Where to Intervene? Benchmarking Fairness-Aware Learning on Differentially Private Synthetic Tabular Data

Vinícius Gabriel Angelozzi
Centre Inria de l'Université Grenoble Alpes
verona.projects@tutanota.com

Héber H. Arcolezi
ÉTS Montréal, Inria Grenoble
heber.hwang-arcolezi@etsmtl.ca

Abstract

Machine learning models are increasingly deployed in high-stakes domains, raising concerns about both privacy and fairness. Differential Privacy (DP) has become a gold standard for privacy-preserving data analysis, while fairness-aware mechanisms aim to mitigate discrimination against underrepresented groups. However, these objectives can conflict: DP often amplifies disparities across demographic groups, and little is known about whether established fairness interventions remain effective under DP constraints. In this work, we present, to our knowledge, the first systematic evaluation of fairness interventions on differentially private synthetic tabular data. Our benchmark centers on the *Adaptive Iterative Mechanism (AIM)*, identified as the state-of-the-art marginal-based DP synthesizer in recent KDD & VLDB 2025 tutorials by Cormode et al. [20]. We thus evaluate fairness interventions across four datasets, multiple group fairness metrics, and three categories of mitigation strategies (pre-processing, in-processing, and post-processing) under a wide range of privacy budgets. We compare four pipeline configurations: (*Baseline*) training on original data; (*DP-only*) training on DP synthetic data; (*Fair-only*) applying fairness mechanisms on original data; and (*DP+Fair*) combining fairness mechanisms with DP synthetic data. Our results demonstrate that while DP alone can degrade both utility and fairness, applying fairness interventions can partially restore equitable outcomes. Among them, *post-processing methods tend to provide more stable fairness–utility trade-offs across privacy budgets and synthesizers*, achieving strong fairness improvements while preserving competitive utility relative to other intervention stages. We release all code, data, and experimental artifacts in an open-source repository (https://github.com/vinicius-verona/dp-fair-intervention-benchmark) to ensure full reproducibility and to support future research on the privacy-fairness-utility trade-off.

Keywords

Differential Privacy, Synthetic Data, Algorithmic Fairness, Privacy–Fairness Trade-off.

1 Introduction

Synthetic tabular data generation is increasingly adopted by the research community [35, 36, 39], regulators [18, 29], and industry [1, 2] as a promising approach to facilitate data sharing and downstream analysis while reducing disclosure risks. These methods aim to approximate the empirical distribution of real datasets and generate artificial records that preserve key statistical properties. However, synthetic data does not guarantee privacy or fairness by default. On the one hand, without formal protections, generative models may memorize or leak sensitive information from the training data, making them vulnerable to membership inference, attribute inference, or reconstruction attacks [30, 33, 53, 60]. Moreover, synthetic data can replicate or even amplify societal biases present in the data, leading to unfair outcomes in downstream predictive models [9].

Differential privacy (DP) [23, 24] provides rigorous, quantifiable protection by bounding the influence of any single record on model training or statistics. DP-based synthetic data generators [44, 55, 62] address the privacy gap by injecting calibrated noise during training or sampling. Most recent benchmarks have therefore concentrated on the *utility* of DP synthetic data, evaluating predictive performance across generative models, tasks, and privacy levels [32, 51, 52, 54]. Yet, beyond aggregate utility, the statistical distortions introduced by formal privacy guarantees may interact with subgroup structure in non-trivial ways.

Beyond utility, however, the interplay between DP and fairness has received far less attention, and is often more complex [28, 59]. When DP is applied directly to model training (*e.g.*, via DP-SGD [4] or PATE [47]), prior work has shown that noise reduces overall accuracy but disproportionately harms minority or underrepresented groups [8]. This phenomenon, often described as *DP disparate impact*, raises the question of whether similar effects also emerge in models trained on DP synthetic data.

Recent studies confirm that they do. Work on *DP synthetic data* [14, 31, 48] has shown that privacy mechanisms and data characteristics interact in subtle ways, often leading to heterogeneous fairness outcomes across groups. However, these evaluations are largely observational: they measure disparities (*e.g.*, statistical parity or subgroup accuracy) but do not examine whether *fairness-aware learning mechanisms* remain effective when applied to DP synthetic data.

This gap motivates our central research question: ***"Where should one intervene in the machine learning (ML) pipeline to mitigate unfairness under DP synthetic data?"*** Should interventions target the data distribution (pre-processing), the training procedure (in-processing), or the model outputs (post-processing)? More fundamentally, how do formal differential privacy guarantees reshape the stability, effectiveness, and relative competitiveness of these intervention strategies across privacy budgets?

The question is timely. Several commercial platforms now offer *DP synthetic data* as a product [30, Table 1], including vendors such as Tumult Labs [1] and YData [2]. These platforms are already being used in sensitive domains such as healthcare and finance. As fairness concerns grow, both ethically and legally (*e.g.*, EU AI

Proceedings on Privacy Enhancing Technologies 2026(4), 53–82

https://doi.org/10.56553/popets-2026-0110

Act [25]), developers may need to integrate fairness-aware interventions *without redesigning entire DP generation pipelines*.

Contributions. Building on these observations, we conduct a structured benchmarking study of fairness mitigation strategies in DP synthetic data pipelines. Our goal is to characterize when and how fairness-aware mechanisms can counteract disparate impact introduced by differential privacy, while preserving predictive utility under explicit privacy budgets. To ensure relevance and rigor, we center our study on the *Adaptive Iterative Mechanism (AIM)* [44], recognized as the current *state-of-the-art differentially private synthesizer for tabular data* in recent *KDD & VLDB 2025 tutorials by Cormode et al. [20]* and *utility-driven* benchmarks [32, 51, 52, 54]. For completeness, we also report results with the *Maximum Spanning Tree (MST)* [43] synthesizer, winner of the NIST Differential Privacy Synthetic Data Challenge, in Appendix B.3. By focusing on widely adopted marginal-based DP generators, our study isolates the interaction between formal privacy guarantees and downstream fairness mitigation. In summary, our main contributions are:

- We present, to our knowledge, the first systematic evaluation of fairness interventions applied to classifiers trained on DP synthetic tabular data, explicitly analyzing how differential privacy reshapes fairness–utility trade-offs. An overview of our benchmark pipeline is shown in Figure 1.
- We evaluate pre-, in-, and post-processing fairness interventions under varying privacy budgets, revealing how the effectiveness and stability of these mechanisms change when applied to DP synthetic rather than original data.
- Using multiple real-world datasets, three ML classifiers, and two state-of-the-art DP synthesizers, we characterize how DP alters fairness–utility dynamics and identify conditions under which interventions are more or less effective at mitigating unfairness.
- We release all code, datasets, and experimental artifacts to support reproducibility and enable future research at the intersection of privacy and fairness (see Section 4.7 and the **GitHub**: https://github.com/vinicius-verona/dp-fair-intervention-benchmark).

Findings. Our results show that fairness interventions applied to models trained on DP synthetic data can partially recover fairness degradation induced by DP noise, although the effect is metric-, budget-, and stage-dependent. Across datasets, privacy budgets, utility metrics, and learning models, Reweighing (RW) [15], Reject Option Classification (ROC) [37], and Equalized Odds Post-Processing (EqOdds) [34] emerge as the most reliable interventions. RW provides stable fairness improvements, especially for parity-oriented metrics, but often incurs measurable utility loss. In contrast, ROC and EqOdds tend to offer the strongest fairness–utility trade-offs, frequently appearing on or near the Pareto frontier, particularly for EOD- and SPD-oriented analyses. In-processing mechanisms, especially EGR, preserve utility close to DP-only but achieve more limited disparity reductions in this DP synthetic setting.

2 Related Work

Privacy and fairness. The interaction between privacy and fairness has been widely studied [3, 7, 8, 16, 26, 32, 41, 42, 56]. A consistent finding is that differentially private learning (*i.e.*, through DP-SGD [4] or PATE [47]) can exacerbate disparities across demographic groups, a phenomenon often referred to as the *disparate impact of DP*. For instance, [8] shows that DP-SGD reduces overall accuracy but disproportionately harms minority groups. Surveys such as [28, 59] further highlight that privacy and fairness objectives may align in some regimes but conflict in others, depending on the data distribution and model class. These studies primarily analyze fairness when DP is applied directly to model training.

DP synthetic data and utility. The release of *DP synthetic data* has motivated a growing line of work benchmarking the *utility* of different generative models. Early studies compare marginal-based synthesizers (*e.g.*, AIM [44], MST [43]) with deep generative approaches (*e.g.*, DP-GANs [55, 58]), showing that marginal-based methods often yield higher utility on tabular datasets [32, 51, 52, 54]. These results position marginal-based models as competitive baselines for structured domains.

DP synthetic data and fairness. More recent work evaluates fairness explicitly. [14] compares four DP synthesizers across multiple datasets and privacy budgets, finding that most degrade fairness but that MST behaves more favorably than GAN-based alternatives. Similarly, [48] analyzes fairness and utility metrics for both GAN-based and marginal-based synthesizers, showing that the latter tend to preserve subgroup accuracy and often maintain or improve group fairness metrics. The work in [31] provides a fine-grained analysis of subgroup disparities in DP synthetic data, showing that DP can either amplify or mitigate imbalance depending on the model, and that classifiers trained on such data exhibit reduced performance for minority groups. These studies demonstrate that privacy-preserving data generation can reshape fairness properties, but they remain largely observational.

Positioning of our work. Prior works on fairness in DP synthetic data [14, 31, 48] primarily measure how differential privacy affects fairness metrics after data release. They do not systematically analyze how *fairness-aware intervention mechanisms* behave when deployed on top of DP synthetic data under varying privacy budgets. In particular, the interaction between intervention stages (pre-, in-, or post-processing) and formal privacy constraints has not been jointly characterized in a unified empirical setting. *Our work addresses this gap* by studying fairness mitigation strategies applied *after* DP synthetic data generation, explicitly isolating the downstream effects of formal privacy guarantees on fairness–utility trade-offs. Rather than designing jointly fair-and-private synthesizers (*e.g.*, [38]), we focus on the practical and increasingly common deployment scenario in which DP synthetic data is produced first, and fairness considerations are addressed subsequently. This separation allows us to characterize how established fairness mechanisms behave under fixed privacy budgets, and to identify which intervention stages remain stable or degrade under privacy constraints.

3 Preliminaries

In this section, we briefly review about generative models, differential privacy, and fairness-aware learning.

3.1 Generative Models and Synthetic Data

Generative models aim to approximate the distribution of real data and to produce artificial records that preserve its statistical properties. Let $D = \{x_i\}_{i=1}^{n}$, with $x_i \in \mathcal{X}$, denote the original dataset drawn *i.i.d.* from an unknown data-generating distribution p_{data}. A

generative model learns a parameterized distribution p_θ intended to approximate p_{data}. Synthetic records are then generated by sampling $\tilde{x}_j \sim p_\theta$, yielding a synthetic dataset $\tilde{D} = \{\tilde{x}_j\}_{j=1}^m$ that is statistically similar to D. This synthetic dataset is released in place of the original data for downstream analysis, with the goal of enabling utility while mitigating disclosure risk.

While many approaches exist for building generative models, in this work we focus on marginal-based methods [43, 44]. These synthesizers are rooted in Bayesian network formulations that decompose the joint distribution of tabular data into lower-dimensional marginals and conditional dependencies. Such approaches have consistently shown strong performance on structured tabular data [32, 54], where feature dependencies can be effectively captured by explicit probabilistic modeling. In contrast, deep generative models, while highly successful for image or text synthesis, often face challenges in faithfully modeling heterogeneous tabular data.

3.2 Differential Privacy

Differential privacy is a property of randomized mechanisms that limits how much the output distribution can change when a single individual's record is modified. Intuitively, it enables learning about the population while revealing little about any one person [24]. We adopt the standard ε-DP notion [23].

Definition 1 (ε-Differential Privacy). *Let $\mathcal{X}$ be the data domain and let datasets $D, D' \in \mathcal{X}^n$ be* neighbors *(written $D \sim D'$) if they differ in exactly one individual's record. Let $\mathcal{O}$ denote the output space of a randomized mechanism. A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ satisfies ε-DP if for all measurable $S \subseteq \mathcal{O}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in S].$$

When $\varepsilon = 0$, outputs for neighboring datasets are identically distributed, implying that the mechanism's output cannot depend on any single record. Larger ε permits greater sensitivity to an individual, weakening privacy. Thus, choosing ε entails a privacy-utility trade-off. We now recall the post-processing property of differential privacy[1], which is central to our deployment model.

Proposition 1 (DP Post-Processing Property [24]). *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ be an ε-differentially private mechanism, and let $f : \mathcal{O} \to \mathcal{O}'$ be any (possibly randomized) mapping that does not access the original dataset D. Then the composed mechanism $f \circ \mathcal{M}$ is also ε-differentially private.*

This property ensures that once a DP synthetic dataset $\tilde{D} = \mathcal{M}(D)$ is released, any downstream learning procedure or fairness intervention that operates solely on $\tilde{D}$ constitutes pure DP post-processing and cannot weaken the privacy guarantee with respect to the original training records D.

3.3 Fairness-Aware Learning

Fairness-aware learning aims to incorporate fairness criteria into predictive models. Let X denote features, $A \in \{0, 1\}$ a protected attribute, $Y \in \{0, 1\}$ the label, and $h : \mathcal{X} \to \{0, 1\}$ (or score $s : \mathcal{X} \to [0, 1]$) the predictor. Models may exhibit biased behavior for diverse reasons: some biases are intrinsic to the data (data-to-model bias), while others emerge from missing representative samples or from limitations and objectives of the learning algorithm [9, 49]. To address these issues, algorithmic interventions are typically organized by *where* they act in the pipeline:

- **Pre-processing.** Methods that transform the training data D into $\hat{D}$ (*e.g.*, reweighting or data transformation) to reduce dependence on the protected attribute A while preserving task utility under the defined fairness notion.
- **In-processing.** Methods that modify the learning algorithm or objective. Let $\mathcal{L}(h; D)$ denote a task-specific empirical loss function evaluated on dataset D, and let $\tau \geq 0$ be a user-defined fairness tolerance. In-processing approaches either solve a constrained optimization problem $\min_h \mathcal{L}(h; D)$ s.t. $\Delta_{\text{fair}}(h) \leq \tau$, or incorporate fairness directly into the objective via a penalty term, $\mathcal{L}(h; D) + \lambda\, \Omega_{\text{fair}}(h)$, where $\Omega_{\text{fair}}(h)$ measures unfairness and $\lambda \geq 0$ controls the fairness-utility trade-off.
- **Post-processing.** Methods that adjust predictions of a trained model (*e.g.*, group-specific thresholds, calibration, or randomized decisions) to satisfy target constraints with minimal utility loss, treating the model as a black box.

4 Benchmark Design

In this section, we first describe the overall benchmark structure and experimental configurations, followed by details on datasets and prediction tasks, the DP generative models used, the fairness mechanisms evaluated, as well as the model training procedure, including data preprocessing, classifier configuration, and protocol for stability and reproducibility.

4.1 Overview and Experimental Configuration

Figure 1 illustrates the overall benchmark pipeline. Starting from the original dataset, we construct four prediction pipelines, corresponding to the configurations described in the following. Each pipeline follows the same three stages, namely, data pre-processing, model training, and inference, but differs in whether DP is applied to the data and whether fairness interventions are applied before, during, or after training. This setup enables a systematic comparison of privacy, fairness, and utility across intervention points.

(1) **`Baseline`.** A standard machine learning model trained directly on the original (non-private, non-fair) training data D. This serves as the *reference point* to evaluate the impact of privacy and fairness interventions.
(2) **`DP-only`.** The original training data D is replaced with *differentially private synthetic data* $\tilde{D}$. No fairness mechanism is applied. This setting isolates the effect of differential privacy on model performance and fairness metrics.
(3) **`Fair-only`.** A fairness intervention is applied to the original training data, to the learning algorithm, or to the model output, without incorporating any differential privacy. Specifically, we consider three families of interventions (see Section 3.3): *pre-processing*, *in-processing*, and *post-processing*. This setting quantifies the effect of fairness mechanisms in isolation.
(4) **`DP+Fair`.** Privacy and fairness interventions are combined. A fairness mechanism is applied either (i) to the DP synthetic data

[1] For clarity, we use "DP post-processing" when referring to the privacy property, and "fairness post-processing" when referring to mitigation strategies.

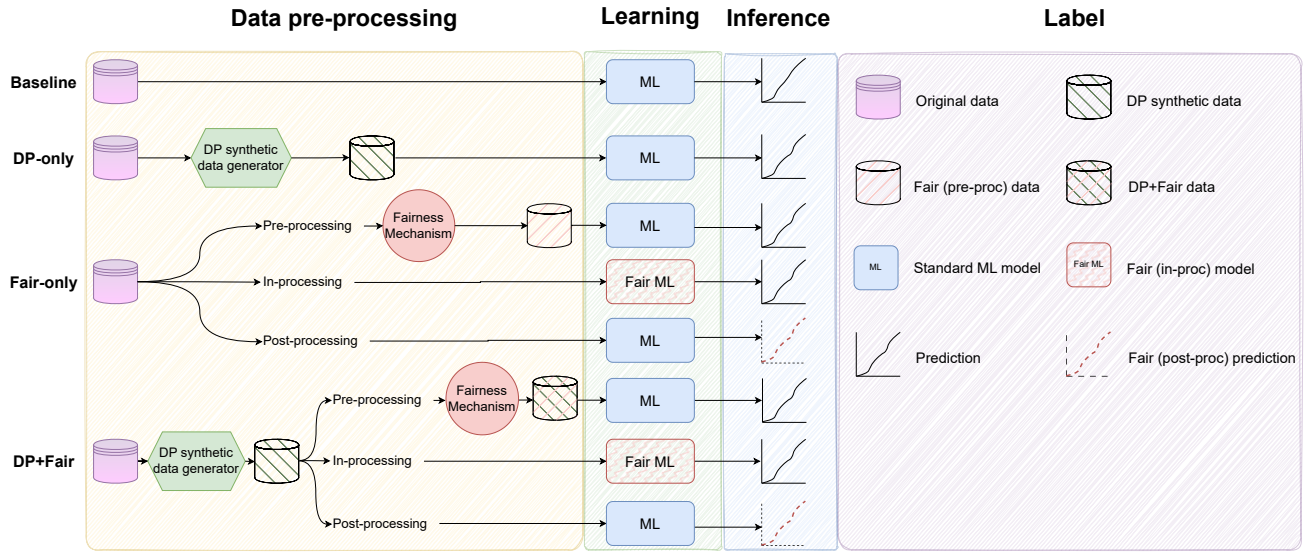


Figure 1: Overview of our benchmark design. We evaluate fairness-aware learning mechanisms applied at three intervention stages, pre-processing, in-processing, and post-processing, across four configurations: (1) `Baseline` (original data, no fairness intervention), (2) `DP-only` (DP synthetic data, no fairness intervention), (3) `Fair-only` (original data with fairness intervention), and (4) `DP+Fair` (fairness interventions on DP synthetic data). This design systematically captures the isolated and combined effects of privacy and fairness interventions.

before training (pre-processing), (ii) during model training (in-processing), or (iii) to the model predictions (post-processing). This setting evaluates whether fairness mechanisms remain effective when operating on DP synthetic data, and whether they can mitigate the fairness degradation introduced by DP.

4.2 Data and Task

Dataset. We conduct our benchmark on four open datasets widely used in fairness research:

- **Adult** [11] (UCI Census Income) contains $n = 47,621$ individuals and the goal is to predict whether a person's income exceeds \$50K/year based on 10 demographic and occupational attributes. *Gender* is used as the protected attribute for fairness evaluation.
- **COMPAS** [6] includes $n = 5,050$ defendants and the goal is to predict recidivism risk based on 7 criminal history and demographic attributes. *Race* is used for fairness evaluation.
- **ACSIncome** [21] extends Adult with richer socioeconomic features from the U.S. Census American Community Survey. We select the Utah state subset with $n = 16,337$ individuals. We set the income threshold at the median value (> 38K/year) and use *gender* as the protected attribute for fairness evaluation.
- **BiasOnDemand** [10] is a synthetic dataset generator designed to benchmark fairness and bias under controlled conditions. We use it to simulate data distributions with known levels of group imbalance and label bias. In total, 6 bias configurations were tested across 3 categories (see Table 3 in Appendix A: imbalance, historical bias, and measurement bias. By default, we present the results with `Config 5` in the full paper. In total, we generate $n = 30,000$ samples, and the goal is to predict the binary value Y conditioned on 2 attributes. The configurations studied are set in a way to: (i) isolate bias on target; (ii) add imbalance to the dataset; (iii) isolate bias to a feature; (iv) add bias to a feature and increase feature correlation and dependence. Furthermore, the values of each configuration parameter were set by experimenting with different values until high values of fairness metrics were achieved, indicating strong bias.

