Are Fairness Mitigation Rankings Stable? An Empirical Cross-Dataset, Cross-Metric Comparison of Bias-Mitigation Pipelines on Tabular Classifiers

Shubhra Mittal
Microsoft
shmitt@microsoft.com

Gaurav Naresh Mittal
Independent Researcher
gaurav.n.mittal@gmail.com

***Abstract*—Algorithmic fairness mitigations are now routine in industry Responsible-AI work, but the literature that guides their use is fragmented: different studies recommend different techniques, on different datasets, under different metrics. We run a controlled comparison of six bias-mitigation pipelines — sample reweighing [1], post-hoc threshold optimization [2], and the exponentiated-gradient reductions method [3], with the latter two evaluated under both Demographic Parity and Equalized Odds constraints, plus an unmitigated baseline — on four tabular fairness datasets (UCI Adult [4], [5], ProPublica COMPAS [6], UCI German Credit [5], and ACS Income [7]) and two base classifiers (logistic regression and gradient-boosted trees). We score each pipeline on accuracy, Demographic Parity difference, and Equalized Odds difference, and then ask a question the per-pipeline numbers leave open: how stable is the ranking of mitigations as we change the dataset, the base model, and the fairness metric being optimized? Using pairwise Kendall's τ [8] over 10 random seeds, we find that rankings carry over cleanly between base classifiers (mean $\tau = 0.65$, 95% CI [0.61, 0.69]) but far less so between datasets (0.31, [0.23, 0.39]), and not at all between fairness metrics (0.02, [−0.04, 0.07], an interval that contains zero). The cross-dataset number hides a sharp split: rankings transfer between similarly-biased datasets (Adult ↔ COMPAS, 0.38, [0.29, 0.47]) but are no better than random for any pair involving the already-fair German Credit data (0.08, [−0.07, 0.23]). The choice of metric can do real harm: on Adult, every Demographic-Parity-constrained mitigation raises the Equalized Odds difference, by as much as 3× over the baseline, whereas the Equalized-Odds-constrained ones lower both metrics at once — the impossibility result of Chouldechova [9] and Kleinberg et al. [10] showing up in practice. The takeaway for practitioners is simple: a recommendation tuned on one dataset, base model, or metric should not be carried over to another without re-checking it on the target setting. We release the code and configuration needed to reproduce everything.**

I. Introduction

As machine learning has moved into consequential decisions — credit, hiring, criminal justice, healthcare — algorithmic fairness has grown into both an academic field and an industry practice [11]. A practitioner today has no shortage of tools: a large catalog of fairness mitigation algorithms (interventions somewhere in the modeling pipeline that aim to reduce group disparity), and an equally large catalog of fairness metrics for judging them.

All of this choice has not produced consolidated guidance. Work on mitigation is split across at least three axes: the family of intervention (pre-processing, in-processing, or post-processing); the fairness definition being enforced, usually Demographic Parity or Equalized Odds; and the base classifier and dataset used for benchmarking. As a result, a claim that "mitigation X is best" usually depends on choices a practitioner cannot easily carry over to their own problem.

We ask one narrow question: how stable is the ranking of mitigation algorithms across the things a practitioner does not control? Holding the protocol fixed — one toolkit, default hyperparameters, fixed train/test splits — we measure how much the ordering of mitigations shifts as we change the dataset, the base classifier, and the fairness metric being optimized.

We make three contributions:

1) A **controlled, reproducible comparison** of six mitigation pipelines across four tabular fairness datasets and two base classifiers, using a single industry-standard toolkit, Fairlearn [12]. The set of mitigations is constructed to include one representative from each of the three intervention families, with constraint-supporting methods evaluated under both Demographic Parity and Equalized Odds.
2) A **ranking-stability analysis** quantifying how much the "best" mitigation moves when datasets, base models, and fairness metrics change, summarized through pairwise Kendall's τ correlations across conditions.
3) **Conditional, practitioner-facing guidance** grounded in the empirical results: which mitigation to reach for first under which circumstances.

Our central finding, measured over 10 random seeds, is that base classifier is the only axis along which mitigation rankings carry over cleanly (mean Kendall's $\tau = 0.65$, 95% CI [0.61, 0.69]). Across datasets (0.31, [0.23, 0.39]), and especially across the fairness metric (0.02, [−0.04, 0.07]), the rankings come apart, with both intervals well below the cross-model one. The dataset effect is uneven: rankings move freely between similarly-biased datasets such as Adult and COMPAS ($\tau = 0.38$, [0.29, 0.47]) but are no better than random for any pair that includes the already-fair German Credit data ($\tau = 0.08$, [−0.07, 0.23]). The metric effect can be actively harmful. On Adult, every Demographic-Parity-

constrained mitigation pushes the Equalized Odds difference up relative to the baseline, by as much as 3×, while the Equalized-Odds-constrained ones bring both metrics down together. That is the impossibility result of Chouldechova [9] and Kleinberg et al. [10] showing up in practice, and it bites hardest in workflows where the fairness criterion is settled late, after a mitigation has already been picked.

