Baltic J. Modern Computing, Vol. 13 (2025), No. 1, pp. 96–127
https://doi.org/10.22364/bjmc.2025.13.1.06

Advancing Equal Opportunity Fairness and Group Robustness through Group-Level Cost-Sensitive Deep Learning

Modar SULAIMAN, Nesma Talaat Abbas MAHMOUD, Kallol ROY

University of Tartu, Institute of Computer Science, Tartu, Estonia

modar.sulaiman@ut.ee, nesma.mahmoud@ut.ee, kallol.roy@ut.ee

ORCID: 0000-0002-7322-078X, ORCID: 0000-0002-5618-2953,
ORCID: 0000-0002-6557-2689

Abstract. Machine learning systems deployed in high-stakes decision-making scenarios increasingly face challenges related to fairness, spurious correlations, and group robustness. These systems can perpetuate or amplify societal biases, particularly affecting protected groups defined by sensitive attributes such as race or age. This paper introduces a novel cost-sensitive deep learning approach at different group levels that simultaneously addresses these interconnected challenges. Thus, our research uncovers a fundamental synergy between group robustness and group fairness. By developing a technique that enhances group fairness, we also improve the model's group robustness to spurious correlations. This approach encourages the model to focus on causally relevant features rather than misleading associations. We propose a comprehensive methodology that specifically targets group-level class imbalances, a crucial yet often overlooked aspect of model bias. By incorporating different misclassification costs at the group level, our approach, Group-Level Cost-Sensitive Learning (GLCS), provides a principled framework for handling both dataset-wide and group-specific class imbalances using different constraints in an optimization framework. Through targeted interventions for underrepresented subgroups, we demonstrate simultaneous improvements in equal opportunity fairness and worst-group performance, ensuring similar true positive rates across demographic groups while strengthening overall group robustness. Extensive empirical evaluation across diverse datasets (CelebA, UTKFace, and CivilComments-WILDS) demonstrates that our method effectively mitigates performance disparities and promotes more equitable outcomes without sacrificing overall model accuracy. These findings present evidence that addressing fundamental data distribution issues at the group level can naturally lead to fairer and more robust machine learning systems. Our work has significant implications for the ethical deployment of machine learning in critical domains such as healthcare, finance, and criminal justice, offering a practical path toward more equitable and reliable automated decision-making systems.

Keywords: Fairness in Machine Learning, Group Robustness, Cost-Sensitive Deep Learning, Bias.

1 Introduction

Machine learning systems have become increasingly prevalent in high-stakes decision-making scenarios, from lending and hiring to healthcare and criminal justice (Barocas et al., 2023; Chouldechova and Roth, 2020). However, these automated systems can perpetuate or amplify existing societal biases, leading to discriminatory outcomes against protected groups defined by sensitive attributes such as race or gender (Mehrabi et al., 2021). This has prompted extensive research in fair machine learning, which aims to develop algorithms that maintain high predictive performance while ensuring equitable treatment across different demographic groups (Du et al., 2020). Various fairness metrics and mitigation strategies have emerged, including statistical parity (Dwork et al., 2012), equal opportunity (Hardt et al., 2016), and individual fairness (Dwork et al., 2012). These approaches fall into three categories: pre-processing techniques that modify training data, in-processing methods that incorporate fairness constraints during model training, and post-processing approaches that adjust model outputs (Caton and Haas, 2024). Despite these advances, achieving fairness while maintaining model performance remains challenging, due to multiple competing criteria (Kleinberg et al., 2016). Moreover, the context-dependent nature of fairness requires careful consideration of domain-specific requirements and societal implications when designing and deploying fair ML systems (Selbst et al., 2019).

A fundamental challenge in achieving fairness lies in the presence of spurious correlations, where machine learning models inadvertently learn misleading associations between features and outcomes that do not reflect true causal relationships. These correlations arise from various sources, including sampling biases, historical data imbalances, or coincidental patterns in training datasets. As a result, the impact of spurious correlations extends beyond mere performance issues, as models that rely on spurious features can inadvertently perpetuate or amplify existing societal biases, leading to discriminatory outcomes that disproportionately affect minority groups. The concept of group robustness (which involves training models to achieve strong performance across all predefined groups within the dataset), measured by the minimum accuracy across all groups (worst-accuracy), is also not immune to these challenges. Models trained with standard empirical risk minimization (ERM) often exhibit poor performance on underrepresented groups due to both geometric and statistical skew on the input training data.

This paper investigates the fundamental synergy between group robustness and group fairness in machine learning models. We demonstrate that our approach, designed to enhance the group fairness metric, also boosts group robustness by ensuring consistent performance across all subgroups. This addresses the common problem of models exploiting spurious patterns that unfairly impact minority groups. We propose a novel Group-Level Cost-Sensitive Framework (GLCS) to address these challenges. Our proposed GLCS framework leverages cost-sensitive deep learning (Khan et al., 2017; Zhou and Liu, 2005) and addresses class imbalance challenges by explicitly incorporating misclassification costs at the group level into the learning process. Our proposed methodology differs fundamentally from conventional techniques of random oversampling, undersampling, or synthetic minority oversampling (SMOTE) by modifying the underlying learning objective rather than manipulating the dataset distribution.

The main contributions of the paper are laying the foundations at the intersection of cost-sensitive deep learning, group fairness, and group robustness in machine learning. The key contributions are outlined as follows:

1. **Novel Group-Level Cost-Sensitive Framework (GLCS):** We introduce a pioneering cost-sensitive deep learning framework that addresses group-level class imbalances, enabling more nuanced handling of demographic disparities in machine learning systems.
2. **Enhanced Fairness and Robust Performance Mechanisms:** The proposed novel cost-sensitive optimization technique GLCS mitigates performance disparities on diverse datasets (CelebA, UTKFace, and CivilComments-WILDS) by strategically balancing group-level representations, thereby improving equal opportunity fairness without compromising overall model accuracy.
3. **Comprehensive Empirical Validation:** Our extensive experimental results validate the generalizability and effectiveness of our approach, showcasing consistent improvements in both group robustness and group fairness, with a particular focus on equal opportunity.

The paper is structured as follows: Section 2 presents a comprehensive review of related work in group fairness, group robustness, fairness and class imbalance, threshold optimization, and cost-sensitive learning. Section 3 establishes the necessary preliminaries and theoretical foundations. Section 4 introduces our Group-Level Cost-Sensitive (GLCS) framework, detailing its mathematical formulation and implementation. Section 5 describes the datasets and baselines employed in our experimental evaluation, while Section 6 outlines our evaluation metrics. Section 7 details the experimental setup and implementation details. Finally, Section 8 presents our results and discusses their implications for group fairness and group robustness in machine learning.

2 Related Works:

Fairness in Machine Learning. The debiasing techniques strategically categorized into three primary approaches: pre-processing, in-processing, and post-processing methods (Wan et al., 2023). While pre-processing and post-processing techniques offer pragmatic interventions, our research specifically focuses on in-processing debiasing methods, which have garnered substantial scholarly attention for their sophisticated approach of embedding fairness constraints directly. These intrinsic fairness techniques, pioneered by seminal works from (Dwork et al., 2012; Hashimoto et al., 2018; Kearns et al., 2018), represent a paradigm shift towards algorithmically engineered fairness. (Zafar et al., 2019) studies constrained optimization by incorporating fairness measures as regularisation terms or constraints.

Cost-Sensitive Learning. Cost-sensitive learning adaptively weighs the importance of different classes during the training process. This is typically achieved through the modification of the loss function in neural networks. The approach is effective in real-world applications of medical diagnosis, fraud detection, or rare event prediction, where misclassification costs are inherently asymmetric. Recent developments in this area have

introduced several innovative methodologies. (Zhou and Zhang, 2016) employed cost-sensitive learning to mitigate the problem of misclassifications of minority or critical classes. The class-balanced loss function (Cui et al., 2019) addresses the long-tailed distribution problem by introducing a weighting factor that is inversely proportional to the effective number of samples. Margin-based approaches (Cao et al., 2019) focus on enhancing the decision boundary's quality by incorporating cost-sensitivity into the margin requirements. Additionally, (Sangalli et al., 2021) uses constrained optimization to train neural networks to improve neural network performance on critical and under-represented classes.

Fairness and Class Imbalance. The intricate relationship between fairness and class imbalance has emerged as a critical research domain in machine learning, with scholars developing sophisticated methodologies to address simultaneous challenges of bias mitigation and distributional disparities. (Dablain et al., 2022) introduced Fair Over-Sampling (FOS), a pioneering approach that simultaneously addresses class imbalance and protected feature bias by generating synthetic minority class instances while encouraging classifiers to minimize reliance on sensitive attributes. Complementing this work, (Hirzel and Ram, n.d.) developed Orbis, an adaptable oversampling algorithm capable of fine-tuned optimization across fairness and accuracy dimensions. (Yan et al., 2020) critically demonstrated how conventional balancing techniques can inadvertently exacerbate unfairness, introducing a novel fair class balancing method that enhances model fairness without explicit sensitive attribute manipulation. (Tarzanagh et al., 2023) advanced this discourse through a tri-level optimization framework incorporating local, fair, and class-balanced predictors, theoretically demonstrating improved classification and fairness generalization. (Subramanian et al., 2021) further expanded these investigations by evaluating long-tail learning methods across sentiment and occupation classification domains, empirically validating fairness enforcement techniques' effectiveness in mitigating demographic biases and class imbalance. (Shui et al., 2022) contributed a principled bilevel objective approach, demonstrating an innovative method for developing fair predictors that simultaneously manage group sufficiency and generalization error.

Group Robustness. Recent machine learning research has developed sophisticated methods to address the problem of group robustness. (Sagawa et al., 2019) introduced Group Distributionally Robust Optimization (Group-DRO), which optimizes a soft version of the worst-group loss. (Liu et al., 2021) proposed Just Train Twice (JTT), a method that employs a two-stage training strategy: Initially, a standard ERM model is trained for several epochs. In the subsequent stage, a refined model is trained by upweighting the training examples that the initial ERM model misclassified. Complementing these approaches, Kirichenko et al.(Kirichenko et al., 2022) demonstrated through Deep Feature Reweighting (DFR) that simple last layer retraining can match or surpass state-of-the-art methods on spurious correlation benchmarks with significantly reduced computational complexity. Building upon this insight, (Qiu et al., 2023) developed Automatic Feature Reweighting (AFR), which retrains the last layer of ERM-trained model with a weighted loss that upweights minority group examples by emphasizing instances where the ERM model performs poorly.