Task. All datasets are cast as binary classification tasks. This choice reflects the dominant and most fundamental and mature experimental setting in the fair machine learning literature, where the majority of group fairness definitions (*e.g.*, Statistical Parity [22], Equal Opportunity [34]) and corresponding mitigation mechanisms (*e.g.*, Calibrated Equalized Odds Post-processing [50], and Reweighing [15]) are formally defined and/or empirically validated for binary labels and binary protected attributes. Moreover, our benchmark is implemented on top of AIF360 [12], which natively supports binary classification and binary protected attributes for all provided fairness interventions, enabling consistent and reproducible comparisons across methods. In addition, some of these provided interventions do not offer the same support to multi-label classification. Exponentiated Gradient Reduction [5] and Grid Search Reduction [5], for example, require an estimator as a parameter, and have the following constraint: "labels y and predictions returned by *predict(X)* are either 0 or 1". Such restrictions, together with the necessity to standardize our experiments and the fact that multi-label classifications can be reduced to a binary classification problem, lead us to proceed with a first experimental setup based on the binary classification tasks. Accordingly, we standardize continuous features, binarize the protected attribute $A \in \{0, 1\}$, and encode

the target variable as binary $Y \in \{0, 1\}$. Further details on dataset preprocessing are provided in Appendix A.

4.3 DP Generative Models

To generate differentially private synthetic data, we focus on two *marginal-based synthesizers*, AIM [44] and MST [43], described hereafter, both implemented in the SmartNoise library (https://docs.smartnoise.org/). Our choice is motivated by recent utility-oriented benchmarks [32, 51, 52, 54], which consistently show that marginal-based models outperform deep generative approaches such as DP-GANs [55] on tabular data. Whereas deep generative models excel in unstructured domains like images, they often struggle with the heterogeneous feature types and sparsity typical of structured tabular data. By contrast, marginal-based synthesizers explicitly model low-dimensional marginals and conditional dependencies, leading to higher fidelity and stronger predictive performance in downstream classification tasks. We therefore adopt AIM and MST as representative DP synthetic data generators.

- **Adaptive Iterative Mechanism (AIM) [44].** AIM is a state-of-the-art differentially private synthetic data generation algorithm. It follows a select-measure-generate paradigm: it iteratively selects informative sets of marginal queries, measures them under differential privacy by adding calibrated noise, and uses Private-PGM [45] to infer an approximate joint distribution from the noisy measurements. Synthetic records are then sampled from this inferred distribution.
- **Maximum Spanning Tree (MST) [43].** MST also follows a select-measure-generate paradigm. It constructs a dependency graph over features using privately measured pairwise correlations, selects a maximum spanning tree to determine which low-dimensional marginals to measure, and then measures these marginals under DP. As in AIM, the resulting noisy marginals are passed to Private-PGM [45], which estimates an approximate joint distribution and samples synthetic records from it.

Our focus on AIM and MST is, therefore, intentional rather than exhaustive. These methods represent the family of marginal-based DP synthesizers that have repeatedly been shown to be competitive for structured tabular data, and they provide a controlled setting for studying how DP synthetic data affects downstream fairness interventions. Accordingly, our conclusions should be interpreted as applying primarily to high-utility marginal-based DP synthetic data pipelines, rather than to all possible DP generative models.

4.4 Fairness Mechanisms

To evaluate whether fairness-aware learning mechanisms remain effective on DP synthetic data, we consider representative methods from the three main categories of interventions: *pre-processing*, *in-processing*, and *post-processing*. These methods have been widely studied in the fairness literature and are implemented in open-source libraries such as AIF360 [12], making them suitable benchmarks for our study.

We deliberately rely on AIF360-based methods because they provide standardized, widely used implementations of canonical fairness interventions across the three main stages of the pipeline. This choice allows us to compare intervention stages under a common experimental protocol, while avoiding additional confounding factors due to heterogeneous implementations, incompatible assumptions, or method-specific engineering choices. Our goal is therefore not to exhaust the full space of fairness algorithms, but to evaluate representative and reproducible mechanisms that make the stage-level comparison meaningful.

Pre-Processing. These approaches modify the training data before model learning to reduce dependence on the protected attribute.

- **Reweighing (RW) [15].** RW relies on resampling and computing weights for the input samples to decrease discrimination.
- **Disparate Impact Remover (DIR) [27].** This method transforms the original dataset to reduce disparate impact between privileged and unprivileged subgroups. It first detects whether disparate impact exists, then removes the dependence of unprotected features on the protected feature, and finally adjusts the distributions of unprotected features so that both privileged and unprivileged groups have similar distributions.
- **Learning Fair Representations (LFR) [61].** LFR creates a probabilistic mapping from a data representation in a given input space to a new representation that reduces the ability to identify protected subgroups while preserving task-relevant information. The objective is to approximate statistical parity across groups while balancing fairness with the predictive utility of the data.

In-Processing. These methods modify the training procedure itself, typically by solving constrained optimization problems that balance accuracy with fairness.

- **Exponentiated Gradient Reduction (EGR) [5].** This method computes an iterative approximation of the saddle point of a Lagrangian by minimizing the classification loss and maximizing the penalty for fairness violation. The idea behind this computation is the same as a zero-sum game with two players.
- **Grid Search Reduction (GSR) [5].** This method shares similar ideas with EGR, but relies on brute force. Essentially, it builds a grid of Lagrangian multipliers and exhaustively searches for the best solution considering the fairness-accuracy trade-off.

Post-Processing. These methods operate on the outputs of a trained model, adjusting predictions to satisfy fairness constraints.

- **Reject Option Classification (ROC) [37].** This method operates on the model's prediction outputs, adjusting decision boundaries to favor fair outcomes in regions of uncertainty. It aims to reduce discrimination by flipping labels for samples near the decision boundary – particularly when such adjustments enhance fairness for protected groups.
- **Equalized Odds Post-Processing (EqOdds) [12, 34].** This method formulates a linear program to learn probabilities with which the output labels will be changed to satisfy equalized odds constraints while maintaining classification fidelity.
- **Calibrated Equalized Odds Post-processing (CEOP) [50].** This method operates similarly to EqOdds, enforcing parity in error rates – false positive rate and false negative rate remain similar across the protected groups. Additionally, CEOP introduces the concern for performing such tasks while maintaining calibration, *i.e.*, ensuring that the predicted scores remain interpretable as true outcome probabilities.

4.5 Metrics

Privacy levels. We vary the privacy budget across a broad range, $\varepsilon \in \{0.05, 0.1, 0.25, 0.5, 0.75, 1, 2, 3, 5, 10, 15, 20\}$, covering high-, moderate-, and low-privacy regimes. Small values of ε provide stronger theoretical privacy guarantees; for example, at $\varepsilon = 0.05$, the likelihood ratio between neighboring datasets is bounded by $e^{0.05} \approx 1.05$. Larger values such as $\varepsilon = 10$ or 20 correspond to weaker theoretical guarantees, but are included for comparability with prior DP synthetic data benchmarks [31, 32] and because DP mechanisms can still empirically reduce practical attack success compared to non-DP synthetic data [40].

Utility metrics. We report two standard metrics, namely, accuracy and F1-score (harmonic mean of precision and recall), computed on the held-out test set.

Fairness metrics. Let $A \in \{0, 1\}$ denote the binary protected attribute, $Y \in \{0, 1\}$ the true label, and $\hat{Y} \in \{0, 1\}$ the predicted label. We evaluate group fairness using three standard metrics:

Definition 2 (Model Accuracy Difference (MAD)). *Difference in overall classification accuracy between groups:*

$$MAD = \Pr[\hat{Y} = Y \mid A = 0] - \Pr[\hat{Y} = Y \mid A = 1].$$

Definition 3 (Equal Opportunity Difference (EOD) [34]). *Difference in true positive rates across groups:*

$$EOD = \Pr[\hat{Y} = 1 \mid Y = 1, A = 0] - \Pr[\hat{Y} = 1 \mid Y = 1, A = 1].$$

Definition 4 (Statistical Parity Difference (SPD)). *Difference in positive prediction rates across groups:*

$$SPD = \Pr[\hat{Y} = 1 \mid A = 0] - \Pr[\hat{Y} = 1 \mid A = 1].$$

All three metrics are defined such that a value of 0 corresponds to perfect parity between groups, while larger deviations indicate increasing disparity.

4.6 Model Training

All datasets are split into training (60%), calibration (20%), and test (20%) subsets using a fixed partition. Following the standard setting in the privacy-preserving machine learning literature [30, 60], only the training split is treated as the protected dataset. In the `DP-only` and `DP+Fair` configurations, models are trained on DP synthetic data generated from this real training split, whereas in the `Baseline` and `Fair-only` configurations, models are trained directly on the real training split. In all configurations, final utility and fairness metrics are computed on the real held-out test set. The calibration and test splits are disjoint from the protected training database and are never used during DP synthesis.

The calibration set is only used to tune post-processing fairness mechanisms, while the test set remains unseen during training and calibration, serving exclusively for final evaluation (including post-processing applied at inference). Before any fairness post-processing calibration, a model trained on DP synthetic data is a downstream use of the DP output and therefore preserves the DP guarantee with respect to the protected training split. After calibration, this guarantee still holds for the original training records used by the synthesizer. However, in the main protocol, the calibration records themselves are not protected by this DP guarantee because fairness post-processing methods are calibrated using real held-out data. This setup follows a common evaluation protocol in DP synthetic data and privacy attack research, while ensuring consistent comparison across the `Baseline`, `DP-only`, `Fair-only`, and `DP+Fair` configurations. To assess whether this calibration protocol affects our conclusions, we further report a DP-compliant calibration ablation in Appendix B.1, where fairness post-processing is calibrated using DP-based calibration records.

Classifier. We use a set of three classifiers across all experiments: Extreme Gradient Boosting (**XGBoost**) [17], Logistic Regression (**LR**) [46], and Random Forest (**RF**) [13]. These models were selected to cover complementary learning paradigms commonly used for tabular data. XGBoost is a widely adopted tree-based ensemble method with strong predictive performance and scalability. LR serves as a classical linear baseline, providing a well-understood and interpretable reference point. RF represents bagging-based ensembles of decision trees, capturing non-linear decision boundaries while remaining robust and widely adopted in practice. To maintain consistency across configurations, we use the `binary:logistic` objective and keep all hyperparameters as close to their default values as possible, unless otherwise stated in Appendix A.5.

Stability. Since DP mechanisms, train/calibration/test splits, synthetic data generation, and classifier training all involve randomness, we repeat each experiment with 20 independent random seeds. In the Pareto trade-off plots used in Section 5, each point corresponds to the mean across seeds for a fixed configuration, namely dataset, classifier, privacy budget, and fairness intervention. To quantify variability across runs, we additionally compute 95% confidence intervals using Student's t-distribution. Some methods (*e.g.*, LFR in Figure 3) exhibit substantially higher variability across seeds, resulting in intervals that visually dominate the plots; for readability, we omit these intervals from the corresponding figures. These statistics are used to assess the robustness and reproducibility of the observed fairness–utility trade-offs.

4.7 Reproducibility and Extensibility

We implement our benchmark on top of the open-source SmartNoise library and the AIF360 fairness toolkit [12]. Our framework integrates DP synthesizers, fairness interventions, and evaluation pipelines in a modular fashion, making it straightforward to add new datasets, generative models, or classifiers. More precisely, our benchmark is organized around two main components: (1) *DP synthetic data generation*, and (2) *execution of experiments*. This design provides a clear separation between data generation and downstream evaluation, making the benchmark both reproducible and easily extensible.

- **Synthetic data generation.** This module integrates existing DP synthesizers (AIM, MST) and can be extended with new generators and different data pre-processors. Users can also add additional datasets to the generation pipeline with minimal configuration.
- **Experimental execution.** This module runs end-to-end experiments, including model training, fairness interventions, and metric evaluation, and can be extended by modifying the base classifier.

Together, these two modules ensure that new datasets, synthesizers, classifiers, interventions, or metrics can be integrated with minimal effort. Moreover, the provided scripts support end-to-end reproduction of the experiments, as well as regeneration of the figures and tables from the corresponding experimental logs. The full codebase, along with documentation and configuration files, is available in the following GitHub repository https://github.com/vinicius-verona/dp-fair-intervention-benchmark. Lastly, to facilitate installation and reuse, the benchmark is also distributed as the open-source Python package BenchmarkDPFair through PyPI at https://pypi.org/project/BenchmarkDPFair/, and can be installed using `pip install BenchmarkDPFair.`

5 Results and Analysis

We now analyze the empirical behavior of fairness interventions applied to DP synthetic data, with a particular focus on *where in the learning pipeline* fairness mechanisms are most effective under privacy constraints. Throughout this section, we report results obtained with AIM, which is currently regarded as the state-of-the-art DP synthesizer for tabular data, as emphasized in recent benchmarks [32, 51, 52, 54] and KDD & VLDB 2025 tutorials [20]. Furthermore, we present the results acquired with the XGBoost algorithm and discuss the other classifiers in Figure 4 and Appendix B.3. For completeness, additional results using MST are also provided in Appendix B.3, with the remaining classifiers; unless otherwise stated, the qualitative trends observed with AIM are consistent across ML models.

Our evaluation spans the four datasets described in Section 4.2—three real-world benchmarks (**Adult**, **COMPAS**, and **ACSIncome**) and one synthetic dataset (**BiasOnDemand**)—and compares fairness interventions applied at the *pre-processing*, *in-processing*, and *post-processing* stages. For each dataset, we analyze trade-offs between privacy, utility, and fairness under the four experimental settings of Figure 1: `Baseline`, `DP-only`, `Fair-only`, and `DP+Fair`.

Figures 2 and 3 present a detailed view of the fairness–utility trade-offs induced by each fairness intervention across all datasets. Specifically, we report: (i) **accuracy (ACC)** versus fairness metrics, and (ii) **F1-score** versus fairness metrics, where F1-score captures the harmonic mean of precision and recall and provides a complementary perspective in the presence of class imbalance. Each subfigure corresponds to one dataset and plots utility against the three group fairness metrics (MAD, EOD, and SPD). Points correspond to different privacy budgets ε, with marker size encoding the magnitude of ε (larger markers indicate weaker privacy guarantees), while hollow markers denote the `Fair-only` setting ($\varepsilon = \infty$)[2]. This visualization highlights how both utility and fairness evolve as privacy constraints tighten, and how different intervention mechanisms respond to DP noise.

Taken together, these figures provide a method-level view of how fairness interventions interact with DP synthetic data across privacy budgets. The `Fair-only` configuration serves as an upper-bound reference for achievable fairness and utility in the absence of DP noise, while `DP-only` isolates the impact of privacy alone. The `DP+Fair` configurations illustrate the extent to which individual fairness mechanisms can mitigate DP-induced disparities. In the following subsections, we analyze pre-processing (Section 5.1), in-processing (Section 5.2), and post-processing (Section 5.3) interventions separately, before discussing our results in Section 6.

5.1 Pre-Processing

We first analyze pre-processing interventions, namely Reweighing (RW) [15], Disparate Impact Remover (DIR) [27], and Learning Fair Representations (LFR) [61]. These methods operate directly on the training data distribution and therefore constitute the earliest point of intervention in the pipeline.

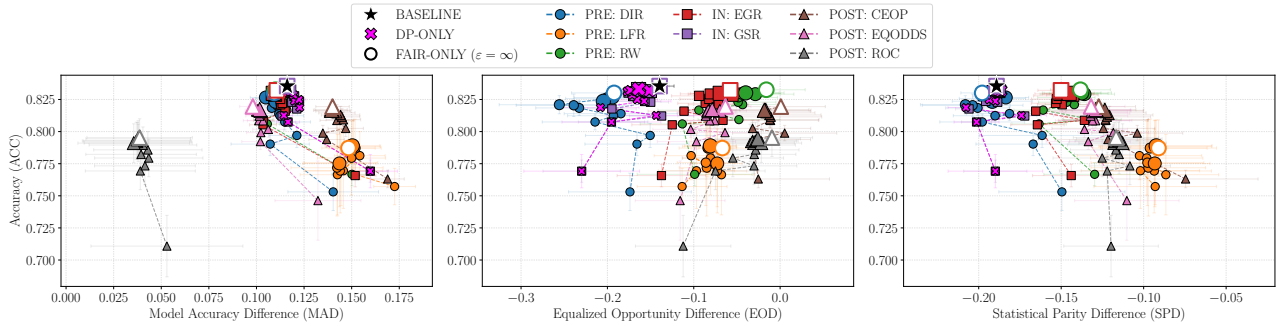
Across datasets, Figures 2 and 3 show that some pre-processing methods under the `DP+Fair` configuration generally shift predictions away from the `DP-only` region toward the corresponding `Fair-only` region in the fairness–utility space. This indicates that pre-processing is capable of correcting not only the bias present in the `Baseline` model, but also a substantial portion of the additional disparities introduced by DP synthetic data. However, these fairness improvements consistently come at the cost of reduced predictive performance, visible as a downward shift in both accuracy and F1-score across privacy budgets.

An additional observation is that pre-processing benefits are only weakly sensitive to the privacy budget ε. While increasing ε improves the overall utility of DP synthetic data, the relative position of pre-processing methods in the fairness–utility space remains largely stable, suggesting that pre-processing corrections saturate quickly and do not scale strongly with weaker privacy.

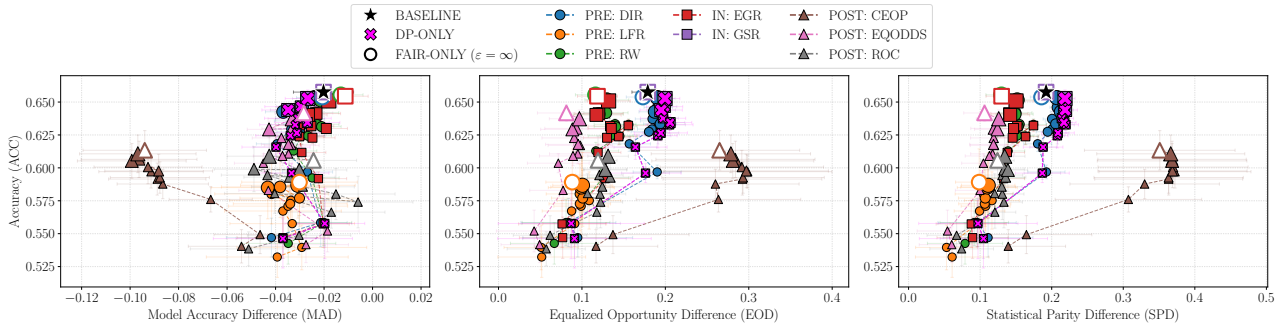
Reweighing. RW reduces group disparities relative to `DP-only`, particularly for SPD, across datasets. In the trade-off plots, RW trajectories shift toward lower disparity while remaining near the `DP-only` utility envelope, indicating moderate fairness gains with limited utility loss. Although RW incurs a systematic decrease in accuracy and F1-score compared to `Baseline` and `Fair-only`, this degradation is less pronounced than for other pre-processing methods. RW therefore provides the most stable and predictable fairness improvements among pre-processing interventions, serving as the strongest baseline in this category. Its gains are most pronounced for SPD, while improvements in error-based metrics such as EOD remain limited, suggesting that RW primarily corrects marginal distributional bias under DP synthetic data.