The remainder of the paper is structured as follows. Section II reviews related work on fairness definitions, mitigation algorithm families, and prior comparative studies. Section III formalizes the methodology and defines the ranking-stability metric. Section IV describes the experimental setup. Section V presents results. Section VI discusses implications, offers an explanation of the observed (in)stability, and gives conditional recommendations. Section VII acknowledges limitations, and Section VIII concludes.

II. Background and Related Work

A. Fairness definitions

Two structurally different definitions of group fairness now dominate the literature. **Demographic Parity** [13], [14] requires that the rate at which a model assigns the positive outcome be equal across protected groups; this is also known as statistical parity or independence. **Equalized Odds** [2] instead requires that the true-positive rate and false-positive rate be equal across groups — i.e., that model errors be distributed symmetrically conditional on the true label. A third family, calibration-based definitions [9], [10], requires that predicted probabilities be equally well-calibrated across groups.

Chouldechova [9] and Kleinberg et al. [10] independently established the foundational impossibility result: except in degenerate cases (equal base rates or perfect classifiers), Demographic Parity, Equalized Odds, and calibration cannot simultaneously hold. This impossibility frames the practitioner question we address: given that one cannot satisfy all fairness criteria at once, which mitigation should be chosen, and how sensitive is that choice to the dataset and the metric being prioritized?

B. Mitigation algorithm families

Mitigation algorithms are conventionally grouped into three families based on the stage of the pipeline at which they intervene.

Pre-processing methods modify the training data. Representative examples include the sample reweighing of Kamiran and Calders [1], the disparate-impact remover of Feldman et al. [15], and learned fair representations [16]. The model itself is left untouched; only the empirical distribution it trains on is altered.

In-processing methods modify the learning objective. Examples include adversarial debiasing [17], constraint-based optimization [18], and the reductions framework of Agarwal et al. [3], which casts fairness-constrained classification as a sequence of cost-sensitive learning problems solved via exponentiated gradient. The output is a randomized classifier that satisfies a chosen fairness constraint up to a specified tolerance.

Post-processing methods modify the trained classifier's outputs. The best known is the threshold optimizer of Hardt et al. [2], which adjusts the per-group decision threshold of an already-trained classifier to satisfy a chosen fairness criterion. The base model is left untouched; only the decision rule changes.

We evaluate one representative from each of the pre- and post-processing families along with the in-processing exponentiated-gradient reduction. For the two methods that admit a choice of fairness constraint (Threshold Optimizer and Exponentiated Gradient), we evaluate both Demographic Parity and Equalized Odds variants. This yields six mitigation configurations (including the baseline) per (dataset, base classifier) pair.

C. Benchmark datasets

We use three datasets that are standard in tabular fairness research and add a fourth, more recent one. **UCI Adult Income** [4], [5] has about 48,000 records; the task is to predict whether annual income exceeds $50,000, with sex as the protected attribute. **ProPublica COMPAS** [6] has about 6,000 records and predicts two-year recidivism, with race as the protected attribute; its limitations are well documented [19], but it remains a common reference point. **UCI German Credit** [5] has 1,000 records on consumer credit risk, with sex (read off the marital-status field) as the protected attribute. **ACS Income** [7] is a larger, more recent income task built from US Census American Community Survey microdata, the "Retiring Adult" successor to UCI Adult, again with sex as the protected attribute; we take the 2018 California one-year sample and subsample it to 45,000 records to match Adult. The four differ widely in size, base rate, and group balance. Three of them — Adult, COMPAS, and ACS Income — start from a substantially biased baseline, while German Credit is already close to parity, and that contrast turns out to matter for our results.

D. Fairness toolkits

Two open-source toolkits are widely used in industry. **Fairlearn** [12], maintained by Microsoft, provides implementations of reweighing, the threshold optimizer, and the reductions approach behind a consistent API on top of scikit-learn [20]. **AIF360** [21], maintained by IBM, provides a broader catalog of metrics and mitigations including learned fair representations and adversarial debiasing. We standardize on Fairlearn because it exposes all of the mitigations under comparison through a single, consistent API on top of arbitrary scikit-learn estimators, and because it is widely adopted in industry Responsible-AI practice. Because the same toolkit is applied identically to every pipeline, this choice does not advantage any one mitigation, and the underlying algorithms are standard and also available in other libraries such as AIF360 (we note a related affiliation disclosure in the Acknowledgments).