Threshold Optimization. Threshold optimization in classification represents a sophisticated computational domain, with seminal works (Lipton et al., 2014; Koyejo et al., 2014; Sanchez, 2016) systematically exploring methodological approaches for determining optimal decision boundaries. Research has advanced through receiver operating characteristic (ROC) curve analysis for identifying optimal operating points (Freeman and Moisen, 2008), cost-sensitive threshold adjustment techniques that explicitly incorporate domain-specific loss functions and contextual constraints into the threshold selection process (Robles et al., 2020), unified theoretical frameworks (Hernández-Orallo et al., 2012) offering comprehensive computational strategies for threshold optimization that transcend traditional binary classification paradigms, and probabilistic methodologies for adaptive threshold determination which significantly enhancing the precision and reliability of predictive models across diverse computational domains (Kazemi et al., 2023), thereby providing a comprehensive approach to optimizing classification thresholds with nuanced consideration of performance, constraints, and contextual requirements. For a more detailed explanation of the threshold optimization method employed in our experiment, please refer to Section 7.3.

3 Mathematical Preliminaries

Definition 1 (Group Fairness). Group fairness is a fundamental concept in machine learning and algorithmic decision-making, particularly relevant when the outcomes affect individuals from different demographic or social groups. The aim of group fairness is to ensure that model predictions are equitable across groups defined by sensitive attributes such as gender, race, or religion.

Definition 2 (Equal Opportunity). It requires that a model achieves the same true positive rate (TPR) for different subgroups when considering only instances with a positive label (Hardt et al., 2016). Formally, it is defined as:

$$P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1),$$

where $\hat{Y}$ represents the predicted outcome, S is the sensitive attribute, and Y is the true label. This condition ensures that individuals from different groups who are actually positive (i.e., have a positive true label) have an equal probability of being classified as positive by the model.

Definition 3 (Equalized Odds). It extends the concept of Equal Opportunity by requiring that both the true positive rate (TPR) and the false positive rate (FPR) be equal across different groups (Hardt et al., 2016). It can be expressed as:

$$P(\hat{Y} = y|S = 0, Y = y) = P(\hat{Y} = y|S = 1, Y = y), \quad y \in \{0, 1\}.$$

This metric ensures that the model's performance is consistent across groups in terms of both correctly identifying positives and avoiding false positives. By examining these metrics, we aim to provide a thorough evaluation of fairness in our models, ensuring they treat all groups equitably.

Definition 4 (Group Robustness). It focuses on maintaining consistent and fair model performance across all subgroups (Liu et al., 2021; LaBonte et al., 2024). Group robustness optimizes strategies to focus on: (i) Identifying and mitigating spurious correlations. (ii) Optimizing performance for worst-performing groups. (iii) Maintaining high overall accuracy while improving minority group performance.

Definition 5 (Augmented Lagrangian Method (ALM)). The Augmented Lagrangian Method (ALM), also known as the method of multipliers, is a powerful optimization technique that bridges the gap between constrained and unconstrained optimization problems. Introduced by Bertsekas (1976).

Definition 6 (Class-based Partitioning). The dataset M can be partitioned into positive (critical) and negative classes as follows:

$$\begin{aligned} P &= \{x_i^p\}_{i=1}^{|P|} \quad \text{(positive class samples)} \\ N &= \{x_i^n\}_{i=1}^{|N|} \quad \text{(negative class samples)} \end{aligned} \tag{1}$$

where $|P| < |N|$, indicating P represents the minority class (Sangalli et al., 2021).

Definition 7 (Protected Attribute-based Partitioning). Let $s \in \{0, 1\}$ denote the protected attribute (e.g., gender or race). We partition the dataset M into two disjoint subsets based on this attribute:

$$\begin{aligned} Z_1 &= \{x_i^{s_1}\}_{i=1}^{|Z_1|} \quad \text{(group with } s = 1\text{)} \\ Z_0 &= \{x_i^{s_0}\}_{i=1}^{|Z_0|} \quad \text{(group with } s = 0\text{)} \end{aligned} \tag{2}$$

where in our discription we select $|Z_1| < |Z_0|$, establishing Z_0 as the majority group (non-protected group) and Z_1 is the minority group (protected group).

The further partitioning of each protected attribute group (Intersectional Subgroups) based on the true labels $y \in \{0, 1\}$ gives the following definitions.

Definition 8 (Protected Group Partitioning (Z_1)).

$$\begin{aligned} Z_{1,1} &= \{x_i^{s_1,y_1}\}_{i=1}^{|Z_{1,1}|} \quad \text{(positive class, } y = 1\text{)} \\ Z_{1,0} &= \{x_i^{s_1,y_0}\}_{i=1}^{|Z_{1,0}|} \quad \text{(negative class, } y = 0\text{)} \end{aligned}$$

where $|Z_{1,1}| < |Z_{1,0}|$, indicating $Z_{1,1}$ is the minority class within Z_1.

Definition 9 (Non-Protected Group Partitioning (Z_0)).

$$\begin{aligned} Z_{0,1} &= \{x_i^{s_0,y_1}\}_{i=1}^{|Z_{0,1}|} \quad \text{(positive class, } y = 1\text{)} \\ Z_{0,0} &= \{x_i^{s_0,y_0}\}_{i=1}^{|Z_{0,0}|} \quad \text{(negative class, } y = 0\text{)} \end{aligned}$$

where $|Z_{0,1}| < |Z_{0,0}|$, indicating $Z_{0,1}$ is the minority class within Z_0.

Group and Subgroup Size Relationship In our theoretical framework, while we initially establish the notational convention that $|Z_1| < |Z_0|$, $|Z_{1,1}| < |Z_{1,0}|$, and $|Z_{0,1}| < |Z_{0,0}|$, we acknowledge that group and subgroup size relationships can exhibit significant variation across different experimental contexts and datasets. Specifically, the relative size constraints may be inverted in certain scenarios, such that $|Z_{1,0}| < |Z_{1,1}|$, and/or $|Z_{0,0}| < |Z_{0,1}|$. Therefore, a critical preliminary step when applying the Group-Level Cost-Sensitive Deep Learning (GLCS) framework is to rigorously characterize and distinguish between minority and majority groups/subgroups to ensure appropriate methodological implementation.

4 Proposed Method

In this paper, building upon the work of (Sangalli et al., 2021), we propose an innovative method (GLCS) formulated as a constrained optimization system for achieving equal opportunity in classification. While (Sangalli et al., 2021) focused on imbalanced dataset classification using constraints (3a) and (3b), we extend their framework by introducing additional constraints (3c) and (3d) to explicitly enforce equal opportunity across protected groups expressed as:

$$\min_{\theta} F(\theta) \quad \text{subject to:} \tag{3a}$$

$$\sum_{k=1}^{|N|} \max\Big(0, -(f_\theta(x_j^p) - f_\theta(x_k^n)) + \delta\Big) = 0, \quad \forall j \in \{1, \ldots, |P|\} \tag{3b}$$

$$\sum_{k=1}^{|Z_{1,0}|} \max\Big(0, -(f_\theta(x_l^{s_1,y_1}) - f_\theta(x_k^{s_1,y_0})) + \delta\Big) = 0, \quad \forall l \in \{1, \ldots, |Z_{1,1}|\} \tag{3c}$$

$$\sum_{k=1}^{|Z_{0,0}|} \max\Big(0, -(f_\theta(x_r^{s_0,y_1}) - f_\theta(x_k^{s_0,y_0})) + \delta\Big) = 0, \quad \forall r \in \{1, \ldots, |Z_{0,1}|\} \tag{3d}$$

where: $f_\theta(\cdot) : \mathcal{X} \to [0, 1]$ is the DNN's output probability function, θ represents the DNN parameters and $\delta > 0$ is the margin parameter. The above constraints enforces three levels of discrimination prevention by (i) ensuring separation between positive and negative classes (ii) enforcing class separation within the protected and non-ptotected groups (iii) optimizing overall AUC performance. This hierarchical constraint system simultaneously addresses both class imbalance and equal opportunity objectives, ensuring consistent performance across all subgroups while maintaining strong overall classification performance. Subsequently, we derive an equivalent unconstrained form of the above constrained system defined in equations (3a)-(3d) using the augmented Lagrangian method (ALM):

$$\mathcal{L}_\mu(\theta,\lambda) = F(\theta) + \underbrace{\frac{\mu_1}{2|P||N|}\sum_{j=1}^{|P|} q_j^2 + \frac{1}{|P||N|}\sum_{j=1}^{|P|}\lambda_j q_j}_{\text{Global Class Separation}} + \underbrace{\frac{\mu_2}{2|Z_{1,1}||Z_{1,0}|}\sum_{l=1}^{|Z_{1,1}|} q_l^2 + \frac{1}{|Z_{1,1}||Z_{1,0}|}\sum_{l=1}^{|Z_{1,1}|}\lambda_l q_l}_{\text{Class Separation Within Protected Group}} + \underbrace{\frac{\mu_3}{2|Z_{0,1}||Z_{0,0}|}\sum_{r=1}^{|Z_{0,1}|} q_r^2 + \frac{1}{|Z_{0,1}||Z_{0,0}|}\sum_{r=1}^{|Z_{0,1}|}\lambda_r q_r}_{\text{Class Separation Within Non-Protected Group}} \tag{4}$$

where the constraint violations q are defined as:

$$q_j = \sum_{k=1}^{|N|} \max\left(0, -(f_\theta(x_j^p) - f_\theta(x_k^n)) + \delta\right) \qquad \text{(global)}$$

$$q_l = \sum_{k=1}^{|Z_{1,0}|} \max\left(0, -(f_\theta(x_l^{s_1,y_1}) - f_\theta(x_k^{s_1,y_0})) + \delta\right) \qquad \text{(protected)}$$

$$q_r = \sum_{k=1}^{|Z_{0,0}|} \max\left(0, -(f_\theta(x_r^{s_0,y_1}) - f_\theta(x_k^{s_0,y_0})) + \delta\right) \qquad \text{(non-protected)}$$

Here, $\mu_1, \mu_2, \mu_3 > 0$ are penalty coefficients for quadratic terms; $\lambda_j, \lambda_l, \lambda_r$ are Lagrange multipliers for positive samples in respective groups and $\delta > 0$ is the margin parameter. This unconstrained formulation facilitates (i) asymmetric treatment by different handling of positive and negative classes in M, Z_1, and Z_0, reflecting their relative importance, (ii) performance focus by prioritizing reduction of False Positive Rate (FPR) at high True Positive Rate (TPR) regions for all groups.

