Neural Collapse-Inspired Debiased Representation Learning for Min-Max Fairness

Shenyu Lu
lu876@purdue.edu
Purdue University
West Lafayette, IN, USA

Junyi Chai
chai28@purdue.edu
Purdue University
West Lafayette, IN, USA

Xiaoqian Wang*
joywang@purdue.edu
Purdue University
West Lafayette, IN, USA

ABSTRACT

Although machine learning algorithms demonstrate impressive performance, their trustworthiness remains a critical issue, particularly concerning fairness when implemented in real-world applications. Many notions of group fairness aim to minimize disparities in performance across protected groups. However, it can inadvertently reduce performance in certain groups, leading to sub-optimal outcomes. In contrast, Min-max group fairness notion prioritizes the improvement for the worst-performing group, thereby advocating a utility-promoting approach to fairness. However, it has been proven that existing efforts to achieve Min-max fairness exhibit limited effectiveness. In response to this challenge, we leverage the recently proposed "Neural Collapse" framework to re-examine Empirical Risk Minimization (ERM) training, specifically investigating the root causes of poor performance in minority groups. The layer-peeled model is employed to decompose a network into two parts: an encoder to learn latent representation, and a subsequent classifier, with a systematic characterization of their training behaviors being conducted. Our analysis reveals that while classifiers achieve maximum separation, the separability of representations is insufficient, particularly for minority groups. *This indicates the sub-optimal performance in minority groups stems from less separable representations, rather than classifiers.* To tackle this issue, we introduce a novel strategy that incorporates a frozen classifier to directly enhance representation. Furthermore, we introduce two easily implemented loss functions to guide the learning process. The experimental assessments carried out on real-world benchmark datasets spanning the domains of Computer Vision, Natural Language Processing, and Tabular data demonstrate that our approach outperforms existing state-of-the-art methods in promoting the Min-max fairness notion.

CCS CONCEPTS

• **Computing methodologies** → *Regularization.*

KEYWORDS

Min-max fairness, Neural collapse, Representation learning

*Corresponding author. This work was partially supported by the EMBRIO Institute, contract #2120200, a National Science Foundation (NSF) Biology Integration Institute, and NSF IIS #1955890, IIS #2146091, IIS #2345235.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671902

ACM Reference Format:
Shenyu Lu, Junyi Chai, and Xiaoqian Wang. 2024. Neural Collapse-Inspired Debiased Representation Learning for Min-Max Fairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671902

1 INTRODUCTION

Machine learning models are trained with Empirical risk minimization (ERM) objectives that aim to optimize the average performance. Numerous studies [2, 5, 18] report unsatisfactory performance across specific demographics from ERM training. This issue is particularly critical in high-stakes fields such as loan approvals, admissions processes, and facial recognition, where fairness across diverse demographics is critical [7]. An unfair machine-learning model can be extremely harmful. For example, when biased facial recognition technology is utilized in crime surveillance, it can lead to unfair treatment and discrimination of certain demographic groups. This, in turn, may result in serious consequences such as false accusations and wrongful arrests [48].

Various fairness definitions have been introduced in the literature, which can generally be categorized into group fairness [13, 17] and individual fairness [21, 24]. In this work, we focus on group fairness. Existing criteria for group fairness concentrate on mitigating disparities in performance across different demographic groups. Abernethy et al. [1] highlight a critical concern with this equity criterion, noting that it may inadvertently compromise the performance of majority groups in an attempt to achieve parity, potentially diminishing overall system performance. Mittelstadt et al. [31] illustrate the "leveling down" problem in fairness, noting that efforts to achieve parity may unintentionally result in decreased performance across all demographic groups. Zietlow et al. [57] point out that when implementing fairness notions based on parity, such as equalized odds, fairness methods lead to a decrease in accuracy across all groups. This degradation in performance is more pronounced in groups that originally exhibited better performance.

Contrary to the pursuit of parity across demographic groups, the concept of Min-max fairness requires a model to optimize for the minimum utility among protected groups [26]. Simply degrading the performance of the majority group fails to fulfill the criteria of Min-max fairness. Min-max fairness can be a preferable notion to enforce when performance improvements for any group are more desirable than achieving parity [41].

A wide range of algorithms have been proposed for fairness in the Min-max framework [12, 18, 34, 40]. Hashimoto et al. [18] proposed using distributionally robust optimization (DRO) to optimize for the worst-case groups. Although it appealing that without using demographics, it runs a risk of optimizing on outliers. Adversarially

Reweighted Learning (ARL) addresses this issue through adversarial training. However, this approach is reported to introduce instability in the training process and necessitates an auxiliary network, which complicates the optimization procedure. [1, 12, 40] proposed to directly optimize the worst performing group to enhance Min-max fairness. However, empirical studies assert that these works do not consistently outperform ERM with regard to accuracy in the worst-performing group [15, 35]. Singh et al. [41] conducted theoretical analysis suggesting that directly optimizing for the worst-case group is less effective in enhancing the worst-group performance.

In light of these challenges, there is a critical need for a new approach to address Min-max fairness from an alternative perspective. Beyond attributing the poor performance of minority groups to the ERM's focus on average performance, it is crucial to investigate the intrinsic causes of biased behavior in ERM training. The phenomenon of "Neural Collapse" has been a groundbreaking revelation in understanding the intricacies of ERM training [32]. We follow the protocol established in neural collapse, utilizing the layer-peeled model that decomposes a L layer neural network ϕ into representation $\boldsymbol{h}$ and a linear classifier $\boldsymbol{W}$, which can be expressed as follows [9, 14, 50]:

$$\phi(\boldsymbol{x}) = \underbrace{\boldsymbol{W}}_{\text{Classifier}\,\boldsymbol{W}} \underbrace{\sigma(\boldsymbol{W}_{L-1}\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x}+b_1)+b_{L-1})}_{\text{Representation}\;\boldsymbol{h}}, \tag{1}$$

Neural Collapse offers a compelling framework for analyzing the behavior of classifiers and representations in neural networks. However, the scope of existing research on neural collapse is in the context of class imbalance [8, 14, 42]. Notably, network performance varies significantly between class imbalance and group[1] imbalance settings. Motivated by the compelling aspects of Neural Collapse, our study pioneers the study of representation and classifier characteristics in scenarios of group imbalance, a critical factor contributing to unfairness.

We systematically re-examine the behavior of representation and classifiers. Our findings reveal that the similarity of representations from minority groups is consistently higher than that of majority groups, indicating that the representations learned via ERM are predominantly influenced by sensitive attributes. Furthermore, we observe that the classifiers adhere to the simplex equiangular tight frame, indicating maximal separation in the space. However, the separation of the representations does not exhibit a comparable level of separability. In other words, to solely attain the objective of accurate classification, a well-separated classifier diminishes the necessity for highly separable representations. Furthermore, we examine a method that directly optimizes for the worst-performing group, a common technique in addressing Min-max fairness. Our analysis reveals that this approach results in marginal improvements in the separability of representations, suggesting that such a method is not optimal for effectively enhancing Min-max fairness.

Expanding upon these insights, we introduce a novel method that does not directly optimize for the Min-max fairness objective. Instead, our method focuses on elevating the worst-case performance by promoting the separability of the representations of minority groups. To achieve this, we employ frozen classifiers, which are aligned with the majority groups within each class. During the training stage, the parameters of the classifier are not updated. We further introduce two easily implementable loss functions: group cross-entropy and group mean alignment loss. These objective functions guide the training procedure, aiming to enhance the fairness of the model and maintain its utility. We benchmark our method across various domains, including computer vision, natural language processing, and tabular data.

The contributions of this study are:

- We present a systematic study of neural collapse phenomena in the context of group-imbalanced datasets, characterizing the behavior of both representations and classifiers in this setting.
- We embark on the application of neural collapse theory to enhance debiased representation, developing a method that effectively addresses fairness issues within the Min-max fairness paradigm.
- We conduct extensive experiments that demonstrate the superior performance of our method in comparison to state-of-the-art approaches.