E. Prior comparative studies

Several earlier studies have compared fairness mitigations, and each leaves an opening for what we do here. Friedler et al. [22] compared four mitigations on four datasets and found that the choice of algorithm mattered about as much as the choice of dataset; their study, though, predates the reductions-based methods of Agarwal et al. [3] that are now standard in production toolkits. Biswas and Rajan [23] tried seven mitigations on three datasets and found no clear winner, but did not check whether the ranking holds up across metrics. Hort et al. [24] is a recent survey rather than a controlled head-to-head comparison. A separate line of work shows that the benchmarks themselves are shaky: the group-disparity metrics used to rank models are themselves noisy, statistically biased estimators, so a single measured value can misstate which model is fairer [25], and benchmark suites can over-represent particular failure modes [7]. The common thread is that a comparison should measure how stable the ranking is, not just crown a single winner.

F. Positioning of this paper

Our paper sits where these three lines meet: comparative fairness studies, critiques of fairness benchmarks, and practitioner guidance. We do not propose a new algorithm or metric. What we offer is a controlled, reproducible measurement of how brittle the choice of mitigation really is, plus conditional guidance that follows from it. We hold to one industry-standard toolkit and to default hyperparameters on purpose, so that the gaps we report between mitigations match what a practitioner would see when reaching for these methods off the shelf.

III. Methodology

A. Research questions

We organize the study around one preliminary question, two primary research questions that correspond to the two stability axes named in the title, and one supplementary analysis:

- **RQ1 (preliminary).** Does a single "winner" exist at all — that is, does any one mitigation dominate across datasets when accuracy and a fixed fairness metric are jointly considered? A negative answer is what motivates studying the stability of rankings rather than searching for one best method.
- **RQ2 (cross-dataset stability).** How stable is the ranking of mitigations when the dataset changes, with the base classifier and the fairness metric held fixed?
- **RQ3 (cross-metric stability).** How stable is the ranking when the fairness criterion being prioritized changes? We examine this general question through one concrete and widely used case study: a change from Demographic Parity to Equalized Odds.
- **Supplementary analysis (cross-model stability).** How stable is the ranking when only the base classifier changes? This axis is not part of the title question; it serves as a reference against which cross-dataset and cross-metric stability are judged, since a change of base classifier is the change a practitioner would expect to be least disruptive.

TABLE I. Dataset Summary

Dataset	n	Positive class	Protected attribute	Majority group share
UCI Adult [4]	45,222	income > \$50k	sex (Male / Female)	67.5% Male
ProPublica COMPAS [6]	5,278	recidivism within 2y	race (African-American / Caucasian)	60.5% African-American
UCI German Credit [5]	1,000	good credit	sex (via marital status)	69.0% Male
ACS Income [7]	45,000	income > \$50k	sex (Male / Female)	52.8% Male

B. Datasets

Table I summarizes the four datasets used.

We follow the standard preprocessing conventions: for COMPAS we apply the ProPublica filtering criteria [6]; for German Credit we derive sex from the `personal_status` attribute; for Adult we drop records with missing values and binarize the target; for ACS Income we use the folktables `ACSIncome` task on the 2018 California one-year sample, subsample to 45,000 records with a fixed seed, and keep all features as the numeric Census codes folktables provides.

C. Base classifiers

We evaluate two base classifiers, both from scikit-learn [20] with default hyperparameters:

- **Logistic Regression** (`LogisticRegression(max_iter=2000)`). A linear model representative of the simplest baseline a practitioner would consider.
- **Gradient-Boosted Trees** (`GradientBoostingClassifier()`). A non-linear ensemble representative of a production-quality tabular model.

We do not tune hyperparameters, on purpose. What we want to isolate is the effect of the mitigation, and the numbers we report should be the ones a practitioner gets when reaching for an off-the-shelf method.

D. Mitigation algorithms

We evaluate six pipelines:

1) **Baseline.** No mitigation; the base classifier is fit on the unmodified training data.
2) **Reweighing** [1]. Sample weights are computed such that the joint distribution of (group, label) in the training data becomes the product of its marginals; the base classifier is fit with these weights.
3) **Threshold Optimizer — DP** [2]. The base classifier is fit normally; per-group decision thresholds are then learned on the training set to satisfy a Demographic Parity constraint.

