# ARE FREQUENCY-BALANCED DATASETS OPTIMAL FOR EQUITABLE CLASS PERFORMANCE?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Resampling is a widely used approach for mitigating class-wise performance disparities in frequency-imbalanced datasets, typically by reallocating training data according to class frequency. However, this framing treats representation as the primary basis for deciding how much data each class should receive, while classes may also differ substantially in learning difficulty. Hardness-aware resampling and reweighting have shown that allocating more training resources to harder classes can improve class-wise fairness, but existing evidence comes predominantly from frequency-imbalanced settings, where the effects of correcting representation and accounting for hardness are difficult to disentangle. It therefore remains **unclear whether hardness provides a useful basis for data allocation in frequency-*balanced* settings**. We investigate this question by deliberately allocating more training data to harder classes. We find that this reallocation reduces the hard-versus-easy recall gap by approximately 1–4 percentage points, depending on the dataset, while maintaining comparable overall performance. We further show that the effectiveness of oversampling depends on the quality of the added samples, with hardness emerging as a strong estimate for downstream task utility. These results demonstrate that **resampling can be useful even when there is no representation imbalance to correct**, and motivate greater attention to class-wise fairness in frequency-balanced settings.

## 1 INTRODUCTION

Class bias, the phenomenon where machine learning classifiers exhibit disparate performance across classes, remains a fundamental challenge in fairness-aware learning and safety-critical applications (Zhu et al., 2026; Li et al., 2025; Katare et al., 2025; Brzezinski et al., 2024; Rahman & Davis, 2013; Khushi et al., 2021; Zong et al., 2023; Tasci et al., 2022). Under standard empirical risk minimization (ERM), each training sample contributes equally to the objective, causing classes with more training examples to exert greater influence on optimization and often resulting in substantially poorer performance on minority (tail) classes (Hashimoto et al., 2018). Resampling is consequently one of the most widely used families of approaches for mitigating class-wise disparities in frequency-imbalanced datasets. By over- and/or undersampling classes, these methods modify the frequency with which different classes contribute to training, typically moving the data towards equal class representation.

However, class frequency is not the only characteristic that can determine how difficult a class is to learn. Empirical evidence has shown that classes with higher intrinsic dimension (Ansuini et al., 2019; Pope et al., 2021; Ma et al., 2024), Lebesgue measure (Ma et al., 2023a), and curvature (Kienitz et al., 2022; Kaufman & Azencot, 2023; Ma et al., 2023b) tend to yield lower accuracy. This suggests that equal class frequency does not necessarily imply equal training requirements: harder classes may require more training resources to achieve performance comparable to easier classes (Narayanan & Niyogi, 2009; Pukowski & Lu, 2024). Indeed, substantial class bias persists even in a frequency-balanced setting (see Fig. 1). Hardness-aware learning methods have therefore been developed to identify and prioritize difficult classes, and hardness-based resampling (HBR) and reweighting (HBW) have been shown to improve class-wise performance beyond approaches based on frequency alone (Sinha et al., 2020; 2022). These findings suggest that class hardness may provide an alternative basis for allocating training resources across classes.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

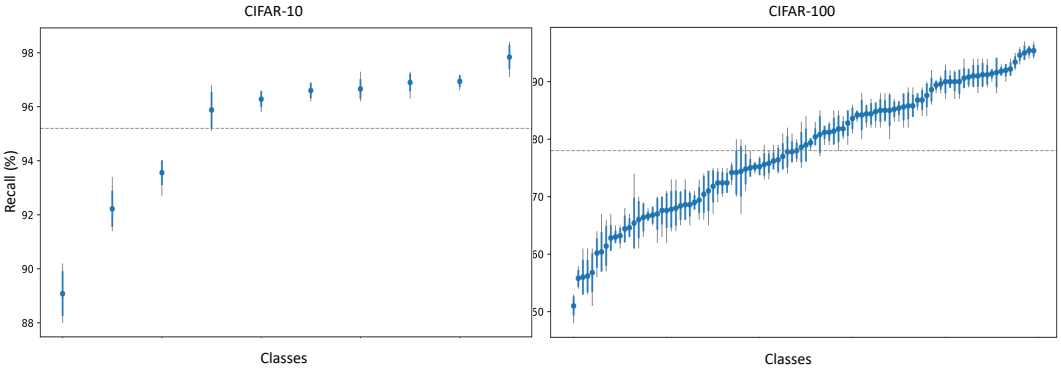


Figure 1: Training five ResNet18 networks on CIFAR-10 and CIFAR-100 reveals that class bias persists despite the lack of frequency imbalance. This paper investigates the effect of Hardness-Based Resampling (HBR) under fixed training resources on fairness and overall performance, to see if gains from addressing hardness imbalance outweigh the negative effects of resampling.

Yet existing evidence for hardness-aware allocation comes predominantly from frequency-imbalanced settings (He et al., 2008; Lin et al., 2017; Sinha et al., 2020; 2022). This creates an important ambiguity: when allocating additional training data to hard, underrepresented classes improves their performance, it is unclear how much of the improvement arises from correcting their representation deficit and how much results from allocating resources according to their greater difficulty. Testing hardness-based allocation in a frequency-balanced setting addresses this ambiguity, but raises a different question: **is the benefit of allocating resources according to hardness sufficient to justify the costs of resampling?** After all, resampling is not inherently beneficial. Undersampling can discard potentially informative training examples, whereas oversampling can introduce redundant or low-quality samples. Consequently, deliberately reallocating data in a balanced dataset is only worthwhile if the benefits of hardness-aware allocation compensate for these costs.

We therefore investigate whether hardness-guided resampling can improve class-wise equity when the initial training data are frequency-balanced. We use Area Under the Margin (AUM) to estimate instance hardness, aggregate hardness at the class level, and reallocate training resources by oversampling harder classes and undersampling easier ones. Across two case studies, HBR consistently reduces class-wise recall disparities while maintaining comparable overall performance. Across two complementary case studies, we find that hardness-based resampling reduces class-wise recall disparities while maintaining comparable overall performance. We further show that the effectiveness of oversampling depends on the quality of the added samples, with hardness emerging as a strong estimate for downstream task utility. Our findings show that hardness can provide a basis for reallocating training resources beyond frequency-imbalanced settings, and that the hardness of additional samples is an important determinant of their utility for reducing class-wise disparities.

To summarize our contributions are as follows:

- We show that frequency balance does not necessarily provide an appropriate allocation of training resources for equitable class-wise performance, and that hardness-aware resampling can improve this allocation even when the initial dataset is frequency-balanced.
- We show that the hardness of oversampled data is an important determinant of its utility for improving class-wise equity.

## 2 Background

**Frequency imbalances.** Frequency imbalance refers to differences in sampling density across a dataset and can occur both between and within classes. Between-class imbalance, where classes contain different numbers of samples, is well studied (Leevy et al., 2018; Kaur et al., 2019). Within-class imbalance refers to non-uniform sampling across regions of a class manifold, resulting in

disjuncts of different sizes and potentially different learning difficulty (Holte et al., 1989; Japkowicz, 2001). Although both forms have been studied, the literature has predominantly focused on between-class imbalance (Leevy et al., 2018; Ren et al., 2023).

**Hardness imbalance.** Hardness imbalance refers to non-uniform distribution of learning difficulty across samples and classes. This happens due to differences in cardinalities (Lorena et al., 2018; 2019), intrinsic dimension (Ansuini et al., 2019; Pope et al., 2021; Ma et al., 2024), Lebesgue measure (Ma et al., 2023a), and curvature (Kienitz et al., 2022; Kaufman & Azencot, 2023; Ma et al., 2023b) across and within classes, as well as due to class overlap and other data-, and sampling-related issues (Lorena et al., 2018; 2019).

Due to the multifaceted nature of hardness imbalance, directly measuring it is not possible with current methodologies, necessitating obtaining hardness estimates from heuristic signals such as a model's confidence or forgetting events (Seedat et al., 2024). These signals are used to identify problematic samples (Pleiss et al., 2020; Agarwal et al., 2022; Yang et al., 2024), guide data pruning (Toneva et al., 2018; Paul et al., 2021), or dynamically adjust training algorithms (Lin et al., 2017; Soviany et al., 2022; Yuan et al., 2025). In this work we employ Area Under the Margin (**AUM**) (Pleiss et al., 2020) as our hardness estimator, as it was shown to produce state-of-the-art results (Seedat et al., 2024). It aggregates the difference between the confidence (softmax probability) in true class $z_y^{(t)}(x)$ and that of the strongest competing class throughout training across all $T$ training epochs:

$$\text{AUM}(x, y) = \frac{1}{T} \sum_{t=1}^{T} \left( z_y^{(t)}(x) - max_{i \neq y} z_i^{(t)}(x) \right). \tag{1}$$

**Undersampling.** Hardness-based pruning has gained recognition for its ability to identify and remove data samples with minimal impact on model performance (Toneva et al., 2018; Paul et al., 2021; Maini et al., 2022). Sorscher et al. (2022) provide particularly compelling evidence of its effectiveness, showing that it can break beyond power-law scaling of error with respect to dataset size. Since data pruning has the same purpose as undersampling—removing a specified number of samples with minimal negative impact on performance—we also use this approach in our work.