2 RELATED WORKS

2.1 Min-max Fairness

Min-max fairness is an important concept in group fairness which aims to minimize the maximum error or harm experienced by a group [30, 43]. It's a crucial concept in scenarios where achieving equality is a priority, but doing so without causing undue harm is equally important. Methods on Min-max fairness generally focus on optimizing over the maximum error rate over all the groups. Diana et al. [12] proposed to relax the problem as a mini-max game, where the empirical average of play converges to an optimal solution. Shekhar et al. [40] proposed an active sampling framework to rectify the training process. Abernethy et al. [1] proposed a sampling and reweighting framework to update the model based on the worst-off group at each iteration. Yang et al. [51] formulated the problem as a zero-sum game and introduce a stochastic gradient method to approximate the solution. However, recent work has demonstrated both empirically and theoretically the minor difference between ERM and their specialized Min-max fairness methods [41, 58], where both converge to the same optimum under recoverable group information.

2.2 Robust optimization and spurious correlation

Spurious correlation refers to the misleading heuristics that work for training examples but do not generalize well to new, unseen data [38]. It is typically formulated as the group-wise covariate shift on testing data regarding spurious features. While the context of spurious correlation and Min-max fairness differs, their fundamental concepts, to minimize over the worst-group error, appear to coincide with each other. Sagawa et al. [38] formulated the task as a distributionally robust optimization problem with dynamical weight assignment during optimization. Idrissi et al. [20] proposed to eliminate the spurious correlation in training data by

[1]We define a group as the combination of target label and sensitive attribute, detailed discussion presented in section 3.1.

subsampling and reweighting over groups. Error-based reweighting methods have also been proposed as group-agnostic rectification of spurious correlation [27, 56]. Yao et al. [52] proposed a mix-up technique for learning invariant predictors via selective augmentation. More recently, last-layer retraining has been proposed to amend biased representation [22, 25]. However, compared with representation learning, last-layer retraining has been shown to suffer from inherent insufficiency in mitigating spurious correlations under various scenarios [37, 53]. Our approach diverges from existing methods that address spurious correlations, which predominantly utilize techniques such as reweighting and sampling. Instead, our method promotes the induction of neural collapse (NC) properties within the representation, recognized as an optimal state for classification tasks [50].

2.3 Neural collapse

Papyan et al. [32] first uncovers a phenomenon known as "neural collapse", highlighting the distinct properties of representations and classifiers in deep learning when training to the ending phase. It is observed that networks trained on class-balanced datasets exhibit an emergent simple and symmetric geometry in both the representation and its corresponding classifiers [32]. The neural collapse phenomenon ($\mathcal{NC}$) can be described as:

- $\mathcal{NC}1$: The representation of the same class collapses into their class mean.
- $\mathcal{NC}2$: Class mean vectors maximize their separability within the representation space.
- $\mathcal{NC}3$: The class mean vector aligns with its corresponding classifier vector.
- $\mathcal{NC}4$: The model's predictions, based on logits, can be simplified to the identification of the nearest class centers.

A follow-up study [14] extends the analysis of the neural collapse phenomenon to scenarios with class imbalance. This research identifies a "minority collapse", a distinct phenomenon where classifiers for minority classes converge more closely as the level of imbalance increases. Thrampoulidis et al. [42] introduced a novel asymmetric geometric model to delineate the behavior of representations and classifiers within the context of class imbalance regimes.

3 MOTIVATION

In this section, we investigate the reason for the suboptimal performance in the minority groups. We first examine the characteristics of the representations and classifiers within the regime of neural collapse. We then investigate the potential limitation associated with directly optimizing the Min-max fairness objective.

3.1 Preliminary

Setting. We consider the problem in a $K-$ class classification setting. A dataset is in the form of $(\boldsymbol{x}, a, y)$, where $\boldsymbol{x} \in \mathcal{X}$ is the input feature, $a \in \mathcal{A}$ is the binary sensitive attribute, and $y \in \mathcal{Y}$ is the target label. Define groups $g_{y,a} \in \mathcal{G}$ based on the Cartesian product of $\mathcal{Y}$ and $\mathcal{A}$. Within each class, we define two groups based on the sensitive attribute a: the smaller group refers to the "minority group" and is denoted by superscript ^{min}; The larger group refers to the "majority group" and is denoted by the superscript ^{maj}. We denote $\boldsymbol{W} = [\boldsymbol{w}_0, ..., \boldsymbol{w}_{K-1}] \in \mathbb{R}^{d\times K}$ as the weight matrix of classifiers for K classes, $\boldsymbol{h} \in \mathbb{R}^d$ denotes a representation vector, $\overline{\boldsymbol{h}_{g_{y,a}}}$ denotes the mean representation vector for the group $g_{y,a}$.

We characterize the representation and the classifier based on properties $\mathcal{NC}1$ to $\mathcal{NC}3$. We omit the discussion on $\mathcal{NC}4$ as it is inferred from $\mathcal{NC}1$ to $\mathcal{NC}3$ [50]. We consider the following metric to delineate the corresponding $\mathcal{NC}$ property:

- For $\mathcal{NC}1$, we compute group variance in each group:

$$s_{y,a} = tr\Big(\frac{1}{|g_{y,a}|}\sum_i ((\boldsymbol{h}_i - \overline{\boldsymbol{h}_{g_{y,a}}})(\boldsymbol{h}_i - \overline{\boldsymbol{h}_{g_{y,a}}})^\top\Big), s.t. \boldsymbol{h}_i \in g_{y,a}$$

  A small $s_{y,a}$ indicates the representations within the same group converge towards their corresponding group mean.
- For $\mathcal{NC}2$, we compute cosine similarity in class vectors:

$$d_{\text{cls-sim}} = sim(\overline{\boldsymbol{h}_{g_{0,\cdot}}}, \overline{\boldsymbol{h}_{g_{1,\cdot}}})$$

  A small $d_{\text{cls-sim}}$ indicates a high degree of separability among class mean representations.
- For $\mathcal{NC}3$, to characterize the unique scenario presented by group imbalances, we calculate three key metrics: Majority Cosine Similarity, Minority Cosine Similarity, and Classifier Cosine Similarity.

$$d_{\text{maj-sim}} = sim(\overline{\boldsymbol{h}^{maj}_{g_{0,\cdot}}}, \overline{\boldsymbol{h}^{maj}_{g_{1,\cdot}}})$$

$$d_{\text{min-sim}} = sim(\overline{\boldsymbol{h}^{min}_{g_{0,\cdot}}}, \overline{\boldsymbol{h}^{min}_{g_{1,\cdot}}})$$

$$w_{\text{sim}} = sim(\boldsymbol{w}_0, \boldsymbol{w}_1)$$

  A small $d_{\text{maj-sim}}$ value indicates enhanced separability within the majority group's mean representations. A small $d_{\text{min-sim}}$ reflects increased separability within the minority group's mean representations. A small w_{sim} suggests high separability among classifiers.

We first conduct experiments on the Waterbirds dataset [38], a synthetic dataset created by combining images of birds [44] and images of places [55]. The dataset is categorized into four distinct groups: waterbirds on water background ($g_{0,0}$), waterbirds on land background ($g_{0,1}$), landbirds on water background ($g_{1,0}$), and landbirds on land background ($g_{1,1}$). The sizes of these groups are as follows: 3498, 184, 56, and 133, respectively. According to the definition, the majority groups are $g_{0,0}$ and $g_{1,1}$, the minority groups are $g_{0,1}$ and $g_{1,0}$. Following previous work [22, 25, 38], we employ a ResNet-50 model pre-trained on ImageNet-1K and train to 100 % accuracy on the training set.