5 Datasets and Baselines

Our empirical evaluation leverages three widely-recognized datasets in fairness-aware machine learning research: CelebA (Liu et al., 2015) and UTKFace (Zhang et al., 2017) for facial attribute analysis and demographic fairness assessment, and CivilComments-WILDS (Koh et al., 2021) for evaluating group robustness under distribution shifts. These datasets were selected for their diverse data modalities and comprehensive demographic annotations, enabling rigorous evaluation of both algorithmic fairness and group robustness. The facial analysis datasets present unique challenges through their demographic distributions and attribute correlations, while CivilComments-WILDS offers extensive toxic comment classifications across varied demographic groups. This diverse dataset selection facilitates thorough validation of our proposed technique across multiple domains and fairness criteria.

5.1 CelebA Dataset

The CelebFaces Attributes (CelebA) dataset (Liu et al., 2015) comprises 202,599 celebrity images with 40 binary attribute annotations, establishing itself as a benchmark dataset in fairness-aware machine learning research (Han et al., 2024). Following the Fair Fairness Benchmark (FFB) preprocessing protocol (Han et al., 2024), we focus on the binary classification task of "Wavy Hair" (y) prediction with "Gender" (s) as the protected attribute. The dataset is split into training (80%, 162,770 samples), validation (10%, 19,867 samples), and test sets (10%, 19,962 samples). Table 1 presents the distribution statistics across gender groups ($s = 1$ for male, $s = 0$ for female) and target attributes ($y = 1$ for wavy hair presence), providing crucial insights into potential data distribution biases.

Target Attribute Distribution (Wavy Hair)		
Positive Class ($y = 1$)		51,982
Negative Class ($y = 0$)		110,788
Protected Attribute Distribution (Gender)		
Male ($s = 1$)		68,261
Female ($s = 0$)		94,509
Intersectional Distribution		
Male with Wavy Hair	$P(s = 1 \mid y = 1)$	9,762
Male without Wavy Hair	$P(s = 1 \mid y = 0)$	58,499
Female with Wavy Hair	$P(s = 0 \mid y = 1)$	42,220
Female without Wavy Hair	$P(s = 0 \mid y = 0)$	52,289

Table 1. CelebA Dataset Statistics and Demographic Distribution

5.2 UTKFace Dataset

The UTKFace dataset (Zhang et al., 2017) contains over 20,000 facial images annotated with age, gender, and ethnicity attributes, making it particularly suitable for investigating intersectional fairness in facial analysis tasks. The dataset exhibits balanced distributions across major demographic factors, with 12,661 young and 11,044 old subjects, and near-equal gender representation (12,391 male, 11,314 female). Table 2 reveals notable age-gender interactions, with males showing higher representation in older age groups (6,854 vs. 5,537) and females in younger groups (7,124 vs. 4,190). Following (Han et al., 2024), images were standardized to 48x48 pixels with 3 color channels and partitioned into training (18,964 samples), validation (2,371 samples), and test sets (2,370 samples), enabling robust evaluation of algorithmic fairness across demographic intersections.

Demographic	Age Group		Total
Group	Old (y = 1)	Young (y = 0)	Sample
Male	6,854	5,537	12,391
Female	4,190	7,124	11,314
Total	11,044	12,661	23,705

Table 2. Demographic Distribution of Age Categories in UTKFace Dataset

Split	Number of Comments	Distribution (%)
Training	269,038	59.79
Validation	45,180	10.04
Test	133,782	29.73
Total	450,000	100.00

Table 3. Data Distribution in CivilComments-WILDS Dataset

5.3 CivilComments-WILDS Dataset

The CivilComments-WILDS dataset (Koh et al., 2021), derived from (Borkan, Dixon, Sorensen, Thain and Vasserman, 2019), contains 450,000 online comments annotated for toxicity and eight demographic identity mentions (gender (male, female), sexual orientation (LGBTQ), race (black, white), and religion (Christian, Muslim, or other)). This dataset is particularly valuable for studying group robustness due to potential spurious correlations between demographic mentions and toxicity labels. Following (Koh et al., 2021), we define 16 overlapping groups—(a, toxic) and (a, non-toxic) for each demographic identity a. Table 3 illustrates the distribution of comments across the dataset splits. Our analysis using the Empirical Risk Minimization (ERM) approach identified comments mentioning Christian identity as the worst-performing group, which we subsequently designated as the sensitive attribute in our GLCS framework. The effectiveness of our proposed GLCS method is assessed using worst-group accuracy metrics, detailed in Section 6, with implementation specifics discussed in Sections 5.5 and 7.5.

5.4 Baselines for group Fairness with CelebA and UTKFace Datasets

We compare our proposed GLCS method against the following baselines for group fairness on CelebA and UTKFace datasets: Empirical Risk Minimization (ERM) (Vapnik, 1991) and DiffEopp (Chuang and Mroueh, 2021; Hardt et al., 2016). ERM is a foundational machine learning technique that aims to minimize the empirical risk on the training dataset (Vapnik, 1991). ERM focuses on optimizing the performance of the model on the observed data, often without considering fairness constraints. On the other hand, DiffEopp is a gap regularization method to address the equal opportunity criterion. DiffEopp ensures that the true positive rates are equal across different demographic groups. In this paper, we utilize the implementation of DiffEopp as provided in the Fair Fairness Benchmark (FFB) (Han et al., 2024). These baselines, ERM and DiffEopp, were selected primarily to thoroughly assess the performance of our GLCS

framework in terms of both predictive accuracy and equal opportunity fairness. This comprehensive evaluation allows us to clearly demonstrate the advantages and trade-offs of our approach.

5.5 Baseline Methods for Group Robustness in CivilComments-WILDS Dataset

In our comprehensive investigation of group robustness, we evaluate several state-of-the-art methods for mitigating performance disparities across demographic groups. Our baseline approaches encompass a range of sophisticated techniques: (ERM), Just Train Twice (JTT) (Liu et al., 2021); Deep Feature Reweighting (DFR) (Kirichenko et al., 2022); Automatic Feature Reweighting (AFR) (Qiu et al., 2023); and Group Distributionally Robust Optimization (Group-DRO) (Sagawa et al., 2019).

6 Metrics

In this section, we discuss different metrics used in our experiments to validate the efficacy of our proposed GLCS approach. The metrics are classified mainly into (i) Threshold-Agnostic Performance Metrics (ii) Threshold-Dependent Performance Metrics (iii) Group Fairness Metrics (iv) Nuanced Metrics (v) Group Robustness Metric.

Threshold-Agnostic Performance Metrics. In the evaluation of binary classification models, several threshold-agnostic performance metrics provide comprehensive insights into machine learning models behavior and efficacy. *Precision-Recall Area Under the Curve (PR-AUC)*, *Receiver Operating Characteristic Area Under the Curve (ROC-AUC)*, *Brier score*, and *AUC-PR Gain* are such pivotal metrics. PR-AUC is important in scenarios with class imbalance, as it focuses on the positive class and is less affected by a large number of true negatives. ROC-AUC measures the model's ability to discriminate between classes by plotting the true positive rate against the false positive rate at various threshold settings. Brier Score measures the mean squared difference between the predicted probability and the actual outcome. AUC-PR Gain gives an improvement over to traditional PR analysis by introducing normalized gain metrics that enable more meaningful model comparisons (Flach and Kull, 2015).

Threshold-Dependent Performance Metrics. Classification metrics in binary prediction tasks inherently depend on the chosen decision threshold. This dependency becomes particularly crucial in cost-sensitive learning scenarios and imbalanced datasets, where optimal thresholds may vary significantly across different models. Our comprehensive analysis of threshold optimization techniques, detailed in Section 7.3, addresses these challenges. This methodological approach ensures equitable model comparisons while reflecting real-world operational requirements. The systematic examination of threshold optimization not only strengthens the validity of our experimental results but also contributes to the broader discourse on performance evaluation in group fairness and robustness assessment. The prominent threshold-dependent performance metrics used in our experiments are the following (i) Balanced Accuracy (ii) F1 Score (iii) Matthews Correlation Coefficient (MCC) (iv) Precision (v) Recall.

Group Fairness Metrics. We utilized different group fairness metrics to evaluate our technique. We used the fairness metrics that are implemented in FAIR FAIRNESS BENCHMARK (FFB) (Han et al., 2024). FFP specifically designed to evaluate different in-processing debiasing methods. In our experiments, we have used the following metrics: (i) Equality of Opportunity (eopp) (ii) Demographic Parity (dp) (iii) p-Rule (prule) (iv) Equalized Odds (eodd) (iv) ROC AUC Parity (aucp) (v) Balance for Negative Class (bfn) (vi) Balance for Positive Class (bfp) (vii) Area Between CDF Curves (abcc). (Han et al., 2024).

Fairness Metrics Notation. For the group fairness metrics we adopt a systematic notation that distinguishes between probability-based and threshold-based evaluations. When utilizing output probability estimates, we denote demographic parity, equal opportunity, equalized odds, and p-Rule as *dpe*, *eoppe*, *eodde*, and *prulee*, respectively. Conversely, when evaluating binary predictions derived from threshold-based classification, these metrics are denoted as *dp*, *eopp*, *eodd*, and *prule*. This notational convention aligns with the experimental framework and code implementation used in FFB (Han et al., 2024), providing consistency in metric interpretation across probability and binary domains.

Nuanced Metrics. Nuanced metrics subgroup-AUC, BPSN-AUC, and BNSP-AUC (Borkan, Dixon, Li, Sorensen, Thain and Vasserman, 2019; Borkan, Dixon, Sorensen, Thain and Vasserman, 2019) provide a threshold-agnostic assessment in machine learning models. Those metrics are used to identify various types of biases. They divide the data into two subgroups: (i) one representing groups which contains both positive and non-positive elements and (ii) another representing a background group. Specifically, the subgroup-AUC, BPSN-AUC, and BNSP-AUC were used to measure the bias minimization performance of the model for individual identity subgroups on datasets, CelebA, UTKFace.

Worst-Group Accuracy Metric. It is defined as the lowest accuracy observed across all the groups. A higher worst-group accuracy value suggests the machine learning models are less likely to mistakenly associate demographic identities with toxicity (Koh et al., 2021).

