitzpatrick-17k and ISIC 2019 datasets under varying values of α, which controls the trade-off between similarity and utility in the data weighting function. Specifically, $\alpha = 0$ indicates that only similarity is considered, while $\alpha = 1$ means that only utility determines the weight. The results show that, on the Fitzpatrick-17k dataset, the light-skin model achieves optimal performance at $\alpha = 0.4$, while the dark-skin model peaks at $\alpha = 0.6$, suggesting that similarity plays a more dominant role in the light-skin model. For the ISIC 2019 dataset, both the young and old models perform best at $\alpha = 0.5$, indicating the value of balancing both factors.

Across both datasets, models trained using only similarity information ($\alpha = 0$) consistently outperform those using only utility ($\alpha = 1$). This highlights the central importance of similarity in guiding sample selection for group-specific modeling, as it directly captures the distributional alignment between samples and their target group. Utility, meanwhile, also contributes to model effectiveness, but serves more as a complementary signal modulating the relative influence of samples based on their estimated informativeness. Together, the two dimensions provide a flexible and principled basis for weighting data in group-specific model training.

5.2.5. RQ5: Ablation study of the combined distance

To further validate the design of the combined distance, we conduct an ablation study that systematically examines the contribution of its utility and similarity components. While the previous analysis varied the trade-off parameter α, it remained unclear whether both components are individually necessary and whether alternative formulations could provide comparable benefits. Ablation is therefore crucial to verify that our design is not only effective but also essential. We select three representative alternatives to compare against our definitions. For similarity, we consider feature centroid distance (FCD), which measures the distance between group feature means, and maximum mean discrepancy (MMD) (Yan et al., 2017), which captures higher-order distributional differences. For utility, we adopt the logit gap (LG) (Wani et al., 2024), a common measure of sample difficulty based on the margin between predicted class probabilities. These alternatives provide meaningful baselines to test the robustness of our design choices.

Table 5 presents the results of this study. When only utility or only similarity is used, performance drops across both datasets and groups, showing that neither component alone is sufficient. Replacing our similarity with FCD or MMD also leads to weaker results, as these definitions fail to align group distributions as effectively as our approach. Likewise, substituting our utility with logit gap produces inferior performance, indicating that our informativeness signal provides a stronger foundation. In contrast, the full method SPARE, which combines our definitions of both utility and similarity, consistently achieves the best balance of precision, recall, and F1 across groups. Together, these comparisons confirm that both components are indispensable and that the particular design choices in SPARE are critical to its effectiveness. The ablation results thus provide strong evidence for the necessity of the combined distance in enabling robust group-specific modeling.

5.2.6. RQ6: resource-performance trade-off

Our framework trains separate models for different demographic groups to better capture group-specific representations. While experimental results demonstrate that this strategy yields significant performance improvements, training fully independent models for each group may be impractical in scenarios with a large number of groups or constrained computational resources. Notably, deep neural networks often

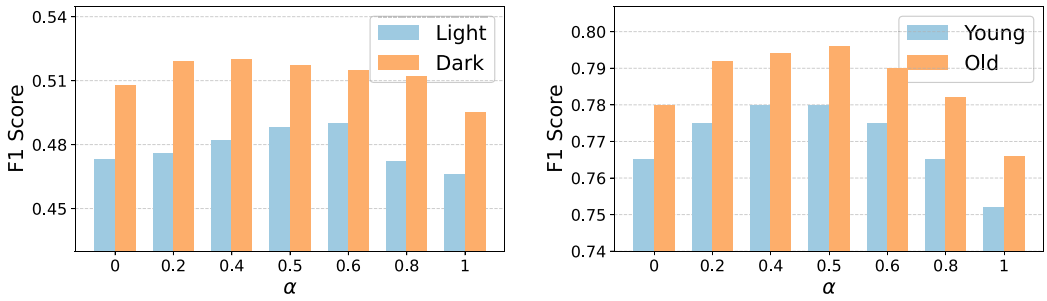


Fig. 5. Performance variation with different α values for the (a) light and dark skin models on Fitzpatrick-17k dataset and (b) young and old models on ISIC 2019 dataset.

Table 5
Ablation study of the combined distance on Fitzpatrick-17k and ISIC 2019 datasets.