3.2 Group imbalanced neural collapse observation

3.2.1 Characteristics of ERM training. The test set results for the Waterbirds dataset are presented in Fig 1. Specifically, Fig 1(a) illustrates the cosine similarity for representations and classifiers, Fig 1(b) shows the group variance.

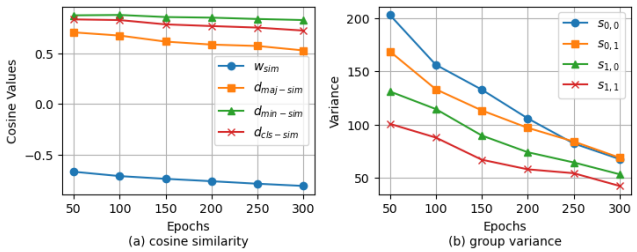

Figure 1: A characteristic plot depicting the behavior of representations and classifiers within the Waterbirds dataset.

From Fig 1(b), we observe that as the number of training epochs increases, there is a noticeable decrease in the within-group variability of the representations. We extend the concept of $\mathcal{NC}$ 1 to scenarios with group imbalances, referring to it as the within-group variability of the representation collapse.

From Fig 1(a), we observe a decreasing trend in w_{sim}, $d_{\text{cls-sim}}$, $d_{\text{maj-sim}}$, and $d_{\text{min-sim}}$. This trend corresponds to an increase in the angular separation among classifiers, as well as between the mean vectors of different classes for both majority and minority samples. An increase in angular separation is advantageous for enhancing feature distinction. However, we find that $\boldsymbol{w}_{sim}$ is much smaller than $d_{\text{cls-sim}}$. The $\boldsymbol{w}_{sim}$ values approach -1, indicating the classifier is progressively moving towards a state that maximizes the pairwise angular separation between the two classes, i.e., a simplex equiangular tight frame [32]. In contrast, the level of separation of representations is much lower (as $d_{\text{cls-sim}}$ is much above -1). It fails to satisfy $\mathcal{NC}2$ and $\mathcal{NC}3$. Importantly, we note that $d_{\text{maj-sim}}$ is lower than $d_{\text{min-sim}}$, suggesting that the separability of the minority samples is less than that of the majority group. This indicates that the representation is highly effected by the sensitive attribute.

To determine the prevalence of this phenomenon in settings characterized by group imbalances, we performed identical tests on both the CelebA [28] and UTK datasets [54]. Details on the datasets used are provided in Appendix A. The results are presented in Fig 2 and 3.

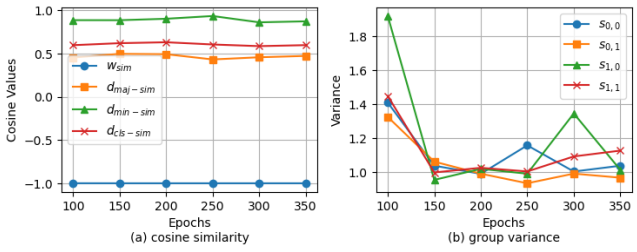

Figure 2: A characteristic plot depicting the behavior of representations and classifiers within the CelebA dataset.

From Fig 2 and Fig 3, we observe a trend consistent with that identified in the Waterbirds dataset. We summarize our findings:

- Classifiers enhance their separability to effectively discriminate between classes.

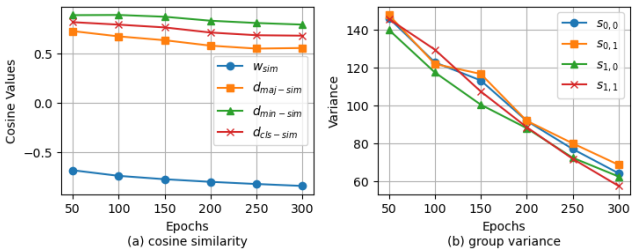

Figure 3: A characteristic plot depicting the behavior of representations and classifiers within the UTK dataset.

- Within the same class, the mean vectors of the majority and minority groups are misaligned.
- Across different classes, the cosine similarity between the mean vectors of the majority groups is lower compared to that of the minority groups.

To elucidate our findings, we visualize the mean vectors of representations and their corresponding classifiers in 3D space. We reduce the dimensionality of the representation from $\mathbb{R}^{2048}$ to $\mathbb{R}^3$ in the learning stage. We compute the mean of representations and classifiers in the test set. The resulting plot is presented in Fig 4. It is observed that the similarity within minority groups is higher than in the majority group, and the classifiers are distinctly separated.

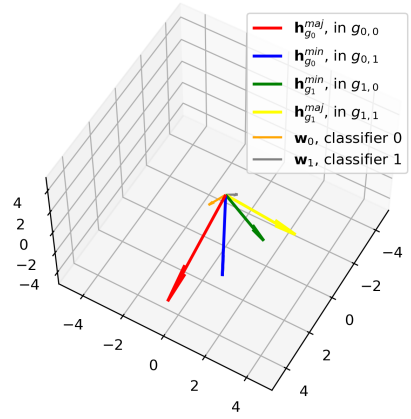

Figure 4: Visualization of representations and classifiers in Waterbirds dataset.

3.2.2 Examine Min-max fairness approach. We evaluate the effectiveness of the existing Min-max fairness strategies in the context of the neural collapse regime. We employ the Min-max Stochastic Gradient Descent method [1] on the Waterbirds dataset. For comprehensive details on the implementation, refer to Section 5.1. We show the result in Fig 5. For clarity, we juxtapose the results of [1] with those from ERM training on the same graph for direct comparison. For ERM, the cosine similarity for the majority group is depicted in red, and for the minority group in purple. Observations indicate that, despite [1] narrowing the gap between cosine similarities within minority and majority groups, the curves for

the Min-max fairness approach (representing $d_{\text{maj-sim}}$, $d_{\text{min-sim}}$) largely coincide with the curve for the minority group under ERM training (ERM $d_{\text{min-sim}}$ in Fig 5). This highlights a limitation in the separability of representations achieved through Min-max training, particularly compromising the majority groups' separability.

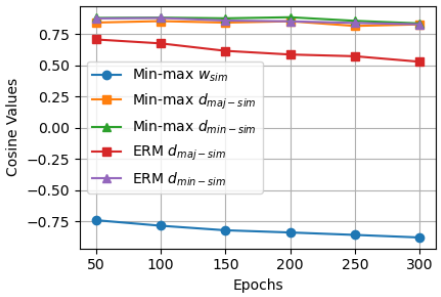

Figure 5: A characteristic plot for the Min-max fairness approach [1].

4 METHOD

We now introduce our method to learn a debiased representation to improve the worst performance group. We present an approach inspired by the neural collapse, utilizing a frozen-classifier technique that emphasizes the separability of the minority group. Subsequently, we propose two loss functions specifically designed to optimize the network's effectiveness.

4.1 Frozen classifiers

As illustrated in Section 3.2, the representation derived from ERM does not exhibit the $\mathcal{NC}3$ property, indicating a strong misalignment between the classifier vectors and the mean vectors of the representations. Furthermore, we observe that the optimization process for linear classifiers converges more rapidly than that for representation vectors. Based on these findings, we freeze the classifier parameters, thereby concentrating on the optimization of the representation.