Disparate Impact Remover. DIR exhibits limited corrective power under DP synthetic data. In Figures 2 and 3, DIR points largely overlap with the `DP-only` region across fairness metrics, indicating minimal reduction in disparity. In some cases, DIR slightly worsens both fairness and utility relative to the `Baseline`. We attribute this behavior to unmet preprocessing assumptions required by DIR [27], which restrict its ability to decorrelate protected attributes from features. While DIR can be effective when these conditions are satisfied, our benchmark enforces a uniform preprocessing pipeline; under these conditions, DIR remains consistently less effective than RW for DP synthetic data.

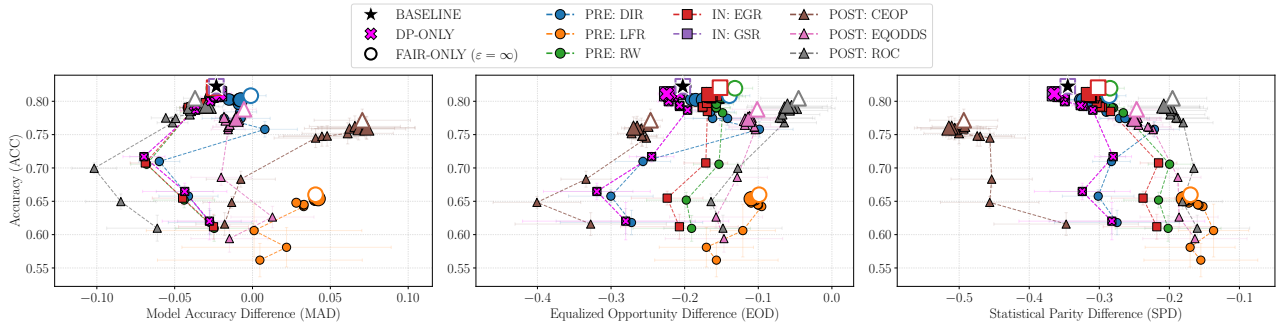
[2]Here, $\varepsilon = \infty$ is used only as a visual shorthand for the non-private fair setting, where fairness interventions are applied to models trained on real data without DP noise; it should not be misperceived with the `Baseline` setting, which is also non-private but does not apply any fairness intervention.



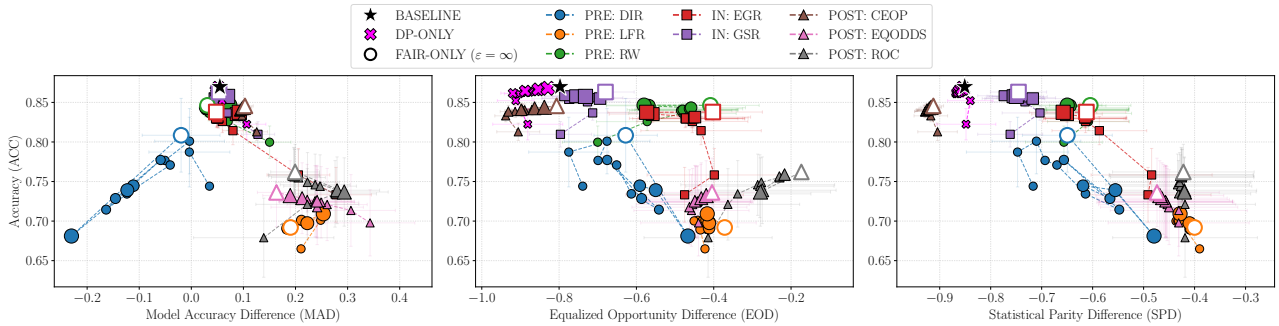
(a) Adult.



(b) COMPAS.

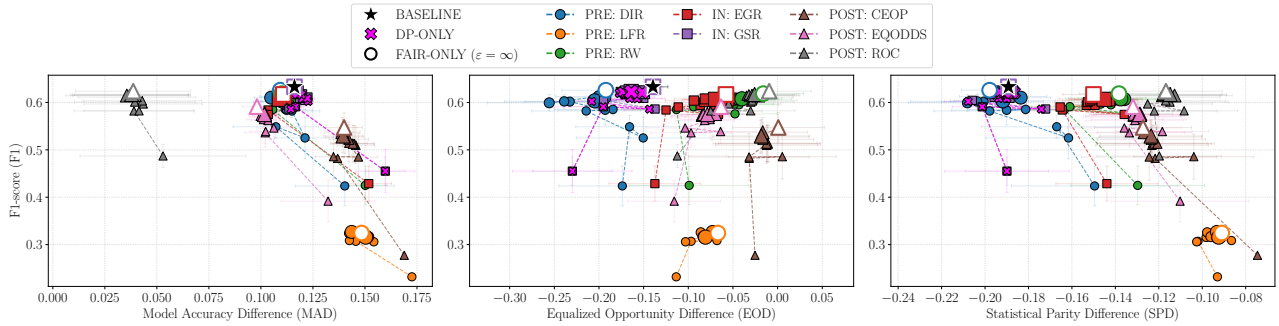


(c) ACSIncome.

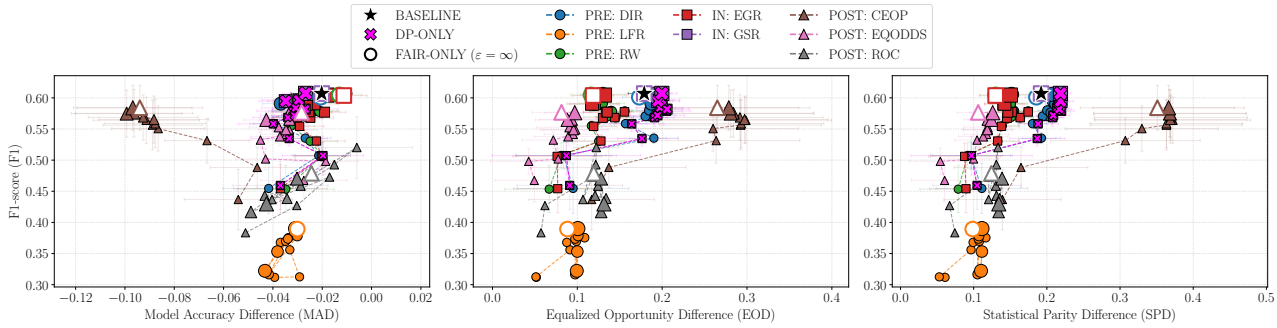


(d) BiasOnDemand (Configuration 5).

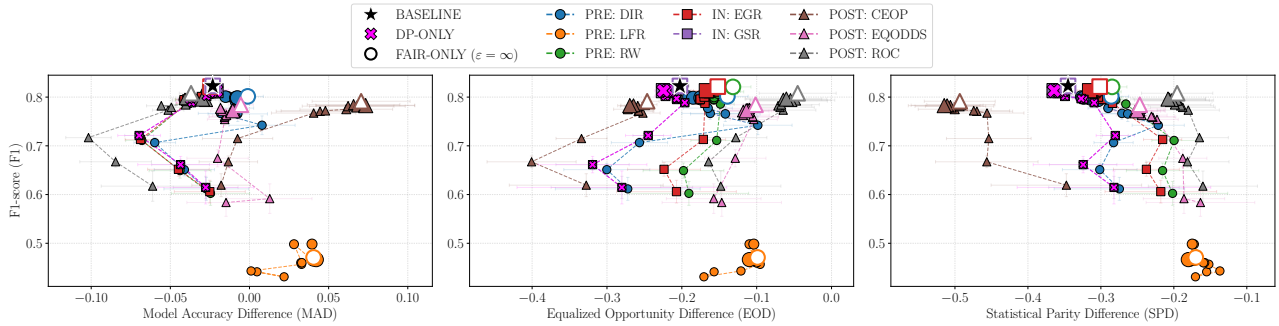
Figure 2: Accuracy–fairness trade-off under the AIM synthesizer across four datasets. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



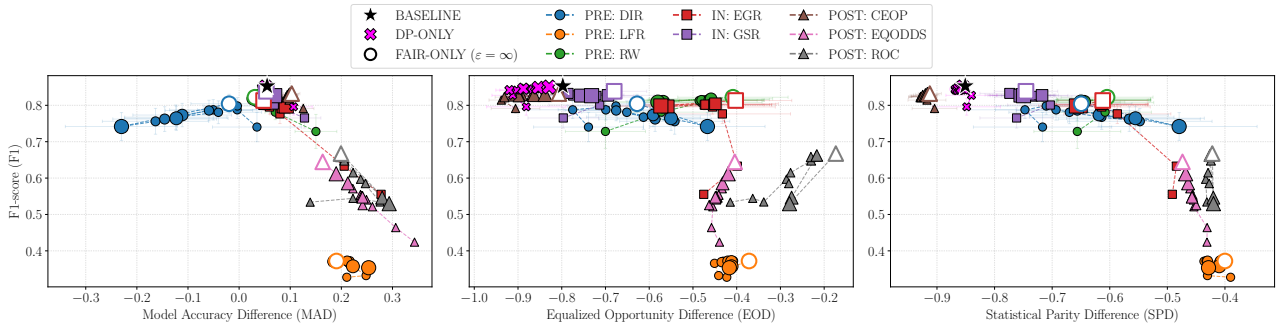
(a) Adult.



(b) COMPAS.



(c) ACSIncome.



(d) BiasOnDemand (Configuration 5).

Figure 3: F1-score–fairness trade-offs under the AIM synthesizer across four datasets. Each subfigure reports F1-score versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.

Learning Fair Representations. LFR applies the most aggressive transformation among pre-processing methods. In the Pareto plots of Figures 2 and 3, LFR consistently attains the lowest disparity values among pre-processing methods across datasets, often pushing SPD and EOD close to zero for all privacy budgets. This behavior confirms that LFR is effective at enforcing strong group-level parity even under DP synthetic data. However, these fairness gains come at a substantial utility cost: both accuracy and F1-score are systematically reduced, placing LFR in a low-utility region of the trade-off space relative to other interventions. In practice, this makes LFR less suitable when maintaining predictive performance is a priority, despite its strong parity guarantees. Beyond this consistent utility degradation, LFR can also exhibit sensitivity to DP noise in certain settings. In particular, for some datasets (notably COMPAS) and specific ε values or random seeds, the learned representations occasionally collapse one of the target classes, preventing successful downstream model training. These rare but impactful failures are documented in Appendix B.3. Overall, while LFR reliably enforces strong fairness, it does so at the expense of both predictive performance and robustness.

Summary. Overall, pre-processing interventions under `DP+Fair` consistently reduce group disparities but do so by trading off predictive performance. RW provides the most stable balance between fairness improvement and utility preservation, DIR remains largely indistinguishable from `DP-only`, and LFR achieves the strongest parity guarantees at the cost of pronounced utility degradation and increased sensitivity to DP noise.

5.2 In-Processing

We next analyze in-processing interventions, namely Exponentiated Gradient Reduction (EGR) and Grid Search Reduction (GSR) [5]. Unlike pre-processing methods, in-processing approaches enforce fairness constraints during model training and therefore operate directly on classifiers trained on DP synthetic data.

Across datasets, Figures 2 and 3 reveal two consistent patterns under the experimental conditions considered in this study. First, compared to pre- and post-processing, in-processing interventions achieve more moderate levels of bias correction. While both EGR and GSR improve fairness metrics relative to `DP-only`, the resulting reductions in disparity remain bounded and generally do not reach the `Fair-only` region of the trade-off space. Second, these fairness improvements are obtained while largely preserving predictive performance: accuracy and F1-score under `DP+Fair` remain close to the corresponding `DP-only` levels across privacy budgets. Taken together, these observations indicate that, under our unified preprocessing pipeline, model configurations, and hyperparameter settings, in-processing methods offer a conservative but stable trade-off, favoring utility preservation while delivering limited yet consistent fairness gains. We note that alternative data representations or task-specific tuning could potentially alter this balance; however, such adaptations fall out of the scope of our work, which intentionally enforces comparable conditions across methods.

Exponentiated Gradient Reduction. Among in-processing approaches, EGR consistently provides the most favorable balance between fairness improvement and utility preservation. In the Pareto plots, EGR trajectories move toward lower disparity relative to `DP-only`, particularly for SPD and EOD, while remaining closely aligned with the `DP-only` utility envelope. This pattern is observed across all datasets and privacy budgets, indicating that EGR is able to mitigate a non-negligible portion of DP-induced disparities without incurring substantial additional loss in accuracy or F1-score. However, the extent of fairness correction achieved by EGR remains bounded: while it improves upon `DP-only`, EGR does not fully converge toward the `Fair-only` region of the trade-off space. Compared to pre- and post-processing methods, its corrective power is more moderate, reflecting a design choice that prioritizes stability and the preservation of utility when training on DP synthetic data.

Grid Search Reduction. In contrast, GSR exhibits little to no systematic benefit over `DP-only`. Across datasets, GSR points largely overlap with the `DP-only` region in the fairness–utility space, indicating negligible disparity reduction. While isolated improvements can be observed for specific datasets (*e.g.*, BiasOnDemand in Figures 2d and 3d), these gains do not generalize and remain inconsistent across privacy budgets. Overall, GSR fails to provide reliable fairness improvements under DP synthetic data.

Summary. Overall, in-processing interventions yield modest and inconsistent fairness gains under DP synthetic data while largely preserving utility. EGR emerges as the only in-processing method that consistently improves fairness relative to `DP-only`, but its impact remains limited compared to pre- and post-processing approaches. GSR, by contrast, remains largely indistinguishable from `DP-only` across datasets and privacy regimes. These findings suggest that in-processing methods offer a conservative trade-off: they preserve predictive performance, but are insufficient when substantial bias mitigation is required.

5.3 Post-Processing

We finally analyze post-processing interventions, namely Reject Option Classification (ROC) [37], Equalized Odds Post-Processing (EqOdds) [12, 34], and Calibrated Equalized Odds (CEOP) [50]. These methods intervene *after* training, adjusting model outputs to satisfy fairness criteria and therefore do not depend on retraining on DP synthetic data.

Across datasets, the method-level Pareto plots (Figures 2 and 3) consistently show that post-processing achieves the strongest bias correction *for a given level of utility* under DP synthetic data, particularly for EOD and SPD, and does so more reliably than pre- and in-processing. For instance, in the stage-level summaries, POST points systematically move toward lower disparity compared to `DP-only` (most clearly along EOD and SPD), while maintaining accuracy or F1-score within a bounded degradation relative to `DP-only` across privacy budgets. This behavior highlights a distinctive advantage of post-processing in the DP synthetic setting: because it operates on predictions, it can exploit residual predictive signal that remains after DP synthesis, and translate it into substantial fairness gains without requiring modifications to the (potentially noisy) training distribution.

At the same time, our results indicate that post-processing outcomes are method-dependent. ROC and EqOdds consistently shift models toward substantially lower disparities (often approaching

the best regions reached by any intervention stage), whereas CEOP exhibits more heterogeneous behavior across datasets and metrics. In particular, CEOP can preserve higher utility in some settings, but its fairness correction is less consistent and can introduce metric-specific trade-offs (*e.g.*, improving one notion while leaving another largely unchanged), making it comparatively less reliable as a default choice under DP synthetic data.

Reject Option Classification. ROC provides the most consistent fairness correction among post-processing methods. In the Pareto plots, ROC points typically move toward substantially lower EOD and SPD compared to `DP-only`, frequently placing ROC among the best-performing methods in terms of disparity reduction across datasets. This improvement is generally obtained with a moderate utility trade-off: ROC may reduce accuracy and/or F1-score relative to `DP-only`, but remains competitive compared to other intervention stages, and its fairness gains are stable across ε.

Equalized Odds Post-Processing. EqOdds performs comparably to ROC in terms of fairness correction, consistently moving solutions toward lower disparities across MAD, SPD, and EOD. In several datasets, EqOdds provides a favorable balance in the trade-off space, achieving strong reductions in disparity while keeping utility close to `DP-only`. Overall, EqOdds is a robust post-processing option whose behavior is stable across privacy regimes.

Calibrated Equalized Odds Post-Processing. CEOP exhibits the most variable behavior among post-processing methods. In the Pareto plots, CEOP can maintain relatively high utility in some settings, but its fairness improvements are less systematic than ROC and EqOdds and can depend on the dataset and the fairness notion considered. In particular, CEOP may yield mixed outcomes across metrics (*e.g.*, limited movement on EOD or SPD compared to ROC/EqOdds), and in some cases, its corrections are not aligned across fairness notions. This variability suggests that CEOP, at least under the uniform configurations used in our benchmark, is less reliable as a default post-processing choice for DP synthetic data.

Summary. Overall, post-processing is the most effective and practically reliable intervention stage under DP synthetic data: it consistently achieves the largest reductions in group disparities (especially EOD and SPD) across privacy budgets and datasets. ROC and EqOdds are the most robust options, repeatedly delivering strong fairness improvements with bounded utility loss, whereas CEOP is more sensitive to dataset and metric choice and can yield heterogeneous trade-offs. These results suggest that *post-processing (particularly ROC and EqOdds) emerges as the most reliable strategy for mitigating bias when training on DP synthetic data, offering the best fairness–utility trade-offs among the evaluated interventions.*

6 Discussion

The results in Section 5 allow us to revisit the central research question of this work: *where should one intervene in the ML pipeline to mitigate unfairness when training on DP synthetic data?*

Where to intervene. Our benchmark reveals that not all points of intervention are equally effective under DP synthetic data. Pre-processing methods (RW, DIR, LFR) can substantially reduce group disparities, but they typically trade fairness improvements for utility degradation. This is expected, as these methods alter the input distribution itself, often at the expense of the signal needed for accurate classification. In-processing methods (EGR, GSR) offer a more conservative trade-off: they largely preserve utility close to `DP-only`, but achieve only bounded reductions in disparity. In particular, EGR yields consistent (yet moderate) improvements, whereas GSR is often indistinguishable from `DP-only`. Post-processing stands out: ROC and EqOdds consistently achieve the strongest disparity reductions (notably for EOD and SPD) while keeping accuracy and F1-score within a bounded degradation relative to `DP-only` across privacy budgets. Overall, ***Overall, the Pareto-front analysis suggests that post-processing is the strongest intervention stage*** in terms of fairness–utility trade-offs.

Mechanism-specific insights. Several consistent patterns emerge across datasets, privacy budgets, and classifiers when using AIM as the DP synthesizer. RW reliably reduces group disparities, particularly SPD, but this improvement is typically accompanied by a measurable loss in predictive utility relative to `DP-only`. LFR enforces the strongest parity guarantees, often driving SPD and EOD close to zero, but does so at a substantial utility cost and exhibits sensitivity to DP noise, occasionally leading to unstable training outcomes. DIR contributes little under the uniform preprocessing pipeline adopted in our benchmark, with performance largely overlapping the `DP-only` region. Among in-processing methods, EGR offers a stable compromise: it consistently improves fairness metrics relative to `DP-only` while largely preserving accuracy and F1-score. However, these gains remain bounded and do not approach the `Fair-only` region. GSR, by contrast, fails to provide reliable or systematic fairness improvements across datasets and privacy regimes. Post-processing methods are the most robust overall. ROC and EqOdds consistently appear on or near the Pareto frontier of fairness–utility trade-offs, delivering substantial reductions in EOD and SPD with bounded utility degradation relative to `DP-only`. CEOP exhibits more heterogeneous behavior, with outcomes that depend on the dataset and fairness metric considered. Taken together, these results identify ***ROC and EqOdds as the most reliable mechanisms for balancing privacy, fairness, and utility under DP synthetic data generated by AIM***.

Why post-processing outperforms other interventions. Post-processing mechanisms tend to outperform other interventions under DP synthetic data *in our benchmark* because they act at the decision stage, directly on model predictions rather than on noisy training data. Methods such as ROC and EqOdds adjust decision thresholds or predicted probabilities, making them comparatively robust to distortions introduced by DP synthesis. Unlike pre- and in-processing approaches that must learn fairness corrections from perturbed features and labels, post-processing leverages residual separability in the model's output space to realign group outcomes with bounded changes in accuracy and F1-score. This mechanism-level explanation is consistent with the empirical patterns observed in the Pareto-front analysis.