**Oversampling.** Oversampling can take several forms. One can simply replicate existing data (random oversampling), apply data augmentation to create synthetic variants (Ahn et al., 2023; Vu et al., 2024), or interpolate between minority class samples (e.g., SMOTE (Chawla et al., 2002) and its variants such as Borderline-SMOTE (Han et al., 2005), ADASYN(He et al., 2008), or DeepSMOTE (Dablain et al., 2022)). However, random oversampling only duplicates existing samples, while interpolation-based methods generate samples constrained by existing observations and may not capture the full underlying distribution. To overcome this limitation, recent works have explored generative models (e.g., Generative Adversarial Networks (Goodfellow et al., 2020), or diffusion models (Ho et al., 2020)) as a way to synthesize novel samples that better capture the true class distribution. These generative approaches have increasingly become a focal point for addressing data scarcity (Ali-Gombe & Elyan, 2019; Zheng et al., 2020; Marchesi et al., 2023). Importantly, most existing evaluations of resampling methods focus on frequency-imbalanced datasets.

## 3 METHODOLOGY

### 3.1 EXPERIMENTAL DESIGN

We define Hardness-Based Resampling (HBR) as a data preprocessing operation that breaks the data balance by resampling classes based on their hardness, changing the number of samples in class $c$ from the original $a_c$ to some $b_c$ that is proportional to the true hardness of that class, $\hat{h}_c$. It is composed of two operations: 1) undersampling; and 2) oversampling. Undersampling works by removing the $a_c - b_c$ easiest samples from classes where $b_c < a_c$. Meanwhile, oversampling is performed by expanding the training set by $b_c - a_c$ samples. In a balanced setting, which we consider in this work, $a_i = a_j$ for all $(i, j) \in \{1, ..., k\}^2$. Naturally, ground-truth hardness $\hat{h}_c$ is inaccessible, necessitating reliance on hardness estimates. In this work, we use AUM to obtain

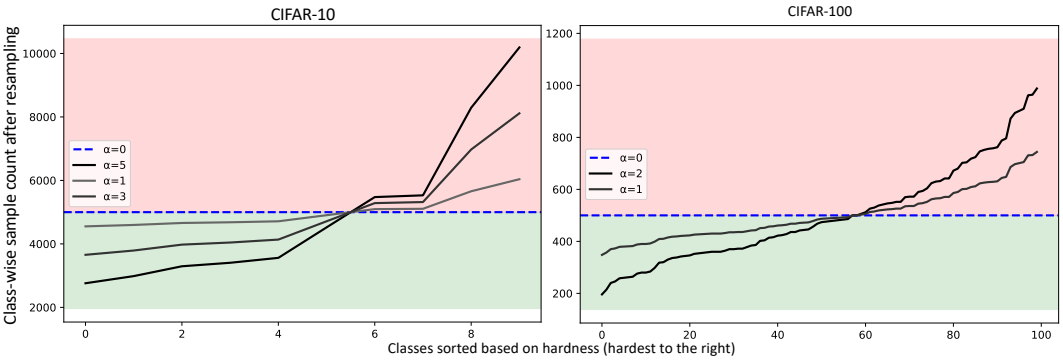


Figure 2: Visualization of the degree of the frequency imbalance introduced through HBR. Adjusting the $\alpha$ parameter allows us to have a finer control over the severity of the introduced imbalance.

instance-level hardness estimates $h_i$, which we later convert to class-level ones through averaging:

$$H_c = \frac{\sum_{i=1}^{n} h_i \times \mathbf{1}_{\{y_i=c\}}}{\sum_{i=1}^{n} \mathbf{1}_{\{y_i=c\}}}, \tag{2}$$

where $\mathbf{1}_{\{y_i=c\}}$ is an indicator function that equals 1 if $y_i = c$ and 0 otherwise, and $n$ is dataset's cardinality. Intuitively, the resampling ratios $b_c$ should be based on those class-level estimates as follows:

$$b_c = \left\lfloor \frac{H_c}{\sum_{c'} H_{c'}} \sum_{c} a_c \right\rfloor. \tag{3}$$

Since Equation 3 does not allow for modifying the degree of the introduced frequency imbalance we update the number of samples after resampling as follows:

$$b'_c = a_c + \lfloor \alpha \, (b_c - a_c) \rfloor \tag{4}$$

with $b'_c$ being the updated sample count for class $c$ and $\alpha$ a scalar that controls the degree of the introduced frequency imbalance. In Figure 2 we show how different $\alpha$ values used in our experiments affect the degree of the introduced frequency imbalance.

### 3.2 Datasets

We evaluate our approach on CIFAR-10 and CIFAR-100 using ensembles of ResNet-18 models adapted for low-resolution images. We choose these datasets and architecture because they are established benchmarks in hardness-aware learning (Toneva et al., 2018; Pleiss et al., 2020; Paul et al., 2021; Maini et al., 2022; Mindermann et al., 2022; Kaufman & Azencot, 2023), while their relatively small scale enables the extensive multi-seed experimentation required for reliable comparisons. Detailed experimental settings are provided in Appendix B, and the paired $t$-test used to assess statistical significance is described in Appendix D.

#### 3.2.1 Designing case studies.

To reiterate, we investigate whether the fairness objective can be met by performing HBR in frequency-balanced setting. This raises three key challenges: (1) oversampling techniques may struggle to generate genuinely new, i.i.d. samples; (2) undersampling removes potentially valuable information; and (3) determining the resampling ratio requires class-level hardness estimates, whose quality is difficult to ascertain. While the third challenge remains a fundamental limitation of HBR, or any hardness-aware method for that matter, the second can be mitigated through hardness-based undersampling, as described in Sec. 2. We therefore focus our case studies on the first challenge and investigate HBR under two complementary settings. Case Study 1 provides access to held-out samples that serve as a simulated oracle for new draws from the underlying distribution, thereby

isolating the effect of HBR from the difficulty of generating new samples. Case Study 2 considers the conventional resampling setting without access to such samples.

**Case Study 1: Access to holdout set.** We assume that CIFAR-10 and CIFAR-100 were obtained through i.i.d. sampling from an underlying distribution $\mathcal{X}$. More specifically, for each class $c$, $\mathcal{D}_c$ consists of i.i.d. samples drawn from class-conditional distribution $\mathcal{X}_c$. We consider a sub-dataset $\mathcal{D}' = \bigcup_{c=1}^{C} \mathcal{D}'_c$, where $\mathcal{D}'_c \subset \mathcal{D}_c$ and $|\mathcal{D}'_c| = a' < a$ for each class $c$. The remaining samples, $\mathcal{D}_c^{\text{ho}} := \mathcal{D}_c \setminus \mathcal{D}'_c$, form a holdout set. Under the i.i.d. assumption, randomly selecting $a'$ samples without replacement from $\mathcal{D}_c$ yields a subset with the same distribution as $a'$ independent draws from $\mathcal{X}_c$, while the remaining observations form an independent sample from $\mathcal{X}_c$, independent from the selected subset. We therefore treat $\mathcal{D}_c^{ho}$ as a *simulated oracle for new draws from* $\mathcal{X}_c$. We refer to this hypothetical situation, in which dataset construction was stopped early, as *Scenario A*. To simulate it, we randomly select $a'$ samples from each class without replacement. The remaining samples constitute the disjoint holdout set $\mathcal{D}^{ho}$. We then apply HBR to $\mathcal{D}'$, which we denote by $\mathcal{D}^{CLP}$ because it is obtained through class-level pruning, allowing HBR to draw additional samples from $\mathcal{D}^{ho}$ when oversampling.

**Case Study 2: No access to holdout set.** Here we apply HBR to the original dataset $\mathcal{D}$. Thus, unlike Case Study 1, no class-level pruning is performed before resampling, and HBR has no access to the holdout samples.

### 3.2.2 Hardness-Based Resampling

For undersampling we ranked samples within each easy class by AUM, removing the easiest ones until the desired sample count was achieved. For oversampling we employ three strategies: (1) random oversampling, (2) SMOTE, and (3) synthetic samples generated using EDM (Karras et al., 2022). For Case Study 1, we use the holdout set instead of EDM-generated samples to avoid information leakage, since the EDM was trained on the full CIFAR training set, including samples unavailable in the pruned dataset $\mathcal{D}^{CLP}$. Since the relationship between synthetic sample hardness and downstream model performance is not yet well understood, we use three subsets sampled from the pool of 1 million synthetic images generated by the EDM trained by Wang et al. (2023) and released publicly by the authors for oversampling:

1. rEDM, where synthetic samples are drawn randomly from the specified classes
2. aEDM, where samples are drawn from the set of $2b_c$ synthetic images whose hardness is closest to the class average (based on AUM)
3. hEDM, where samples are drawn from the set of $2b_c$ hardest synthetic images of class $c$

We use a deterministic seed scheme to disentangle randomness from dataset construction (resampling and pruning) and model training. Each setting therefore consists of independently generated dataset replicates and matched model initializations, enabling reproducible paired comparisons between baseline and HBR (details in C).