7 Experimental Setting

7.1 Neural Network Architecture

For our experimental framework, we adopt the ResNet-18 architecture (He et al., 2016) as the backbone network, following the implementation detailed in Fair Fairness Benchmark (FFB) (Han et al., 2024). This architecture serves as the foundation for all experiments conducted on the CelebA and UTKFace image datasets with ERM, GLCS and DiffEopp. For the CivilComments-WILDS dataset, our proposed GLCS framework leverages the BERT model (Devlin, 2018). The baseline comparisons, including ERM, DFR, Group-DRO, JTT, and AFR, maintain consistency with the implementations specified in (Qiu et al., 2023) for the CivilComments-WILDS experiments, ensuring a fair comparative analysis.

7.2 Early Stopping Criterion

The challenge of determining optimal stopping criteria in fair machine learning is particularly complex due to the inherent trade-offs between multiple competing objectives: group robustness, model utility, and fairness metrics. This critical aspect of training has received limited attention in the existing literature. Han et al. (2024) proposed a deterministic stopping strategy in their Fair Fairness Benchmark (FFB) framework based on learning rate decay. In contrast, Sulaiman et al. (2024) employed a more empirical approach in their work as follows: (1) monitor model performance on the validation set. (2) evaluate multiple metrics simultaneously (utility and fairness metrics). (3) Stop training when a satisfactory trade-off is achieved within early epochs. In our experiments, we follow the approach proposed by Sulaiman et al. (2024).

7.3 Classification Thresholds in the Experimental Datasets:

Binary classification in deep learning confronts significant challenges when applied to imbalanced datasets, where conventional threshold-setting strategies fail to capture nuanced distributional complexities. For example, CelebA dataset exemplifies this critical challenge, presenting a stark class distribution disparity with 32% positive instances (51,982 samples) against 68% negative instances (110,788 samples), systematically challenging traditional machine learning paradigms. Our empirical analysis reveals the inherent limitations of the standard 0.5 threshold, which presupposes uniform class representation—a premise fundamentally misaligned with real-world data characteristics. By recalibrating the classification threshold to approximately 0.32, we demonstrate a principled approach to mitigating class imbalance that enhances minority class sensitivity and improves overall predictive performance. Moreover, our Group-Level Cost-Sensitive (GLCS) approach introduces a sophisticated probabilistic framework that recalibrates class boundaries, fundamentally challenging traditional binary classification paradigms. Unlike conventional methods, our approach explicitly accommodates group-level heterogeneity by implementing constraint mechanisms that transform class separation strategies across distinct demographic or feature-based subgroups. By developing a flexible threshold optimization strategy, we enable a more granular and contextually responsive machine learning model that can adjust its decision boundaries to reflect the intricate complexities of real-world data representations. Therefore, the selection of an optimal classification threshold necessitates a sophisticated multi-dimensional analysis that integrates statistical techniques such as receiver operating characteristic (ROC) curve evaluation, F1 score maximization, and precision-recall curve assessment. Domain-specific considerations fundamentally modulate threshold selection—for instance, medical diagnostics prioritize sensitivity to minimize false negatives, while cybersecurity applications might emphasize precision to mitigate false positive risks.

7.3.1 Implementation and Empirical Methodology. Our rigorous threshold optimization framework is underpinned by the sophisticated `binclass-tools` package[1]

[1] https://github.com/lucazav/binclass-tools

on both CelebA and UTKFace datasets, a comprehensive computational toolkit designed for advanced binary classification analysis. The implementation follows a meticulously structured empirical protocol that systematically addresses the complexities of threshold optimization across diverse machine-learning models. We commence by training models using multiple approaches, including established baselines and our proposed GLCS, which enables a comprehensive comparative analysis. The methodology involves generating nuanced probability distributions for each model, allowing for granular insight into predictive performance characteristics. Leveraging the `binclass-tools` package, we use it to determine optimal classification thresholds. Our evaluation protocol rigorously assesses model performance using these optimized thresholds, employing consistent and theoretically grounded evaluation criteria to ensure methodological integrity. This systematic approach not only facilitates a fair and comprehensive comparison across different methodological approaches but also maintains a principled framework for cost-sensitive learning and group fairness.

7.4 Calibrating Neural Networks.

Deep neural networks have demonstrated remarkable discriminative performance across various tasks; however, their probability estimates often lack proper calibration, potentially leading to overconfident or underconfident predictions. A well-calibrated model should produce probability estimates that reflect true empirical frequencies—for instance, among predictions with confidence of 0.8, approximately 80% should be correctly classified. Calibration is particularly crucial in high-stakes applications where reliable uncertainty quantification is essential for decision-making processes.

In our experiments, we employ Temperature Scaling (Guo et al., 2017), a simple yet effective post-processing calibration technique that can be applied to the logits while preserving the model's discriminative capabilities. Given our model's output probabilities $p_i \in [0, 1]$, we first convert these to logits through the inverse sigmoid function: $z_i = \log(\frac{p_i}{1-p_i})$. Temperature scaling then modifies these logits by introducing a temperature parameter $T > 0$, and the calibrated probabilities are computed as follows:

$$\hat{p}_i = \sigma(z_i/T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \quad T > 0$$

The optimal temperature parameter T^* is learned by minimizing the negative log-likelihood (NLL) on a held-out validation set. This approach is particularly advantageous as it maintains the model's ranking performance metrics (e.g., AUC-ROC) due to the monotonic nature of the temperature scaling operation. Furthermore, the optimization of a single parameter T reduces the risk of overfitting compared to more complex calibration methods, while effectively addressing both under-confidence and over-confidence in model predictions. Our experimental results demonstrate that this calibration method successfully improves the reliability of probability estimates while maintaining the model's discriminative performance and fairness properties. This is particularly important as well-calibrated probabilities enable more reliable decision-making processes and better interpretation of model confidence.

7.5 GLCS Methodology with CivilComments-WILDS Dataset

Our comprehensive research methodology systematically addresses group fairness challenges in the CivilComments-WILDS dataset through a carefully designed experimental protocol. We commenced by conducting an initial baseline evaluation using Empirical Risk Minimization (ERM) to train our model, which enabled us to comprehensively assess the baseline performance and identify group-specific disparities, critically revealing the Christian demographic group as experiencing the most significant accuracy degradation (Section 8.3.2). Leveraging these insights, we subsequently applied our proposed approach (GLCS) with a targeted intervention focused on improving fairness specifically for the underperforming Christian group. To validate the effectiveness of our approach, we performed a rigorous comparative analysis, benchmarking our GLCS method against state-of-the-art baseline techniques commonly employed for enhancing group robustness (Section 5.5). This methodical approach allows us to systematically demonstrate the potential of our proposed method in mitigating group-based performance disparities within complex machine learning fairness challenges and improving group robustness.

7.6 Hyperparameters

In this section, we elucidate the hyperparameter configurations employed across different datasets, highlighting our methodical approach to parameter selection.

CelebA and UTKFace Datasets. For the GLCS framework proposed in Section 4, we selected hyperparameters based on the unique characteristics of each dataset. On the CelebA dataset, characterized by substantial intersectional distribution disparities (as evidenced in Table 1), we employed $\mu_1 = \mu_2 = \mu_3 = 0.5$ and $\delta = 0.2$. Conversely, the UTKFace dataset, exhibiting minimal group distribution variations (detailed in Table 2), warranted a more nuanced approach with $\mu_1 = \mu_2 = \mu_3 = \epsilon$ and $\delta = 0.1$, where $\epsilon \approx 0$. For comparative methods, namely ERM and DiffEopp, we consistently utilized the hyperparameters established in the Fair Fairness Benchmark (FFB) (Han et al., 2024) across both datasets.

CivilComments-WILDS Dataset. In the context of the CivilComments-WILDS dataset, we configured the GLCS method with $\mu_1 = \mu_2 = \mu_3 = 1$ and $\delta = 0.2$. For comparative methods, we utilized the hyperparameters for Just Train Twice (JTT) as specified in (Liu et al., 2021) and those for ERM, AFR, DFR, and Group-DRO following (Qiu et al., 2023). Consistent with prior work (Qiu et al., 2023; Liu et al., 2021), we applied the standard 0.5 threshold for metric evaluation across all methods.

8 Results and Analysis

We present a comprehensive analysis of our proposed GLCS approach compared with baseline methods on three datasets: CelebA, UTKFace, and CivilComments-WILDS. This comparison enables us to assess the efficiency of GLCS and Calibrated GLCS (after applying Temperature Scaling to our GLCS model) in achieving a balanced trade-off

between model performance and fairness objectives. The following subsections explain our findings for each dataset.

8.1 Experimental Evaluation on CelebA Dataset

8.1.1 Analysis of Threshold-Agnostic Performance Metrics. Our experimental evaluation on the CelebA dataset reveals notable performance variations across the different methodologies (Table 4). The ERM approach demonstrates superior performance across all metrics, achieving the highest ROC AUC (0.8576), AUC-PR Gain (0.7911) and PR AUC (0.7913), while maintaining the lowest Brier Score (0.1475). The DiffEopp method shows a considerable performance decline, with ROC AUC, AUC-PR Gain and PR AUC dropping to 0.7815, 0.6008 and 0.6011, respectively, though maintaining a relatively competitive Brier Score of 0.1861. Both GLCS and Calibrated GLCS exhibit identical discriminative capabilities, with ROC AUC of 0.8443, AUC-PR Gain of 0.7369 and PR AUC of 0.7371, positioning them as intermediate solutions between ERM and DiffEopp. However, they differ significantly in their calibration performance, with Calibrated GLCS achieving a substantially better Brier Score (0.2044) compared to standard GLCS (0.3349). The Calibrated GLCS emerges as a particularly promising approach, demonstrating robust discriminative capabilities while significantly enhancing probability calibration compared to its uncalibrated variant. Furthermore, both Calibrated GLCS and GLCS exhibit comparable fairness characteristics across various fairness metrics (discussed in Section 8.1.5). This comprehensive evaluation suggests that these methods achieve an optimal balance between maintaining competitive predictive performance and satisfying fairness constraints, positioning them as viable solutions for applications where both accuracy and fairness are crucial considerations. The empirical evidence particularly favors the Calibrated GLCS variant. It preserves the fairness properties of the base GLCS while providing more reliable probability estimates as shown by its improved Brier Score and AUC-PR Gain.