Synthesizer and classifier effects. The relative effectiveness of fairness interventions depends on both the DP synthesizer and, to a lesser extent, the classifier. In the main paper, we focus on

AIM, which retains sufficient predictive utility for fairness interventions to operate meaningfully. Under AIM, the qualitative stage-level trends remain stable across the three classifiers considered, as shown in Figure 4 and further detailed in Appendix B.3. Although absolute utility and disparity values vary across model families, pre-, in-, and post-processing interventions occupy comparable regions of the fairness–utility space. LR exhibits larger sensitivity on the Adult dataset, particularly for recall-sensitive fairness metrics, which is consistent with its linear hypothesis class and different baseline error structure compared to tree-based models (*i.e.*, XGBoost and RF). Importantly, these quantitative differences do not overturn the overall stage-level conclusion.

At the same time, the achievable fairness–utility trade-offs depend on the synthetic data generator. With AIM, several interventions move `DP+Fair` outcomes toward their `Fair-only` counterparts, although the gap to `Fair-only` is not always fully closed and depends on the intervention stage and fairness metric. In contrast, results with MST reveal a different limitation: while fairness interventions under `DP+Fair` can reduce disparities, the corresponding utility levels remain substantially below those achieved by `Fair-only`, even as the privacy budget increases. This persistent gap suggests that, under MST, DP synthetic data imposes a structural ceiling on achievable utility that fairness interventions cannot overcome. As a result, improvements in fairness must be interpreted relative to a constrained utility regime, rather than as steps toward the non-private fair optimum. Practitioners should therefore consider not only *where* to intervene, but also whether the chosen DP synthesizer preserves sufficient utility headroom for fairness interventions to approach non-private performance.

Ablations and robustness checks. We further assess whether two methodological choices affect our conclusions. First, to verify that the advantage of post-processing is not driven by access to a real held-out calibration set, we repeat the post-processing experiments using DP-compliant calibration data. As reported in Appendix B.1, the qualitative conclusions remain unchanged: ROC and EqOdds continue to provide the strongest fairness–utility trade-offs. Second, to examine whether the more limited gains of in-processing methods are due to default hyperparameter choices, we conduct a hyperparameter-sensitivity analysis for EGR and GSR. The results, reported in Appendix B.2, show that tuning can improve some local trade-offs but does not alter the overall stage-level conclusion: in-processing remains more conservative than post-processing under DP synthetic data.

Implications. The broader implication of our findings is that fairness interventions are not uniformly transferable from non-private to DP settings. While post-processing remains reliable, pre- and in-processing methods are more fragile under DP-induced distributional shifts. Moreover, the utility–fairness trade-off interacts strongly with the synthesizer: mechanisms that are effective with AIM may have a limited effect under MST. For practitioners, this suggests two guidelines. First, when utility preservation is a priority, combining a high-utility DP synthesizer such as AIM with post-processing interventions (in particular ROC or EqOdds) yields the most favorable trade-offs. Second, when enforcing stronger parity constraints is critical, and some utility loss is acceptable, pre- or in-processing methods such as RW or EGR may be considered, provided their costs and stability limitations are carefully evaluated.

Key takeaway and statistical validation. Overall, our benchmark shows that although DP amplifies fairness–utility trade-offs when training on synthetic data, carefully chosen intervention strategies can still recover part of the fairness loss while preserving predictive utility. To complement the Pareto-front analysis, we statistically validate two central claims using one-sided paired Wilcoxon signed-rank tests [19] across the four main datasets, three classifiers (XGBoost, RF, and LR), and AIM as the DP synthesizer.

- **Claim 1: `DP+Fair` recovers fairness degradation caused by `DP-only`.** For each fairness metric $m \in \{\text{MAD}, \text{EOD}, \text{SPD}\}$, we pair configurations with the same dataset, classifier, privacy budget ε, seed, and DP synthesizer, and compute
$$d_i = |m^{(i)}_{\text{DP+Fair}}| - |m^{(i)}_{\text{DP-only}}|.$$
We test the one-sided alternative that $d_i < 0$, where negative values indicate that the fairness intervention reduces the absolute disparity observed under `DP-only`. As shown in Table 1, `DP+Fair` significantly reduces EOD and SPD, with confidence intervals entirely below zero, while MAD does not improve globally. Thus, Claim 1 is supported for opportunity- and parity-based group fairness metrics, while the MAD result shows that fairness recovery under `DP+Fair` is metric-dependent.

Metric	n	$\bar{d}$	95% CI	p-value	Win rate
MAD	20,998	0.0204	[0.0194, 0.0213]	1.000	40.0%
EOD	20,998	-0.1185	[-0.1208, -0.1162]	$< 10^{-16}$	72.6%
SPD	20,998	-0.0879	[-0.0898, -0.0859]	$< 10^{-16}$	74.4%

Table 1: Paired Wilcoxon signed-rank analysis for Claim 1 under AIM across the four main datasets and three classifiers. Negative $\bar{d}$ indicates that `DP+Fair` reduces the absolute fairness gap relative to `DP-only`.

- **Claim 2: POST provides stronger EOD/SPD-oriented fairness–utility trade-offs than PRE/IN.** To jointly compare utility and fairness, we use a weighted Euclidean distance to the ideal point $(U = 1, |m| = 0)$. For each dataset, classifier, seed, privacy budget ε, utility metric, and fairness metric, we compute
$$S^{(i)}_a = \sqrt{w_U \left(1 - U^{(i)}_a\right)^2 + w_F \left|m^{(i)}_a\right|^2}, \tag{1}$$
where $U \in \{\text{ACC}, \text{F1}\}$, $m \in \{\text{EOD}, \text{SPD}\}$, and $w_U = w_F = 0.5$. The square root reflects the Euclidean-distance interpretation; lower values indicate better fairness–utility trade-offs. For each intervention stage s, we define $S^{(i)}_s = \min_{a \in s} S^{(i)}_a$, and compare POST against PRE and IN through
$$d_i = S^{(i)}_{\text{POST}} - S^{(i)}_{\text{OTHER}}, \qquad \text{OTHER} \in \{\text{PRE}, \text{IN}\}.$$
We again test the one-sided alternative that $d_i < 0$. Negative values indicate that POST is closer to the ideal fairness–utility point. As shown in Table 2, POST, defined as the best score among ROC and EqOdds, significantly outperforms both PRE and IN for EOD/SPD-oriented trade-offs.

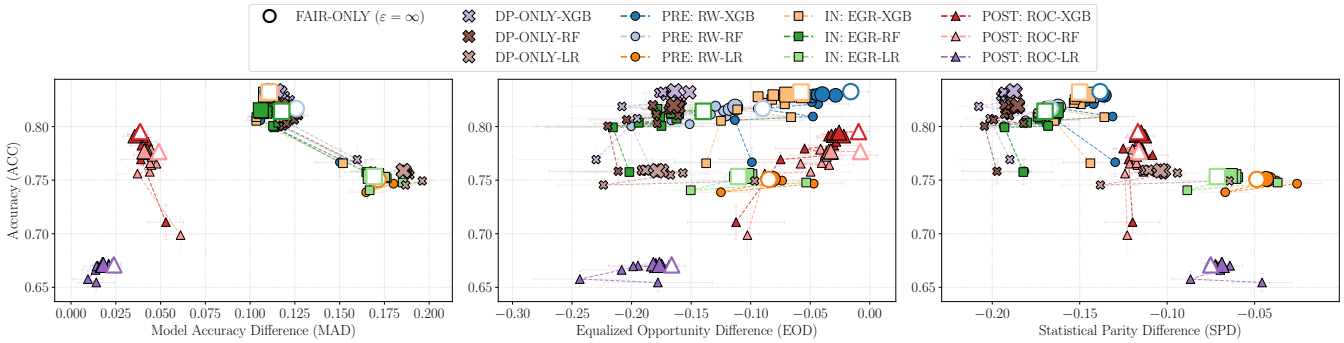


Figure 4: Accuracy–fairness trade-offs under the AIM synthesizer across the Adult dataset and all tested ML models (XGBoost, Random Forest, and Logistic Regression). Each subfigure reports accuracy across three fairness metrics (MAD, EOD, SPD) for representative interventions from each stage. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$).

Comparison	n	$\bar{d}$	95% CI	p-value	Win rate
POST vs PRE	11,088	-0.0199	[-0.0207, -0.0190]	$< 10^{-16}$	72.8%
POST vs IN	11,088	-0.0294	[-0.0303, -0.0286]	$< 10^{-16}$	79.2%

Table 2: Paired Wilcoxon signed-rank analysis for Claim 2 under AIM across the four main datasets and three classifiers. The analysis is restricted to EOD/SPD-oriented trade-offs, with POST defined as the best score among ROC and EqOdds. Negative $\bar{d}$ indicates that POST achieves a lower scalarized fairness–utility score than the compared stage.

These tests support the main EOD/SPD-oriented conclusions while also showing that the findings remain criterion-dependent, since MAD does not improve globally, and POST superiority is established under the scalarized EOD/SPD trade-off in Equation (1).

7 Conclusion, Limitations, and Perspectives

Concluding remarks. We introduced, to our knowledge, the first systematic benchmark of fairness-aware learning on *DP synthetic tabular data*, covering multiple fairness intervention stages (pre-, in-, and post-processing), four datasets, a wide range of privacy budgets, and three different ML classifiers. Our main analysis centers on AIM [44], a state-of-the-art marginal-based DP synthesizer for tabular data, with additional results for MST [43] reported in Appendix B.3. Across configurations, we find that while DP alone generally reduces utility and can exacerbate group disparities, *fairness interventions can partially recover fairness under DP synthetic data*. Within the evaluated scope of AIF360-based group fairness mechanisms and marginal-based DP synthesizers, *post-processing methods tend to provide the most stable fairness–utility trade-offs under AIM*, particularly at moderate privacy budgets. In particular, methods such as ROC [37] and EqOdds [34] frequently reduce group disparities while incurring bounded utility loss relative to `DP-only` training. These results provide *actionable guidance* on where and how to intervene in DP-synthetic learning pipelines, while highlighting that both the intervention stage and the DP synthesizer shape the attainable privacy–fairness–utility trade-off.

Limitations. Our study focuses on *tabular, binary classification* with a *binary protected attribute*. While this represents the most widely studied and standardized setting in the fairness literature [9, 28, 49, 59], extensions to multi-class tasks, regression problems, and multi-valued or intersectional protected attributes are important to cover a broader range of real-world scenarios. Such extensions, however, would require substantial adaptations of fairness metrics, mitigation strategies (which currently rely on the AIF360 framework and its binary attribute/label formulations), and evaluation protocols, and are therefore left for future work. Moreover, we restrict our evaluation to *marginal-based* DP synthetic data generators, as prior benchmarks [32, 48, 54] show that they consistently outperform *deep generative* DP synthesizers on tabular data under realistic privacy budgets. As a result, our conclusions may not directly transfer to settings where lower-utility or domain-specific DP generative models are employed. Regarding learning algorithms, we evaluate *three representative base classifiers* (XGBoost, RF, and LR) using default hyperparameters, unless otherwise stated, to ensure controlled and reproducible comparisons. Other model families or extensively tuned configurations may interact differently with both DP noise and fairness interventions. Finally, our fairness evaluation focuses on *group-level accuracy and error disparity metrics*. Broader notions of fairness, such as full Equalized Odds, counterfactual fairness, or individual fairness, are not considered and remain important directions for extending this study.

Future directions. Building on this benchmark, several extensions remain open. These include evaluating *additional families of DP synthesizers* (*e.g.*, DP-GANs [55]), extending to *multi-task, multi-label, and multi-class settings* with richer and potentially intersecting protected attributes, and assessing a wider range of *learning algorithms* and *hyperparameter regimes*. Future work should also expand beyond AIF360-based interventions by considering newer fairness mechanisms, broaden the metric suite to include *calibration-based*, *individual*, or *counterfactual fairness* criteria, and compare post-hoc mitigation strategies with approaches that jointly enforce fairness and privacy during synthetic data generation (*e.g.*, [38]). To facilitate progress along these directions, we release all code and experimental artifacts (see Section 4.7).

Acknowledgments

This work was partially supported by the French National Research Agency (ANR) research grants (ANR-24-CE23-6239, ANR-23-IACL-0006) and by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (RGPIN-2026-04945).

AI-Generated Content Acknowledgement

The authors acknowledge the use of ChatGPT (OpenAI, GPT-5 model) and Claude (Anthropic, Sonnet-4.6 model) to assist with language-related improvements, including grammar correction, spelling, formatting consistency, and refinement of phrasing for clarity and readability. Moreover, the authors acknowledge the use of the aforementioned models as auxiliaries to build automation shell scripts to facilitate artifact generation. All conceptual, technical, and experimental contributions originate solely from the authors.

References

[1] 2024. Tumult Labs. https://www.tmlt.io/differentially-private-synthetic-data.
[2] 2024. YData. https://ydata.ai/products/synthesizer.html.
[3] Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. 2025. On the Alignment of Group Fairness with Attribute Privacy. In *International Conference on Web Information Systems Engineering*. Springer, 333–348. https://doi.org/10.1007/978-981-96-0567-5_24
[4] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318
[5] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html
[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.
[7] Héber H Arcolezi, Mina Alishahi, Adda-Akram Bendoukha, and Nesrine Kaaniche. 2025. Fair Play for Individuals, Foul Play for Groups? Auditing Anonymization's Impact on ML Fairness. In *ECAI 2025*. IOS Press, 1009–1018. https://doi.org/10.3233/FAIA250909
[8] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems* 32 (2019).
[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
[10] Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1002–1013. https://doi.org/10.1145/3593013.3594058
[11] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.
[12] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.
[13] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[14] Blake Bullwinkel, Kristen Grabarz, Lily Ke, Scarlett Gong, Chris Tanner, and Joshua Allen. 2022. Evaluating the fairness impact of differentially private synthetic data. *arXiv preprint arXiv:2205.04321* (2022).
[15] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. 13–18. https://doi.org/10.1109/ICDMW.2009.83
[16] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028
[17] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[18] CNIL. 2026. AI system development: CNIL's recommendations to comply with the GDPR. https://www.cnil.fr/en/ai-system-development-cnils-recommendations-comply-gdpr.
[19] William Jay Conover. 1999. *Practical nonparametric statistics*. john wiley & sons.
[20] Graham Cormode, Shripad Gade, Samuel Maddock, and Enayat Ullah. 2025. Synthetic Tabular Data: Methods, Attacks and Defenses. *Proc. VLDB Endow.* 18, 12 (Sept. 2025), 5448–5450. https://doi.org/10.14778/3750601.3750692
[21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).
[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255
[23] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.
[24] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[25] European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
[26] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*. 15–19.
[27] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311
[28] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. 2022. Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. IJCAI. https://doi.org/10.24963/ijcai.2022/766
[29] FSA. 2023. Feedback Statement on Synthetic Data Call for Input. https://www.fca.org.uk/publications/feedback-statements/fs23-1-feedback-statement-synthetic-data-call-for-input.
[30] Georgi Ganev and Emiliano De Cristofaro. 2025. The Inadequacy of Similarity-Based Privacy Metrics: Privacy Attacks Against "Truly Anonymous" Synthetic Datasets. In *2025 IEEE Symposium on Security and Privacy (SP)*. 4007–4025. https://doi.org/10.1109/SP61157.2025.00218
[31] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. 2022. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*. PMLR, 6944–6959.
[32] Georgi Ganev, Kai Xu, and Emiliano De Cristofaro. 2024. Graphical vs. Deep Generative Models: Measuring the Impact of Differentially Private Mechanisms and Budgets on Utility. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 1596–1610. https://doi.org/10.1145/3658644.3690215
[33] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. 2023. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proceedings on Privacy Enhancing Technologies* 2 (2023), 312–328.
[34] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
[35] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. 2024. SoK: Privacy-Preserving Data Synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*. 4696–4713. https://doi.org/10.1109/SP54263.2024.00002
[36] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257* (2022).
[37] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*. IEEE, 924–929.
[38] Soyeon Kim, Yuji Roh, Geon Heo, and Steven Euijong Whang. 2025. PFGuard: A Generative Framework with Privacy and Fairness Safeguards. In *The Thirteenth International Conference on Learning Representations*.
[39] Tongyu Liu, Ju Fan, Guoliang Li, Nan Tang, and Xiaoyong Du. 2024. Tabular data synthesis with generative adversarial networks: design space and optimizations. *The VLDB Journal* 33, 2 (2024), 255–280.

[40] Andrew Lowy, Zhuohang Li, Jing Liu, Toshiaki Koike-Akino, Kieran Parsons, and Ye Wang. 2024. Why does differential privacy with large epsilon defend against practical membership inference attacks? *arXiv preprint arXiv:2402.09540* (2024).
[41] Karima Makhlouf, Heber H Arcolezi, Sami Zhioua, Ghassen Ben Brahim, and Catuscia Palamidessi. 2024. On the impact of multi-dimensional local differential privacy on fairness. *Data Mining and Knowledge Discovery* 38, 4 (2024), 2252–2275. https://doi.org/10.1007/s10618-024-01031-0
[42] Karima Makhlouf, Tamara Stefanović, Héber H. Arcolezi, and Catuscia Palamidessi. 2024. A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results. In *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*. 1–16. https://doi.org/10.1109/CSF61375.2024.00039
[43] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978* (2021).
[44] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. 2022. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.* 15, 11 (July 2022), 2599–2612. https://doi.org/10.14778/3551793.3551817
[45] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. 2019. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*. PMLR, 4435–4444.
[46] Kevin P Murphy. 2023. *Probabilistic machine learning: Advanced topics*. MIT press.
[47] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. 2018. Scalable Private Learning with PATE. In *International Conference on Learning Representations*.
[48] Mayana Pereira, Meghana Kshirsagar, Sumit Mukherjee, Rahul Dodhia, Juan Lavista Ferres, and Rafael de Sousa. 2024. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE* 19, 2 (Feb. 2024), e0297271. https://doi.org/10.1371/journal.pone.0297271
[49] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (Feb. 2022), 44 pages. https://doi.org/10.1145/3494672
[50] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf
[51] Zhaozhi Qian, Rob Davis, and Mihaela Van Der Schaar. 2023. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. *Advances in neural information processing systems* 36 (2023), 3173–3188.
[52] Lucas Rosenblatt, Bernease Herman, Anastasia Holovenko, Wonkwon Lee, Joshua Loftus, Elizabeth McKinnie, Taras Rumezhak, Andrii Stadnik, Bill Howe, and Julia Stoyanovich. 2023. Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy. *Proc. VLDB Endow.* 16, 11 (July 2023), 3178–3191. https://doi.org/10.14778/3611479.3611517
[53] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler
[54] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2021. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238* (2021).
[55] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
[56] Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Joslin Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. 2021. Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv preprint arXiv:2106.12576* (2021).
[57] U.S. Census Bureau. 2023. 2023 ACS 1-Year Public Use Microdata Sample (PUMS) Code Lists. https://www2.census.gov/programs-surveys/acs/tech_docs/pums/code_lists/ACSPUMS2023CodeLists.xls. Accessed: 2025-09-06.
[58] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).
[59] Kai Yao and Marc Juarez. 2025. SoK: What Makes Private Learning Unfair?. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning*. 841–857. https://doi.org/10.1109/SaTML64287.2025.00052
[60] Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. 2025. The DCR delusion: measuring the privacy risk of synthetic data. In *European Symposium on Research in Computer Security*. Springer, 469–487. https://doi.org/10.1007/978-3-032-07884-1_24
[61] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[62] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42, 4, Article 25 (Oct. 2017), 41 pages. https://doi.org/10.1145/3134428

A Classifiers, Datasets, & Data Pre-Processing

This section provides complementary details to support the reproducibility of the experimental results presented in the main paper. We first describe the data pre-processing pipeline applied to each dataset, followed by additional results for alternative classifier and synthesizer configurations. The data treatment pipeline is structured into three stages: (i) data cleaning, (ii) feature encoding, and (iii) binary transformation. For some datasets, feature domain compression is additionally applied, following the recommendations of [44]. When used, this compression is explicitly described under **Feature encoding** for the corresponding dataset.