4) **Threshold Optimizer — EO.** As above, with the constraint replaced by Equalized Odds.
5) **Exponentiated Gradient — DP** [3]. The base classifier is repeatedly retrained under a sequence of reweighted cost-sensitive problems, yielding a randomized classifier that satisfies a Demographic Parity constraint up to the algorithm's default tolerance.
6) **Exponentiated Gradient — EO.** As above, with the constraint replaced by Equalized Odds.

All six pipelines are implemented using Fairlearn [12] with default arguments unless noted.

E. Fairness metrics

We report two group-fairness metrics. Let $\hat{y} \in \{0, 1\}$ denote the model prediction, y the true label, and $A \in \{0, 1\}$ the protected attribute.

Demographic Parity Difference (DP-Δ). The absolute difference in positive prediction rate between groups:

$$\text{DP-}\Delta = |\Pr(\hat{y} = 1 \mid A = 0) - \Pr(\hat{y} = 1 \mid A = 1)| \,. \quad (1)$$

Equalized Odds Difference (EO-Δ). The larger of the absolute true-positive-rate gap and the absolute false-positive-rate gap:

$$\text{EO-}\Delta = \max\left(|\text{TPR}_0 - \text{TPR}_1|,\ |\text{FPR}_0 - \text{FPR}_1|\right). \quad (2)$$

A value of zero is perfectly fair under the respective definition. Both metrics are computed on the held-out test set using the implementations in Fairlearn [12].

F. Ranking stability

For each (dataset, base classifier, fairness metric) condition we compute the ranking of the six mitigation pipelines from fairest (lowest metric value) to least fair. We summarize the stability of these rankings across conditions using pairwise Kendall's τ [8], a non-parametric measure of rank correlation that takes the value $+1$ when two rankings agree perfectly, 0 when they are uncorrelated, and -1 when they are reversed. We report:

- **Cross-dataset τ (RQ2):** mean Kendall's τ between rankings that share the same base classifier and metric but differ in dataset.
- **Cross-metric τ (RQ3):** mean Kendall's τ between rankings that share dataset and base classifier but differ in fairness metric.
- **Cross-model τ (supplementary):** mean Kendall's τ between rankings that share dataset and metric but differ in base classifier.

If cross-model τ comes out high while cross-dataset and cross-metric τ stay low, as we find it does, then a practitioner can carry a mitigation ranking from one base classifier to another but not from one dataset, or one fairness metric, to the next. We report each τ as a 10-seed mean with a 95% confidence interval (Section V-C), so the comparison across axes rests on intervals rather than single-run point values.

IV. Experimental Setup

All experiments use a fixed reference seed (SEED = 42), repeated over nine further seeds (0–8) for the multi-seed statistics, a 70/30 stratified train/test split, and a fixed preprocessing pipeline consisting of standard scaling of numeric features and one-hot encoding of categorical features. The protected attribute is left in the feature matrix, so each base classifier can use group membership directly; we keep it because that is the default Fairlearn usage and because every pipeline sees the same base-model inputs, so the choice cannot favor one mitigation over another. The mitigations additionally receive the protected attribute through Fairlearn's dedicated `sensitive_features` argument. Adult is loaded via `fetch_openml("adult",version=2)`; German Credit via `fetch_openml("credit-g",version=1)`; COMPAS is downloaded from the ProPublica GitHub repository and filtered following the original analysis [6]; ACS Income is loaded with the folktables package (2018 California one-year ACS) and subsampled to 45,000 records with a fixed seed.

Library versions: Python 3.11, scikit-learn [20], Fairlearn $\geq$ 0.13 [12], folktables, pandas, numpy, and scipy; exact pinned versions are listed in `requirements-lock.txt`. All experiments run on a single laptop CPU; no GPU is required. Wall-clock time for the full 48-pipeline sweep is about ten minutes of compute (plus dataset download), dominated by the gradient-boosted Exponentiated-Gradient pipelines.

The complete experimental code is released alongside this paper as `01_experiment.py`, with the post-hoc Kendall's τ analysis in `02_analysis.py`, the seed-42 columns of the results tables (Tables II–III) regenerated from `results/results.csv` by `03_make_tables.py`, the cross-metric ranking-inversion figure produced by `04_inversion_plot.py`, and the 10-seed analysis behind Table IV and the 10-seed columns of Tables II–III in `05_multiseed.py`. Every table and figure in this paper is generated directly from the released `results.csv` (and `results/multiseed_summary.csv` and `results/multiseed_perpipeline_summary.csv` for the multi-seed statistics), so the reported numbers cannot drift out of sync with the data. The code, pinned configuration, and generated `results/` are available at https://github.com/shubmittal/fairness-mitigation-comparison.