Following the definitions in section3.1, the majority group for each class can be expressed as $g_y^{maj} = g_y,\ s.t.\ |g_y| = \max_a(|g_{y,a}|)$. We compute the mean vector from each majority group:

$$\overline{\boldsymbol{h}_y^{maj}} = \frac{1}{|g_{y,\cdot}^{maj}|} \sum_i \boldsymbol{h}_i, s.t.\ \boldsymbol{h}_i \in g_{y,\cdot}^{maj}, y \in \mathcal{Y} \tag{2}$$

The frozen parameters of the classifier, $\boldsymbol{W} = [\boldsymbol{w}_0, .., \boldsymbol{w}_{K-1}]$, are assigned by:

$$\boldsymbol{w}_k := \frac{\overline{\boldsymbol{h}_k^{maj}}}{||\overline{\boldsymbol{h}_k^{maj}}||_2}, k \in \{0, .., K-1\} \tag{3}$$

Remark 1 (Focusing on the minority group): We consider the cross entropy loss function:

$$\mathcal{L}_{CE}(\boldsymbol{h}, \boldsymbol{W}) = -\log(\frac{\exp(\boldsymbol{h}^\top \boldsymbol{w}_y)}{\sum_{k=1}^K \exp(\boldsymbol{h}^\top \boldsymbol{w}_k)}) \tag{4}$$

Assigning the classifier's weights to align with the mean vector of the majority group results in higher losses for minority groups, compared with those for majority groups, due to their misalignment with the classifier's orientation, as illustrated in Fig 4. Consequently, this misalignment guides the update process to concentrate on the minority group primarily.

Remark 2 (Maintaining separability): We take the gradient of CE loss w.r.t. $\boldsymbol{h}$:

$$-\frac{\partial \mathcal{L}_{CE}}{\partial \boldsymbol{h}} = (1 - \frac{\exp(\boldsymbol{h}^\top \boldsymbol{w}_y)}{\sum_{j=1}^K \exp(\boldsymbol{h}^\top \boldsymbol{w}_j)})\boldsymbol{w}_y - \sum_{k \neq y}^K \frac{\exp(\boldsymbol{h}^\top \boldsymbol{w}_k)}{\sum_{j=1}^K \exp(\boldsymbol{h}^\top \boldsymbol{w}_j)} \boldsymbol{w}_k \tag{5}$$

Equation (5) shows the update direction of $\boldsymbol{h}$, explicitly incorporating a negative sign in the gradient term for illustrative clarity. In the first component, $(1 - \frac{\exp(\boldsymbol{h}^\top \boldsymbol{w}_y)}{\sum_{j=1}^K \exp(\boldsymbol{h}^\top \boldsymbol{w}_j)})\boldsymbol{w}_y$, the coefficient associated with $\boldsymbol{w}_y$ is strictly positive, which directs the adaptation of $\boldsymbol{h}$ towards the classifier $\boldsymbol{w}_y$. For the second component, $-\sum_{k \neq y}^K \frac{\exp(\boldsymbol{h}^\top \boldsymbol{w}_k)}{\sum_j \exp(\boldsymbol{h}^\top \boldsymbol{w}_j)} \boldsymbol{w}_k$, the coefficients associated with each $\boldsymbol{w}_k$ are strictly negative, thereby steering $\boldsymbol{h}$ away from the corresponding classifiers $\boldsymbol{w}_k$.

Equation (3) defines each classifier $\boldsymbol{w}$ as the mean representation of the majority group within its respective class. Consequently, the representation vector $\boldsymbol{h}$ will be oriented towards $\boldsymbol{h}_y$, the mean vector of the majority group in class y, while pushing away itself from $\boldsymbol{h}_k$, the mean vector of the majority group in any other class k, where $k \neq y$. In other words, for each sample $\boldsymbol{h}$, it will converge within its class and separate from other classes. As a result, this mechanism leads to closer representations within each class and enhances the separation between different classes.

4.2 Group Cross Entropy

Given that ERM training primarily focuses on average performance, we adopt a group-wise tracking approach to assess the optimization progress within each group. We have formulated a group-wise loss, defined as follows:

$$\mathcal{L}_{g_{y,a}} = \frac{1}{|g_{y,a}|} \sum_i \mathcal{L}(\phi(\boldsymbol{x}_i), y_i), s.t. \boldsymbol{x}_i \in g_{y,a}, \tag{6}$$

where $\phi(\cdot)$ is a neural network with a backbone and a linear classifier. We conducted experiments on the Waterbirds dataset utilizing the ERM training objective to train ResNet50 models for 20 epochs. We monitored the $\mathcal{L}g_{y,a}$ in the training set and have depicted the resulting trends in Fig 6.

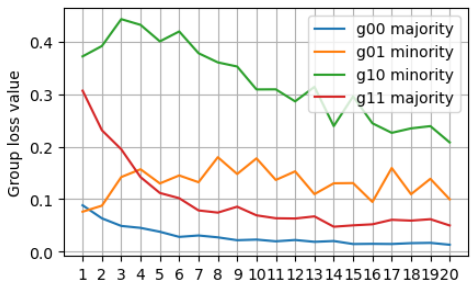

Figure 6: Group-wise loss in the training stage.

Analysis indicates that during ERM training, the loss function for the majority group exhibits rapid convergence and stabilizes at a notably lower magnitude in comparison to the minority group. This phenomenon underscores the inherent challenges associated with optimizing models to adequately address the needs of minority groups.

In response to the challenges associated with optimizing performance for minority groups, we propose the direct optimization of the Group Cross Entropy (GCE) Loss to enhance the optimization process by accommodating group-specific characteristics more effectively. The GCE Loss is formulated with the following objective:

$$\mathcal{L}_{GCE} = \sum_{g_{y,a} \in \mathcal{G}} \frac{1}{|g_{y,a}|} \sum_{i} \mathcal{L}_{CE}(\phi(\boldsymbol{x}_i), y_i), \boldsymbol{x}_i \in g_{y,a} \tag{7}$$

where $\mathcal{L}_{CE}(\cdot)$ is the cross-entropy loss. $\phi(\boldsymbol{x})$ is the prediction and y is the label.

4.3 Group mean alignment

Alignment between the representation and the classifier leads to improved task performance. In our approach, where the classifier is fixed to favor majority groups, our goal is to promote the convergence of the minority group's mean vector towards that of the majority group. This approach seeks to achieve a more balanced representation, enhancing overall system performance. In alignment with this objective, we introduce the "group mean alignment loss" defined as:

$$\mathcal{L}_{align} = \sum_{y=0}^{K-1} \mathcal{L}_{MSE}(\overline{\boldsymbol{h}_{g_{y,0}}}, \overline{\boldsymbol{h}_{g_{y,1}}}), \tag{8}$$

4.4 Algorithm

We summarize our method in Algorithm 1 and 2.

Algorithm 1 Classifier weights assigning

Input: Input feature $\boldsymbol{x}$, label y, sensitive attribute a. Backbone model $f_\theta(\cdot)$.
Output: The frozen classifier $q_w(\cdot)$.

1: Identify the majority groups: $g_y^{maj} = g_y, s.t. \ |g_y| = \max_a(|g_{y,a}|)$.
2: Compute mean vectors from the majority groups:

$$\overline{\boldsymbol{h}_y^{maj}} = \frac{1}{|g_{y,\cdot}^{maj}|} \sum_i f_\theta(\boldsymbol{x}_i), s.t. \ \boldsymbol{x}_i \in g_{y,\cdot}^{maj}, y \in \mathcal{Y}$$

3: Assign the value to each classifier $\boldsymbol{w}_k := \frac{\overline{\boldsymbol{h}_k^{maj}}}{||\overline{\boldsymbol{h}_k^{maj}}||_2}, k \in \{0, .., K-1\}$.

return $q_w = [\boldsymbol{w}_0, ..., \boldsymbol{w}_{K-1}]$

Algorithm 2 De-biased representation learning

Input: Input feature $\boldsymbol{x}$, label y, sensitive attribute a in the training set, epochs E, learning rate η.
Output: The parameter of the backbone model θ

1: Define $g_{y,a} = \{\boldsymbol{x}|y, a\}, g_{y,a} \in \mathcal{G}$
2: **for** epoch to E **do**:
3: Assigning the weight of classifier $q_w(\cdot)$ by Algorithm 1.
4: Compute group cross-entropy loss:

$$\mathcal{L}_{GCE} = \sum_{g_{y,a} \in \mathcal{G}} \frac{1}{|g_{y,a}|} \sum_{i} \mathcal{L}_{CE}(q_w(f_\theta(\boldsymbol{x}_i)), y_i), \boldsymbol{x}_i \in g_{y,a}$$