## 3.3 Measuring degree of class bias

We quantify class bias through disparities in per-class recall. Let $r_c := \text{Rec}(M, \mathcal{D}_c^{\text{test}})$ denote the recall of model $M$ on class $c$, and let $\mathbf{r} = (r_1, \ldots, r_C)$ denote the vector of class-wise recalls, where $C$ is the total number of classes. We use several complementary measures. The maximum gap

$$\Delta_m = \max_c r_c - \min_c r_c \tag{5}$$

measures the difference between the highest- and lowest-recall classes. The quantile gap

$$\Delta_q = \frac{1}{k} \sum_{i=1}^{k} r_{(C+1-i)} - \frac{1}{k} \sum_{i=1}^{k} r_{(i)} \tag{6}$$

measures the difference between the average recall of the $k$ highest- and lowest-recall classes, where $r_{(i)}$ denotes the $i$-th smallest element of $\mathbf{r}$ and $C$ is the number of classes. We set $k = 2$ for CIFAR-10 and $k = 10$ for CIFAR-100. The hardness-based equality gap

$$\Delta_{he} = \frac{1}{|\mathcal{E}|} \sum_{c \in \mathcal{E}} r_c - \frac{1}{|\mathcal{H}|} \sum_{c \in \mathcal{H}} r_c \tag{7}$$

measures the difference in mean recall between the easy classes $\mathcal{E}$ and hard classes $\mathcal{H}$ identified by HBR, where $\mathcal{E}$ and $\mathcal{H}$ are the classes targeted for undersampling and oversampling, respectively. We additionally report the standard deviation $\Delta_\sigma$ and median absolute deviation $\Delta_{\text{MAD}}$ of class-wise recall. Lower values of all $\Delta$ measures indicate smaller class-wise recall disparities. We report class-averaged recall and precision alongside these measures to distinguish changes in class bias from changes in overall performance.

## 4 RESULTS

In this section, we analyze the results of both case studies. Figure 3 presents the changes in the fairness metrics, defined in Section 3.3, for the first case study, while Figure 4 shows the corresponding results for the second case study. In Appendix E, we report the outcomes of the Student's t-test and include further analysis.

**Sample quality is crucial for HBR.** We first observe a consistent decrease in recall-based class bias resulting from HBR. In the first case study, we find statistically significant decreases in $\Delta_m$, $\Delta_q$, $\Delta_{he}$, and $\Delta_\sigma$ when using the holdout set for oversampling. Specifically, across all pruning rates, HBR decreases the recall gap between the average hard and easy class ($\Delta_{he}$) by approximately $0.01$ on CIFAR-10 and $0.04$ on CIFAR-100. Similar trends appear in the second case study when only the hEDM samples are used for oversampling: $\Delta_{he}$ decreases from $0.043$ to $0.029$ on CIFAR-10 and from $0.165$ to $0.138$ on CIFAR-100. Using rEDM and aEDM samples also frequently reduces recall gaps, although less consistently. In contrast, random oversampling and SMOTE generally yield smaller or statistically insignificant changes, and occasionally worsen the class bias. The importance of high-quality synthetic samples is particularly evident in Figure 3 on CIFAR-100, where we see that changing from SMOTE or random oversampling to holdout set almost doubles the changes in recall-based metrics.

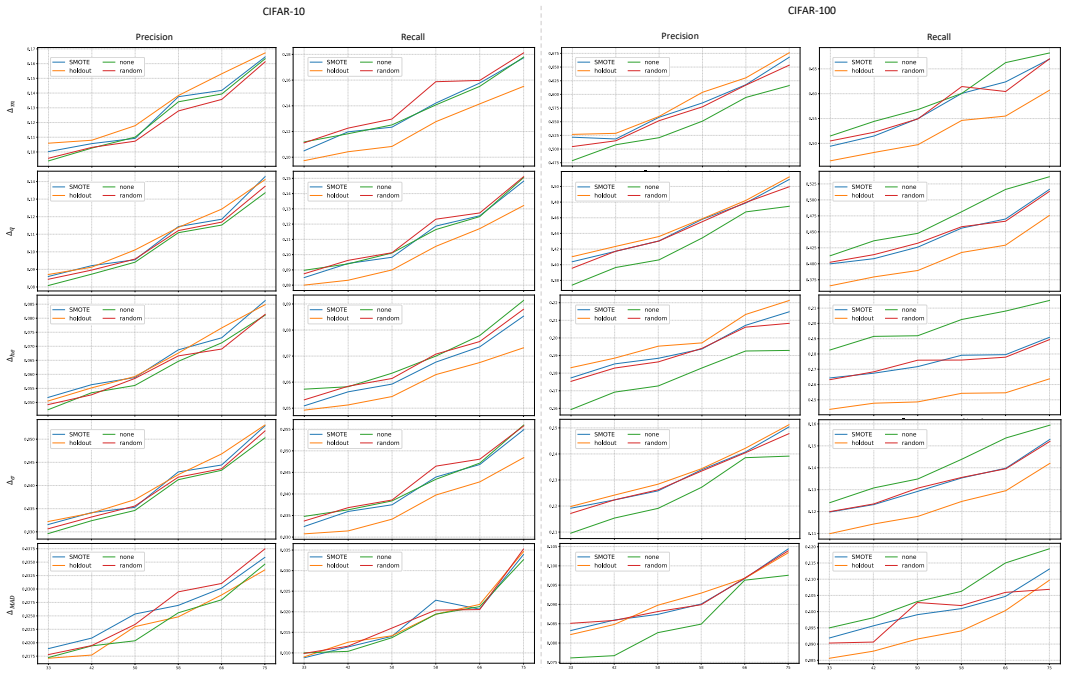


Figure 3: **Case Study 1:** Analyzing changes in various fairness metrics on pruned versions of CIFAR-10 and CIFAR-100 reveals that HBR reduces the recall gap across classes at the cost of precision gap (lower $\Delta$ values indicate smaller class gaps). Improvements are consistent across pruning rates (X-axis) and rely on the quality of samples used for oversampling, as indicated by the poor performance of random- and SMOTE-based oversampling. We report means over four independently pruned datasets $\mathcal{D}^{CLP}$ and four models per $\mathcal{D}^{CLP}$; standard deviations are omitted for clarity.

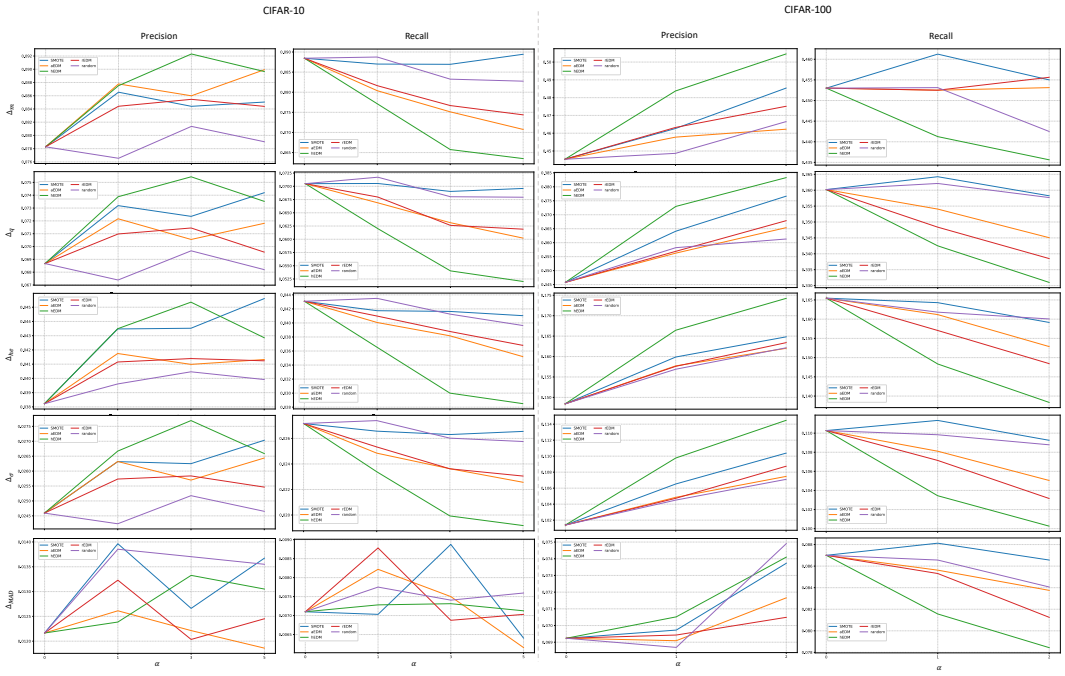


Figure 4: **Case Study 2:** We notice that, in most cases, the higher the degree of the introduced frequency imbalance (X-axis) the larger the effect on fairness when high-quality synthetic samples are available. We expect the existence of some threshold $\alpha$ beyond which the drawback from too heavy frequency imbalance starts to overwhelm the benefits of addressing hardness imbalance.

These observations yield two important insights. (i) Unlike standard class imbalance, where simple oversampling succeeds by increasing class cardinality and thus gradient impact, hardness imbalance requires informative samples that populate low-density and/or high-complexity regions—simple replication or interpolation cannot achieve this. (ii) Our results suggest that sample hardness is informative of downstream utility of generative data. Within the synthetic pool considered in our experiments, harder samples (hEDM) consistently provide greater reductions in class-wise recall disparities than different selection methods. This suggests that hardness could serve as a useful signal for selecting synthetic data according to their downstream utility, complementing commonly used measures of generative quality such as FID and IS. From this point onward we will focus on results on holdout set or hEDM in our analysis, unless specified otherwise.