Metric	ERM	DiffEopp	GLCS	Calibrated GLCS
ROC AUC ↑	0.8576	0.7815	0.8443	0.8443
PR AUC ↑	0.7913	0.6011	0.7371	0.7371
Brier Score ↓	0.1475	0.1861	0.3349	0.2044
AUC-PR Gain ↑	0.7911	0.6008	0.7369	0.7369

Table 4. Invariant Performance Metrics for different methods on Celeb-A Dataset

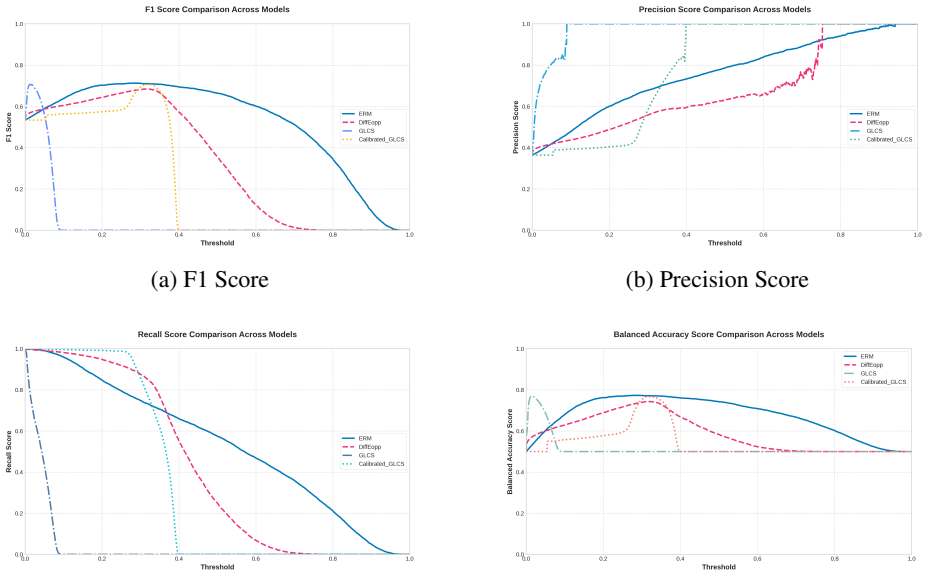
8.1.2 Analysis of Threshold-Dependent Performance Metrics. Our experimental results in Table 5 demonstrate notable performance variations across different methodological approaches. The baseline ERM achieves the highest F1 score (0.7132) at a threshold of 0.281. Moreover, ERM achieves balanced accuracy (0.7737), establishing a strong performance benchmark. The Calibrated GLCS shows comparable performance

metrics (balanced accuracy: 0.7715, F1 score: 0.7100), at a threshold of 0.315, while incorporating the proposed constraints. This minimal performance trade-off is particularly noteworthy, as using constraints typically incurs more substantial accuracy penalties, as we see for DiffEopp method. The DiffEopp method exhibits the highest recall (0.8601) but the lowest precision (0.5680), indicating a potential bias toward positive predictions. The GLCS method's very low threshold (0.012) for best F1 score (0.7672) compared to other methods (ranging from 0.281 to 0.315) confirms our hypothesis about probability space compression. This is effectively addressed through calibration as evidenced by the Calibrated GLCS's threshold restoration to 0.315. The Matthews Correlation Coefficient, a particularly robust metric for imbalanced datasets, shows consistent ranking with F1 scores, with ERM and Calibrated GLCS achieving the highest values (0.5341 and 0.5357, respectively).

Method	**Threshold**	**Balanced Accuracy**	**F1 Score**	**Matthews Corr. Coef.**	**Precision**	**Recall**
ERM	0.281	0.7737	0.7132	0.5341	0.6678	0.7652
DiffEopp	0.313	0.7428	0.6842	0.4698	0.5680	0.8601
GLCS	0.012	0.7672	0.7066	0.5161	0.6345	0.7972
Calibrated GLCS	0.315	0.7715	0.7100	0.5357	0.6864	0.7354

Threshold selected for optimal F1 score across different methods.
Table 5. Performance Metrics Comparison on Celeb-A Dataset

8.1.3 Performance Metrics Across Threshold Spectrum. Our comprehensive investigation of threshold sensitivity reveals nuanced performance characteristics across different fairness-aware machine learning models. The ERM model demonstrates significant robustness, maintaining F1 scores above 0.6 across a broad threshold range (0.2–0.8), in stark contrast to DiffEopp and Calibrated GLCS, which exhibit sharp performance degradation beyond their optimal threshold regions (Figure 1a). The base GLCS model displays an exceptionally narrow operational range, suggesting significant compression induced by its fairness constraints. Balanced accuracy analysis (Figure 1d) reveals a consistent peak within the 0.2–0.4 threshold range for all models, except GLCS, with ERM showcasing the most gradual performance decline at higher thresholds. The recall-precision trade-off curves (Figures 1b, 1c) illuminate the models' distinct behavioral patterns: ERM maintains the most balanced transition, while GLCS models exhibit more abrupt shifts, particularly in recall sensitivity. The wide threshold variations, especially in the GLCS approach, provide critical insights into the mechanisms of fairness-aware model design in GLCS framework. The probability space compression appears to stem from simultaneously satisfying multiple group-level fairness constraints, with the interaction between fairness penalties and base loss fun-



(a) F1 Score (b) Precision Score

(c) Recall Score (d) Balanced Accuracy Score

Fig. 1. Performance Metrics Across Threshold Spectrum on CelebA Dataset

damentally reshaping the model's decision boundaries. Our investigation unveils that hyperparameters μ_1, μ_2, μ_3, and δ substantially govern this probability space compression, with incremental parametric adjustments potentially yielding significant shifts in distributional representation based on different thresholds, as we meticulously illustrate in our UTKFace dataset analysis (Section 8.2).

The Calibrated GLCS's restoration of a more standard threshold through temperature scaling demonstrates an elegant solution—effectively "decompressing" the probability space while preserving the model's discriminative and fairness characteristics. This calibration approach not only broadens the model's robust performance region but also provides a promising strategy for maintaining fairness without sacrificing predictive consistency across different threshold values.

8.1.4 Subgroup Performance Analysis using Nuanced Metrics. Our empirical evaluation in Table 6 reveals significant variations in performance across methods and gender subgroups in the CelebA dataset. The baseline ERM demonstrates strong discriminative power with Subgroup-AUC scores of 0.805 and 0.831 for male and female subgroups, respectively, indicating robust within-group classification capabilities. However, the stark contrast between BPSN-AUC (0.958) and BNSP-AUC (0.512) metrics indicates substantial asymmetry in cross-group performance, suggesting potential systematic biases in the model's decision boundary. The DiffEopp method achieves more balanced cross-group metrics (BPSN-AUC: 0.873, BNSP-AUC: 0.615). This enhanced fairness, however, comes at the cost of reduced within-group performance, particularly

for the female subgroup, where Subgroup-AUC drops to 0.714. This trade-off exemplifies the challenging balance between fairness and performance objectives. Notably, our GLCS, including its calibrated variant, maintains strong within-group performance (male: 0.795, female: 0.807) while exhibiting cross-group behavior similar to ERM (BPSN-AUC: 0.956, BNSP-AUC: 0.487). The preservation of high female Subgroup-AUC (0.807) is particularly noteworthy, as it represents only a 2.9% decrease from ERM while incorporating the proposed constraints for fairness in GLCS framework. Our results validate that GLCS effectively maintains discriminative power while working within the fairness framework. The substantial disparity in subgroup sizes (male: 7,715, female: 12,247) adds another dimension to these findings, highlighting the importance of considering demographic imbalance in fairness-aware model development. The analysis underscores the complex interplay between maintaining strong predictive performance and achieving equitable treatment across demographic subgroups.

Method	**Subgroup**	**Subgroup Size**	**Subgroup AUC**	**BPSN AUC**	**BNSP AUC**
ERM	Male	7,715	0.8050	0.9583	0.5118
	Female	12,247	0.8307	0.5118	0.9583
DiffEopp	Male	7,715	0.7976	0.8734	0.6150
	Female	12,247	0.7142	0.6150	0.8734
GLCS & Calibrated GLCS	Male	7,715	0.7954	0.9564	0.4866
	Female	12,247	0.8067	0.4867	0.9564

Table 6. Nuanced Metrics on CelebA Dataset

8.1.5 Empirical Analysis of Fairness Metrics using CelebA Dataset. Our empirical evaluation of algorithmic fairness methodologies reveals critical insights into the performance of ERM, DiffEopp, GLCS, and Calibrated GLCS across multiple fairness dimensions (Table 7). The GLCS approach emerges as a standout winner, consistently demonstrating superior fairness metrics across various evaluation criteria. Notably, GLCS achieves exceptional results in minimizing opportunity disparities (eoppe), with an error rate of 2.84%, significantly outperforming both DiffEopp (4.60%) and the baseline ERM (30.78%). The equalized odds analysis (eodde) further substantiates GLCS's effectiveness, revealing minimal error (4.20%) compared to Calibrated GLCS (12.80%), DiffEopp (18.07%), and ERM (47.25%), which underscores its robust capability in maintaining fairness across positive and negative outcome scenarios.

The Demographic Parity and distribution divergence metrics provide additional validation of GLCS's approach. With a Demographic Parity (dpe) of 2.51%, GLCS significantly surpasses DiffEopp (15.97%) and ERM (29.38%), demonstrating its abil-

ity to ensure equitable prediction distributions across demographic groups. The Calibrated GLCS variant further enhances these results, achieving a p-rule score (prulee) of 72.74% and maintaining minimal Area Between CDF Curves (9.05%), compared to DiffEopp's 15.97% and ERM's 29.38%. The ROC AUC Parity (aucp) analysis reveals remarkably consistent performance, with GLCS and Calibrated GLCS showing minimal disparities (1.13%), in stark contrast to ERM's 2.57% and DiffEopp's 8.34% variations.

The examination of balanced for positive (bfp) class and negative class (bfn) reveals a performance hierarchy. GLCS achieves the most optimal error balance with bfp at 2.84% and bfn at 1.36%, demonstrating minimal classification disparities. Calibrated GLCS maintains strong performance with 5.24% bfp and 7.56% bfn, while DiffEopp shows moderate imbalance (4.60% bfp, 13.47% bfn). The baseline ERM approach exhibits the most significant error disparities, with 30.78% bfp and 16.47% bfn, highlighting the critical importance of fairness-aware methodological interventions.