5: Compute mean alignment loss:

$$\mathcal{L}_{align} = \sum_{y=0}^{K-1} \mathcal{L}_{MSE}(\overline{\boldsymbol{h}_{g_{y,0}}}, \overline{\boldsymbol{h}_{g_{y,1}}}),$$

where $\overline{\boldsymbol{h}_{g_{y,a}}} = \frac{1}{|g_{y,a}|} \sum_i f_\theta(\boldsymbol{x}_i), \boldsymbol{x}_i \in g_{y,a}$.
6: Total loss: $\mathcal{L} = \mathcal{L}_{GCE} + \mathcal{L}_{align}$
7: Update θ by gradient descent:

$$\theta = \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

8: **end for**
return θ

5 EXPERIMENT

In this section, we evaluate our method's performance on benchmark datasets. We delineate the experimental setup, followed by a comprehensive presentation of the results. Subsequently, we provide a GradCam visualization to compare our method with ERM training. We conduct an ablation study to show the effectiveness of each component within our method. Lastly, we examine the optimization process for each group during the training phase and evaluate the quality of the learned representations. Our code is available at https://github.com/lu876/Neural-collapse-inspired-debiased-representation-learning-for-Min-max-fairness

5.1 Experiment setup

Datasets. We study the following five datasets which are well-established as benchmarks for fairness research: Waterbirds [23, 38], CelebA [28], ISIC [6], MultiNLI [45], the UCI Adult [3].

- Waterbirds dataset [23, 38] comprises images where the primary task (y) is to classify the bird type (Landbird or Waterbird), with the sensitive attribute (a) being the habitat (Water or Land background).
- CelebA [28] is a dataset containing $200K$ celebrity faces. Guided by a recent study [16] that identifies datasets suitable for group fairness research, we have adopted their recommendations and, in accordance with established protocols [4, 33, 36], selected 'attractiveness' as the task label (y) and 'male' as the sensitive attribute (a) in our use of the CelebA dataset. An ethical statement on the use of this task is provided in a subsequent section of our paper.
- ISIC [6] is a skin cancer diagnosis dataset. We follow [47], employing the version of the dataset released in 2018. The

task is to predict $y = \{\text{Benign, Malignant}\}$ and sensitive attribute $a = \{\text{patch, without patch}\}$.
- MultiNLI [45] is a text multi-class classification dataset that the task is to predict $y = \{\text{entailment, neutral, contradiction}\}$. The sensitive attribute $a = \{\text{negation, no negation}\}$.
- The UCI Adult [3] dataset is a tabular dataset comprising 48842 samples. Its primary task (y) is to predict whether an individual's income exceeds $50K$ per year. We select a = gender as the sensitive attribute.

Metrics. We evaluate all methods from two perspectives: Utility and Fairness. For utility, we compute the average accuracy (*Acc*). For fairness, we use the Min-max fairness criteria to report the worst group accuracy (*WGA*). To ensure a comprehensive evaluation, the Best Group Accuracy (*BGA*) is also reported. Additionally, we show the disparity between the highest and the lowest group accuracies to emphasize fairness considerations. This is quantified by $\Delta\text{Acc} = |\max_{g\in G} Acc_g - \min_{g\in G} Acc_g|$, offering an understanding of performance gaps among different groups. For the ISIC dataset, the test set is significantly imbalanced, with a deficiency in samples that are both malignant and exhibit a patch. We follow [47] to utilize ROC AUC as the metric for performance evaluation.

Baselines. We compare our proposed algorithm against several state-of-the-art methods for Min-max fairness. Apart from ERM, these methods employ diverse strategies to address fairness concerns. Specifically:

- Min-max Stochastic gradient descent (MMSGD) [1] employs active sampling to optimize min-max fairness. At each iteration, it specifically targets optimization for the worst-performing group.
- MinimaxFair (MMF) [12] employs a method based on multiplicative weights update to achieve Min-max fairness. The weights are determined using an exponential weights algorithm, with respect to the errors of each group.
- Hilbert-Schmidt Independence Criterion (HSIC) [16] is designed to learn a representation that minimizes the HSIC between the representation itself and the sensitive attributes.
- Subsampling large groups (SUBG) [20] involves equalizing group sizes by subsampling each group to match the size of the smallest group.
- Group DRO (GDRO) [38]: By leveraging group information, GDRO employs the distributionally robust optimization, which allows for the dynamic amplification of the weight assigned to the worst-group loss during the optimization process.

Implementations. We implement all methods on a single NVIDIA RTX-3090 GPU. Each method is independently trained three times, and we report the mean and standard deviation of the results.

For the Waterbirds, CelebA, and ISIC datasets, we employ a ResNet-50 model [19] pre-trained on ImageNet [10], as provided in the torchvision library [29]. The MultiNLI dataset is processed using a pre-trained BERT base model [11], sourced from Huggingface [46]. For the UCI Adult dataset, we utilize a two-layer multilayer perceptron with hidden layers configured to {32, 64} nodes and employing ReLU activation functions.

A consistent figuration of hyperparameters is adopted across all evaluated methods. Specifically, for the CelebA and Waterbirds datasets, we employ the SGD optimizer with a learning rate and weight decay both set to 10^{-3}, training with a batch size of 32 over 50 epochs [22]. The ISIC dataset was also trained with SGD, but with training extended to 100 epochs. For the MultiNLI, the AdamW optimizer is utilized with a learning rate of 3×10^{-5} and no weight decay, fine-tuning for 10 epochs with a batch size of 64. For the Adult dataset, the Adam optimizer is applied with a learning rate of 10^{-3}, a batch size of 256, and training for 50 epochs.

Model-specific hyperparameters are selected based on configurations reported in the literature or officially released code for each dataset. In cases where specific implementations for a dataset are absent, we tune the hyperparameters to optimize the accuracy of the worst-performing group within the validation set, ensuring a consistent comparison standard across all methods evaluated.

5.2 Results

We report the results within the computer vision domain in Tables 1 2, 3, within the NLP domain in Table 4, and within the tabular domain in Table 5. The best results are highlighted in **bold** and the second best are underlined for clarity.

Table 1: Results of WaterBirds Dataset.

	WGA↑	BGA↑	Δ Acc ↓	ACC ↑
ERM	81.41 ± 0.51	**99.44 ± 0.17**	18.03	91.84 ± 0.49
HSIC	80.53 ± 1.11	99.41 ± 0.23	18.88	90.76 ± 1.31
SUBG	88.99 ± 0.85	92.43 ± 0.95	**3.44**	91.15 ± 0.72
MMSGD	87.92 ± 1.25	93.57 ± 1.93	5.65	90.16 ± 0.36
MMF	84.84 ± 0.70	93.57 ± 0.98	8.73	90.25 ± 0.80
GDRO	88.61 ± 0.60	94.15 ± 0.50	5.54	91.13 ± 0.24
Ours	**90.14 ± 0.08**	94.66 ± 0.35	4.52	**92.04 ± 0.20**

Table 2: Results of CelebA Dataset.

	WGA↑	BGA↑	Δ Acc ↓	ACC ↑
ERM	70.05 ± 0.85	87.34 ± 1.37	17.29	**81.94 ± 0.94**
HSIC	64.41 ± 5.99	**88.93 ± 1.92**	24.52	79.93 ± 1.40
SUBG	74.13 ± 3.38	83.37 ± 2.24	9.24	79.63 ± 0.32
MMSGD	75.01 ± 1.01	86.01 ± 0.50	11.00	80.45 ± 0.23
MMF	77.19 ± 0.65	81.40 ± 0.94	**4.21**	79.58 ± 0.35
GDRO	76.40 ± 1.53	82.27 ± 0.55	5.87	79.17 ± 1.11
Ours	**78.26 ± 1.25**	83.23 ± 1.12	4.97	80.19 ± 0.36

In our evaluation of vision tasks, we observed that our method consistently outperforms benchmarked methods in terms of worst-group accuracy, while maintaining comparable average accuracy. Although reweighting techniques like GDRO and MMF improve worst-group accuracy, our method outperforms these in both worst-group accuracy (WGA) and best-group accuracy (BGA), thereby providing a more comprehensive enhancement in performance relative to these reweighting strategies.