	Fitzpatrick-17k				ISIC 2019			
	Skin Tone	Precision	Recall	F1	Age	Precision	Recall	F1
Utility Only	Light	0.479	0.483	0.466	Young	0.731	0.793	0.752
	Dark	0.520	0.522	0.495	Old	0.777	0.748	0.766
Similarity Only	Light	0.487	0.482	0.473	Young	0.730	0.798	0.765
	Dark	0.530	0.536	0.508	Old	0.781	0.776	0.780
Utility + FCD	Light	0.481	0.485	0.468	Young	0.741	0.796	0.759
	Dark	0.518	0.528	0.496	Old	0.783	0.765	0.772
Utility + MMD	Light	0.486	0.482	0.471	Young	0.745	0.798	0.762
	Dark	0.522	0.528	0.498	Old	0.780	0.783	0.776
Similarity + LG	Light	0.501	0.492	0.477	Young	0.764	0.797	0.771
	Dark	0.525	0.531	0.507	Old	0.807	0.763	0.781
SPARE	Light	0.508	0.499	0.488	Young	0.768	0.803	0.780
	Dark	0.534	0.542	0.517	Old	0.809	0.785	0.796

Table 6
Performance comparison across different sharing layers on Fitzpatrick-17k and ISIC 2019 datasets.

	Fitzpatrick-17k				ISIC 2019			
	Skin Tone	Precision	Recall	F1	Age	Precision	Recall	F1
Full Sharing	Light	0.467	0.468	0.449	Young	0.718	0.786	0.743
	Dark	0.512	0.511	0.490	Old	0.764	0.765	0.758
Main Sharing	Light	0.484	0.484	0.473	Young	0.738	0.788	0.756
	Dark	0.519	0.521	0.494	Old	0.779	0.773	0.774
Half Sharing	Light	0.497	0.487	0.479	Young	0.756	0.796	0.771
	Dark	0.534	0.542	0.517	Old	0.783	0.780	0.785
No Sharing	Light	0.508	0.499	0.488	Young	0.768	0.803	0.780
	Dark	0.534	0.542	0.517	Old	0.809	0.785	0.796

learn low-level, generic features (e.g., edges and textures) in the early layers. This observation raises a natural question: Is it necessary to train entirely separate models for each group, or can early layers be shared without substantially sacrificing performance?

To address this, we conducted additional experiments to explore the impact of sharing early network layers across groups. Using ResNet-18 as the backbone, we evaluated four configurations on two datasets:

(1) Full sharing: a fully shared model with no group-specific components;
(2) Main sharing: a model where only the final layer is group-specific;
(3) Half sharing: a partially specialized model in which approximately half of the layers are group-specific; and
(4) No sharing: fully group-specific models. The results, presented in Table 6, show that performance is lowest when all groups share the entire model. As more group-specific layers are introduced, performance consistently improves, reaching its highest point with fully separated models.

These findings highlight a trade-off exists between performance and computational efficiency. While fully specialized models offer the best performance, they require proportionally more resources. In resource-constrained environments, sharing early layers among groups provides a practical compromise, enabling competitive performance while significantly reducing the computational burden.

6. Discussion and conclusion

Ensuring fairness in medical AI remains a complex and actively debated challenge. Performance disparities across demographic groups are particularly concerning in clinical contexts, where diagnostic decisions have direct and potentially serious implications for patient outcomes. If left unresolved, these inequities may erode trust in AI-assisted diagnosis among both clinicians and patients. Existing fairness-aware

algorithms have made progress in addressing these gaps. However, most of these works rely on a shared, group-agnostic model and seek to equalize outcomes through implicit resource reallocation. This approach often improves performance for underrepresented groups at the expense of reducing accuracy for those already well-served. While this trade-off may be acceptable in low-risk domains, in healthcare-even small drops in accuracy can have serious clinical consequences. As a result, whether fairness should be achieved by compromising performance for any group remains an open and pressing question in high-stakes clinical applications.

In this work, rather than seeking fairness by trading off accuracy, we advocate for fairness through maximizing group-specific performance. While this approach may demand additional computational resources, we argue that such investment is justified in domains like healthcare, where precision and reliability are essential. To operationalize this idea, we propose SPARE-a sample reweighting algorithm that enhances group-specific model performance by selectively incorporating out-of-group training samples. SPARE estimates the utility of each candidate sample and its distributional similarity to the target group, balancing performance gain with robustness to distribution shift. Empirical results across multiple medical datasets demonstrate that SPARE significantly improves performance for target groups while preserving fairness metrics comparable to state-of-the-art baselines. These findings suggest that SPARE may serve as a practical complement to existing fairness interventions, especially in clinical applications where model reliability must extend across diverse patient populations.