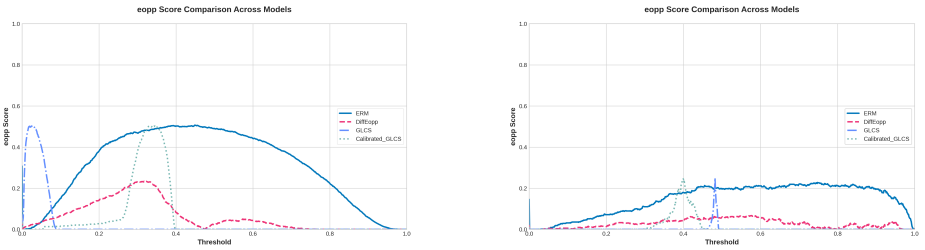
Metric	ERM	DiffEopp	GLCS	Calibrated GLCS
p-Rule (prulee) ↑	31.67	56.76	22.68	72.74
Equal Opportunity (eoppe) ↓	30.78	4.60	2.84	5.24
Equalized Odds (eodde) ↓	47.25	18.07	4.20	12.80
Demographic Parity (dpe) ↓	29.38	15.97	2.51	9.05
Balance for Positive Class (bfp) ↓	30.78	4.60	2.84	5.24
Balance for Negative Class (bfn) ↓	16.47	13.47	1.36	7.56
ROC AUC Parity (aucp) ↓	2.57	8.34	1.13	1.13
Area Between CDF Curves (abcc) ↓	29.38	15.97	2.51	9.05

Table 7. Calculate various fairness metrics with CelebA Datase

Threshold Sensitivity Analysis of Equal Opportunity. Figure 2a illustrates the Equal Opportunity (eopp) score variations across different classification thresholds for ERM, DiffEopp, GLCS, and Calibrated GLCS models. The analysis reveals several notable patterns: ERM exhibits the highest sensitivity to threshold selection, with eopp scores peaking at approximately 0.5 around the 0.4 threshold and gradually declining toward both extremes. In contrast, both GLCS and Calibrated GLCS demonstrate remarkable stability across most threshold values, maintaining consistently low eopp scores (< 0.1) except for a brief spike in GLCS at very low thresholds (< 0.1). DiffEopp shows intermediate performance with moderate threshold sensitivity, reaching a maximum eopp score of approximately 0.25 around 0.35 threshold. Notably, Calibrated GLCS exhibits a localized increase in eopp score around the 0.35 threshold region but quickly returns to stable performance. This comprehensive analysis suggests that GLCS and Calibrated GLCS provide more robust and threshold-invariant fairness guarantees compared to

traditional ERM and DiffEopp approaches, making them more reliable choices for applications requiring consistent fairness across different operating points.

8.1.6 Performance and Equal Opportunity Trade-Off on CelebA Dataset. Our comprehensive experimental evaluation unveils intricate trade-offs between predictive performance and fairness metrics across heterogeneous threshold configurations on CelebA Dataset. To ensure a rigorous and fair comparative analysis, we employ threshold-agnostic metrics: AUC-PR Gain and the threshold-agnostic Equal Opportunity Difference (eoppe) metric. The proposed GLCS and Calibrated GLCS approaches demonstrate remarkable fairness characteristics, consistently exhibiting substantially lower Equal Opportunity Difference scores. Specifically, the Calibrated GLCS achieved an eoppe of 5.24, while the GLCS method realized an eoppe of 2.84, in stark contrast to the Empirical Risk Minimization (ERM) baseline (eoppe = 30.78) and the DiffEopp approach (eoppe = 4.60). Notably, these improved fairness metrics are attained without compromising predictive performance. Both GLCS variants maintained competitive AUC-PR Gain scores (0.7369), comparable to DiffEopp (0.6008) and ERM (0.7911). This empirical evidence suggests that the proposed GLCS methodologies offer a principled approach to mitigating discriminatory outcomes while preserving high-fidelity predictive precision.



(a) eopp Score on CelebA (b) eopp Score on UTKFace

Fig. 2. eopp Metric Across Threshold Spectrum on CelebA and UTKFace

8.2 Experimental Evaluation on UTKFace Dataset

8.2.1 Analysis of Threshold-Agnostic Performance Metrics Our experimental evaluation on the UTKFace dataset reveals interesting performance patterns across the different methodologies (Table 8). The ERM approach maintains its superior performance across all metrics, achieving the highest ROC AUC (0.9015), AUC-PR Gain (0.8983) and PR AUC (0.8993), while demonstrating the lowest Brier Score (0.1266). Although DiffEopp shows slightly reduced performance compared to ERM, it still maintains strong discriminative capabilities with ROC AUC of 0.8754, AUC-PR Gain of 0.8599 and PR AUC of 0.8609, along with a competitive Brier Score of 0.1459. Both GLCS

and Calibrated GLCS exhibit identical discriminative performance, with ROC AUC of 0.8866, AUC-PR Gain of 0.8741 and PR AUC of 0.8752, positioning them between ERM and DiffEopp in terms of predictive capability. However, they differ in their calibration performance, with Calibrated GLCS achieving a better Brier Score (0.2276) compared to standard GLCS (0.2410). The Calibrated GLCS emerges as a particularly promising approach, demonstrating robust discriminative capabilities while enhancing probability calibration compared to its uncalibrated variant.

Metric	ERM	DiffEopp	GLCS	Calibrated GLCS
ROC AUC ↑	0.9015	0.8754	0.8866	0.8866
PR AUC ↑	0.8993	0.8609	0.8752	0.8752
Brier Score ↓	0.1266	0.1459	0.2410	0.2276
AUC-PR Gain ↑	0.8983	0.8599	0.8741	0.8741

Table 8. Invariant Performance Metrics for different methods on UTKFace Dataset

8.2.2 Analysis of Threshold-Dependent Performance Metrics. Our experimental results in Table 9 demonstrate notable performance variations across different methodological approaches based on the threshold for best F1 score for each approach. The baseline ERM achieves superior performance across multiple metrics, including the highest balanced accuracy (0.8069), F1 score (0.8010), and precision (0.7606) at a threshold of 0.406. The Calibrated GLCS demonstrates remarkably competitive performance (balanced accuracy: 0.8039, F1 score: 0.7997) at a threshold of 0.385. This minimal performance trade-off is particularly noteworthy, as some constraints typically incur more substantial accuracy penalties, as evidenced by the DiffEopp method's performance. While DiffEopp exhibits the highest recall (0.8614), it shows the lowest performance across other metrics, including precision (0.7248) and F1 score (0.7873), suggesting a potential bias toward positive predictions. The GLCS method achieves intermediate performance levels (balanced accuracy: 0.7950, F1 score: 0.7925) with a notably higher threshold (0.479) compared to other methods. Importantly, the Matthews Correlation Coefficient, which is particularly robust for imbalanced datasets, confirms the relative performance ordering, with ERM and Calibrated GLCS achieving the highest values (0.6128 and 0.6075, respectively). These results suggest that Calibrated GLCS effectively maintains strong predictive performance while incorporating the proposed constraints, making it a promising approach for applications requiring both accuracy and fairness considerations.

8.2.3 Performance Metrics Across Threshold Spectrum. Our examination of threshold sensitivity across models (Figure 3) reveals several distinctive behavioral patterns and performance characteristics. The F1 score analysis (Figure 3a) demonstrates that ERM and DiffEopp models maintain robust performance across a broad threshold range

Metric	ERM	DiffEopp	GLCS	Calibrated GLCS
Balanced Accuracy ↑	0.8069	0.7881	0.7950	0.8039
F1 Score ↑	0.8010	0.7873	0.7925	0.7997
Matthews CC ↑	0.6128	0.5782	0.5908	0.6075
Precision ↑	0.7606	0.7248	0.7364	0.7510
Recall ↑	0.8460	0.8614	0.8578	0.8551
Best Threshold	0.406	0.452	0.479	0.385

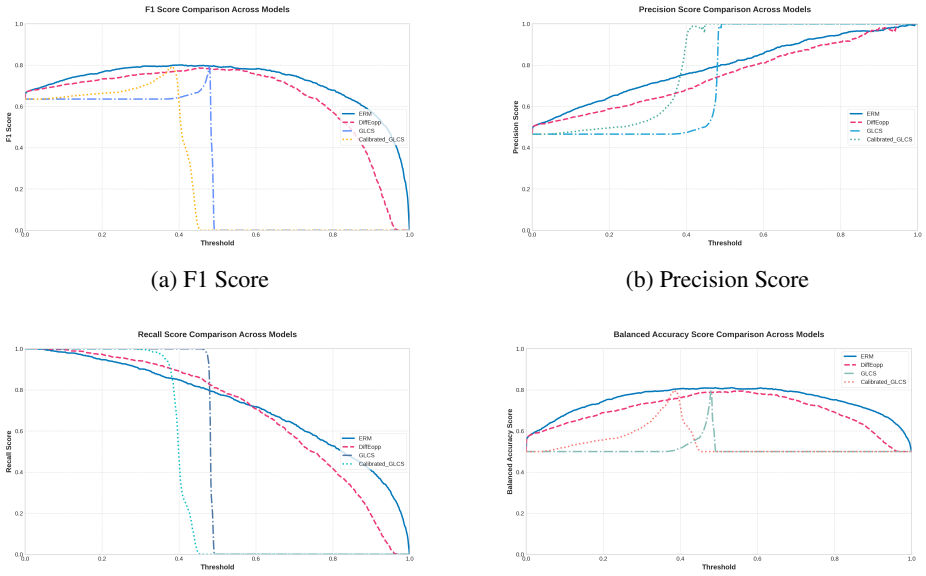
Table 9. Threshold for best F1 score for different methods on UTKFace Dataset with the corresponding variant performance metrics

(0.2-0.6), consistently achieving F1 scores above 0.6. While both GLCS and Calibrated GLCS exhibit comparable peak performance, they show more pronounced degradation outside their optimal threshold regions (approximately 0.37-0.5). The base GLCS displays a distinctive sharp performance spike near the 0.48 threshold, indicating a highly concentrated probability distribution. In terms of balanced accuracy trends (Figure 3d), the ERM baseline achieves the highest overall performance, reaching and maintaining a balanced accuracy of approximately 0.8 across the threshold range of 0.4-0.6, with graceful degradation at extreme thresholds. DiffEopp demonstrates comparable but slightly lower performance, maintaining balanced accuracy scores around 0.75-0.78 in the optimal range (0.4-0.6), though showing more pronounced degradation at higher thresholds compared to ERM. The GLCS and Calibrated GLCS methods exhibit notably different behaviors: Calibrated GLCS shows a gradual improvement up to threshold 0.4, reaching a peak of approximately 0.78, followed by an abrupt performance drop, while GLCS maintains a constant lower performance (around 0.5) before displaying a sharp, localized spike to 0.8 at threshold around 0.48.