**Quality of EDM-generated samples is insufficient.** Analyzing the results across case studies reveals further nuances. On CIFAR-10, the improvements from HBR are remarkably consistent across both case studies: $\Delta_m$ and $\Delta_q$ improve by $\approx 0.01$, while $\Delta_{he}$ and $\Delta_\sigma$ improve by $\approx 0.005$. This means that on CIFAR-10, the EDM-generated images enable bias reductions comparable to those achieved with the holdout set, suggesting they are of sufficient quality for this task. However, this parity disappears on CIFAR-100. While the first case study (holdout set) achieves a $\Delta_{he}$ reduction of $\approx 0.04$, the second case study (hEDM) only achieves $\approx 0.02$ for $\alpha = 1$. Because of the limited cardinality of the holdout set, we are forced to set $\alpha = 1$ in Case Study 1, and so when comparing the two settings we have to consider only $\alpha = 1$.

We attribute this discrepancy to a shortfall in synthetic informativeness rather than a saturation of the resampling method itself. While it is true that the reduction in $\Delta_{he}$ introduced by HBR decreases with the pruning rate, the holdout set at $33\%$ pruning already produces a reduction in $\Delta_{he}$ slightly larger than that obtained with hEDM on the full dataset. Considering the monotonic trend in Case Study 1 with respect to the pruning rate, it is hence natural to extrapolate that the holdout set at $0\%$ pruning would outperform hEDM. This suggests that EDM is not yet able to produce samples with downstream utility comparable to that of real samples. This becomes even more apparent when we consider that we compared random holdout samples with the hardest available synthetic samples: if the holdout set were larger and we could also perform hardness-based selection, the difference

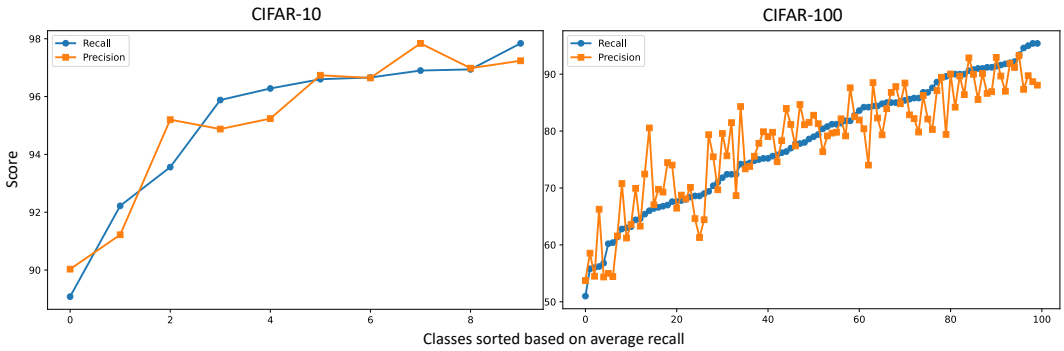


Figure 5: Precision and recall, with classes ordered by average recall. Both measures are strongly aligned in our experiments, indicating that *any attempts in reducing recall gap will lead to increase in precision gap and vice versa* due to the precision-recall trade-off occurring at class level.

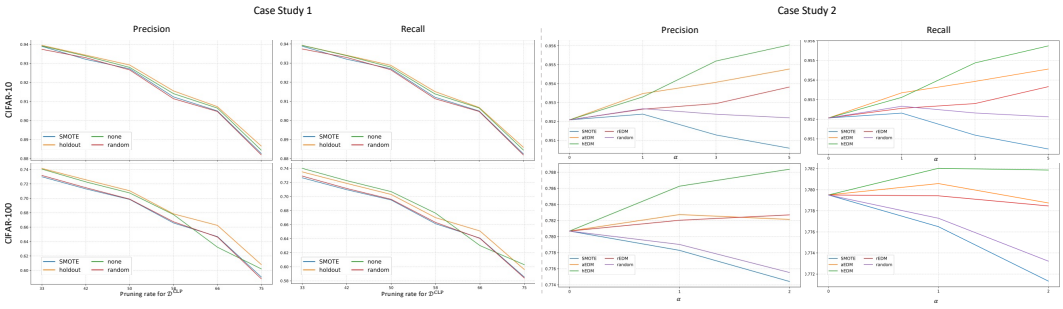


Figure 6: Analyzing the changes to macro precision and recall we notice that the reduction of class bias, reported in Figures 3 and 4, is not accompanied by a significant reduction in overall performance. In fact, we notice that in some cases the average metrics slightly increase (here higher value indicates higher overall performance).

between real and synthetic data utility could be even larger. There is therefore substantial room to improve the downstream utility of synthetic data. Our earlier results show that hardness provides a useful signal for distinguishing more and less useful synthetic samples, suggesting that steering generative models toward harder samples may be a promising direction for closing this gap.

**HBR worsens precision gaps.** Addressing recall-based class bias was the main objective of introducing HBR. However, it is also useful to study the effects that this method has on precision gaps (Fig. 5) and macro-metrics (Fig. 6). We find that the reduction of recall gaps is paired with an increase in precision gaps. As shown in Appendix F, Figures 7–14, oversampling hard classes increases the number of positive predictions for those classes, leading to higher counts of both true positives (TP) and false positives (FP), while false negatives (FN) and true negatives (TN) decrease. This explains the observed increase in recall for hard classes and decrease for easy classes, and the consequential reduction of recall gaps. Naturally, the resulting changes in precision are opposite to those observed for recall. When combined with the observed increase in precision gaps, this suggests that recall and precision gaps are aligned across classes: classes with high recall also tend to have high precision, and vice versa. We verify this empirically on CIFAR-10 and CIFAR-100 (with recall measured as an average over five ResNet18 models), observing strong correlations—Pearson correlations of 0.9395 and 0.9025, and Spearman correlations of 0.9394 and 0.8961 for CIFAR-10 and CIFAR-100, respectively (see Figure 5). This indicates that the precision–recall trade-off cannot be avoided when attempting to reduce class bias. If this correlation were weaker, one could improve both precision and recall gaps simultaneously—for example, by benefiting classes with high recall but low precision. However, under the observed strong alignment, improvements in recall gaps necessarily come at the expense of increased precision gaps.

**Impact on macro values** A key question is whether these fairness gains come at the expense of overall performance. Examining macro precision and recall, we find that in most situations HBR

maintains or even slightly improves macro-level metrics across both datasets (see Figure 6). This observation also explains why we focus on precision and recall rather than derived metrics like F1: because both measures consistently move in the same direction—typically increasing together and rarely diverging—a harmonic mean would simply mirror this joint trend without providing additional insight. The observed reductions in recall gaps therefore translate to a net positive or neutral effect on aggregate performance, rather than a tradeoff. Naturally, aggressive resampling with sufficiently large $\alpha$ would eventually degrade performance, but within the range we examine, HBR achieves its fairness goals without compromising accuracy.

## 5 Conclusion and future work

In this work, we investigated the relationship between class-wise performance and data distribution, uncovering two important insights into the nature of class bias. First, our empirical evaluation on CIFAR-10 and CIFAR-100 demonstrates that dataset-level balance is insufficient to ensure balanced performance: despite being frequency-balanced, both datasets exhibit substantial recall disparities across classes. This shows that aggregate performance metrics can mask systematic class bias. Consequently, we advocate complementing conventional aggregate measures with gap- and dispersion-based metrics that can expose such disparities even in perfectly balanced datasets.

Second, we have shown that a targeted hardness-based resampling strategy can reduce these recall gaps without the corresponding degradation in overall performance. This suggests that improving class-wise performance need not necessarily require sacrificing aggregate performance, and that deliberately introducing controlled frequency imbalance can be an effective way of compensating for differences in data difficulty. At the same time, the effectiveness of HBR depends on both the quality of the oversampled data and the magnitude of the induced imbalance.

Unlike classical frequency imbalance, for which random oversampling can often be effective, hardness imbalance is a more complex phenomenon arising from the geometry and topology of the data distribution. Our empirical findings support this distinction, showing that naive oversampling is insufficient to address disparities driven by differences in class difficulty.

An additional finding concerns the potential role of hardness in evaluating generated data. Our results suggest that hardness may provide an estimate of the downstream utility of generated samples. This is particularly relevant because FID and IS, the commonly used measures for evaluating generative models, primarily capture aspects of distributional fidelity, diversity, and class predictability. While such measures are valuable for assessing distributional fidelity, they do not directly quantify downstream task utility or capture issues such as memorization, as discussed by Ravuri & Vinyals (2019) and Jiralerspong et al. (2023). Our findings suggest that hardness could complement existing distributional metrics by providing information about the utility of generated data, although further investigation is required to establish the generality and reliability of this relationship.

Overall, our findings motivate a shift from viewing dataset balance as a sufficient proxy for balanced performance towards an adaptive framework that explicitly estimates and addresses hardness-based disparities. Further developing such hardness-aware evaluation and resampling strategies may provide a useful direction for improving the consistency and reliability of model performance across classes.

## References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378, 2022.

Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RgUPdudkWlN.