A.1 Adult Dataset

Data Cleaning. First, we start by removing all samples with one or more missing features (null samples) and, consequently, their label. Then, we proceeded to relabel the dataset so that only two labels remained ($\leq 50K$ and $> 50K$). Lastly, we remove the following features from the dataset: (i) *fnlwgt*; (ii) *education-num*; (iii) *capital-loss*; (iv) *capital-gain*. By the end of this step, the remaining dataset contains 10 features and 47, 621 samples.

Feature encoding. The first step is to compress the domain of the dataset (as cardinality would pose a problem for AIM or MST). Such a task was performed by grouping the *ages* into 20 different bins, each with a 5-year range (1-5, 6-10, and so on). Then, we perform the same for the *hours-per-week* feature, using five different sets with a 20-hour range. Lastly, we compress the *native-country* feature by directly mapping each unique country to a continent. The next step in the encoding pipeline is to ensure all non-numeric features are transformed into numerical features. This is done using pandas categorical encoding, which assigns an integer to each unique value per feature.

Binary transformation. The last step of the pipeline is the binary transformation of each sensitive feature. This task is performed once for all sensitive features to facilitate tests with different values for the protected attribute A. For the Adult dataset, this transformation is as follows: (i) maps "Male" *sex* to 1 and others to 0; (ii) maps "White" *race* to 1 and others to 0.

A.2 COMPAS Dataset

Data Cleaning. First, we start by removing all samples where the *race* is neither **African-American** nor **Caucasian**. Then we remove all columns except for (i) *sex*, (ii) *race*, (iii) *age*, (iv) *c_charge_degree*, (v) *prior_counts*, (vi) *score_text*, and (vii) *two_year_recid* (target column). Then, as in Adult, we remove all samples with one or more missing features (null samples) and, consequently, their labels. By the end of this step, the remaining dataset contains 7 features and 5, 050 samples.

Feature encoding. The first step is to ensure all non-numeric features are transformed into numerical features. This is done exactly like in the Adult dataset, using pandas categorical encoding.

Binary transformation. The last step of the pipeline is the binary transformation of each sensitive feature. For the COMPAS dataset, this transformation is as follows: (i) maps "Male" *sex* to 1 and others to 0; (ii) maps "Caucasian" *race* to 1 and others to 0. (iii) maps $age \geq 25$ and $age \leq 45$ to 1 and others to 0.

A.3 ACSIncome Dataset

Data Cleaning. First, we start by filtering the census to the year 2018 and the state of Utah. Then we remove all features except for: (i) *AGEP*, (ii) *COW*, (iii) *SCHL*, (iv) *MAR*, (v) *OCCP*, (vi) *POBP*, (vii) *RELP*, (viii) *WKHP*, (ix) *SEX*, (x) *RAC1P*, and (xi) *PINCP* (target column). Then, as in Adult and COMPAS, we remove all samples with one or more missing features (null samples) and, consequently, their labels. By the end of this step, the remaining dataset contains 11 features and 16, 337 samples.

Feature encoding. Like in Adult, the first step is to compress the domain of large features, such as *OCCP*. This is done by grouping each occupation into 6 categories, following the code list [57]. We then ensure that all non-numeric features are transformed into numerical features. This is done exactly like in the Adult dataset, using pandas categorical encoding.

Binary transformation. The last step of the pipeline is the binary transformation of each sensitive feature. For the ACSIncome dataset, this transformation is as follows: (i) maps "Male" *sex* to 1 and others to 0; (ii) maps "White alone" *race* to 1 and others to 0.

A.4 BiasOnDemand Dataset

BiasOnDemands (BoD) has a simpler pipeline, as it is a fully customised dataset. Using the BoD dataset generator with the configuration described in Table 3, we can skip the data cleaning step and proceed to feature encoding[3].

Feature encoding. This step first discretises all features in the dataset that have continuous values (R, Q). For that, the values are grouped into 5 different bins with integer values representing each bin.

Binary transformation. Essentially, all that needs to be done is to invert the values of the protected attribute A, to follow the pattern of the other datasets.

A.5 Classifiers

In this section, we report any modifications to the default values of the hyperparameters of each classifier, and thereafter, the motivation for each parameter. Both XGBoost and Random Forest are executed in their default settings, with only the random seed being set, for reproducibility. As for the Logistic Regression, we refer to Table 4 to display the parameters. The motivation behind this set of parameters is mostly empirical experimentation. Prior results with logistic regression under default parameters did not reach reasonable utility metrics (compared to the other two), either due to high-dimensionality and correlation of the data, over/under-fitting, or the non-linear structure of the data. Under these conclusions, we experimented with different sets of parameters (via Grid Search optimization) until reaching reasonable utility metrics. The parameters in Table 4 were chosen as an attempt to balance L1 and L2 regularization while improving the reliability of the convergence in high-dimensional and correlated feature spaces.

[3]A description for each parameter can be found here: https://github.com/rcrupiISP/BiasOnDemand/tree/main

B Ablation and Additional Results

This section provides complementary analyses that support the main findings reported in Section 5. We first present two ablation studies targeting key methodological choices: (i) whether the advantage of post-processing methods depends on access to a real held-out calibration set, and (ii) whether the comparatively limited gains of in-processing methods are due to their default hyperparameter settings. We then report additional experimental results across alternative classifier and synthesizer configurations, as well as extended BiasOnDemand settings. Together, these analyses help assess whether the observed fairness–utility–privacy trade-offs are stable across calibration protocols, in-processing tuning choices, model families, DP synthesizers, and controlled bias scenarios.

B.1 Ablation: DP-Compliant Calibration Set for Post-Processing

In the main benchmark (Section 4.1), fairness post-processing methods are calibrated using a real held-out calibration set. As discussed in Section 4.6, this does not affect the DP guarantee with respect to the protected training split used by the synthesizer, but the calibration records themselves are not protected by that guarantee. To assess whether this design choice biases the comparison in favor of post-processing methods, we conduct an ablation in which ROC, EqOdds, and CEOP are calibrated using DP-compliant calibration data rather than non-private calibration records.

In contrast to the methodology described in Section 4.6, the partitioning of training, calibration, and test subsets is performed as follows: (1) The initial dataset D is partitioned into disjoint subsets D_{temp} and D_{test} using an 80-20 split, where $|D_{\text{test}}| = 0.2 \times |D|$. (2) The temporary set D_{temp} undergoes differential privacy (DP) preprocessing to ensure ϵ-DP compliance. (3) Following this processing, D_{temp} is further partitioned into D_{train} and D_{cal} with a 75-25 split ratio, yielding $|D_{\text{train}}| = 0.75 \times |D_{\text{temp}}| = 0.6 \times |D|$ and $|D_{\text{cal}}| = 0.25 \times |D_{\text{temp}}| = 0.2 \times |D|$.

This partitioning yields three final subsets with the following proportions relative to the original dataset: DP-processed training set (60%), DP-processed calibration set (20%), and untouched test set (20%), as in Section 4.6 proportions. We note that for experimental configurations where post-processing calibration is not required (*i.e.*, pre-processing or in-processing fairness interventions), D_{cal} is excluded from the pipeline. For the `Baseline` and `Fair-only` classifiers, all three subsets remain in their original, untouched form, maintaining the same distributional partitioning scheme as described above.

With this experimental setup, we aim to demonstrate that the main qualitative findings in Sections 5 and 6 exhibit robustness under this alternative calibration scheme. Figure 5 presents a comparison of all fairness mechanisms, in which post-processing methods are calibrated using a DP-compliant dataset subject to the same privacy guarantees as the training set (ε-DP). While the influence of DP-calibrated data becomes more pronounced in high-privacy regimes ($\varepsilon < 1$), post-processing methods consistently remain closer to the ideal value of zero across both EOD and SPD metrics, while

Parameter	Config 1	Config 2	Config 3	Config 4	Config 5	Config 6
l_y	10	0	0	0	0	0
l_{my}	0	15	0	0	0	0
thr_supp	1	1	1	1	1	1
l_{hr}	0	0	0	10	0	10
l_{hq}	0	0	0	0	10	0
l_m	0	0	0	0	0	0
p_u	1	1	0.2	1	1	1
l_r	false	false	false	false	false	false
l_o	false	false	false	false	false	false
l_{yb}	0	0	0	0	0	0
l_q	2	2	2	2	10	2
sy	5	5	5	5	5	5
l_{rq}	0	0	0	0	0	10
$l_{my\_non\_linear}$	false	false	false	false	false	false

Table 3: Configuration parameters for BoD data generation.

Parameter	Choice
solver	saga
penalty	elasticnet
l1_ratio	0.5
C	0.8
max_iter	10000

Table 4: Configuration parameters for Logistic Regression classifier.

preserving utility performance comparable to `DP-only`. This behaviour corroborates the trends established in Section 5.3.

Similar to Figure 4 in the main paper, Figure 6 reproduces the multi-classifier analysis across the Adult dataset for three learning algorithms, each paired with a representative fairness intervention corresponding to its respective processing stage. The geometric configuration and relative ordering of mechanisms in Figure 6 remain largely invariant to those of Figure 4. Specifically, these trends persist independently of the calibration set composition, confirming that the adoption of DP-calibrated data does not fundamentally alter the outcome of the fairness correction.

Generally, across the classifiers, ROC and EqOdds continue to provide the strongest overall fairness–utility trade-offs among the evaluated interventions, even when fairness post-processing is calibrated without direct access to non-private calibration records. This indicates that the advantage of post-processing observed in the main experiments is not solely driven by the use of a real held-out calibration set.

B.2 Ablation: Hyperparameter Sensitivity for In-Processing Methods

In Section 4.1, the benchmark evaluates each fairness mechanism using its default hyperparameter configuration. As discussed previously in Section 5.2, several methods yielded only modest and inconsistent improvements in fairness while largely preserving predictive utility. This trend was particularly pronounced for the in-processing mechanisms. To determine the extent to which these results may be influenced by the use of default hyperparameters, we perform an ablation study in which multiple hyperparameter combinations are evaluated for both the EGR and GSR mechanisms. Finally, we compare the configurations that achieve the most favourable utility–fairness trade-off for each fairness metric.

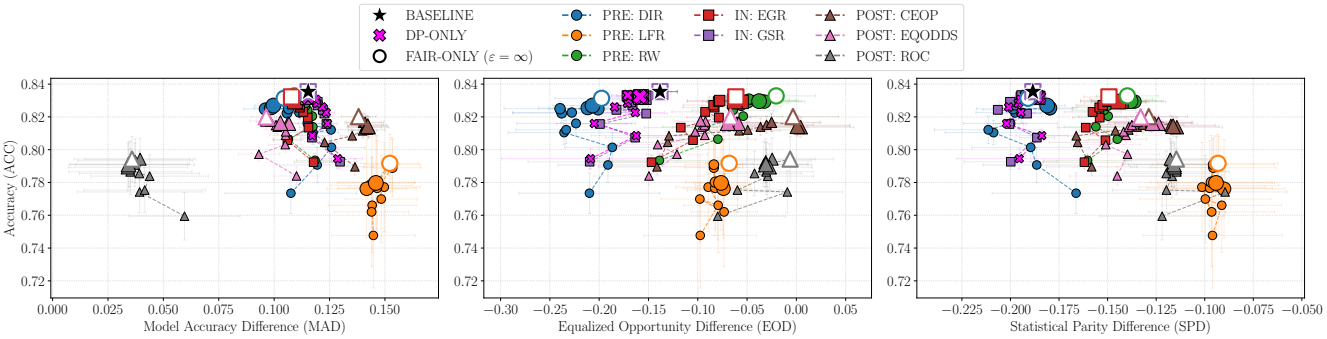
Hyperparameters were selected based on their expected influence on model performance, while keeping the fairness constraints and fairness-violation bounds fixed. For EGR, we vary *max_iter* and *nu*. The parameter *max_iter* determines the maximum number of optimisation iterations, whereas *nu* specifies the convergence threshold. For GSR, we vary *grid_size* and *grid_limit*. The former controls the size of the search space, while the latter determines the range of Lagrange multipliers considered during optimisation.

For each mechanism, we construct a grid of 25 hyperparameter combinations. The full set of candidate values is reported in Table 5. For each fairness mechanism, we then identify the configuration that provides the best utility–fairness trade-off as per Equation (1) and visualise the corresponding results.

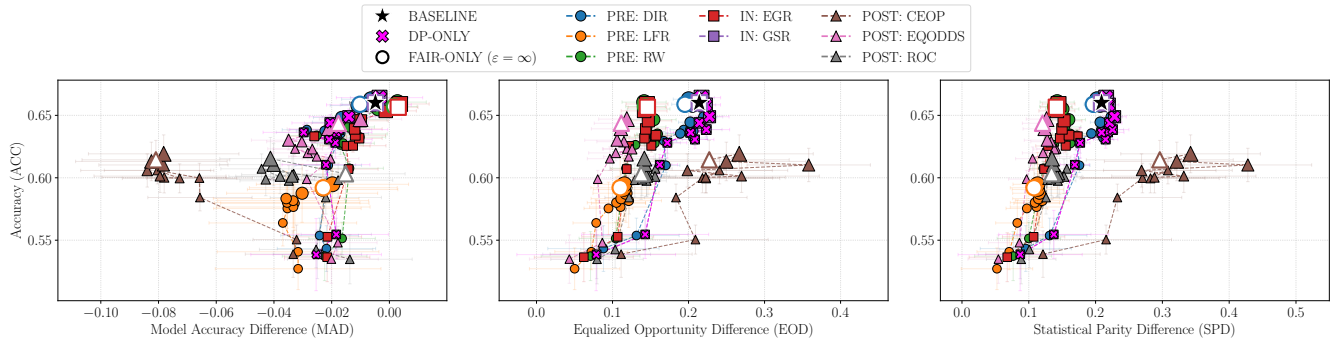
Algorithm	Hyperparameter	Candidate Values
EGR	`max_iter`	{50, 75, 100, 125, 150}
EGR	`nu`	{None, 0.01, 0.05, 0.1, 0.5}
GSR	`grid_size`	{10, 15, 20, 30, 50}
GSR	`grid_limit`	{2.0, 1.0, 1.5, 2.5, 3.0}

Table 5: Hyperparameter grid used for tuning fairness-aware reduction methods.

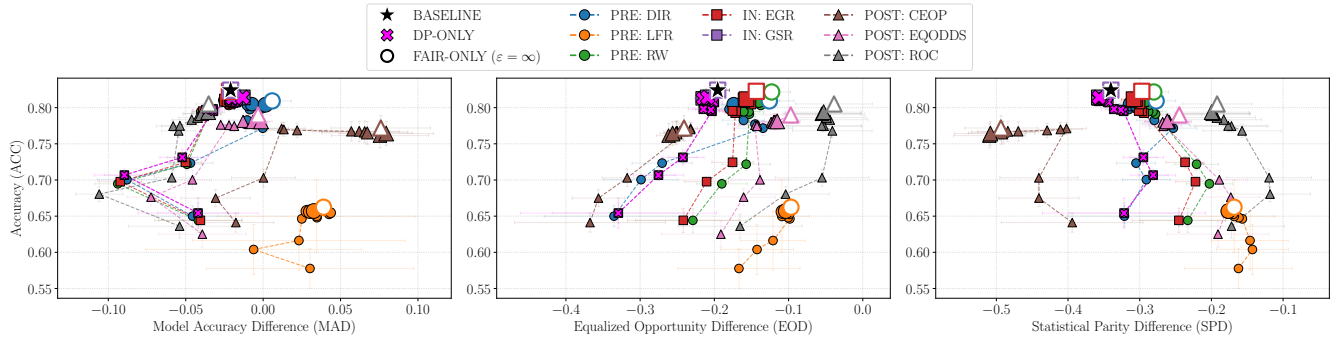
The results presented in Figure 7 indicate that variations in hyperparameter configuration lead to slight changes in the utility–fairness trade-off. However, these differences remain modest and consistent across configurations. This suggests that the trends identified in Section 5.2 are robust and remain largely unchanged even after hyperparameter tuning of the fairness mechanisms.



(a) Adult.



(b) Compas.



(c) ACSIncome.

Figure 5: Accuracy–fairness trade-off under the AIM synthesizer across three real-world datasets (Adult, Compas and ACSIncome) with the XGBoost classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. The post-processing `DP+Fair` settings are calibrated using a DP synthetic calibration set. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.

B.3 Additional Results

In this section, we report additional experimental results that complement those presented in Section 5. Specifically, we consider the following combinations of classifiers and DP synthesizers: (1) XGBoost with MST; (2) Logistic Regression with AIM; (3) Random Forest with AIM, which are reported in Subsections B.3.1, B.3.2, and B.3.3, respectively. In addition, Subsection B.3.4 presents a comparative analysis of the different classifier configurations on the BiasOnDemand dataset. Along with this analysis, we present a summary of the common findings and whether the patterns found are either classifier-dependent or dataset-dependent in Subsection B.3.5.

B.3.1 Results of XGBoost with MST. Figure 8 reports the results obtained with the XGBoost classifier trained on data generated by the MST synthesizer. Overall, post-processing methods, in particular ROC and EqOdds, achieve the most consistent reductions

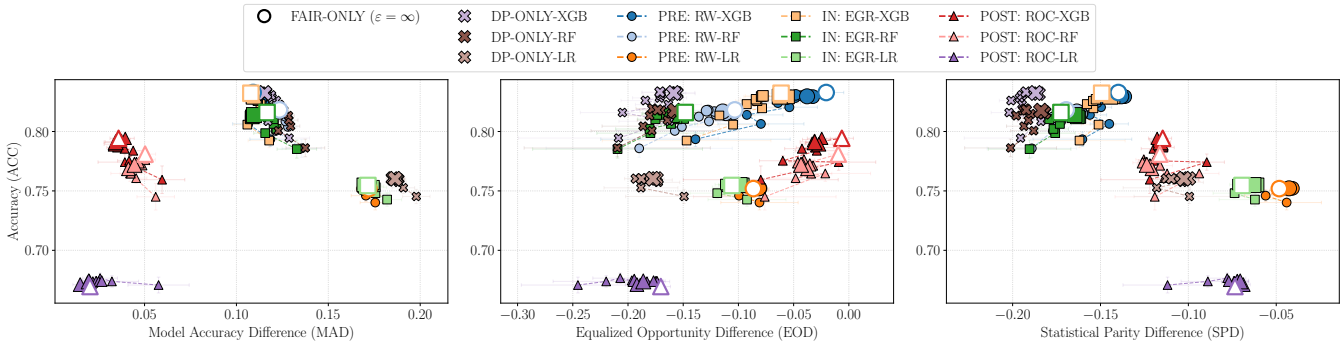


Figure 6: Accuracy–fairness trade-off under the AIM synthesizer across the Adult dataset and three different classifiers (Logistic Regression, Random Forest, and XGBoost). Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for three different fairness mechanisms, across representative interventions from each stage. The post-processing `DP+Fair` settings are calibrated using a DP synthetic calibration set. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$).

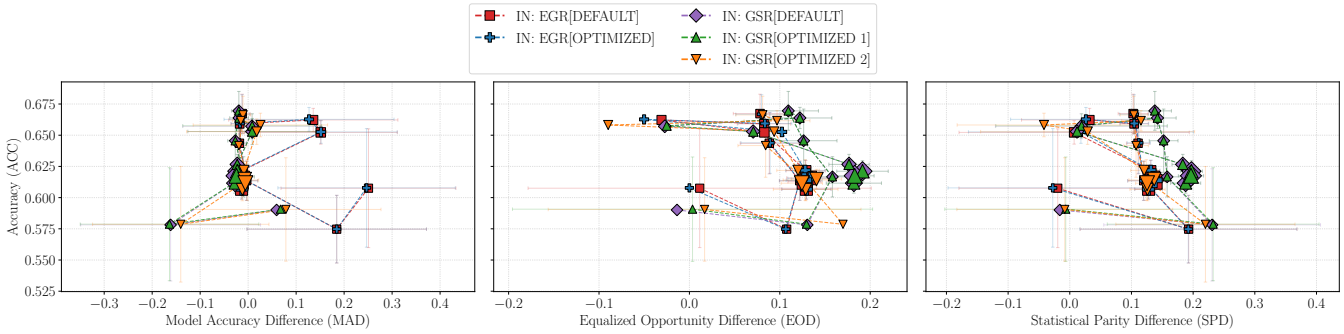


Figure 7: Accuracy–fairness trade-off under the AIM synthesiser across the Compas dataset, two fairness mechanisms and the classifier XGBoost. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for 2 different fairness mechanisms, across 5 hyperparameter combinations. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$).

in group disparities across both EOD and SPD, while maintaining utility relatively close to the `DP-only` baseline. Other mechanisms, such as LFR (pre-processing) and EGR (in-processing), can also reduce disparities, but do so less favorably: LFR consistently incurs substantial utility loss, whereas EGR exhibits more limited and privacy-budget-dependent improvements when compared to post-processing. Among pre-processing methods, Reweighing emerges as the most reliable option in this setting. RW provides noticeable improvements in SPD while keeping accuracy close to the `DP-only` configuration. However, these gains may come at the expense of other fairness notions, such as EOD, as observed for the Adult dataset. Taken together, the XGBoost × MST results highlight two main observations: (i) when predictive utility is a priority, aggressive pre-processing methods such as LFR are not suitable; and (ii) in-processing approaches offer more limited benefits compared to pre- and post-processing, with post-processing methods providing the most favorable and robust fairness–utility trade-offs under this setting.