While we advocate for subgroup-specific performance maximization as a more appropriate paradigm for achieving fairness in medical AI, this work also opens up several avenues for further exploration. One consideration lies in the complexity of demographic structures in real-world populations. In practice, demographic groups are rarely binary; instead, they consist of complex intersections-such as combinations of gender, race, and age-resulting in a potentially vast number of subgroups. Training a dedicated model for each subgroup is infeasible. A promising direction may lie in group clustering-identifying a small number of representative subgroups that capture the key variations across the population, and then applying group-specific optimization at this reduced granularity. Another opportunity for future research concerns scalability. Recent advances in parameter-efficient fine-tuning (Liu et al., 2023, 2024) suggest that full model retraining for each group may not be necessary. Instead, lightweight modules could offer a scalable way to tailor models while filtering harmful out-of-group samples or adapting representations selectively. These techniques may provide practical means to support subgroup performance without incurring prohibitive computational costs. Finally, it is worth noting that while our approach consistently improves both subgroup performance and fairness metrics in experiments, SPARE does not explicitly optimize fairness criteria such as Equal Opportunity or Equalized Odds, and therefore cannot guarantee improvements under these definitions. Encouragingly, we observe that fairness metrics often improve as a byproduct, likely because underrepresented groups benefit disproportionately when subgroup-specific performance is maximized. Future work could investigate the theoretical connection between performance maximization and fairness-gap reduction, potentially providing formal support for when and why such improvements occur.

We view SPARE as an initial step toward this more flexible approach to fairness-one that moves beyond uniformity and allows models to adapt to group-specific needs while maintaining clinical rigor. We hope this work encourages further research into practical fairness strategies that can more effectively support equitable outcomes across diverse patient populations.

CRediT authorship contribution statement

Gelei Xu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization; **Yuying Duan:** Writing – review & editing, Validation, Formal analysis, Conceptualization; **Jun Xia:** Writing – review & editing, Validation; **Ching-Hao Chiu:** Writing – review & editing, Methodology; **Michael Lemmon:** Writing – review & editing, Methodology, Conceptualization; **Wei Jin:** Writing – review & editing, Methodology, Conceptualization; **Yiyu Shi:** Writing – review & editing, Supervision, Resources, Methodology, Investigation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT 4o to improve the readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gelei Xu, Yuying Duan, Jun Xia, Ching-hao Chiu reports financial support was provided by University of Notre Dame. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This project is supported in part by the National Institutes of Health under Grant R01EB033387, and by the National Science Foundation under Grants CNS-2228092 and IIS-2437345.

Appendix A.

A.1. Proof for Eq. (1)

Proof: We have:

$$
\begin{aligned}
& e_0(f_w) - e_0(f^*) \\
=& (e_0(f_w) - e_w(f_w)) + (e_w(f_w) - e_0(f^*)) \\
=& \sum_{\mathcal{X}} \Pr_{D_0}(x = x_i) \cdot l(f_w(x_i), y_i) - \Pr_{D_w}(x = x_i) \cdot l(f_w(x_i), y_i) \\
& + (e_w(f_w) - e_0(f^*)) \\
\leq& \sum_{\mathcal{X}} l(f_w(x_i), y_i) \cdot |\Pr_{D_w}(x = x_i) - \Pr_{D_0}(x = x_i)| + (e_w(f_w) - e_0(f^*)) \\
=& \sum_{\mathcal{X}} l(f_w(x_i), y_i) \cdot |\Pr_{D_w}(x = x_i) - \Pr_{D_0}(x = x_i)| + (e_w(f_w) + e_w(f_w^*)) \\
& + (e_w(f_w^*) - e_0(f^*)) \\
\leq& \sum_{\mathcal{X}} \underbrace{G_i}_{\text{utility}} \underbrace{\cdot |\Pr_{D_w}(x_i) - \Pr_{D_0}(x_i)|}_{\text{similarity}} + \underbrace{\sqrt{\frac{\log(4/\delta)}{2|D_w|}}}_{\text{empirical data size}} + c
\end{aligned}
\tag{A.1}
$$

The first term, G_i, represents the empirical risk of sample x_i. $|\Pr_{D_w}(x = x_i) - \Pr_{D_0}(x = x_i)|$ quantifies the divergence between the initial distribution of group 0, D_0, and the distribution of the mixture samples, D_w. The second term is bounded with probability at least $(1 - \delta)$ by Hoeffding's inequality:

$$
e_w(f_w) + e_w(f_w^*) \leq \sqrt{\frac{\log(4/\delta)}{2|D_w|}},
$$

The last term is a constant c, as the optimal risk is the ground truth and independent of the sample selection.