Adamu Ali-Gombe and Eyad Elyan. Mfc-gan: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212–221, 2019.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinskyi. Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, 18(7):1–18, 2024.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, 34(9): 6390–6404, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

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

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. IEEE, 2008.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Robert C Holte, Liane Acker, Bruce W Porter, et al. Concept learning and the problem of small disjuncts. In *IJCAI*, volume 89, pp. 813–818, 1989.

Nathalie Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian society for computational studies of intelligence*, pp. 67–77. Springer, 2001.

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36:33095–33119, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Dewant Katare, David Solans Noguero, Souneil Park, Nicolas Kourtellis, Marijn Janssen, and Aaron Yi Ding. Analyzing and mitigating bias for vulnerable road users by addressing class imbalance in datasets. *IEEE Open Journal of Intelligent Transportation Systems*, 2025.

Ilya Kaufman and Omri Azencot. Data representations' study of latent image manifolds. In *International Conference on Machine Learning*, pp. 15928–15945. PMLR, 2023.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM computing surveys (CSUR)*, 52(4):1–36, 2019.

Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975, 2021.

Daniel Kienitz, Ekaterina Komendantskaya, and Michael Lones. The effect of manifold entanglement and intrinsic dimensionality on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7160–7167, 2022.

Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

Xinyue Li, Zhenpeng Chen, Jie M Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. Bias behind the wheel: Fairness testing of autonomous driving systems. *ACM Transactions on Software Engineering and Methodology*, 34(3):1–24, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Ana C Lorena, Aron I Maciel, Péricles BC de Miranda, Ivan G Costa, and Ricardo BC Prudêncio. Data complexity meta-features for regression problems. *Machine Learning*, 107(1):209–246, 2018.

Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

Yanbiao Ma, Licheng Jiao, Fang Liu, Yuxin Li, Shuyuan Yang, and Xu Liu. Delving into semantic scale imbalance. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=07tc5kKRIo.

Yanbiao Ma, Licheng Jiao, Fang Liu, Shuyuan Yang, Xu Liu, and Lingling Li. Curvature-balanced feature manifold learning for long-tailed classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15824–15835, 2023b.

Yanbiao Ma, Licheng Jiao, Fang Liu, Lingling Li, Wenping Ma, Shuyuan Yang, Xu Liu, and Puhua Chen. Unveiling and mitigating generalized biases of dnns through the intrinsic dimensions of perceptual manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Pratyush Maini, Saurabh Garg, Zachary Lipton, and J Zico Kolter. Characterizing datapoints via second-split forgetting. *Advances in Neural Information Processing Systems*, 35:30044–30057, 2022.

Raffaele Marchesi, Nicolo Micheletti, Nicholas I-Hsien Kuo, Sebastiano Barbieri, Giuseppe Jurman, and Venet Osmani. Generative ai mitigates representation bias and improves model fairness through synthetic health data. *medRxiv*, pp. 2023–09, 2023.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Hariharan Narayanan and Partha Niyogi. On the sample complexity of learning smooth cuts on a manifold. In *COLT*, 2009.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Pawel Pukowski and Haiping Lu. Investigating the impact of hard samples on accuracy reveals in-class data imbalance. *arXiv preprint arXiv:2409.14401*, 2024.

M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

Zhijun Ren, Tantao Lin, Ke Feng, Yongsheng Zhu, Zheng Liu, and Ke Yan. A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 72:1–35, 2023.

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Dissecting sample hardness: A fine-grained analysis of hardness characterization methods for data-centric AI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=icTZCUbtD6.

Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian conference on computer vision*, 2020.

Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-difficulty based methods for long-tailed visual recognition. *International Journal of Computer Vision*, 130(10):2517–2531, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Erdal Tasci, Ying Zhuge, Kevin Camphausen, and Andra V Krauze. Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets. *Cancers*, 14(12):2897, 2022.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Duc-Quang Vu, Trang TT Phung, Jia-Ching Wang, and Son T Mai. Lcsl: long-tailed classification via self-labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11): 12048–12058, 2024.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International conference on machine learning*, pp. 36246–36263. PMLR, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Suqin Yuan, Runqi Lin, Lei Feng, Bo Han, and Tongliang Liu. Instance-dependent early stopping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P42DbV2nuV.

Ming Zheng, Tong Li, Rui Zhu, Yahui Tang, Mingjing Tang, Leilei Lin, and Zifei Ma. Conditional wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification. *Information Sciences*, 512:1009–1023, 2020.

Beier Zhu, Kesen Zhao, Jiequan Cui, Qianru Sun, Yuan Zhou, Xun Yang, and Hanwang Zhang. Reducing class-wise performance disparity via margin regularization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KfjpyOcPQj.

Yongshuo Zong, Yongxin Yang, and Timothy M. Hospedales. Medfair: Benchmarking fairness for medical imaging. In *ICLR*, 2023. URL https://openreview.net/forum?id=6ve2CkeQe5S.

# A AI USE STATEMENT

In this work, we used generative AI tools for refining hypotheses, providing feedback on research methodology and experiments, implementing methods. We have not used generative AI tools for all other tasks with required disclosure, or those are not applicable to this work. We have reviewed all AI-assisted work. The code has been thoroughly verified and documented, any rewrites proposed by the generative AI tool has been read thoroughly to ensure no hallucination and reduce the amount of hyperbolization. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

# B DETAILED EXPERIMENTAL DESIGN

**Hardness estimation.** HBR relies on hardness estimates at two stages: (i) to compute resampling ratios by extension identifying easy and hard classes (Equation 3), and (ii) to guide pruning of easy classes. Ideally, both stages—and both case studies—would rely on identical hardness estimates. However, this is problematic in *Scenario A* (case study 1), where the subsampling alters the relative hardness of many samples: an originally easy sample may become hard simply due to sampling effects. Conversely, recomputing hardness estimates for each scenario would incur significant computational overhead (especially considering the instability of single seed estimates, and cost of producing multiple-seed estimates). We therefore use a single, consistent set of hardness estimates (AUM) across all settings, while acknowledging that this introduces some noise in *Scenario A*.

We compute AUM values by averaging results from ten baseline models trained on unmodified versions of CIFAR-10 and CIFAR-100. We chose this number as we found the estimates to stabilize at this point. We adopt AUM as our default estimator because the benchmark of Seedat et al. (2024) identified it as one of the best-performing hardness estimators. We compute the resampling ratio based on those hardness estimates. We offset each class-hardness by the hardness of the hardest class (the one with lowest AUM) plus some epsilon if it is negative to ensure sensible resampling ratios. After that we invert AUM to ensure high values correspond to hard classes (to match equation 3)

**Case Study 1.** *Design.* We prune the dataset of interest using thresholds of $33\%$, $42\%$, $50\%$, $58\%$, $66\%$, and $75\%$. Smaller pruning thresholds are omitted because, for these values, the maximum number of samples after resampling ($b_c$) exceeds the available sample count $a_c$ in the holdout set for the hardest classes. After pruning via class-level pruning (**CLP**), which removes a fixed number of random samples from each class, we perform HBR on $\mathcal{D}^{\text{CLP}}$ with $\alpha = 1$.

**Case Study 2.** *Design.* In this case study, we compute resampling ratios using AUM as the proxy for $\hat{h}_c$. We set $\alpha \in \{1, 3, 5\}$ for CIFAR-10 and $\alpha \in \{1, 2\}$ for CIFAR-100. Larger $\alpha$ values are used for CIFAR-10 to compensate for its weaker hardness imbalance relative to CIFAR-100. We find that setting $\alpha = 3$ for CIFAR-100 leads to a significant drop in overall performance although the reduction in class bias is also more substantial. Hence, the optimal $\alpha$ is likely somewhere between 2 and 3 for CIFAR-100, although finding it is beyond the scope of this work.

*Oversampling strategies.* We employ three main oversampling strategies: (1) random oversampling, (2) SMOTE, and (3) synthetic samples generated using EDM (Karras et al., 2022). Since training a diffusion model is computationally prohibitive, we use one million of synthetic samples made publicly available by Wang et al. (2023). This is also why we do not use diffusion models in Case Study 1, as retraining them for each *Scenario A* is significantly beyond the scope of this work. We further define three EDM sub-strategies:

1. **rEDM (random EDM):** We simply pick synthetic images from the target class at random, without considering how hard or easy they are. This serves as a baseline.

2. **aEDM (average EDM):** For each class, we first compute the average confidence of the real training images. We then select synthetic images whose hardness scores are closest to that class average. The idea is to match the typical difficulty level of the class, neither too easy nor too hard.

3. **hEDM (hardest EDM):** We select the synthetic images with the highest hardness scores. These samples are most likely to lie near class boundaries, in low-density or high-complexity regions.

For both aEDM and hEDM, we first pre-select a candidate pool per class: the $2b_c$ synthetic images whose confidence is closest to the class average (for aEDM) or the $2b_c$ hardest images (for hEDM), where $b_c$ is the target number of samples after resampling. During each resampling iteration, we then draw randomly from this pre-selected pool. This ensures a sufficiently large and hardness-targeted supply while maintaining stochasticity across training runs.

Throughout these definitions, we estimate hardness using the confidence of the ten baseline models—the same models used to compute AUM on real data. For real data, we could also use AUM, but for consistency we use confidence for both real and synthetic samples. We deliberately avoid AUM for synthetic samples because AUM relies on the evolution of the margin over training epochs; we lack this learning dynamics information for synthetic data without retraining all ten models from scratch on the combined real-plus-synthetic dataset. Using confidence from the already-trained baseline models provides a practical, consistent, and well-defined proxy for hardness on both real and synthetic data.