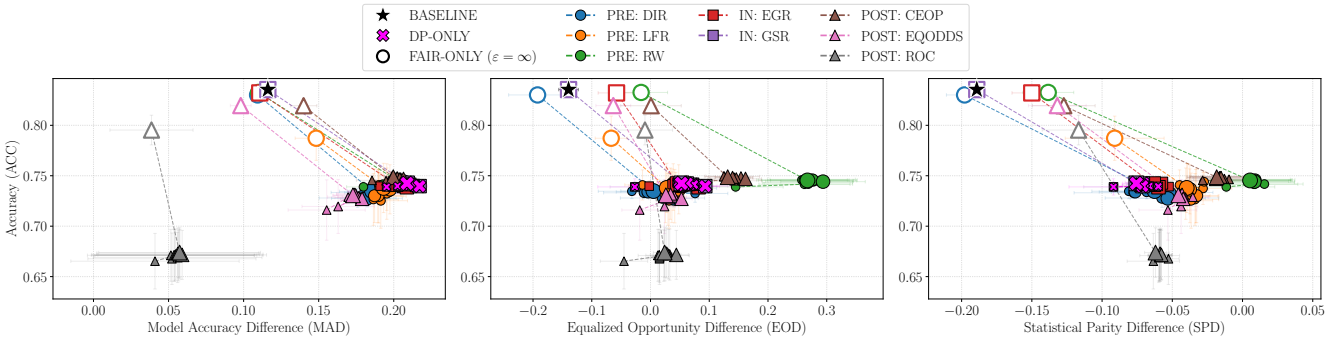
B.3.2 Results of Logistic Regression with AIM. Figure 9 reports the results obtained with Logistic Regression (LR) trained on AIM synthetic data. Compared to the XGBoost-based results in Section 5, some fairness mechanisms exhibit a less favorable accuracy–fairness trade-off on the Adult dataset. At first glance, this suggests a stronger dependence on the choice of classifier. A closer inspection, however, reveals that this behavior is largely driven by differences in *error profiles* rather than a failure of the fairness mechanisms. In particular, the baseline LR model achieves relatively high accuracy but suffers from low recall on the Adult dataset, indicating conservative predictions. Several fairness interventions reduce accuracy while improving recall, leading to a more balanced precision–recall trade-off. This effect is captured by the F1-score results in Figure 10, where the apparent degradation observed under accuracy largely disappears. These observations highlight that the

perceived effectiveness of fairness interventions can depend on the chosen utility metric and on the calibration properties of the underlying classifier. Importantly, despite these metric-dependent effects, the overall qualitative patterns remain consistent with the main results: Reweighing, Exponentiated Gradient Reduction, Reject Option Classification, and Equalized Odds Post-Processing continue to reduce group disparities while maintaining utility close to the `DP-only` baseline. This confirms that the main conclusions that our benchmark extends beyond XGBoost and remain valid for linear classifiers such as Logistic Regression.

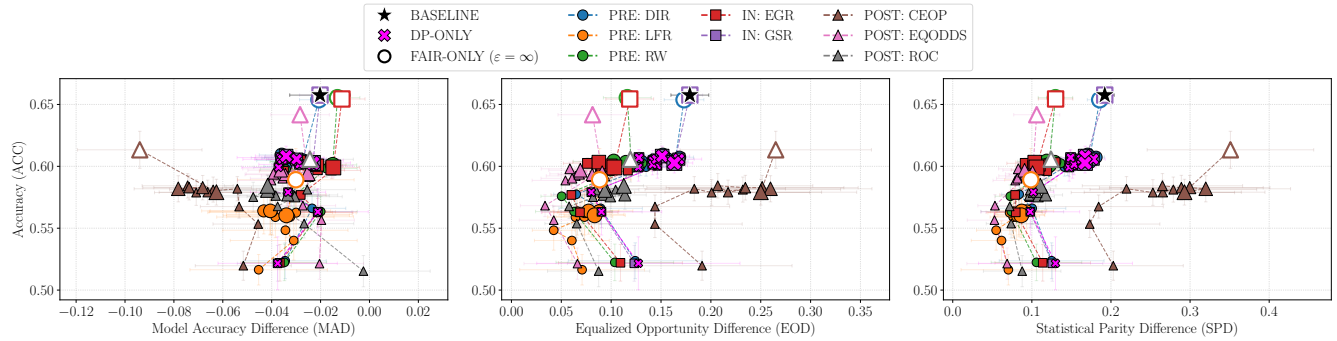
B.3.3 Results of Random Forest with AIM. Figure 11 reports the results obtained using the Random Forest (RF) classifier with the AIM synthesizer. Overall, the trends identified in Section 5 remain consistent under this configuration. In particular, post-processing methods (ROC and EqOdds) continue to offer the most reliable fairness–utility trade-offs: across datasets and privacy budgets, `DP+Fair` models with post-processing closely track the utility of `Baseline` and `DP-only`, while substantially reducing group disparities. The isolated analyses further show that, although models trained on DP synthetic data may initially exhibit worse fairness metrics than the `Baseline`, several mechanisms consistently recover fairness without significant additional utility loss. Specifically, ROC and EqOdds provide the strongest and most stable corrections, while Reweighing and Exponentiated Gradient Reduction also achieve meaningful disparity reductions with moderate utility impact. By contrast, improvements observed for other mechanisms (*e.g.*, GSR) tend to be dataset-specific and do not generalize across domains. Taken together, these results confirm that the qualitative conclusions drawn in the main paper are robust to the choice of classifier: under AIM, post-processing methods—particularly ROC and EqOdds—consistently deliver the best fairness–utility trade-offs, a pattern that holds across datasets and learning algorithms.

B.3.4 Results of Bias on Demand. Figures 12–17 report the results obtained with XGBoost, Logistic Regression, and Random Forest classifiers across the six configurations of the Bias On Demand (BoD) dataset. Analyzing each fairness mechanism separately, several consistent patterns emerge. First, regardless of the induced bias type (imbalance, historical bias, or measurement bias), Reject Option Classification (ROC) consistently corrects group disparities, often bringing them close to zero, while keeping utility loss relatively bounded (on the order of one digit to low double-digit percentage points). Second, Equalized Odds post-processing (EqOdds) follows a similar correction pattern to ROC, but typically incurs a larger utility loss for comparable levels of bias mitigation. Third, Learning Fair Representations (LFR) exhibits the same behavior observed in other datasets: it effectively enforces parity across all bias types, but does so at the cost of the largest utility degradation among the evaluated mechanisms. Fourth, both Exponentiated Gradient Reduction (EGR) and Reweighing achieve comparable fairness–utility trade-offs overall; however, they are more sensitive to the specific type of induced bias and to the classifier used, as their relative performance varies across BoD configurations. Finally, the remaining mechanisms show some ability to reduce bias in specific settings, but their behavior is less stable, with either limited correction capability or high variability across bias types and classifiers.

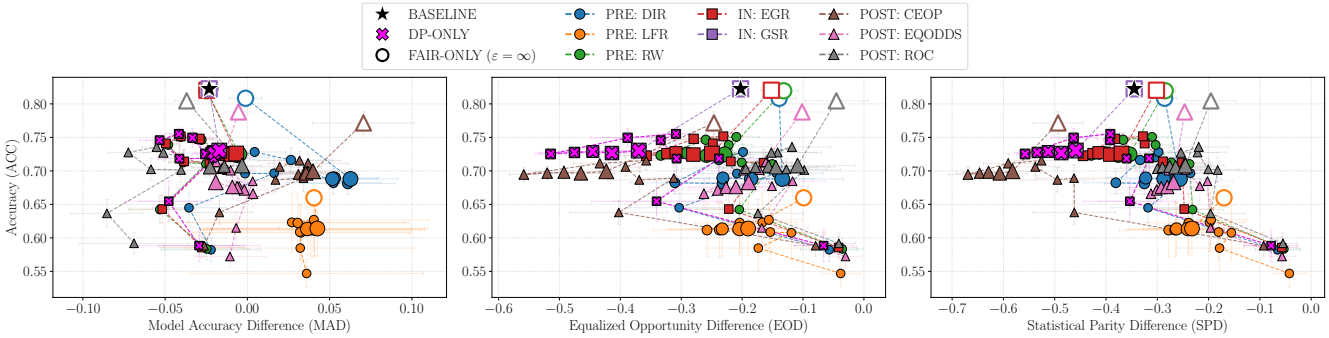
B.3.5 Correlation Between Classifier, Fairness Mechanisms and Dataset. The results presented in this appendix may suggest a dependence of fairness mechanisms on the specific experimental configuration, *i.e.*, the combination of classifier, fairness intervention, and dataset. While such correlations do exist, our analysis shows that they primarily affect the *degree* of bias correction rather than its feasibility. Indeed, several mechanisms exhibit consistent qualitative behavior across classifiers and datasets, with only a few isolated cases deviating from the general patterns observed in the main results. This indicates that these mechanisms remain effective for correcting bias in models trained on *differentially private* synthetic data, even when performance varies across configurations. In particular, the post-processing methods ROC and EqOdds stand out for their robustness, consistently achieving stable and well-balanced fairness–utility trade-offs across all experimental settings.



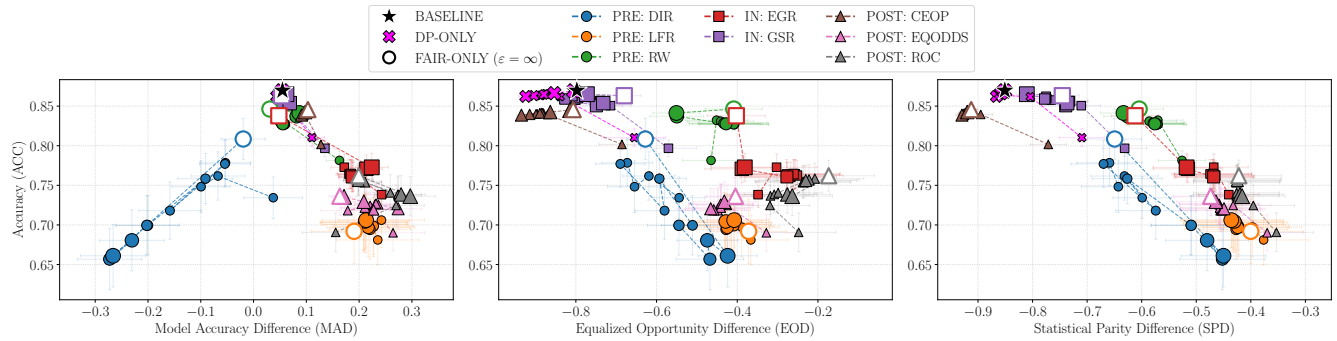
(a) Adult.



(b) COMPAS.

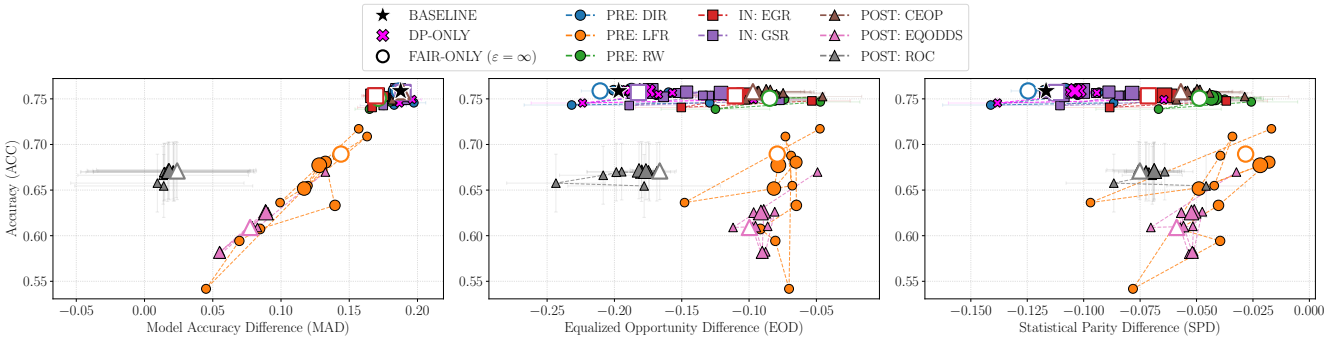


(c) ACSIncome.

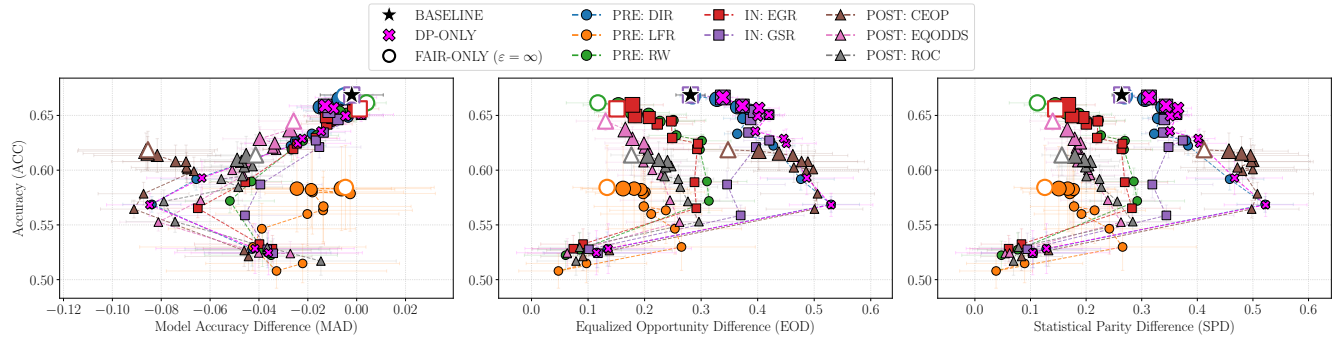


(d) BiasOnDemand (Configuration 5).

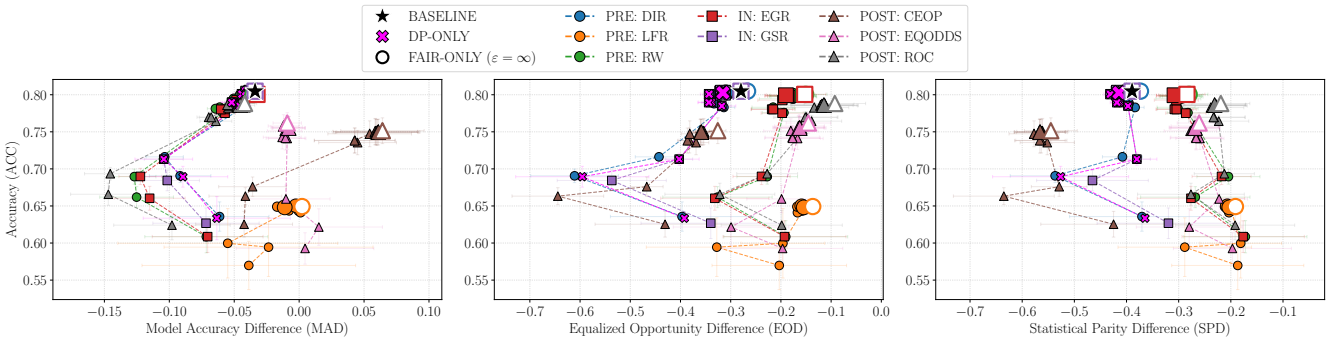
Figure 8: Accuracy–fairness trade-off under the MST synthesizer across four datasets with the XGBoost classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



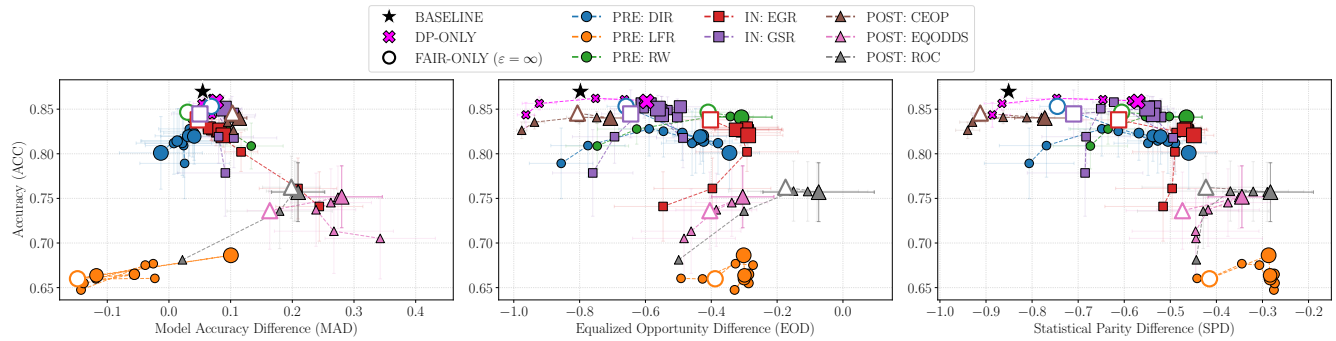
(a) Adult.



(b) COMPAS.

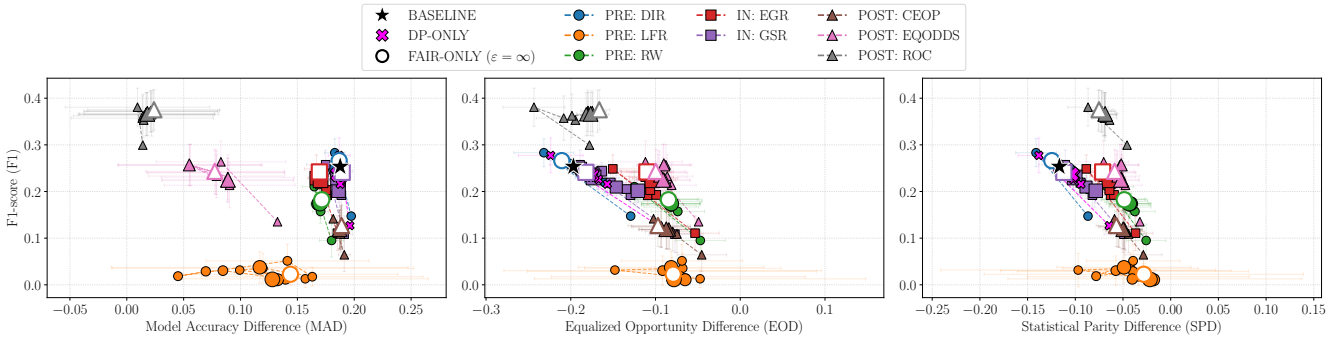


(c) ACSIncome.