This comparative analysis suggests that while fairness-oriented approaches like DiffEopp can achieve near-baseline performance, they may introduce some performance trade-offs, particularly in threshold sensitivity. The distinct behavioral patterns of GLCS variants indicate potential stability challenges in their balanced accuracy maintenance across different classification thresholds.

The recall-precision trade-off analysis (Figures 3b, 3c) reveals that recall curves demonstrate the expected monotonic decrease with increasing threshold, accompanied by corresponding increases in precision. ERM maintains the most gradual transition between these metrics, indicating superior calibration. Both GLCS variants exhibit step-like transitions around their respective optimal thresholds (0.48 for base GLCS, 0.39 for Calibrated GLCS), suggesting binary-like behavior in their predictions. Notably, DiffEopp maintains higher recall but lower precision compared to other models across most thresholds.

8.2.4 Subgroup Performance Analysis using Nuanced Metrics. Our empirical evaluation in Table 10 reveals noteworthy patterns in performance across methods and gender subgroups in the UTKFace dataset. The baseline ERM demonstrates strong discriminative power with Subgroup-AUC scores of 0.8927 and 0.9015 for male and female



(a) F1 Score (b) Precision Score

(c) Recall Score (d) Balanced Accuracy Score

Fig. 3. Performance Metrics Across Threshold Spectrum on UTKFace Dataset

subgroups respectively, indicating robust within-group classification capabilities. The complementary relationship between BPSN-AUC (0.9578) and BNSP-AUC (0.7946) metrics suggests a moderate asymmetry in cross-group performance.

The DiffEopp method achieves 0.8950 for BPSN-AUC (male subgroup) and 0.8446 for BNSP-AUC (female subgroup). This comes with a modest trade-off in within-group performance, with Subgroup-AUC slightly decreasing to 0.8723 and 0.8717 for male and female subgroups, respectively.

Our proposed GLCS approach and its calibrated variant maintain strong Subgroup AUC performance (male: 0.8874, female: 0.8760) while showing cross-group behavior similar to ERM (BPSN-AUC: 0.9521, BNSP-AUC: 0.7654). The relatively balanced subgroup sizes (male: 1131, female: 1239) in the UTKFace dataset provide a more equitable basis for evaluation compared to more imbalanced datasets.

These results demonstrate that while GLCS effectively maintains discriminative power within the fairness framework, the challenge of cross-group prediction asymmetry persists, albeit to a lesser degree than in comparable datasets. This analysis highlights the delicate balance between maintaining strong predictive performance and achieving equitable treatment across demographic subgroups, even in relatively balanced dataset conditions.

8.2.5 Empirical Analysis of Fairness Metrics using UTKFace Dataset. Our comprehensive empirical evaluation of algorithmic fairness on the UTKFace dataset presents a comparative analysis of the four methodologies: ERM, DiffEopp, GLCS, and Cal-

Method	Subgroup	Subgroup AUC	BPSN AUC	BNSP AUC	Size
ERM	Male	0.8927	0.9578	0.7946	1131
	Female	0.9015	0.7946	0.9578	1239
DiffEopp	Male	0.8723	0.8950	0.8446	1131
	Female	0.8717	0.8446	0.8950	1239
GLCS & Calibrated GLCS	Male	0.8874	0.9521	0.7654	1131
	Female	0.8760	0.7654	0.9521	1239

Table 10. Nuanced Metrics for different methods on UTKFace Dataset

ibrated GLCS. The results, presented in Table 11, reveal significant findings across multiple fairness dimensions. The evaluation demonstrates that GLCS achieves exceptional performance in minimizing *Equal Opportunity* disparities (eoppe), with an error rate of 0.23%, significantly outperforming DiffEopp (2.16%) despite the latter being specifically designed for this criterion (Equal Opportunity). Calibrated GLCS maintains strong performance with 1.25% for eoppe, while the baseline ERM exhibits substantially higher disparity (14.61%).

Analysis of the *Equalized Odds* reveals a clear performance hierarchy, with GLCS achieving optimal fairness at 0.65% for eodde, followed by Calibrated GLCS demonstrating strong fairness capability at 3.28%. DiffEopp shows improved performance with 7.09% for eodde, while ERM exhibits the highest disparity at 27.24%. These findings underscore GLCS's superior capability in maintaining fairness across outcome scenarios. Furthermore, GLCS demonstrates superior performance with a *p-rule* score of 98.69%, significantly outperforming both Calibrated GLCS (92.01%), DiffEopp (80.22%) and ERM (61.78%). This metric indicates that GLCS effectively maintains balanced probability distributions for positive and negative outcomes across demographic groups.

The analysis of *Demographic Parity* (dpe) reveals substantial variations, with GLCS achieving remarkable performance with a dpe of 0.62%, significantly surpassing Calibrated GLCS (3.00%), DiffEopp (10.69%) and ERM (22.03%). This demonstrates GLCS's effectiveness in ensuring equitable prediction distributions across demographic groups.

In terms of error rate balance, analysis of *Balance for Positive Class* (bfp) and *Balance for Negative Class* (bfn) shows that GLCS achieves optimal balance (bfp: 0.23%, bfn: 0.42%), while Calibrated GLCS maintains strong balance (bfp: 1.25%, bfn: 2.03%). DiffEopp exhibits moderate imbalance (bfp: 2.16%, bfn: 4.93%), and ERM shows significant disparity (bfp: 14.61%, bfn: 12.63%). Moreover, in terms of ROC AUC Parity (aucp), DiffEopp demonstrates minimal disparities (0.06%) in predictive performance across demographic groups, outperforming ERM (0.88%). Both GLCS and Calibrated GLCS show slightly higher but consistent disparities (1.15%). The Area Between CDF Curves (abcc) metric further validates GLCS's effectiveness, achieving minimal distribution divergence (0.62%), followed by Calibrated GLCS (3.00%). DiffEopp (10.72%) and ERM (22.03%) exhibit substantially higher disparities.

Metric	ERM	DiffEopp	GLCS	Calibrated GLCS
p-Rule (prulee) ↑	61.78	80.22	98.69	92.01
Equal Opportunity (eoppe) ↓	14.61	2.16	0.23	1.25
Equalized Odds (eodde) ↓	27.24	7.09	0.65	3.28
Demographic Parity (dpe) ↓	22.03	10.69	0.62	3.00
Balance for Positive Class (bfp) ↓	14.61	2.16	0.23	1.25
Balance for Negative Class (bfn) ↓	12.63	4.93	0.42	2.03
ROC AUC Parity (aucp) ↓	0.88	0.06	1.15	1.15
Area Between CDF Curves (abcc) ↓	22.03	10.72	0.62	3.00

Table 11. Calculate various fairness metrics with UTKFace Dataset

Threshold Sensitivity Analysis of Equal Opportunity. Figure 2b demonstrates the superior fairness characteristics of GLCS-based approaches compared to alternative methods. While the ERM baseline exhibits persistent unfairness with eopp scores steadily increasing to approximately 0.2 across the 0.4-0.8 threshold range, and DiffEopp showing moderate improvement with scores around 0.05, both GLCS and Calibrated GLCS demonstrate remarkable fairness preservation across most threshold values, maintaining near-zero eopp scores throughout the majority of the threshold spectrum. The tiny elevation in eopp scores around threshold 0.4 for these methods can be interpreted as a controlled trade-off point where the models actively adjust their decision boundaries to maintain long-term fairness stability. This localized behavior suggests a sophisticated fairness optimization strategy, where the models temporarily accept a minor fairness deviation to establish robust equilibrium across the broader threshold range. Particularly noteworthy is how both GLCS variants achieve nearly perfect Equal Opportunity (eopp $\approx$ 0) across extensive threshold regions (0.0-0.35 and 0.45-1.0), demonstrating their ability to maintain consistent fairness guarantees without the continuous fairness drift observed in ERM and DiffEopp. This comprehensive analysis suggests that GLCS-based approaches offer superior fairness preservation through their unique ability to establish and maintain stable equal opportunity metrics across diverse operating conditions.

8.2.6 Performance and Equal Opportunity Trade-Off on UTKFace Dataset. Our experimental evaluation reveals nuanced trade-offs between predictive performance and fairness metrics across varying threshold configurations. To ensure a rigorous and fair comparative analysis, we use threshold-agnostic metrics: AUC-PR Gain and eoppe metric. The proposed GLCS and Calibrated GLCS demonstrate remarkable fairness characteristics, consistently exhibiting substantially lower equal opportunity difference scores (eoppe). Specifically, the Calibrated GLCS achieved an eoppe of 1.25, while the GLCS method realized an eoppe of 0.23, in stark contrast to ERM baseline (eoppe = 14.61) and the DiffEopp approach (eoppe = 2.16). Notably, these improved fairness metrics are attained without compromising predictive performance for our approach.

Both GLCS variants maintained competitive AUC-PR Gain scores (0.8741), comparable to DiffEopp (0.8599) and ERM (0.8983). This empirical evidence suggests that the proposed GLCS methodologies offer a principled approach to mitigating discriminatory outcomes while preserving high-fidelity predictive precision. The results underscore the potential of GLCS methods in domains requiring stringent algorithmic fairness, particularly in high-stakes decision-making contexts where balancing performance and equitable outcomes is paramount.

8.3 Experimental Evaluation on CivilComments-WILDS Dataset

8.3.1 Analysis of Performance and Fairness Metrics. We evaluate our approach on the CivilComments-WILDS dataset using two key performance metrics: ROC AUC and Average Precision (AP), as shown in Table 12. The results demonstrate that both GLCS and ERM achieve comparable performance, with GLCS obtaining an ap of 74.62% and ROC AUC of 94.53%, while ERM achieves an ap of 74.74% and ROC AUC of 94.55%. Further analysis of fairness metrics between GLCS and ERM is presented in Table 13. The evaluation reveals that GLCS demonstrates superior performance across multiple fairness dimensions. Specifically, GLCS achieves exceptional performance in minimizing *Equal Opportunity* disparities, with (eoppe) of 4.91%, substantially outperforming ERM's 8.14%. The analysis of *Equalized Odds* reveals a clear advantage for GLCS, achieving an error rate (eodde) of 5.83% compared to ERM's 8.51%. In terms of *Demographic Parity*, GLCS demonstrates remarkable fairness with a demographic parity error (dpe) of 0.86%, significantly lower than ERM's 2.43%. Furthermore, GLCS achieves superior group fairness with a *p-rule* score of 95.85%, substantially outperforming ERM's 79.56%, indicating more balanced treatment across different demographic groups.