V. Results

Fig. 1 gives the overall accuracy–fairness landscape: for every dataset and metric it plots each (mitigation $\times$ base model) pipeline together with the Pareto frontier. The subsections below quantify the same data — Section V-A under a Demographic Parity constraint, Section V-B under Equalized Odds, and Section V-C the ranking-stability analysis that is the paper's focus.

A. Demographic Parity outcomes (Demographic-Parity-constrained mitigations)

Table II reports accuracy, DP-Δ, and EO-Δ for the four mitigations evaluated under a Demographic Parity constraint.

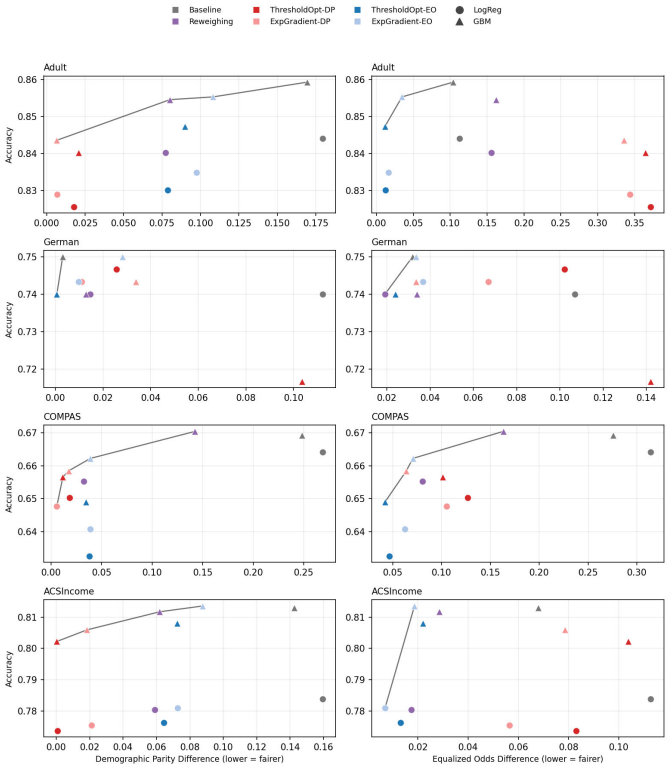


Fig. 1. Accuracy–fairness trade-off with Pareto frontier, by dataset (rows) and fairness metric (columns). Each point is one (mitigation × base model) pipeline — circles are Logistic Regression, triangles Gradient-Boosted Trees. Lower Δ is fairer and higher accuracy is better, so the upper-left envelope (drawn) is Pareto-optimal. Demographic-Parity-constrained pipelines (red) reach the fairest DP-Δ but the worst EO-Δ; Equalized-Odds-constrained pipelines (blue) are the mirror image. Generated by `06_pareto.py`; see `results/plots/pareto_frontier.png`.

The unmitigated baseline serves as the reference. Tables II and III headline the 10-seed mean ± standard deviation of each metric; the reference-seed (42) values are retained alongside because Figs. 1 and 2 plot them. Numbers quoted in this subsection and the next are seed-42 values unless marked as 10-seed means, and we note where the means change the picture.

Three things stand out. **First, ExpGradient-DP and Threshold Optimizer are the two strongest DP-Δ reducers**: at seed 42, ExpGradient-DP has the lowest DP-Δ in four of the eight dataset × model cells and Threshold Optimizer in three (COMPAS/GBM and both ACS Income cells). The eighth is German Credit/GBM, where the untouched baseline is already the fairest model — we come back to this below. Over 10 seeds the same two methods remain the strongest DP-Δ reducers, but the lead shifts toward the Threshold Optimizer, which has the lowest mean DP-Δ in seven of the eight cells (ExpGradient-DP in German Credit/GBM). Reweighing closes less of the DP-Δ gap on the three biased datasets. Accuracy barely moves in most cells, usually by under two percentage points; the one real exception is Threshold Optimizer on German Credit/GBM, which drops 3.3 points at seed 42. These costs line up with earlier reports [22], [23].

Second, and more telling: on Adult, every DP-constrained mitigation makes EO-Δ worse than the baseline — from 0.11 up to 0.16 (Reweighing), 0.37 (ThresholdOpt-DP), and 0.34 (ExpGradient-DP) under logistic regression, and much the same under GBM. The effect is not a seed-42 artifact: the 10-seed means rise from 0.111 (baseline) to 0.144, 0.328, and 0.300 under logistic regression, and from 0.098 to 0.147, 0.332, and 0.299 under GBM. So on Adult the fairest model under Equalized Odds is the one with no mitigation at all. A workflow that applies a DP-constrained mitigation and never re-checks EO-Δ can end up less fair on the criterion the stakeholder actually cares about.