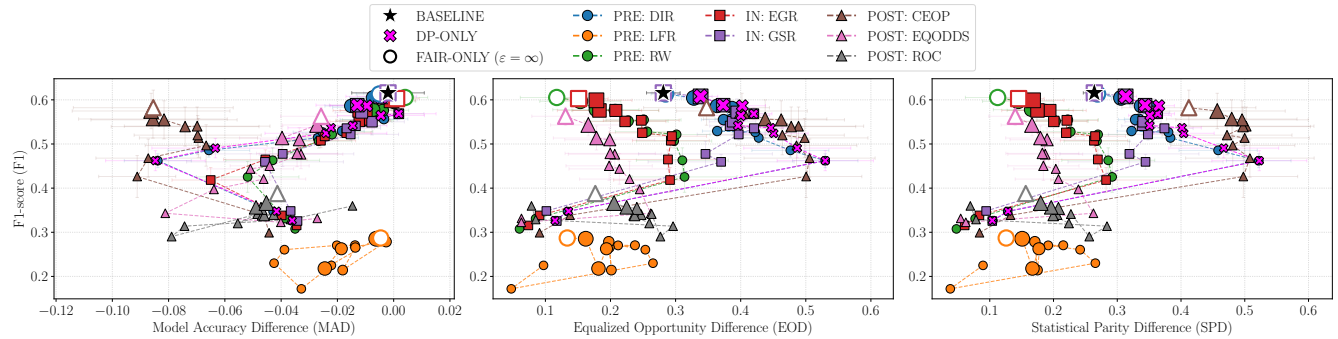


(d) BiasOnDemand (Configuration 5).

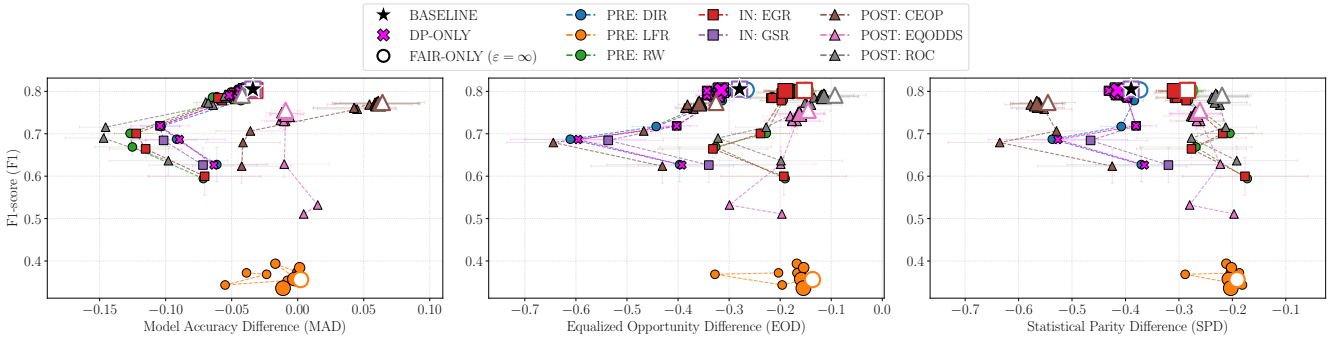
Figure 9: Accuracy–fairness trade-off under the AIM synthesizer across four datasets with the Logistic Regression classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



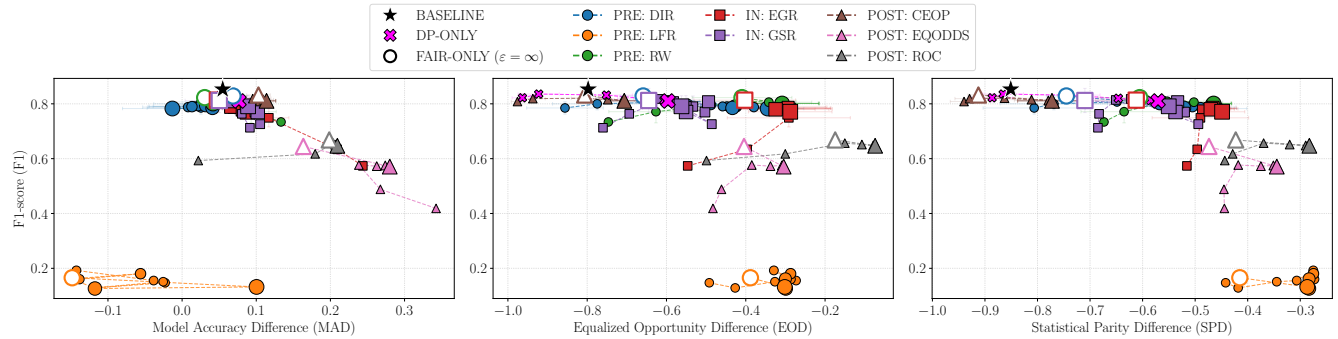
(a) Adult.



(b) COMPAS.

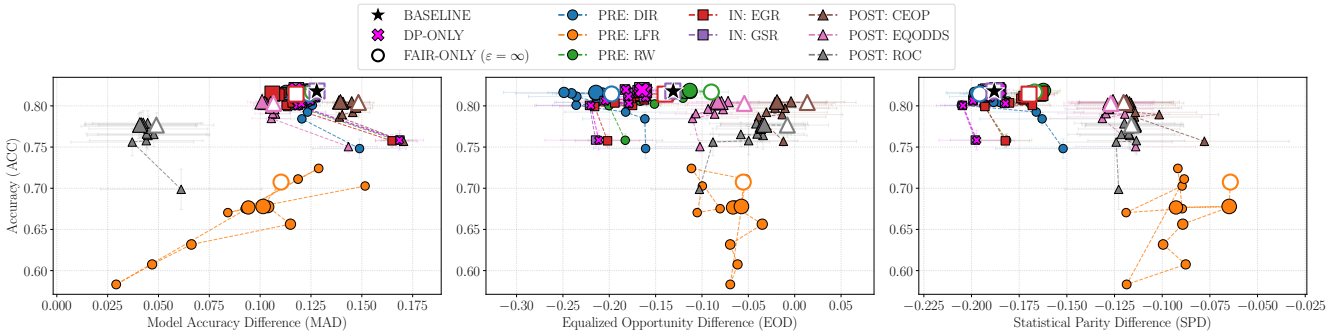


(c) ACSIncome.

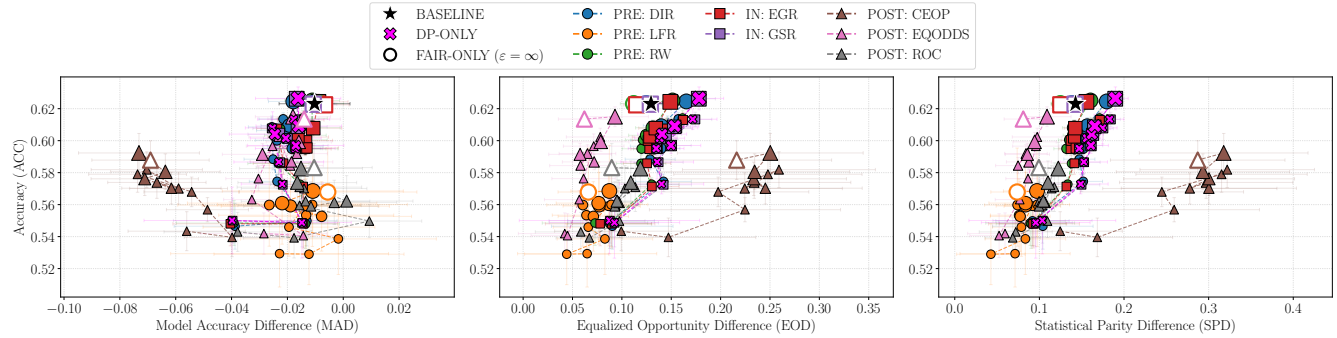


(d) BiasOnDemand (Configuration 5).

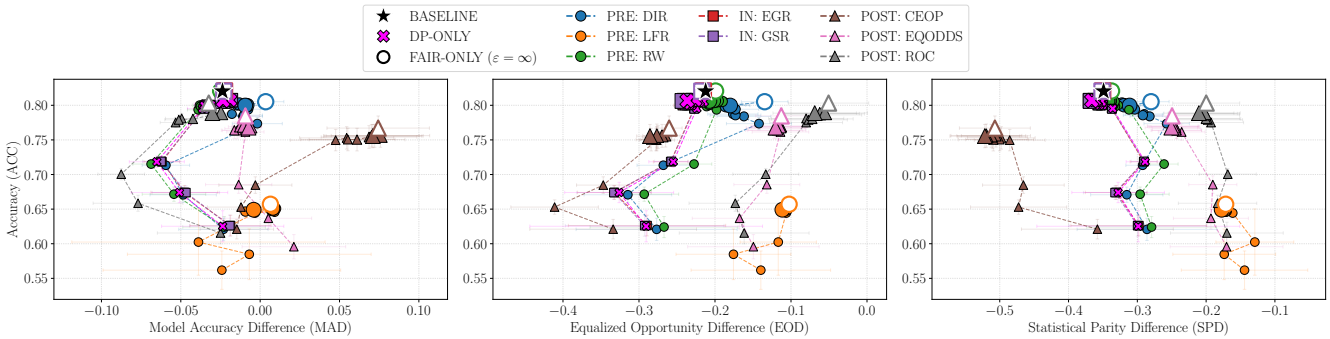
Figure 10: F1–fairness trade-off under the AIM synthesizer across four datasets with the Logistic Regression classifier. The figure reports F1 versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



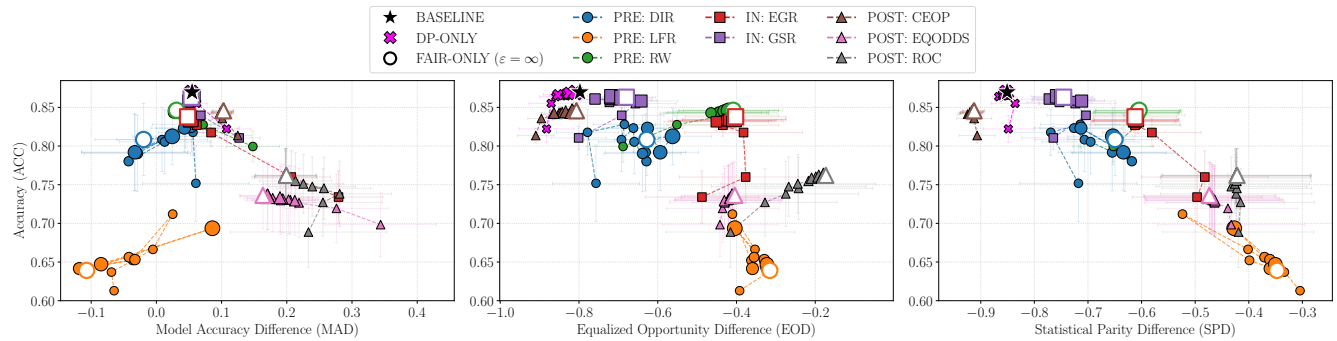
(a) Adult.



(b) COMPAS.

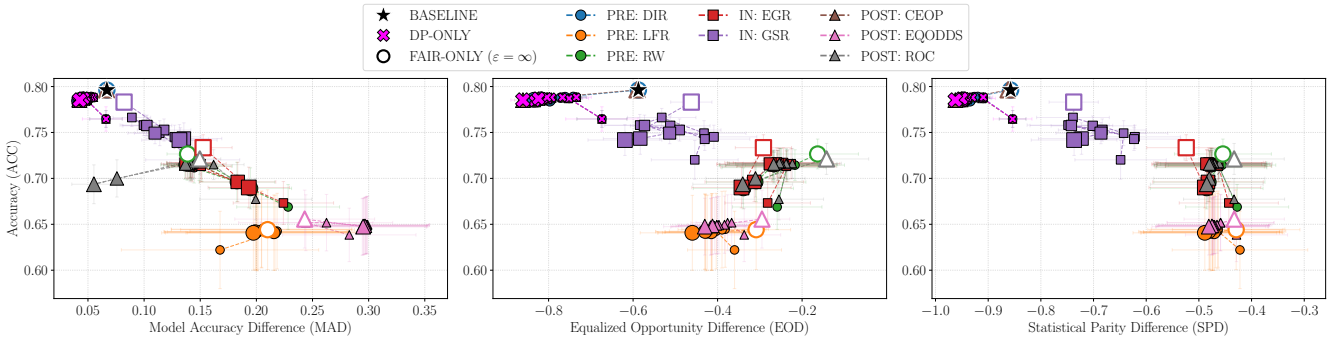


(c) ACSIncome.

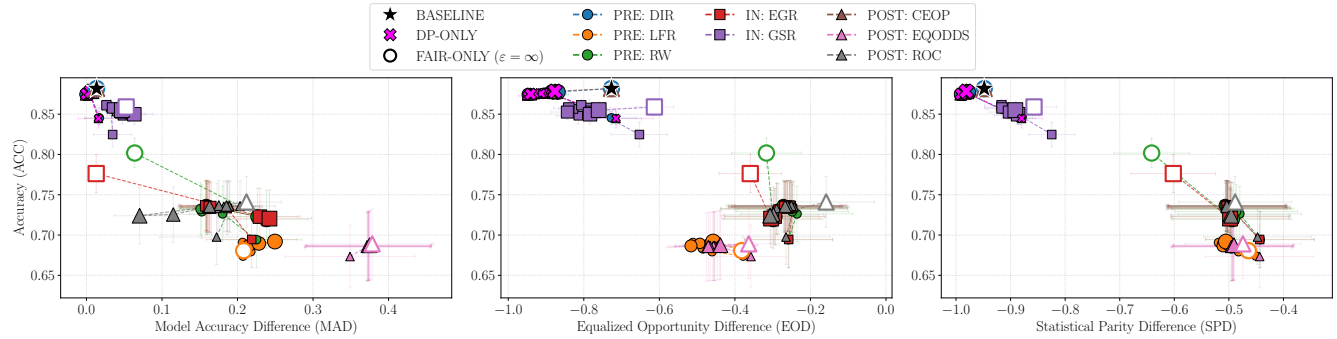


(d) BiasOnDemand (Configuration 5).

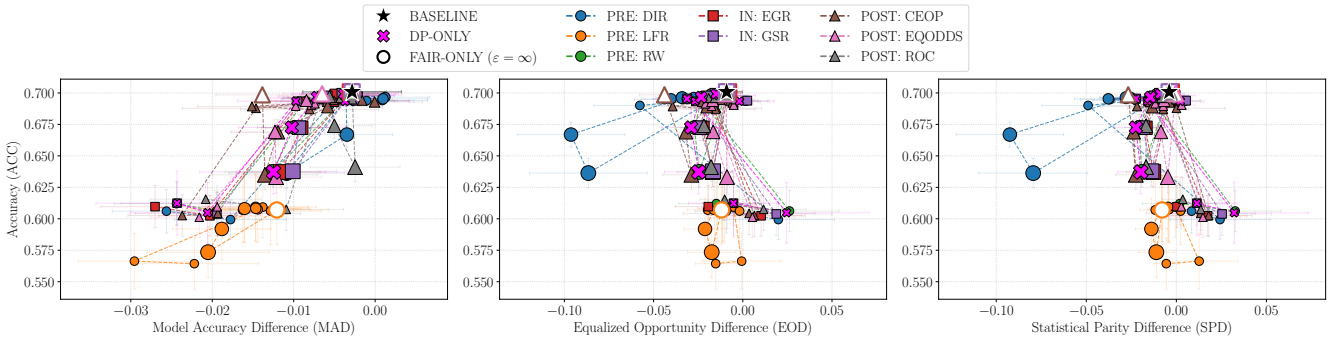
Figure 11: Accuracy–fairness trade-off under the AIM synthesizer across four datasets with the Random Forest classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the ⋆ symbol.



(a) BiasOnDemand Configuration 1.

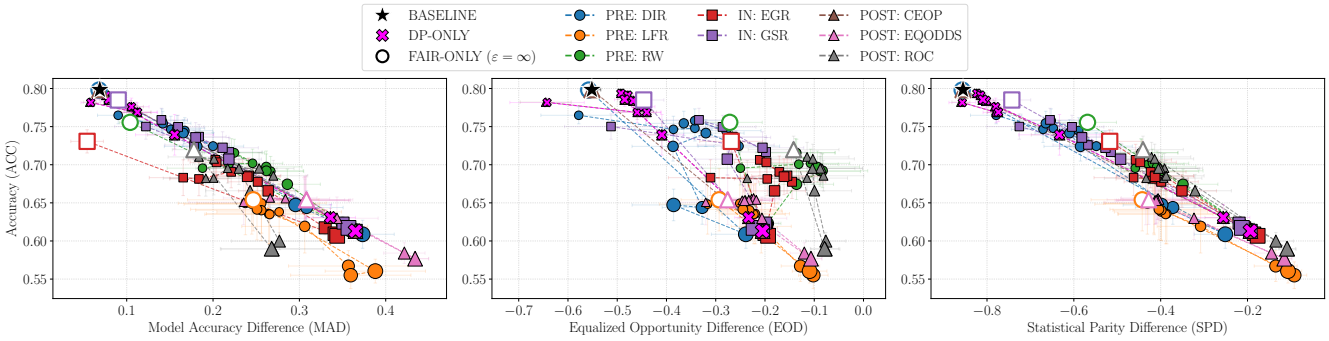


(b) BiasOnDemand Configuration 2.

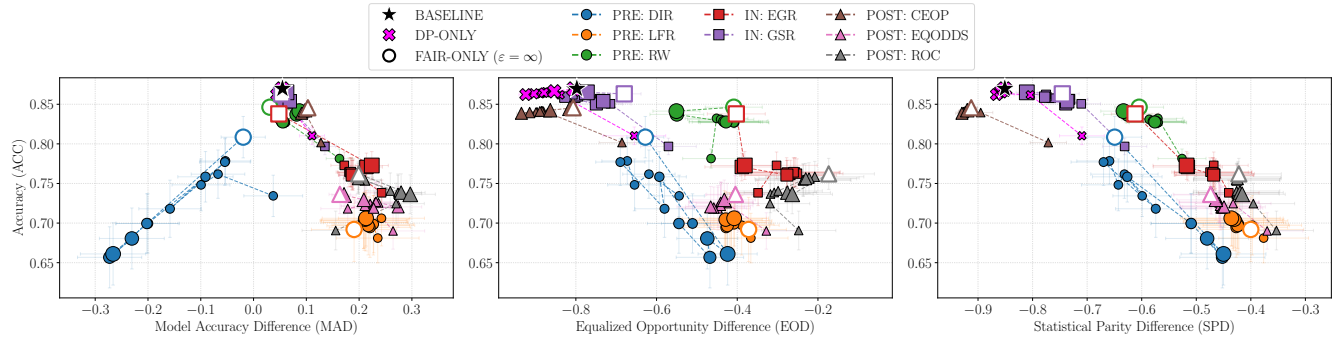


(c) BiasOnDemand Configuration 3.

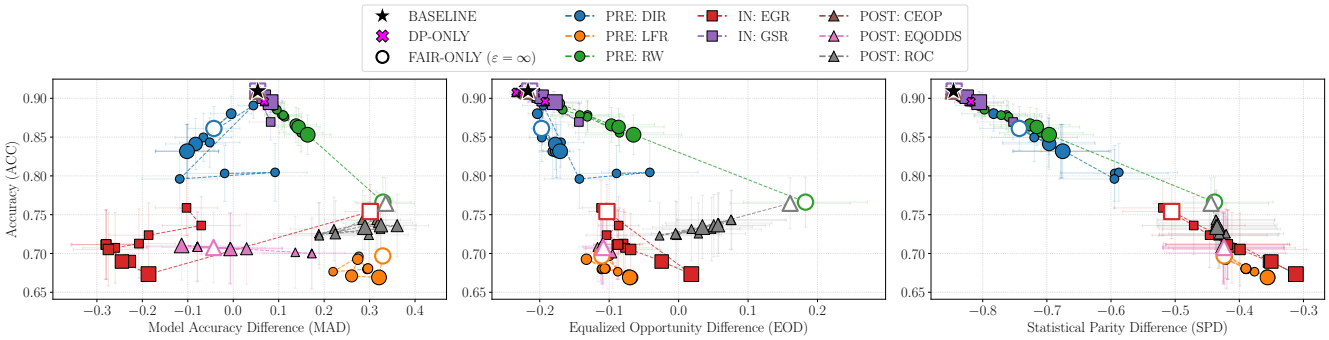
Figure 12: Accuracy–fairness trade-off under the MST synthesizer across configurations 1, 2, and 3 of the BoD dataset with the XGBoost classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



(a) BiasOnDemand Configuration 4.