**Experimental setup** *Model and optimization.* We use a ResNet-18 variant adapted for low-resolution inputs: the original $7 \times 7$ convolution and initial max-pooling are replaced with a single $3 \times 3$ convolution. All models are trained with stochastic gradient descent (SGD) with momentum $0.9$, initial learning rate $0.1$, and weight decay $5 \times 10^{-4}$. The learning rate is multiplied by $0.2$ at epochs $60$, $120$, and $160$. Training runs for $200$ epochs with batch size $128$.

## C Ensuring statistical reliability and reproducibility

Model performance in both case studies depends on stochastic components arising from data pruning and resampling. To ensure robustness and enable paired statistical testing, we adopt a deterministic seed scheme that guarantees reproducibility and one-to-one correspondence between models across datasets.

In Case Study 1, the source of randomness differs between the base and resampled datasets. For models trained on the base pruned datasets ($\mathcal{D}^{CLP}$), randomness originates solely from the pruning process. For models trained on the corresponding resampled datasets, randomness arises from both pruning and resampling. Hence, to ensure robust results we generate four independently pruned datasets for each pruning threshold, each using a distinct pruning seed, and train four independently initialized models on each of them (i.e., four pruned datasets × four models each). Model initialization seeds are computed deterministically as

$$\text{seed}_{\text{model}}(i, j) = 42 + 420{,}000 \cdot i + 42 \cdot j,$$

where $i \in \{0, 1, 2, 3\}$ indexes the dataset replicate and $j \in \{0, 1, 2, 3\}$ indexes the model within that replicate. These model seeds are independent of both the pruning and resampling seeds. Meanwhile, the pruning seed used to obtain $\mathcal{D}^{CLP}$ is set as

$$\text{seed}_{\text{prune}}(i) = 42 \cdot i.$$

In Case Study 2, randomness stems exclusively from resampling. Most oversampling and undersampling methods (except deterministic SMOTE) yield different datasets when initialized with different seeds. Accordingly, we again generate four independently resampled datasets, each controlled by its own resampling seed, and train four independent models per dataset using the same deterministic model-seed scheme as above. This design yields a total of $4 \times 4 = 16$ model-dataset combinations in both case studies.

## D Statistical significance analysis

To quantify the consistency of the observed effects in Case Study 1, we compare the fairness metrics obtained from models trained on the base pruned dataset $D_{\text{CLP}}$ with those obtained from models trained on the corresponding resampled datasets. The experimental design provides a natural pairing: for each dataset replicate and model initialization, the base and resampled models share the

same dataset-replicate and model-initialization indices. We therefore obtain $n = 16$ matched pairs for each experimental condition.

Let $X_i$ denote a fairness metric obtained from the $i$-th model trained on the base dataset $D_{\mathrm{CLP}}$, and let $Y_i$ denote the corresponding metric obtained from the model trained on the resampled dataset. For metrics for which lower values indicate better fairness, we define the paired difference as

$$D_i = X_i - Y_i,$$

so that positive values correspond to a reduction in the measured class disparity. For metrics for which higher values indicate better performance, the difference is defined in the opposite direction,

$$D_i = Y_i - X_i,$$

so that positive values consistently indicate an improvement.

We assess whether the mean paired difference differs from zero using a two-sided paired-samples $t$-test. The null hypothesis is

$$H_0 : \mu_D = 0,$$

while the alternative hypothesis is

$$H_A : \mu_D \neq 0.$$

The test statistic is

$$t = \frac{\bar{D}}{s_D/\sqrt{n}},$$

where

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

is the mean paired difference and

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

is the sample variance of the paired differences. The resulting statistic is compared against a $t$-distribution with $n-1 = 15$ degrees of freedom. Statistical significance is assessed using a threshold of $p < 0.05$.

Because the test is two-sided, statistical significance indicates that the mean paired difference is distinguishable from zero; the sign of the corresponding $t$-statistic determines the direction of the observed effect. Positive values indicate an improvement under the convention defined above, whereas negative values indicate a change in the opposite direction.

The $t$-statistics for Case Study 1 are reported below. Significant results ($p < 0.05$) are bolded.

## E Results of paired t-test and further analysis

**More pronounced frequency imbalance leads to higher fairness gains.** Beyond dataset-specific variations, we also analyze how the severity of the induced imbalance (controlled by $\alpha$) modulates these trends (see Figure 4). Consistent with expectations, we observe that in most cases greater

Table 1: **Case Study 1 (CIFAR-10):** t-statistics for gap- and dispersion-based metrics under different oversampling strategies. Significant results ($p < 0.05$) are **bolded**. Positive values indicate a reduction in the corresponding class-bias metric, whereas negative values indicate an increase.

| Fairness metric | Pruning rate | random oversampling | | SMOTE oversampling | | holdout oversampling | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision | Recall | Precision |
| $\Delta_m$ | 33 | 0.17 | -0.52 | 1.82 | -1.57 | **4.52** | **-3.93** |
| | 42 | -0.93 | -0.12 | -0.36 | -0.59 | **3.24** | -1.33 |
| | 50 | -1.41 | 1.02 | 0.73 | 0.20 | **5.37** | -1.75 |
| | 58 | -2.09 | 1.60 | -0.22 | -0.70 | 1.63 | -0.82 |
| | 66 | -0.77 | 0.98 | -0.50 | -0.65 | **2.51** | **-2.89** |
| | 75 | -0.61 | 0.40 | 0.12 | -0.26 | **3.24** | -0.76 |
| $\Delta_q$ | 33 | 0.89 | -1.38 | 1.77 | **-2.42** | **5.00** | **-3.81** |
| | 42 | -0.91 | -0.76 | -0.11 | -1.45 | **5.49** | -1.11 |
| | 50 | -0.10 | -0.71 | 1.70 | -0.37 | **6.20** | -2.09 |
| | 58 | -1.92 | -0.57 | -0.69 | -1.01 | **2.89** | -0.69 |
| | 66 | -0.74 | -0.45 | -0.16 | -1.40 | **2.29** | **-2.38** |
| | 75 | -0.18 | -1.27 | 0.87 | **-3.54** | **5.08** | **-2.29** |
| $\Delta_{he}$ | 33 | **2.15** | -1.53 | **3.03** | **-2.61** | **5.16** | **-2.90** |
| | 42 | -0.10 | 0.59 | 1.29 | -1.61 | **6.63** | -0.94 |
| | 50 | 1.63 | -1.53 | **2.92** | -1.15 | **6.09** | -1.33 |
| | 58 | -0.33 | -0.76 | 1.31 | -1.87 | **3.72** | -0.96 |
| | 66 | 1.40 | 1.04 | **2.72** | -0.74 | **6.52** | -2.04 |
| | 75 | 1.22 | -0.11 | **2.54** | **-2.16** | **9.22** | -1.58 |
| $\Delta_\sigma$ | 33 | 0.97 | -1.21 | 1.97 | **-2.27** | **4.82** | **-4.64** |
| | 42 | -0.42 | -0.86 | 0.40 | -1.48 | **6.03** | -1.44 |
| | 50 | -0.34 | -1.18 | 1.23 | -0.56 | **5.93** | -1.99 |
| | 58 | -1.89 | -0.56 | -0.41 | -1.51 | **2.46** | -0.72 |
| | 66 | -0.69 | -0.32 | 0.28 | -1.36 | **3.81** | **-2.68** |
| | 75 | 0.14 | -1.60 | 0.91 | **-2.27** | **5.77** | -1.88 |
| $\Delta_{MAD}$ | 33 | 0.12 | -0.48 | 1.17 | -1.71 | 1.07 | 0.16 |
| | 42 | -1.30 | -0.04 | -1.21 | -1.14 | -1.57 | 1.30 |
| | 50 | -1.73 | **-3.85** | -0.26 | **-5.91** | -0.36 | **-2.17** |
| | 58 | -0.74 | -1.92 | -2.08 | -0.76 | -0.00 | 0.35 |
| | 66 | 0.34 | -1.43 | 0.41 | -1.44 | -0.30 | -0.70 |
| | 75 | -1.45 | **-2.49** | -0.92 | -0.72 | -0.98 | 0.64 |

imbalance yields stronger and more statistically significant gains from HBR. Although our experiments do not explicitly identify a turning point, we hypothesize the existence of a threshold $\alpha_t$ beyond which the harm from frequency imbalance outweighs the benefits of mitigating hardness imbalance. Designing an algorithm to efficiently estimate this threshold is beyond the scope of this work.