Third, when the baseline is already close to fair, a mitigation can introduce disparity rather than remove it. At seed 42 the GBM baseline on German Credit sits at DP-Δ $= 0.003$, essentially zero, yet Threshold Optimizer pushes it to 0.104 and costs 3.3 points of accuracy. This is not a flaw in the method: a mitigation is built to enforce a constraint, not to first check whether the constraint is even binding. This particular instance depends on the seed: over 10 seeds the GBM baseline's DP-Δ averages 0.053 ± 0.048, and the German pipelines' mean DP-Δ values differ by less than one standard deviation. What persists is the harm on the other metric: ThresholdOpt-DP raises the mean EO-Δ from 0.086 to 0.144. We return to this seed sensitivity in Section VI-C.

B. Equalized Odds outcomes (Equalized-Odds-constrained mitigations)

If DP-constrained mitigations often make EO-Δ worse, the natural question is whether EO-constrained ones close the EO-Δ gap, and what they cost in accuracy and DP-Δ. Table III runs Threshold Optimizer and Exponentiated Gradient under an Equalized Odds constraint, with the baseline repeated for reference.

Two things follow. The EO-constrained variants do close the EO-Δ gap on the three biased datasets: on Adult, COMPAS, and ACS Income, both ThresholdOpt-EO and ExpGradient-EO bring EO-Δ under 0.10 in all six configurations, and under 0.04 in four of them (both Adult and both ACS Income). More usefully, they also lower DP-Δ relative to the baseline in every one of those configurations. The same holds for the 10-seed means, where EO-Δ stays below 0.07 in every one of these configurations and both metrics improve on the baseline in all of them. So an EO constraint improves both metrics at once, where a DP constraint (Table II) buys a lower DP-Δ by driving EO-Δ up, by as much as 3× on Adult. The practical reading is simple: if you are not yet sure which fairness criterion will matter downstream, Equalized Odds is the safer constraint to enforce.

Why this does not contradict the impossibility theorem. The asymmetry can look like it conflicts with Chouldechova [9] and Kleinberg et al. [10], so it is worth saying why it does not. The theorem rules out satisfying Demographic Parity, Equalized Odds, and calibration exactly and at once when base rates differ; it says nothing against a single mitigation approximately improving more than one metric over a baseline that already

TABLE II. PER-PIPELINE RESULTS UNDER DEMOGRAPHIC-PARITY-CONSTRAINED MITIGATIONS: 10-SEED MEAN ± SD AND REFERENCE-SEED (42) VALUES. LOWER DP-Δ AND EO-Δ ARE FAIRER; BOLD MARKS THE BEST 10-SEED MEAN PER DATASET × BASE CLASSIFIER COLUMN.