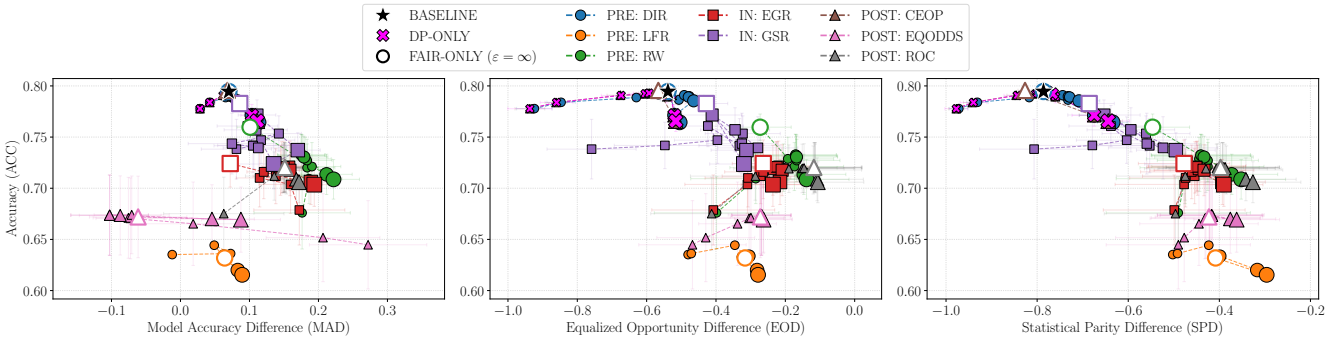


(b) BiasOnDemand (Configuration 5).

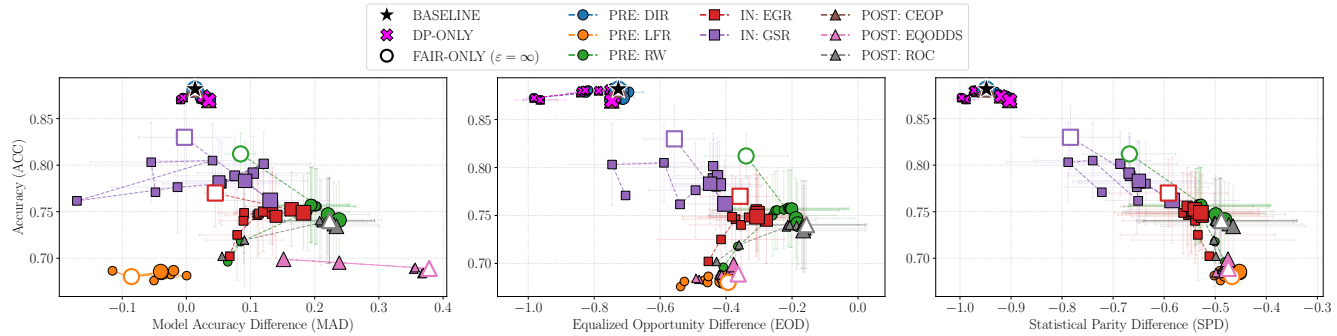


(c) BiasOnDemand Configuration 6.

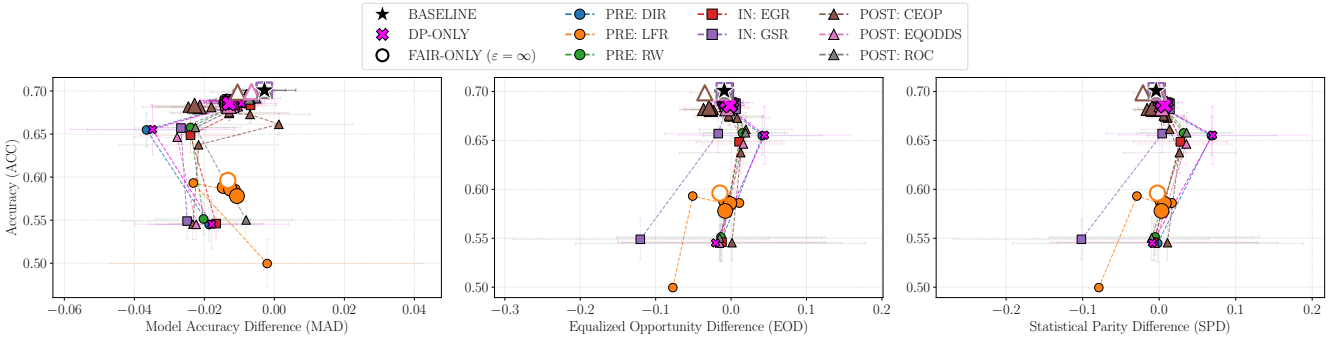
Figure 13: Accuracy–fairness trade-off under the MST synthesizer across configurations 4, 5, and 6 of the BoD dataset with the XGBoost classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



(a) BiasOnDemand Configuration 1.

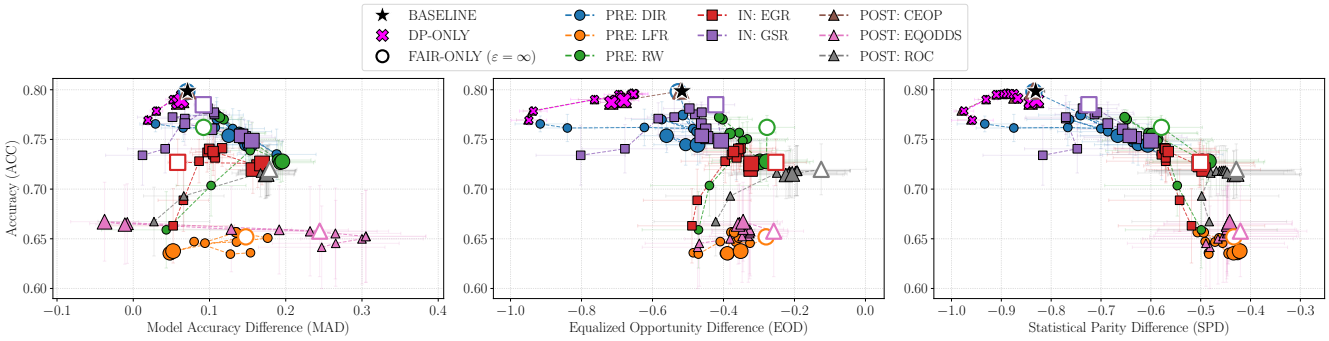


(b) BiasOnDemand Configuration 2.

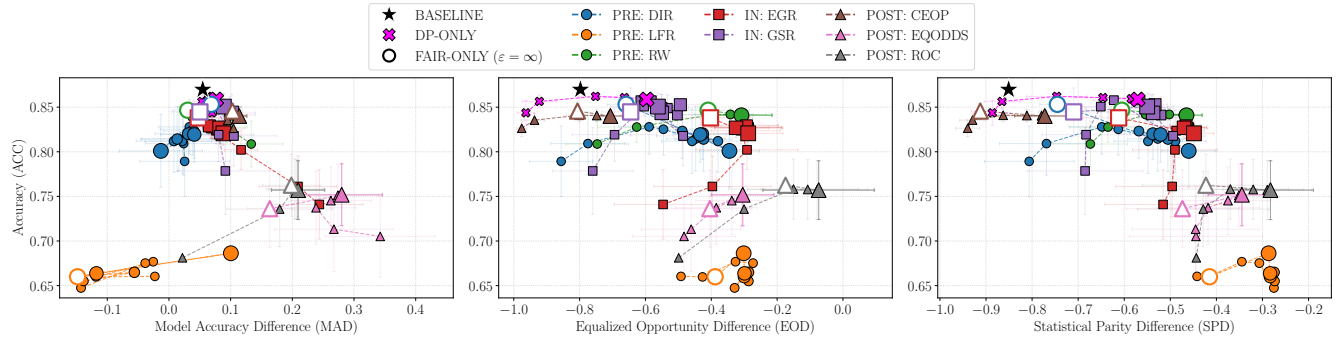


(c) BiasOnDemand Configuration 3.

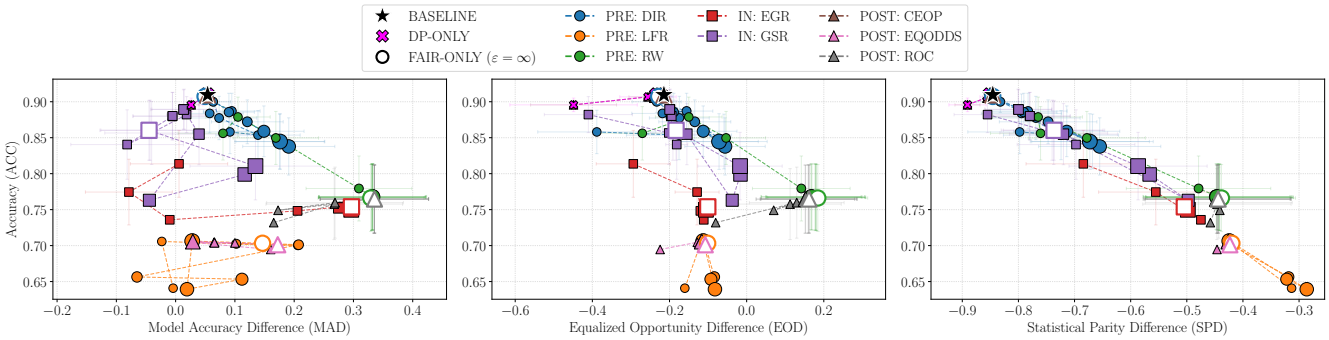
Figure 14: Accuracy–fairness trade-off under the AIM synthesizer across configurations 1, 2, and 3 of the BoD dataset with the Logistic Regression classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



(a) BiasOnDemand Configuration 4.

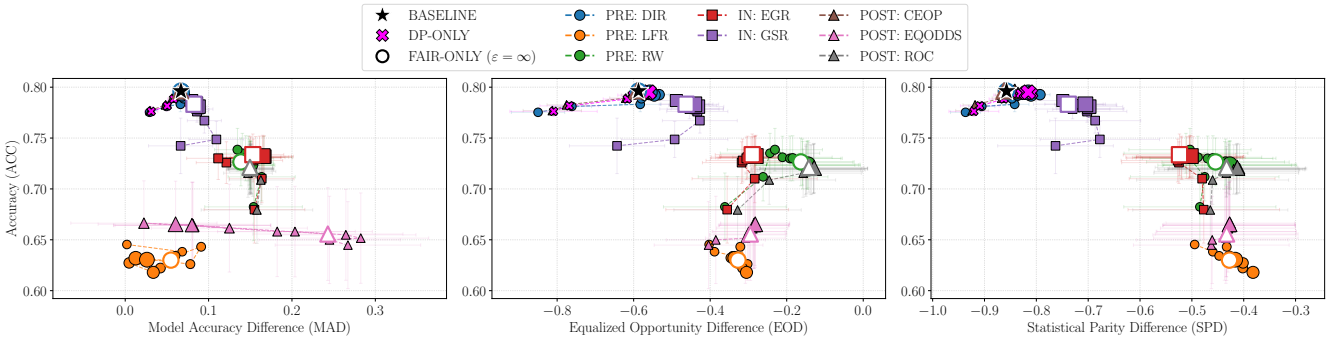


(b) BiasOnDemand (Configuration 5).

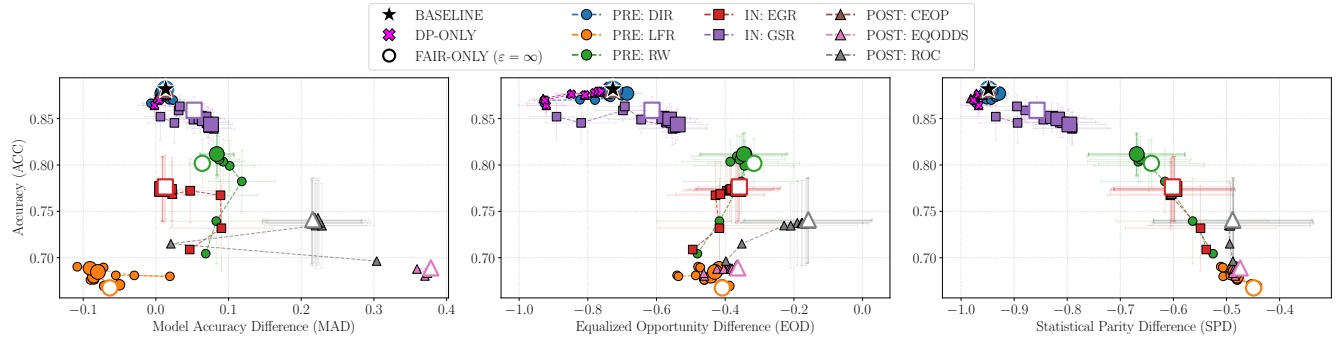


(c) BiasOnDemand Configuration 6.

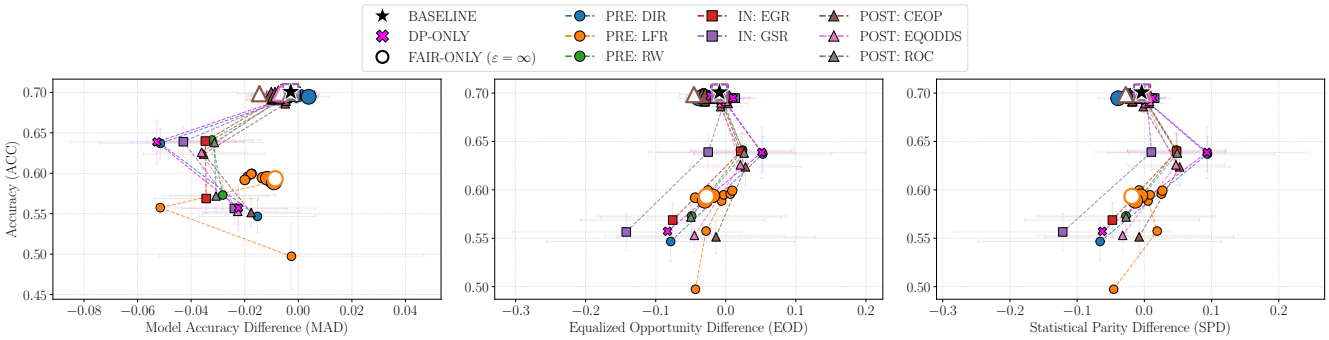
Figure 15: Accuracy–fairness trade-off under the AIM synthesizer across configurations 4, 5, and 6 of BoD dataset with the classifier Logistic Regression. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



(a) BiasOnDemand Configuration 1.

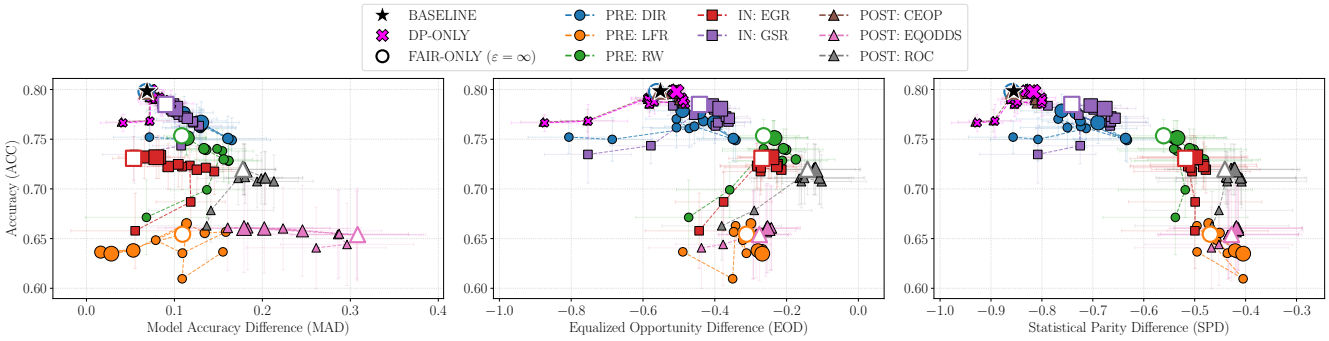


(b) BiasOnDemand Configuration 2.

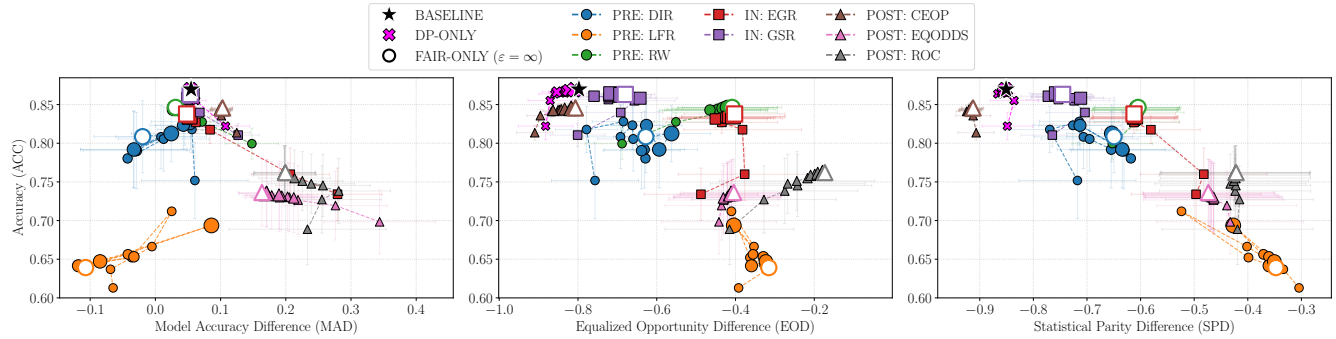


(c) BiasOnDemand Configuration 3.

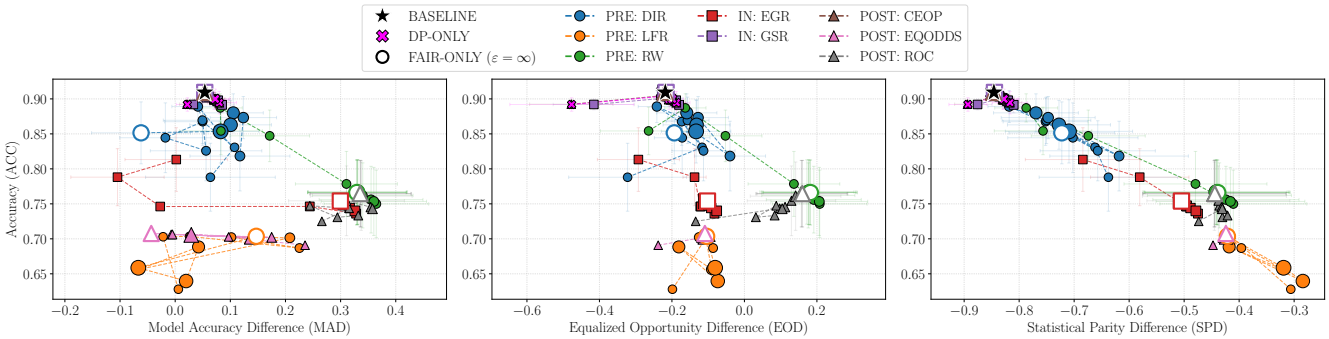
Figure 16: Accuracy–fairness trade-off under the AIM synthesizer across configurations 1, 2, and 3 of the BoD dataset with the Random Forest classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the $\star$ symbol.



(a) BiasOnDemand Configuration 4.



(b) BiasOnDemand (Configuration 5).



(c) BiasOnDemand Configuration 6.

Figure 17: Accuracy–fairness trade-off under the AIM synthesizer across configurations 4, 5, and 6 of the BoD dataset with the Random Forest classifier. Each subfigure reports accuracy (ACC) versus three fairness metrics (MAD, EOD, SPD) for each fairness intervention applied at the pre-, in-, and post-processing stages. Marker size encodes the privacy budget ε for both `DP-only` and `DP+Fair` settings, while hollow markers indicate the `Fair-only` setting ($\varepsilon = \infty$). The `Baseline` setting (no DP and no fairness intervention) is indicated by the ⋆ symbol.