References

Aayushman, Gaddey, H., Mittal, V., Chawla, M., Gupta, G.R., 2024. Fair and accurate skin disease image classification by alignment with clinical labels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 394–404.

Afrose, S., Song, W., Nemeroff, C.B., Lu, C., Yao, D., 2022. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. Commun. Med. 2 (1), 111.

Agiza, A., Neseem, M., Reda, S., 2024. MTLoRA: low-rank adaptation approach for efficient multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16196–16205.

Alfiana, F., Khofifah, N., Ramadhan, T., Septiani, N., Wahyuningsih, W., Azizah, N.N., Ramadhona, N., 2023. Apply the search engine optimization (seo) method to determine website ranking on search engines. Int. J. Cyber IT Serv. Manage. 3 (1), 65–73.

Alvi, M., Zisserman, A., Nellåker, C., 2018. Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0.

Ansari, F., Chakraborti, T., Das, S., 2024. Algorithmic fairness in lesion classification by mitigating class imbalance and skin tone bias. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 373–382.

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D., 2019. Invariant risk minimization. arXiv:1907.02893.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79, 151–175.

Bu, Y., Aminian, G., Toni, L., Wornell, G.W., Rodrigues, M., 2022. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 8673–8699.

Caini, S., Gandini, S., Sera, F., Raimondi, S., Fargnoli, M.C., Boniol, M., Armstrong, B.K., 2009. Meta-analysis of risk factors for cutaneous melanoma according to anatomical site and clinico-pathological variant. Eur. J. Cancer 45 (17), 3054–3063.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (sp). Ieee, pp. 39–57.

Chen, I., Johansson, F.D., Sontag, D., 2018. Why is my classifier discriminatory? Adv. Neural Inf. Process. Syst. 31,3543–3554.

Chiu, C.-H., Chen, Y.-J., Wu, Y., Shi, Y., Ho, T.-Y., 2024. Achieve fairness without demographics for dermatological disease diagnosis. Med. Image Anal. 95, 103188.

Chiu, C.-H., Chung, H.-W., Chen, Y.-J., Shi, Y., Ho, T.-Y., 2023. Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 97–107.

Combalia, M., Codella, N. C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. BCN20000: dermoscopic lesions in the wild. arXiv:1908.02288.

Cortes, C., Mohri, M., Medina, A.M., 2019. Adaptation based on generalized discrepancy. J. Mach. Learn. Res. 20 (1), 1–30.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2018. AutoAugment: learning augmentation policies from data. arXiv:1805.09501.

Dehdashtian, S., Sadeghi, B., Boddeti, V.N., 2024. Utility-fairness trade-offs and how to find them. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12037–12046.

Du, M., Yang, F., Zou, N., Hu, X., 2020. Fairness in deep learning: a computational perspective. IEEE Intell. Syst. 36 (4), 25–34.

Duan, Y., Xu, G., Shi, Y., Lemmon, M., 2025. The cost of local and global fairness in federated learning. Proc. 28th Int. Conf. Artif. Intell. Stat. 258, 4186–4194.

Duda, R.O., Hart, P.E., et al., 1973. Pattern Classification and Scene Analysis. Vol. 3. Wiley New York.

Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M., 2018. Decoupled classifiers for group-fair and efficient machine learning. In: Conference on Fairness, Accountability and Transparency. PMLR, pp. 119–133.

Evgeniou, T., Pontil, M., 2004. Regularized multi–task learning. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 109–117.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (59), 1–35.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Gordon, R., 2013. Skin cancer: an overview of epidemiology and risk factors. In: Seminars in Oncology Nursing. Vol. 29. Elsevier, pp. 160–169.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A Kernel two-sample test. J. Mach. Learn. Res. 13 (1), 723–773.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O., 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1820–1828.

Guo, Y., Jia, Z., Hu, J., Shi, Y., 2024. FairQuantize: achieving fairness through weight quantization for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 329–338.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. 29,3323–3331.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer, New York, NY, USA.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Huang, C., Li, Y., Loy, C.C., Tang, X., 2016. Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5375–5384.

Jain, S., Hamidieh, K., Georgiev, K., Ilyas, A., Ghassemi, M., Madry, A., 2024. Improving subgroup robustness via data selection. Adv. Neural Inf. Process. Syst. 37, 94490–94511.

Jesse, M., Jannach, D., 2021. Digital nudging with recommender systems: survey and future directions. Comput. Hum. Behav. Rep. 3, 100052.