			10-seed mean ± SD			Seed 42		
Dataset	**Base model**	**Mitigation**	**Accuracy**	**DP-Δ**	**EO-Δ**	**Accuracy**	**DP-Δ**	**EO-Δ**
Adult	LogReg	Baseline	**0.848** ± 0.003	0.186 ± 0.006	**0.111** ± 0.020	0.844	0.179	0.113
Adult	LogReg	Reweighing	0.841 ± 0.003	0.087 ± 0.007	0.144 ± 0.021	0.840	0.077	0.156
Adult	LogReg	ThresholdOpt-DP	0.827 ± 0.002	**0.007** ± 0.005	0.328 ± 0.025	0.826	0.018	0.372
Adult	LogReg	ExpGradient-DP	0.830 ± 0.003	0.012 ± 0.007	0.300 ± 0.028	0.829	0.007	0.344
Adult	GBM	Baseline	**0.863** ± 0.002	0.175 ± 0.008	**0.098** ± 0.027	0.859	0.169	0.104
Adult	GBM	Reweighing	0.856 ± 0.003	0.087 ± 0.008	0.147 ± 0.019	0.855	0.080	0.162
Adult	GBM	ThresholdOpt-DP	0.842 ± 0.003	**0.008** ± 0.007	0.332 ± 0.022	0.840	0.021	0.365
Adult	GBM	ExpGradient-DP	0.845 ± 0.002	0.012 ± 0.007	0.299 ± 0.025	0.844	0.007	0.336
German	LogReg	Baseline	0.750 ± 0.018	0.112 ± 0.055	0.147 ± 0.066	0.740	0.112	0.107
German	LogReg	Reweighing	**0.752** ± 0.017	0.060 ± 0.036	0.095 ± 0.054	0.740	0.015	0.019
German	LogReg	ThresholdOpt-DP	0.748 ± 0.016	**0.044** ± 0.035	0.113 ± 0.070	0.747	0.026	0.102
German	LogReg	ExpGradient-DP	0.746 ± 0.020	0.055 ± 0.037	**0.082** ± 0.067	0.743	0.011	0.067
German	GBM	Baseline	0.753 ± 0.015	0.053 ± 0.048	0.086 ± 0.034	0.750	0.003	0.032
German	GBM	Reweighing	**0.756** ± 0.015	0.047 ± 0.039	**0.085** ± 0.037	0.740	0.013	0.034
German	GBM	ThresholdOpt-DP	0.748 ± 0.022	0.062 ± 0.052	0.144 ± 0.060	0.717	0.104	0.142
German	GBM	ExpGradient-DP	0.751 ± 0.016	**0.044** ± 0.026	0.087 ± 0.050	0.743	0.034	0.033
COMPAS	LogReg	Baseline	**0.672** ± 0.011	0.280 ± 0.029	0.307 ± 0.041	0.664	0.268	0.314
COMPAS	LogReg	Reweighing	0.659 ± 0.009	0.027 ± 0.019	**0.081** ± 0.031	0.655	0.032	0.081
COMPAS	LogReg	ThresholdOpt-DP	0.660 ± 0.010	**0.022** ± 0.018	0.099 ± 0.031	0.650	0.018	0.127
COMPAS	LogReg	ExpGradient-DP	0.660 ± 0.013	0.028 ± 0.017	0.083 ± 0.024	0.648	0.006	0.105
COMPAS	GBM	Baseline	**0.681** ± 0.012	0.250 ± 0.020	0.266 ± 0.036	0.669	0.248	0.276
COMPAS	GBM	Reweighing	0.672 ± 0.015	0.109 ± 0.034	0.111 ± 0.045	0.670	0.142	0.163
COMPAS	GBM	ThresholdOpt-DP	0.661 ± 0.011	**0.026** ± 0.013	0.090 ± 0.023	0.657	0.012	0.101
COMPAS	GBM	ExpGradient-DP	0.664 ± 0.011	0.027 ± 0.020	**0.075** ± 0.030	0.658	0.017	0.064
ACSIncome	LogReg	Baseline	**0.785** ± 0.003	0.160 ± 0.005	0.125 ± 0.012	0.784	0.159	0.113
ACSIncome	LogReg	Reweighing	0.782 ± 0.003	0.060 ± 0.004	**0.011** ± 0.005	0.780	0.059	0.017
ACSIncome	LogReg	ThresholdOpt-DP	0.777 ± 0.003	**0.004** ± 0.002	0.074 ± 0.008	0.774	0.001	0.083
ACSIncome	LogReg	ExpGradient-DP	0.779 ± 0.003	0.018 ± 0.004	0.053 ± 0.007	0.775	0.021	0.056
ACSIncome	GBM	Baseline	**0.814** ± 0.003	0.140 ± 0.005	0.085 ± 0.011	0.813	0.143	0.068
ACSIncome	GBM	Reweighing	0.810 ± 0.004	0.059 ± 0.007	**0.021** ± 0.009	0.812	0.062	0.029
ACSIncome	GBM	ThresholdOpt-DP	0.805 ± 0.003	**0.004** ± 0.005	0.085 ± 0.011	0.802	0.000	0.104
ACSIncome	GBM	ExpGradient-DP	0.807 ± 0.004	0.019 ± 0.007	0.067 ± 0.012	0.806	0.018	0.079

fails all three. The mechanism is intuitive. Forcing equal positive-prediction rates (DP) makes the classifier discard label information that happens to track group membership, and when base rates differ that loss shows up directly in the group-conditional error rates EO-Δ measures. Forcing equal group-conditional error rates (EO) lets the classifier keep that structure, so it does not have to distort the overall positive-prediction rate as much. The result, visible in Tables II and III, is that an EO constraint tends to leave the unconstrained metric alone, while a DP constraint can damage it.

Fig. 2 shows this cross-metric ranking inversion on Adult, where it is sharpest.

German Credit is again the exception. At seed 42 its GBM baseline already sits at DP-Δ = 0.003 and EO-Δ = 0.032; only ThresholdOpt-EO beats it on both (DP-Δ = 0.001, EO-Δ = 0.024), and only barely, while several mitigations — ThresholdOpt-DP especially — make things clearly worse. Over 10 seeds every German GBM pipeline except ThresholdOpt-DP lies within one standard deviation of the baseline on both metrics. The same lesson from Section V-A applies: measure the baseline before reaching for a mitigation.