Table 3: Results of ISIC Dataset.

	WGA ↑	BGA ↑	Δ Acc ↓	ROC AUC ↑
ERM	27.97 ± 2.15	97.89 ± 1.54	69.92	59.18 ± 4.06
HSIC	28.17 ± 2.78	96.36 ± 0.05	68.19	58.44 ± 5.09
SUBG	44.71 ± 6.36	97.58 ± 0.86	52.87	66.42 ± 4.13
MMSGD	29.10 ± 2.62	95.76 ± 0.86	66.66	59.30 ± 3.17
MMF	26.72 ± 2.70	**98.79 ± 0.86**	72.07	61.46 ± 4.37
GDRO	35.23 ± 4.98	92.73 ± 2.97	57.50	62.00 ± 2.03
Ours	**61.23 ± 11.32**	94.58 ± 2.54	**33.35**	**83.77 ± 7.82**

Although Han et al. [16] demonstrate that HSIC achieves an optimal balance between utility and fairness in their benchmark tests, we observe that it falls short in enhancing the performance of the worst-performing group. It is important to note that the benchmarking [16] focuses on parity-based notions of fairness (e.g. Equalized odds). This approach does not necessarily lead to practical improvements in Min-max fairness. MMSGD, which explicitly targets the Min-max fairness objective, yields enhancements in worst-group accuracy. Nonetheless, its level of improvement is relatively modest when compared to other methods.

On the ISIC dataset, our method achieves a substantial improvement in performance margin compared to other methods. The ISIC dataset presents two significant challenges: the existence of exceptionally small-sized groups and class imbalance within the training set. These issues often result in the suboptimal performance of most existing methods, particularly in terms of enhancing the worst-performing groups. While reweighting methods like SUBG and GDRO show some improvement, their impact is limited, highlighting their substantial limitations in effectively addressing rare instances in the training set. In contrast, our approach emphasizes the direct optimization of representations, leading to a significant enhancement in performance across a range of data skewness scenarios.

Table 4: Results of MultiNLI Dataset.

	WGA↑	BGA ↑	Δ Acc ↓	ACC ↑
ERM	64.40 ± 2.35	95.13 ± 0.36	30.73	81.14 ± 0.49
HSIC	64.40 ± 4.03	**95.47 ± 0.68**	31.07	**81.34 ± 0.59**
SUBG	67.78 ± 0.49	79.76 ± 2.78	11.98	71.37 ± 0.60
MMSGD	68.80 ± 1.68	92.32 ± 1.49	23.52	80.82 ± 0.32
MMF	70.94 ± 0.23	86.24 ± 4.43	15.30	77.08 ± 1.14
GDRO	76.02 ± 2.16	86.40 ± 2.86	10.38	80.53 ± 0.10
Ours	**77.90 ± 0.58**	85.08 ± 0.99	**7.18**	80.96 ± 0.23

For the NLP task, our method achieves the optimal worst group accuracy and exhibits a balanced group performance as indicated by the ΔAcc metric, outperforming other comparative methods. Our method preserves competitive average accuracy, with a slight decrease of only 0.38 % compared to the highest observed performance. While SUBG improves the worst group accuracy, it significantly reduces the best group accuracy compared to the ERM model, highlighting that simple down-sampling of the training set is a sub-optimal solution.

However, it is important to note a reduction in the best group accuracy within our method. This can potentially be attributed to the alignment of the mean vector. While this alignment offers certain advantages, it may inadvertently impact the performance of the highest-performing group, suggesting a trade-off between overall alignment benefits and the best group performance.

Table 5: Results of Adult Dataset.

	WGA ↑	BGA ↑	Δ Acc ↓	ACC ↑
ERM	57.93 ± 1.18	**95.98 ± 0.35**	38.05	**84.34 ± 0.12**
HSIC	61.10 ± 0.51	95.19 ± 0.46	34.09	84.11 ± 0.24
SUBG	77.67 ± 0.26	86.98 ± 0.15	**9.31**	80.62 ± 0.03
MMSGD	67.26 ± 0.39	92.79 ± 0.69	25.53	83.74 ± 0.22
MMF	77.22 ± 0.66	86.56 ± 0.40	9.34	80.48 ± 0.11
GDRO	77.73 ± 0.06	87.79 ± 0.04	10.06	81.16 ± 0.02
Ours	**78.00 ± 0.30**	87.57 ± 0.38	9.57	81.00 ± 0.14

For the tabular dataset, our analysis reveals that while methods such as GDRO, MMF, and SUBG produce comparable results, our method achieves the optimal worst group accuracy. However, the marginal improvement of less than 1% suggests that our method does not significantly surpass the baseline methods in the tabular setting. This limited enhancement could be partially attributed to the inherent structure of tabular data, which already encodes semantic information in its input features, potentially diminishing the impact of advanced representation learning.

5.3 Visualization

We present visualization results from the test set of the Waterbirds dataset. We employ Grad-CAM [39] to identify and highlight the regions of interest that our model prioritizes. We provide visual comparisons between the standard ERM training approach and our proposed training strategy. For the implementation of Grad-CAM, we utilize the OmniX AI package [49]. The comparative results are illustrated in Fig 7. To provide a more comprehensive evaluation, we include visualization results for the ISIC dataset, which can be found in Appendix C.

Fig 7 illustrates that our approach more accurately focuses on the target object in an image, unlike ERM training, which erroneously emphasizes background features.

5.4 Ablation Study

We investigate the efficacy of two key components within our methodology: the Frozen Classifier and Group Mean Alignment. As the primary objective of optimization, we establish a baseline that utilizes solely the group cross-entropy loss. Subsequently, we conduct a series of ablation studies by removing one or both components to assess their contributions to performance. On removing the frozen classifier, we replace it with a trainable classifier. The results are shown in Tab 6.

From Table 6, we draw the following conclusions: utilizing a frozen classifier contributes to improving the accuracy of the worst-performing group. This is evidenced by a notable decrease in worst group accuracy of 2.03 % when a trainable classifier is employed

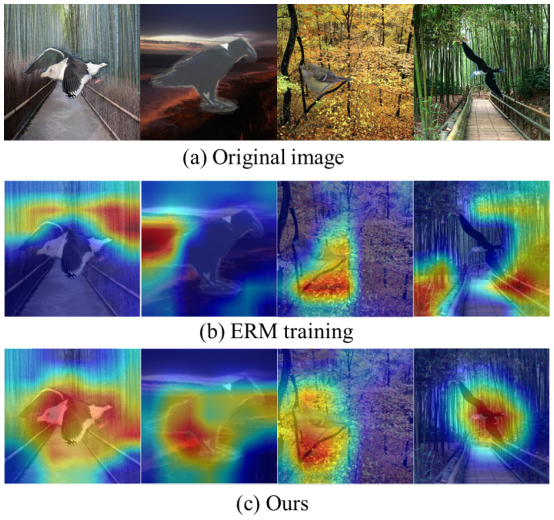

Figure 7: Grad-CAM visualization for exemplary test images.

Table 6: Ablation study to investigate the effects of the major components in our methods. 'FC' denotes Frozen Classifiers, 'GMA' denotes group mean alignment.