Jung, S., Lee, D., Park, T., Moon, T., 2021. Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12115–12124.

Kamiran, F., Calders, T., 2012. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33 (1), 1–33.

Kong, Q., Chiu, C.-H., Zeng, D., Chen, Y.-J., Ho, T.-Y., Hu, J., Shi, Y., 2024. Achieving fairness through channel pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 24–34.

Liang, Y., Chen, C., Tian, T., Shu, K., 2023. Fair classification via domain adaptation: a dual adversarial learning approach. Front. Big Data 5, 1049565.

Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., Zheng, Y., 2023. MOELoRA: an MOE-based parameter efficient fine-tuning method for multi-task medical applications. arXiv:2310.18339.

Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., Zheng, Y., 2024. When MOE meets LLMs: parameter efficient fine-tuning for multi-task medical applications. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, Inc., pp. 1104–1114.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., Datta, A., 2020. Gender bias in neural natural language processing. In: Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday, V. Nigam, T. Ban Kirigin, C. Talcott, J. Guttman, S. Kuznetsov, B. T. Loo, M. Okada (eds.), Springer International Publishing, Cham, Switzerland, pp. 189–202. https://doi.org/10.1007/978-3-030-62077-6_14.

Luo, J., Wu, S., 2022. Adapt to adaptation: learning personalization for cross-silo federated learning. In: IJCAI: Proceedings of the Conference. Vol. 2022, p. 2166.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54 (6), 1–35.

Narayanan, D.L., Saladi, R.N., Fox, J.L., 2010. Ultraviolet radiation and skin cancer. Int. J. Dermatol. 49 (9), 978–986.

Ngxande, M., Tapamo, J.-R., Burke, M., 2020. Bias remediation in driver drowsiness detection systems using generative adversarial networks. IEEE Access 8, pp. 55592–55601.

Pham, T.-H., Zhang, X., Zhang, P., 2023. Fairness and accuracy under domain generalization. arXiv:2301.13323.

Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P., 2021. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 413–423.

Quadrianto, N., Sharmanska, V., Thomas, O., 2019. Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8227–8236.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2022. Dataset Shift in Machine Learning. Mit Press.

Rodolfa, K.T., Lamba, H., Ghani, R., 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nat. Mach. Intell. 3 (10), pp. 896–904.

Roszkowska, E., 2013. Rank ordering criteria weighting methods–a comparative overview. Optim. Stud. Ekonomiczne (5 (65)), pp. 14–33,https://doi.org/10.15290/ose.2013.05.65.02.

Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P., 2019. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. arXiv:1911.08731.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Stan, S., Rostami, M., 2024. Preserving fairness in AI under domain shift. J. Artif. Intell. Res. 81, pp. 907–934.

Sun, B., Saenko, K., 2016. Deep coral: correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14. Springer, pp. 443–450.

Tan, A.Z., Yu, H., Cui, L., Yang, Q., 2022. Towards personalized federated learning. IEEE Trans. Neural Netw. Learn. Syst. 34 (12), pp. 9587–9603.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5 (1), pp. 1–9.

Valera, I., Singla, A., Gomez Rodriguez, M., 2018. Enhancing the accuracy and fairness of human decision making. Adv. Neural Inf. Process. Syst. 31,pp. 1774–1783.

Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., Ren, K., 2022. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10379–10388.

Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O., 2020. Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8919–8928.

Wani, F.A., Bucarelli, M.S., Silvestri, F., 2024. Learning with noisy labels through learnable weighting and centroid similarity. In: 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–9.

Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J., 2022. FairPrune: achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 743–753.

Xu, D., Yuan, S., Zhang, L., Wu, X., 2018. FairGAN: fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE, pp. 570–575.

Xu, Z., Zhao, S., Quan, Q., Yao, Q., Zhou, S.K., 2023. FairAdaBN: mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 307–317.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W., 2017. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2272–2281.

Yuan, Z., Yan, Y., Sonka, M., Yang, T., 2021. Large-scale robust deep AUC maximization: a new surrogate loss and empirical studies on medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3040–3049.

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

Zhang, C., Zhang, L., Ye, J., 2012. Generalization bounds for domain adaptation. Adv. Neural Inf. Process. Syst. 25,pp. 3320–3328.

Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., Ghassemi, M., 2022. Improving the fairness of chest X-ray classifiers. In: Conference on Health, Inference, and Learning. PMLR, pp. 204–233.

Zong, Y., Yang, Y., Hospedales, T., 2022. MEDFAIR: benchmarking fairness for medical imaging. arXiv:2210.01725.