**Changes to class bias are not restricted to the most extreme classes.** We find that the relative magnitudes of changes to class bias do not directly align with their statistical significance. While HBR yields the largest absolute reductions in $\Delta_m$, followed by $\Delta_q$ and $\Delta_{he}$, the corresponding $t$-values show the reverse trend. This discrepancy suggests that improvements in the recall gap across most extreme classes ($\Delta_m$ and $\Delta_q$) are more variable across runs, but also more pronounced. Meanwhile, improvements spread across all hard and easy classes ($\Delta_{he}$) are smaller but more consistent. This can be explained by higher robustness of $\Delta_{he}$ as a fairness metric, as we discussed in Section 3.3. Paired with the fact that $\Delta_{MAD}$ gets reduced on CIFAR-100, this indicates that *the impact of HBR is not confined to the most extreme classes but rather spread across the whole hardness spectrum*. We believe that $\Delta_{MAD}$ increasing on CIFAR-10 as a result of HBR does not contradict these insights but rather is a consequence of the unique hardness spectrum characterizing this dataset. As is visible in Figure 1, while on CIFAR-100 recall decreases almost linearly across classes, CIFAR-

Table 2: **Case Study 1 (CIFAR-100):** t-statistics for gap- and dispersion-based metrics.

| Fairness metric | Pruning rate | random oversampling | | SMOTE oversampling | | holdout oversampling | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision | Recall | Precision |
| $\Delta_m$ | 33 | 0.99 | **-3.02** | 1.95 | **-4.61** | **4.73** | **-3.83** |
| | 42 | **3.32** | -0.62 | **2.33** | -1.05 | **6.40** | **-2.29** |
| | 50 | 1.48 | **-3.43** | 1.44 | **-3.33** | **8.68** | **-3.03** |
| | 58 | -1.14 | -2.12 | -0.00 | -2.01 | **5.09** | **-4.07** |
| | 66 | **6.15** | -1.54 | **3.54** | -1.74 | **11.41** | **-3.54** |
| | 75 | 1.24 | **-2.86** | 1.50 | **-3.91** | **8.86** | **-5.47** |
| $\Delta_q$ | 33 | **2.54** | **-4.99** | **3.18** | **-4.80** | **12.17** | **-6.24** |
| | 42 | **5.09** | **-4.09** | **4.86** | **-4.31** | **10.49** | **-7.44** |
| | 50 | **4.19** | **-4.65** | **4.58** | **-4.32** | **10.98** | **-4.86** |
| | 58 | **4.82** | **-3.57** | **7.45** | **-3.60** | **18.91** | **-4.57** |
| | 66 | **12.54** | -1.97 | **11.23** | **-2.55** | **18.24** | **-3.19** |
| | 75 | **7.72** | **-4.11** | **7.29** | **-6.41** | **21.04** | **-6.22** |
| $\Delta_{he}$ | 33 | **9.10** | **-7.61** | **8.98** | **-7.97** | **17.57** | **-9.58** |
| | 42 | **9.61** | **-6.00** | **7.73** | **-9.19** | **13.90** | **-12.86** |
| | 50 | **8.01** | **-5.01** | **12.06** | **-6.26** | **16.80** | **-8.43** |
| | 58 | **11.04** | **-4.70** | **11.33** | **-2.83** | **23.33** | **-4.06** |
| | 66 | **13.53** | **-4.89** | **11.09** | **-4.61** | **22.57** | **-8.69** |
| | 75 | **20.60** | **-5.75** | **12.13** | **-8.14** | **26.02** | **-10.53** |
| $\Delta_\sigma$ | 33 | **3.41** | **-6.99** | **4.28** | **-6.23** | **15.18** | **-7.25** |
| | 42 | **4.82** | **-6.14** | **4.87** | **-5.85** | **10.20** | **-10.66** |
| | 50 | **7.06** | **-5.26** | **6.67** | **-5.56** | **15.11** | **-6.09** |
| | 58 | **5.76** | **-3.94** | **9.60** | **-3.49** | **16.25** | **-5.09** |
| | 66 | **11.13** | -1.24 | **12.27** | -1.63 | **22.35** | **-2.97** |
| | 75 | **8.36** | **-5.50** | **6.37** | **-7.11** | **19.86** | **-7.82** |
| $\Delta_{MAD}$ | 33 | 2.08 | **-4.47** | 1.37 | **-3.69** | **4.12** | **-3.13** |
| | 42 | **3.35** | **-3.20** | 1.29 | **-3.98** | **4.48** | **-3.04** |
| | 50 | 0.15 | **-3.72** | 1.62 | **-2.37** | **4.81** | **-2.82** |
| | 58 | 1.85 | -1.83 | **4.26** | -1.77 | **4.65** | **-3.04** |
| | 66 | **4.53** | -0.21 | **6.16** | -0.13 | **7.93** | -0.16 |
| | 75 | **7.91** | **-2.64** | **2.95** | **-2.93** | **5.23** | -1.91 |

Table 3: **Case Study 1:** t-statistics for $\Delta_{avg}$ under different oversampling strategies.

| Metric | Pruning rate | CIFAR-10 | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|---|
| | | random | SMOTE | holdout | random | SMOTE | holdout |
| Recall | 33 | **-2.30** | -0.54 | 0.07 | **-12.87** | **-12.14** | **-3.68** |
| | 42 | -1.12 | -1.95 | 0.17 | **-14.20** | **-12.68** | **-3.42** |
| | 50 | -1.54 | -0.91 | 1.43 | **-9.53** | **-9.51** | **-3.57** |
| | 58 | -1.82 | -1.55 | 0.83 | **-14.23** | **-12.74** | **-4.50** |
| | 66 | -1.48 | -1.57 | 0.25 | **5.52** | **6.17** | **14.51** |
| | 75 | -1.41 | -1.46 | 0.95 | **-8.28** | **-10.74** | **-4.46** |
| Precision | 33 | **-2.17** | -0.26 | 0.53 | **-9.65** | **-9.44** | 0.72 |
| | 42 | -1.08 | -1.88 | 0.40 | **-9.26** | **-11.05** | **2.53** |
| | 50 | -1.49 | -0.80 | 2.03 | **-8.02** | **-7.83** | **3.13** |
| | 58 | -1.88 | -1.52 | 0.95 | **-9.89** | **-8.36** | 0.67 |
| | 66 | -1.44 | -1.43 | 0.73 | **7.76** | **8.05** | **19.42** |
| | 75 | -1.34 | -1.32 | 1.38 | **-6.27** | **-7.86** | **4.02** |

10 recall follows a logistic-like trend, with only a few low-performing classes and most classes clustered around similar recall values. This means that on CIFAR-10 the effects of HBR are more related to the hardest classes, and any changes to the $\Delta_{MAD}$ are bound to be insignificant and inconsistent—if most classes already have similar recall, there is little room for median dispersion to change.

## F ADDITIONAL EXPERIMENTAL RESULTS

In this section we include further experiments conducted for Section 4. Specifically, we show the changes to true positives, true negatives, false positives, and false negatives across classes as an effect of HBR for top-5 most extreme classes. These classes were chosen based on the recall averaged over the sixteen models trained on $\mathcal{D}^{CLP}$ with pruning rate of $50\%$. This means that only the five classes with the highest and lowest averaged recalls are displayed, with the former representing the easiest classes and the latter the hardest ones. We decided to use recall instead of AUM to sort classes, as we consider it as ground truth for hardness (just like accuracy is often used as a ground truth for hardness in relevant literature).

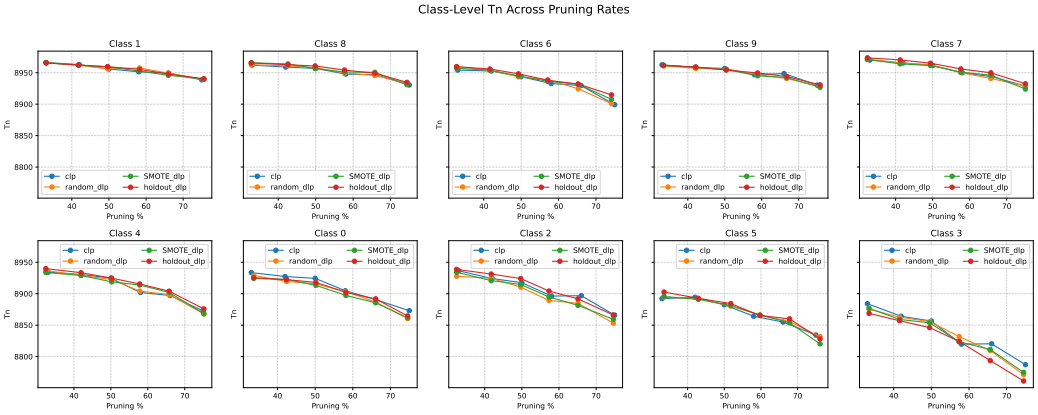


Figure 7: True negatives across classes on CIFAR-10.

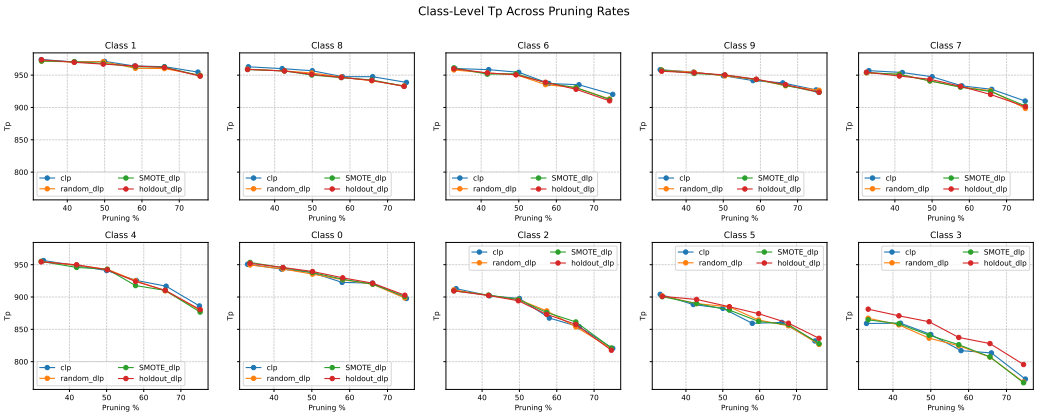


Figure 8: True positives across classes on CIFAR-10.