	WGA↑	BGA↑	Δ Acc ↓	ACC ↑
w/o FC	88.16	**97.73**	9.57	93.56
w/o GMA	89.56	94.41	4.85	91.94
w/o FC,GMA	87.53	97.52	9.99	**93.68**
Full version	**90.19**	94.28	**4.09**	91.85

in the full model. Additionally, the incorporation of Group mean alignment loss further increases the performance of the worst-performing group, underscoring the effectiveness of our proposed method in enhancing the worst group performance.

5.5 Optimization and Representations

Performing on GCE. In section 4.2, we delineate the challenges associated with optimizing for minority groups. To address these difficulties, we introduce the GCE loss, which simplifies the optimization process concerning minority groups. This section presents an experimental analysis to illustrate the trend of group loss across training epochs. We monitor the group-wise loss (introduced in section 4.2) during the training using the GCE loss.

As illustrated in Fig 8, the application of GCE loss leads to more effective and rapid optimization of minority groups compared to traditional cross entropy loss. Furthermore, the relatively balanced loss across each group suggests that GCE loss optimizes all groups uniformly, avoiding an overemphasis on a single group.

Performing on representations. Following Section 3.2, we evaluate our method's quality of representations by comparing the cosine similarity within majority and minority groups after attaining of 100 % training accuracy. Fig 9 presents results using our training method alongside comparative data from ERM training, as sourced

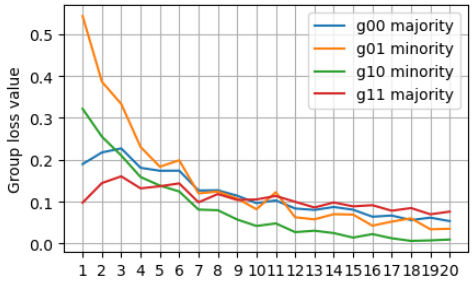

Figure 8: Group-wise loss in the training stage by using ours method.

from Fig 1. It indicates that our method significantly enhances separability in majority groups compared to ERM. We still observe that there is a gap between the curves from the majority group and the minority group. However, the separability of the minority group is close to that of the majority under ERM training. This highlights the effectiveness of our approach in enhancing minority group differentiation.

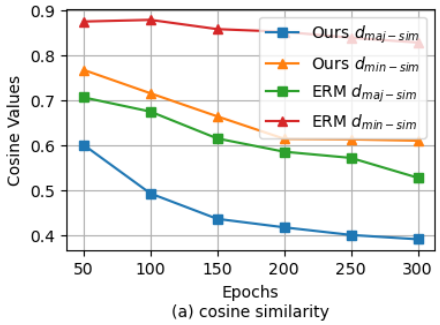

Figure 9: Comparison of representations Learned from Our method versus ERM.

6 CONCLUSION

Min-max fairness is a critical yet challenging aspect of fairness in machine learning, with existing methods achieving only incremental improvements. In this study, we leverage the concept of neural collapse to analyze the behavior of representations and classifiers learned via ERM models. Our analysis reveals that the suboptimal performance for minority groups can be attributed to low separability in their representations. Furthermore, we observe that traditional min-max fairness approaches often yield high similarity across groups, compromising their effectiveness. To address these issues, we leverage the neural collapse property, aligning classifiers with the majority group within each class to facilitate the learning of debiased representations. Additionally, we introduce two simple and efficient loss functions designed to guide the learning process more effectively. We conduct an extensive experiment on five datasets spanning from the Vision, NLP, and tabular data. Our methods effectively improve the worst group accuracy among all datasets and achieve a comparable average accuracy.

REFERENCES

[1] Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. 2022. Active Sampling for Min-Max Fairness. In *International Conference on Machine Learning*. PMLR, 53–65.

[2] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485* (2020).

[3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[4] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. 2022. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems* 35 (2022), 19152–19164.

[5] Junyi Chai and Xiaoqian Wang. 2022. Fairness with adaptive weights. In *International Conference on Machine Learning*. PMLR, 2853–2866.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368* (2019).

[7] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[8] Hien Dang, Tho Tran Huu, Stanley Osher, Nhat Ho, Tan Minh Nguyen, et al. 2023. Neural Collapse in Deep Linear Networks: From Balanced to Imbalanced Data. In *International Conference on Machine Learning*. PMLR, 6873–6947.

[9] Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. 2024. Neural Collapse for Cross-entropy Class-Imbalanced Learning with Unconstrained ReLU Feature Model. *arXiv preprint arXiv:2401.02058* (2024).

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 66–76.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[14] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences* 118, 43 (2021), e2103091118.

[15] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. 2022. Subgroup robustness grows on trees: An empirical baseline investigation. *Advances in Neural Information Processing Systems* 35 (2022), 9939–9954.

[16] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. 2023. FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. *arXiv preprint arXiv:2306.09468* (2023).

[17] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. 2022. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*. PMLR, 336–351.

[21] Christina Ilvento. 2019. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250* (2019).

[22] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *The Eleventh International Conference on Learning Representations*.

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[25] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. 2023. Towards last-layer retraining for group robustness with fewer annotations. *arXiv preprint arXiv:2309.08534* (2023).

[26] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33 (2020), 728–740.

[27] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[29] TorchVision maintainers and contributors. 2016. *TorchVision: PyTorch's Computer Vision library*.

[30] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[31] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default. *arXiv preprint arXiv:2302.02404* (2023).

[32] Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences* 117, 40 (2020), 24652–24663.

[33] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10389–10398.

[34] Thomas Pethick, Grigorios Chrysos, and Volkan Cevher. 2022. Revisiting adversarial training for the worst-performing class. *Transactions on Machine Learning Research* (2022).

[35] Stephen R Pfohl, Haoran Zhang, Yizhe Xu, Agata Foryciarz, Marzyeh Ghassemi, and Nigam H Shah. 2022. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports* 12, 1 (2022), 3254.

[36] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8227–8236.

[37] Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. 2023. Is Last Layer Re-Training Truly Sufficient for Robustness to Spurious Correlations? *arXiv e-prints* (2023), arXiv–2308.

[38] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[40] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. 2021. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems* 34 (2021), 24535–24544.

[41] Harvineet Singh, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara, and Chris Russell. 2023. When do minimax-fair learning and empirical risk minimization coincide?. In *International Conference on Machine Learning*. PMLR, 31969–31989.

[42] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. 2022. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems* 35 (2022), 27225–27238.

[43] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. PMLR, 6373–6382.

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[45] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana). Association for Computational Linguistics, 1112–1122. http://aclweb.org/anthology/N18-1101

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[47] Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. 2023. Discover and Cure: Concept-aware Mitigation of Spurious Correlation. In *ICML*.

[48] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. 2021. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 578–586.

[49] Wenzhuo Yang, Hung Le, Silvio Savarese, and Steven Hoi. 2022. OmniXAI: A Library for Explainable AI. (2022). https://doi.org/10.48550/ARXIV.2206.01612 arXiv:206.01612

[50] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. 2022. Inducing Neural Collapse in Imbalanced Learning: Do We Really Need

a Learnable Classifier at the End of Deep Neural Network? *Advances in Neural Information Processing Systems* 35 (2022), 37991–38002.
[51] Zhenhuan Yang, Yan Lok Ko, Kush R Varshney, and Yiming Ying. 2023. Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11909–11917.
[52] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. 2022. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*. PMLR, 25407–25437.
[53] Haotian Ye, James Zou, and Linjun Zhang. 2023. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 8968–8990.
[54] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.
[56] Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. 2022. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*. PMLR, 27203–27221.
[57] D Zietlow, M Lohaus, G Balakrishnan, M Kleindessner, F Locatello, B Scholkopf, and C Russell. [n. d.]. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In 2022 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10400–10411.
[58] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. 2022. MEDFAIR: Benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725* (2022).