Metric	GLCS	ERM
Average Precision (ap)	74.62	74.74
ROC AUC	94.53	94.55

Table 12. Performance Metrics Comparison on CivilComments-WILDS Dataset

8.3.2 Analysis of Group Robustness. Our experimental results are shown in Tables 14 and 15, the baseline ERM method achieves an Average Accuracy of 92.4% but shows suboptimal performance with a Worst-Group Accuracy of 58.3% (Christian demographic group), indicating significant performance disparities across groups. By incorporating the Christian group as the sensitive feature in our proposed GLCS framework, we observe more balanced performance metrics. Specifically, GLCS achieves an Average Accuracy of 91.3% while substantially improving the Worst-Group Accuracy to 70.7%, representing a remarkable improvement of 12.4 percentage points in

Metric	GLCS	ERM
Demographic Parity Error (dpe) ↓	0.86	2.43
Equality of Opportunity Error (eoppe) ↓	4.91	8.14
Equalized Odds Error (eodde) ↓	5.83	8.51
p-Rule Error (prulee) ↑	95.85	79.56

Table 13. Fairness Metrics Comparison on CivilComments-WILDS Dataset

worst-group performance compared to ERM, with only a modest decrease of 1.1 percentage points in average accuracy. Furthermore, our experimental results demonstrate that GLCS consistently outperforms other robust baselines (DFR, Group-DRO, JTT, and AFR) on the CivilComments-WILDS dataset, establishing a new state-of-the-art in balancing average performance and group robustness on the CivilComments-WILDS dataset for this challenging benchmark.

Group	**#Samples**		**GLCS Method**		**ERM Method**	
	Non-Toxic	**Toxic**	**Non-Toxic**	**Toxic**	**Non-Toxic**	**Toxic**
Male	12,092	2,203	0.898	0.737	0.937	0.647
Female	14,179	2,270	0.912	0.725	0.946	0.640
LGBTQ	3,210	1,216	0.784	0.745	0.880	0.620
Christian	12,101	1,260	0.935	0.707	0.962	0.583
Muslim	5,355	1,627	0.820	0.744	0.903	0.607
Other religions	2,980	520	0.882	0.746	0.935	0.623
Black	3,335	1,537	0.737	0.798	0.856	0.680
White	5,723	2,246	0.760	0.784	0.866	0.660

Table 14. Group Accuracy Comparison and Sample Distribution across GLCS and ERM Methods

Method	Average Accuracy ↑	Worst-Group Accuracy ↑
ERM	0.924	0.583
DFR	0.872	0.701
Group-DRO	0.889	0.699
JTT	0.911	0.693
AFR	0.898	0.687
GLCS (Ours)	0.913	0.707

Table 15. Comparative Performanc of Group Fairness Method

9 Conclusion

This paper introduces a novel method of Group-Level Cost-Sensitive Learning (GLCS) framework, a pioneering approach that addresses critical challenges at the intersection of cost-sensitive learning, group fairness and group robustness in machine learning. By systematically incorporating group-level misclassification costs, we validate our proposed mehtod for mitigating bias while maintaining high model accuracy.

The key contributions of our work extend beyond traditional fairness interventions. We have empirically validated a fundamental synergy between group robustness and group fairness, revealing that targeted optimization strategies can simultaneously enhance model performance across underrepresented subgroups. Our approach fundamentally differs from conventional techniques by modifying the learning objective rather than artificially manipulating dataset distributions, thereby providing a more principled framework for addressing inherent biases in machine learning systems.

Our experimental results across multiple datasets provide compelling evidence of the GLCS framework's effectiveness. By encouraging models to focus on causally relevant features and implement nuanced group-level constraints, we have shown that it is possible to develop machine learning systems that are both more equitable and more robust. The implications of this research are particularly significant for high-stakes decision-making domains such as healthcare, finance, and criminal justice, where algorithmic fairness is paramount. Our work provides a practical pathway toward developing automated systems that can handle complex intersectional data distributions more reliably and ethically.

Future research directions include extending the GLCS framework to additional domains, exploring more sophisticated cost-sensitive optimization techniques, and developing more comprehensive metrics to evaluate group fairness and group robustness. As machine learning continues to play an increasingly critical role in societal decision-making, methodologies like GLCS will be crucial in ensuring that these systems remain both performant and fundamentally fair.

References

Barocas, S., Hardt, M., Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*, MIT press.

Bertsekas, D. P. (1976). Multiplier methods: A survey, *Automatica* **12**(2), 133–145.

Borkan, D., Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L. (2019). Limitations of pinned auc for measuring unintended bias, *arXiv preprint arXiv:1903.02088* .

Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification, *Companion proceedings of the 2019 world wide web conference*, pp. 491–500.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss, *Advances in neural information processing systems* **32**.

Caton, S., Haas, C. (2024). Fairness in machine learning: A survey, *ACM Computing Surveys* **56**(7), 1–38.

Chouldechova, A., Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning, *Communications of the ACM* **63**(5), 82–89.

Chuang, C.-Y., Mroueh, Y. (2021). Fair mixup: Fairness via interpolation, *arXiv preprint arXiv:2103.06503* .

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S. (2019). Class-balanced loss based on effective number of samples, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277.

Dablain, D., Krawczyk, B., Chawla, N. (2022). Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning, *arXiv preprint arXiv:2207.06084* .

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Du, M., Yang, F., Zou, N., Hu, X. (2020). Fairness in deep learning: A computational perspective, *IEEE Intelligent Systems* **36**(4), 25–34.

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

Flach, P., Kull, M. (2015). Precision-recall-gain curves: Pr analysis done right, *Advances in neural information processing systems* **28**.

Freeman, E. A., Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, *Ecological modelling* **217**(1-2), 48–58.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. (2017). On calibration of modern neural networks, *International conference on machine learning*, PMLR, pp. 1321–1330.

Han, X., Chi, J., Chen, Y., Wang, Q., Zhao, H., Zou, N., Hu, X. (2024). FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods, *Proceedings of the International Conference on Learning Representations*.
https://openreview.net/forum?id=TzAJbTClAz

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

Hashimoto, T., Srivastava, M., Namkoong, H., Liang, P. (2018). Fairness without demographics in repeated loss minimization, *International Conference on Machine Learning*, PMLR, pp. 1929–1938.

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

Hernández-Orallo, J., Flach, P., Ferri Ramírez, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss, *Journal of Machine Learning Research* **13**, 2813–2869.

Hirzel, M., Ram, P. (n.d.). Oversampling to repair bias and imbalance simultaneously, *AutoML Conference 2023*.

Kazemi, H. R., Khalili-Damghani, K., Sadi-Nezhad, S. (2023). Estimation of optimum thresholds for binary classification using genetic algorithm: An application to solve a credit scoring problem, *Expert Systems* **40**(3), e13203.

Kearns, M., Neel, S., Roth, A., Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, *International conference on machine learning*, PMLR, pp. 2564–2572.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data, *IEEE transactions on neural networks and learning systems* **29**(8), 3573–3587.

Kirichenko, P., Izmailov, P., Wilson, A. G. (2022). Last layer re-training is sufficient for robustness to spurious correlations, *arXiv preprint arXiv:2204.02937* .

Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores, *arXiv preprint arXiv:1609.05807* .

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I. et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts, *International conference on machine learning*, PMLR, pp. 5637–5664.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics, *Advances in neural information processing systems* **27**.

LaBonte, T., Muthukumar, V., Kumar, A. (2024). Towards last-layer retraining for group robustness with fewer annotations, *Advances in Neural Information Processing Systems* **36**.

Lipton, Z. C., Elkan, C., Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure, *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, Springer, pp. 225–239.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., Finn, C. (2021). Just train twice: Improving group robustness without training group information, *International Conference on Machine Learning*, PMLR, pp. 6781–6792.

Liu, Z., Luo, P., Wang, X., Tang, X. (2015). Deep learning face attributes in the wild, *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738.

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

Qiu, S., Potapczynski, A., Izmailov, P., Wilson, A. G. (2023). Simple and fast group robustness by automatic feature reweighting, *International Conference on Machine Learning*, PMLR, pp. 28448–28467.

Robles, E., Zaidouni, F., Mavromoustaki, A., Refael, P. (2020). Threshold optimization in multiple binary classifiers for extreme rare events using predicted positive data., *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*.

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 preprint arXiv:1911.08731* .

Sanchez, I. E. (2016). Optimal threshold estimation for binary classifiers using game theory, *F1000Research* **5**.

Sangalli, S., Erdil, E., Hötker, A., Donati, O., Konukoglu, E. (2021). Constrained optimization to train neural networks on critical and under-represented classes, *Advances in neural information processing systems* **34**, 25400–25411.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems, *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68.

Shui, C., Xu, G., Chen, Q., Li, J., Ling, C. X., Arbel, T., Wang, B., Gagné, C. (2022). On learning fairness and accuracy on multiple subgroups, *Advances in Neural Information Processing Systems* **35**, 34121–34135.

Subramanian, S., Rahimi, A., Baldwin, T., Cohn, T., Frermann, L. (2021). Fairness-aware class imbalanced learning, *arXiv preprint arXiv:2109.10444* .

Sulaiman, M., Roy, K. et al. (2024). The fairness stitch: A novel approach for neural network debiasing, *Acta Informatica Pragensia* **13**(3), 359–373.

Tarzanagh, D. A., Hou, B., Tong, B., Long, Q., Shen, L. (2023). Fairness-aware class imbalanced learning on multiple subgroups, *Uncertainty in Artificial Intelligence*, PMLR, pp. 2123–2133.

Vapnik, V. (1991). Principles of risk minimization for learning theory, *Advances in neural information processing systems* **4**.

Wan, M., Zha, D., Liu, N., Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey, *ACM Transactions on Knowledge Discovery from Data* **17**(3), 1–27.

Yan, S., Kao, H.-t., Ferrara, E. (2020). Fair class balancing: Enhancing model fairness without observing sensitive attributes, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1715–1724.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification, *Journal of Machine Learning Research* **20**(75), 1–42.

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

Zhou, S., Zhang, Y. (2016). Active learning for cost-sensitive classification using logistic regression model, *2016 IEEE international conference on big data analysis (ICBDA)*, IEEE, pp. 1–4.

Zhou, Z.-H., Liu, X.-Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem, *IEEE Transactions on knowledge and data engineering* **18**(1), 63–77.

Received December 4, 2024 , accepted January 26, 2025