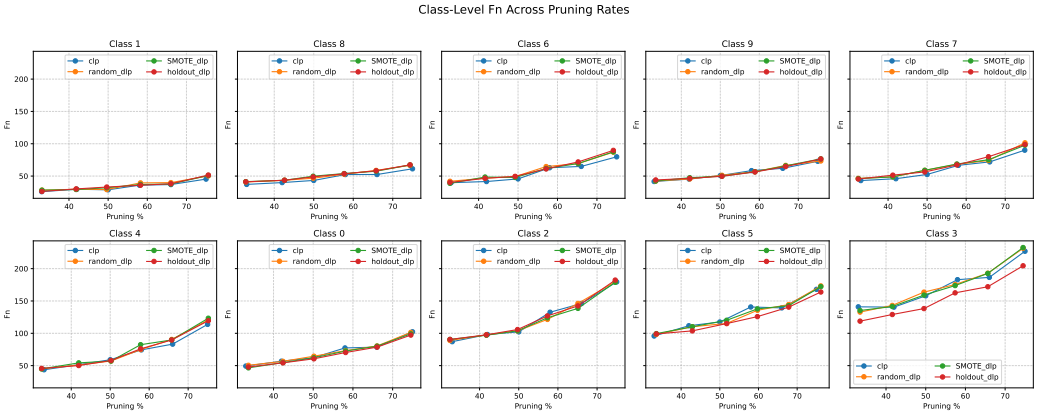


Figure 9: False negatives across classes on CIFAR-10.

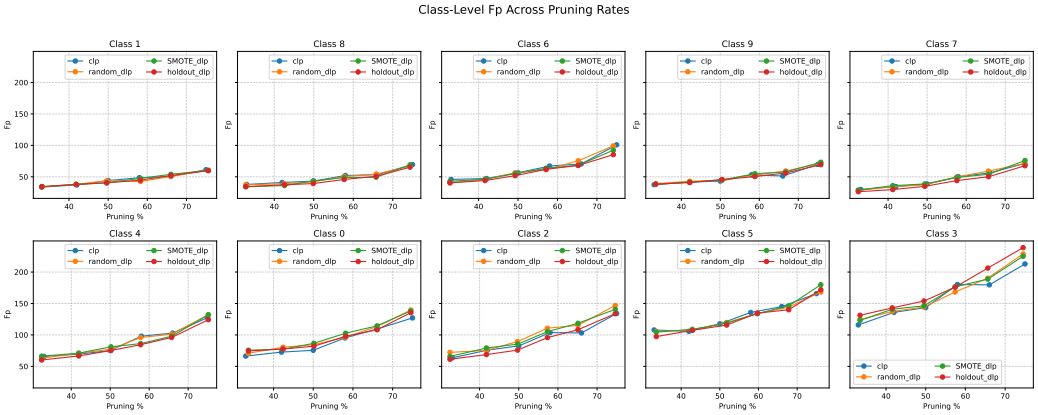


Figure 10: False positives across classes on CIFAR-10.

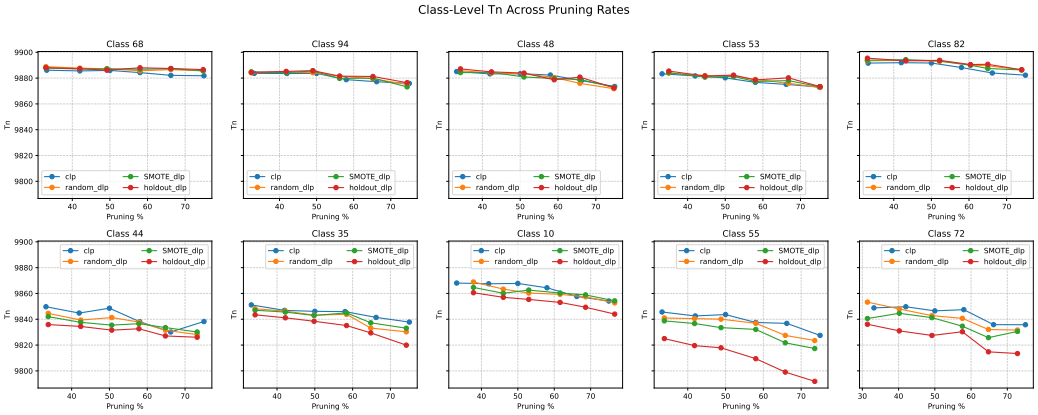


Figure 11: True negatives across classes on CIFAR-100.

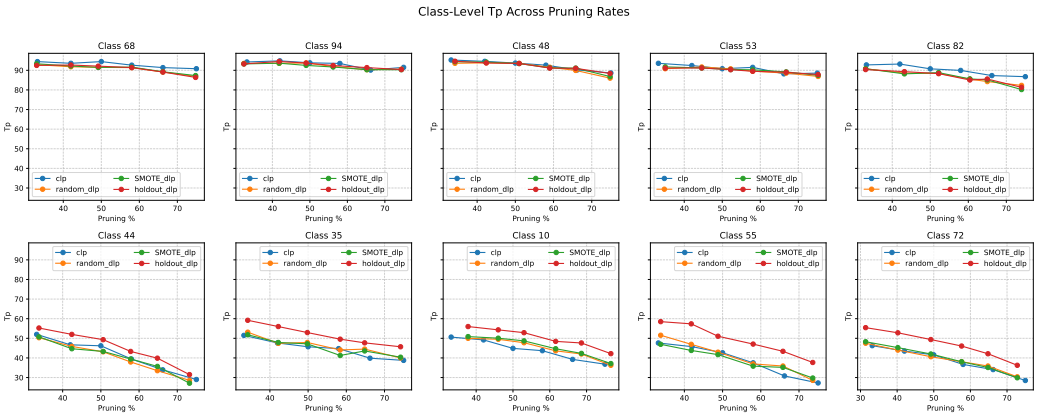


Figure 12: True positives across classes on CIFAR-100.

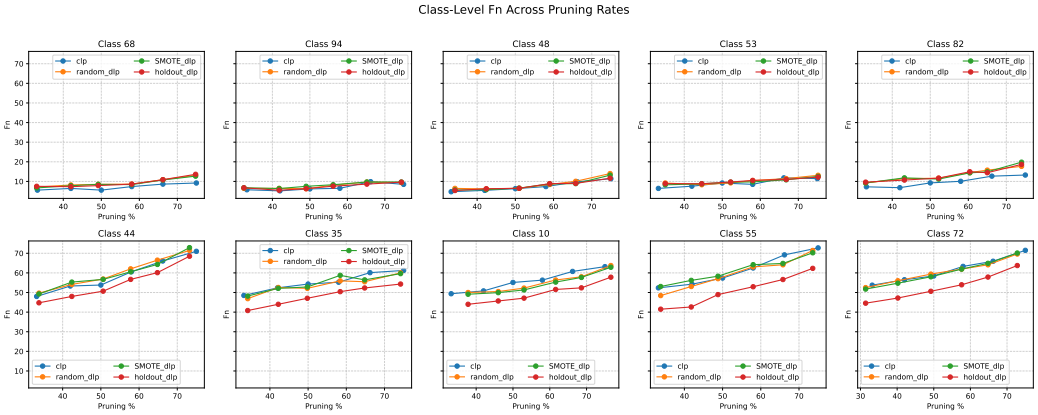


Figure 13: False negatives across classes on CIFAR-100.

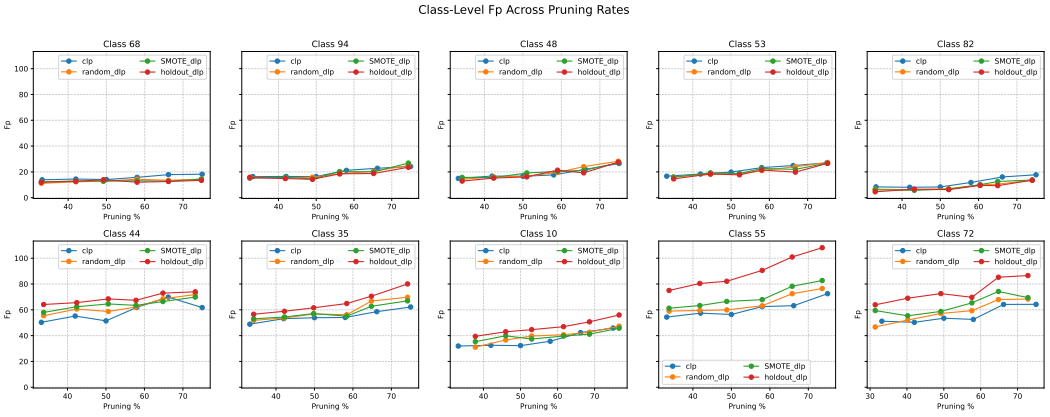


Figure 14: False positives across classes on CIFAR-100.