A DATASETS USED IN SECTION 3.2

The UTKFace dataset [54] is a comprehensive facial dataset annotated with age (0 to 116 years), gender (male and female), and ethnicity (White, Black, Asian, Indian, and Others). For our analysis, we concentrate on gender classification (y = gender) and simplify ethnicity (a = race) by consolidating it into binary categories: White and non-White. We then categorize the data into four groups based on these attributes: White males ($g_{0,0}$), White females ($g_{0,1}$), non-White males ($g_{1,0}$), and non-White females ($g_{1,1}$), with respective sizes of 3801, 651, 482, and 4723. The majority groups are $g_{0,0}$ and $g_{1,1}$, the minority groups are $g_{0,1}$ and $g_{1,0}$.

In the CelebA dataset, the objective is to predict attractiveness with gender as the sensitive attribute, paralleling the task outlined in section 5.1. The dataset is segmented into four groups based on these criteria: non-attractive females ($g_{0,0}$), non-attractive males ($g_{0,1}$), attractive females ($g_{1,0}$), and attractive males ($g_{1,1}$), comprising 29916, 49242, 64581, and 19013 instances, respectively. The majority groups are $g_{0,1}$ and $g_{1,0}$, the minority groups are $g_{0,0}$ and $g_{1,1}$.

B COMPUTATIONAL COST

For a comprehensive evaluation of our methods, we assess performance and training time for the Waterbirds dataset. Experiments are performed on a single NVIDIA RTX 3090 GPU, with precautions taken to ensure no external processes interfere with GPU performance. As shown in Table 7, our method increases training time by 4.8 % but enhances Min-max fairness by a significant margin.

Table 7: Training time in the Waterbirds dataset.

	Training time (s)
ERM	2481.8
Ours	2601.8

C ADDITIONAL VISUALIZATION RESULTS

Additional GradCam visualization results are presented in Fig 10. Our method exhibits a targeted effectiveness on the objective, a distinct advantage not achieved by ERM training. This capability is particularly critical in biomedical and healthcare contexts, where nuanced distinctions can have significant implications.

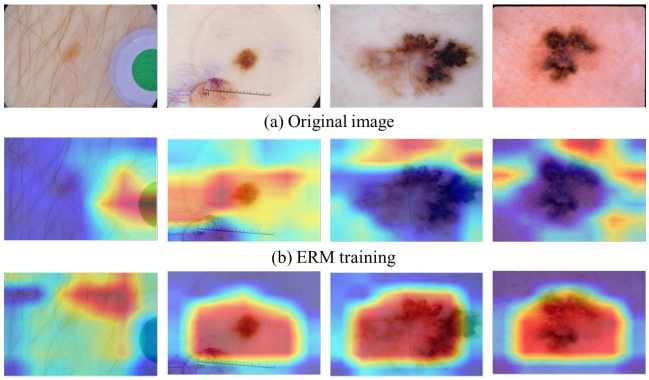

Figure 10: Grad-CAM visualization for exemplary test images in ISIC dataset.

D EVALUATION USING PARITY NOTION FAIRNESS METRICS.

For a comprehensive evaluation, we conducted experiments utilizing a variety of fairness metrics to assess the performance of each method. For the ISIC dataset, due to the lack of sufficient samples of malignant diagnostics with color patches, we are unable to compute the parity notion fairness criterion on this dataset.

In our experiments, we incorporate several fundamental fairness notions: Demographic Parity (DP) [13], Equal Opportunity (EOp) [17], and Equalized Odds (EOd) [17]. The results are presented in Tables 8, 9, 10, and 11. Observations indicate that although our method is primarily designed to optimize for min-max fairness, it also demonstrates enhanced performance under the parity notion fairness criterion.

Table 8: Fairness evaluation in the Waterbirds dataset.

	DP↓	EOp↓	EOd ↓	ACC ↑
ERM	15.87 ± 0.79	15.06 ± 0.26	15.58 ± 0.51	<u>91.84 ± 0.49</u>
HSIC	15.98 ± 2.33	15.94 ± 0.96	15.97 ± 1.83	90.76 ± 1.31
SUBG	**1.06 ± 0.24**	**0.16 ± 0.01**	**0.77 ± 0.14**	91.15 ± 0.72
MMSGD	<u>2.61 ± 1.25</u>	1.56 ± 1.40	<u>2.05 ± 1.28</u>	90.16 ± 0.36
MMF	4.03 ± 1.94	3.01 ± 1.85	3.71 ± 1.82	90.25 ± 0.80
GDRO	5.28 ± 0.58	2.86 ± 1.09	4.41 ± 0.75	91.13 ± 0.24
Ours	3.60 ± 0.27	<u>0.62 ± 0.01</u>	2.54 ± 0.16	**92.04 ± 0.20**

Table 9: Fairness evaluation in the CelebA dataset.

	DP↓	EOp↓	EOd ↓	ACC ↑
ERM	38.53 ± 4.02	15.10 ± 2.97	15.06 ± 3.71	**81.94 ± 0.94**
HSIC	32.92 ± 4.26	12.44 ± 2.82	10.77 ± 3.50	79.93 ± 1.39
SUBG	26.22 ± 2.77	2.97 ± 2.58	3.38 ± 2.16	79.63 ± 0.32
MMSGD	30.69 ± 0.88	4.51 ± 1.43	6.72 ± 0.94	80.45 ± 0.23
MMF	24.03 ± 1.57	2.47 ± 1.93	2.62 ± 0.53	79.58 ± 0.35
GDRO	**21.83 ± 2.11**	2.79 ± 1.83	2.17 ± 1.62	79.17 ± 1.11
Ours	24.16 ± 0.41	**1.68 ± 1.09**	**1.14 ± 0.41**	80.19 ± 0.36

Table 10: Fairness evaluation in the MultiNLI Dataset.

	DP↓	EOp↓	EOd ↓	ACC ↑
ERM	47.66 ± 0.43	15.43 ± 0.77	12.93 ± 0.76	81.14 ± 0.49
HSIC	47.97 ± 0.43	16.04 ± 1.70	13.11 ± 0.75	**81.34 ± 0.59**
SUBG	**32.87 ± 3.28**	10.19 ± 3.20	6.43 ± 0.83	71.37 ± 0.60
MMSGD	46.82 ± 1.11	14.41 ± 1.24	12.45 ± 1.12	80.82 ± 0.32
MMF	40.52 ± 4.07	11.91 ± 3.23	8.37 ± 2.97	77.08 ± 1.14
GDRO	39.36 ± 1.74	6.62 ± 1.46	5.56 ± 0.97	80.53 ± 0.10
Ours	38.37 ± 0.54	**6.29 ± 0.40**	**4.95 ± 0.74**	80.96 ± 0.23

Table 11: Fairness evaluation in the Adults dataset.

	DP↓	EOp↓	EOd ↓	ACC ↑
ERM	19.71 ± 0.69	9.23 ± 0.77	9.15 ± 0.63	**84.34 ± 0.12**
HSIC	19.62 ± 0.18	7.34 ± 0.22	8.22 ± 0.15	84.11 ± 0.24
SUBG	18.54 ± 0.20	3.52 ± 0.42	6.28 ± 0.12	80.62 ± 0.03
MMSGD	**16.05 ± 0.85**	1.73 ± 0.92	**3.98 ± 0.22**	83.74 ± 0.22
MMF	19.19 ± 0.74	**1.55 ± 0.80**	5.44 ± 0.21	80.48 ± 0.11
GDRO	20.50 ± 0.11	1.89 ± 0.14	5.97 ± 0.11	81.16 ± 0.02
Ours	19.44 ± 0.10	1.57 ± 0.48	5.57 ± 0.22	81.00 ± 0.14

E ETHICAL STATEMENT

In our use of the CelebA dataset, the "attractiveness" label serves only as a benchmark for assessing and contrasting the efficacy of existing methods. According to [16], employing this task for evaluating algorithmic fairness is deemed appropriate. We want to emphasize that our objective is not to assess or define individual attractiveness through this dataset. Instead, our goal is to advance the development of machine learning algorithms that are imposed with an awareness of fairness.