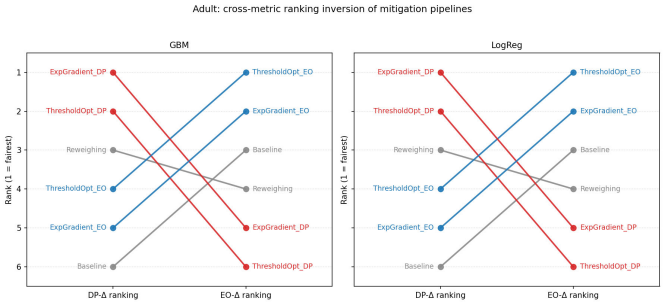


Fig. 2. Cross-metric ranking inversion on the Adult dataset. Each line traces a single mitigation pipeline's rank under Demographic Parity Difference (left axis) versus Equalized Odds Difference (right axis); rank 1 is the fairest. Red lines mark Demographic-Parity-constrained mitigations, blue lines mark Equalized-Odds-constrained mitigations, gray lines mark the unmitigated baseline and sample reweighing. The two DP-constrained pipelines occupy the fairest positions under DP-Δ and the least-fair positions under EO-Δ; the two EO-constrained pipelines are mirror-imaged. Generated by `04_inversion_plot.py`; see `results/plots/Adult_ranking_inversion.png`.

TABLE III. PER-PIPELINE RESULTS UNDER EQUALIZED-ODDS-CONSTRAINED MITIGATIONS: 10-SEED MEAN ± SD AND REFERENCE-SEED (42) VALUES. LOWER DP-Δ AND EO-Δ ARE FAIRER; BOLD MARKS THE BEST 10-SEED MEAN PER DATASET × BASE CLASSIFIER COLUMN. BASELINE ROWS ARE DUPLICATED FROM TABLE II FOR DIRECT COMPARISON.

			10-seed mean ± SD			Seed 42		
Dataset	**Base model**	**Mitigation**	**Accuracy**	**DP-Δ**	**EO-Δ**	**Accuracy**	**DP-Δ**	**EO-Δ**
Adult	LogReg	Baseline	**0.848** ± 0.003	0.186 ± 0.006	0.111 ± 0.020	0.844	0.179	0.113
Adult	LogReg	ThresholdOpt-EO	0.831 ± 0.002	**0.089** ± 0.008	0.027 ± 0.015	0.830	0.079	0.013
Adult	LogReg	ExpGradient-EO	0.836 ± 0.002	0.105 ± 0.007	**0.025** ± 0.016	0.835	0.097	0.016
Adult	GBM	Baseline	**0.863** ± 0.002	0.175 ± 0.008	0.098 ± 0.027	0.859	0.169	0.104
Adult	GBM	ThresholdOpt-EO	0.850 ± 0.002	**0.101** ± 0.008	**0.026** ± 0.022	0.847	0.090	0.012
Adult	GBM	ExpGradient-EO	0.856 ± 0.002	0.117 ± 0.006	0.029 ± 0.018	0.855	0.108	0.035
German	LogReg	Baseline	0.750 ± 0.018	0.112 ± 0.055	0.147 ± 0.066	0.740	0.112	0.107
German	LogReg	ThresholdOpt-EO	0.748 ± 0.017	0.053 ± 0.035	**0.110** ± 0.066	0.743	0.010	0.037
German	LogReg	ExpGradient-EO	**0.751** ± 0.020	**0.052** ± 0.033	0.113 ± 0.058	0.743	0.010	0.037
German	GBM	Baseline	0.753 ± 0.015	0.053 ± 0.048	0.086 ± 0.034	0.750	0.003	0.032
German	GBM	ThresholdOpt-EO	0.747 ± 0.020	**0.044** ± 0.035	**0.084** ± 0.025	0.740	0.001	0.024
German	GBM	ExpGradient-EO	**0.754** ± 0.014	0.050 ± 0.041	0.098 ± 0.041	0.750	0.028	0.033
COMPAS	LogReg	Baseline	**0.672** ± 0.011	0.280 ± 0.029	0.307 ± 0.041	0.664	0.268	0.314
COMPAS	LogReg	ThresholdOpt-EO	0.637 ± 0.008	**0.035** ± 0.018	**0.050** ± 0.030	0.633	0.038	0.047
COMPAS	LogReg	ExpGradient-EO	0.643 ± 0.009	**0.035** ± 0.018	0.055 ± 0.032	0.641	0.039	0.063
COMPAS	GBM	Baseline	**0.681** ± 0.012	0.250 ± 0.020	0.266 ± 0.036	0.669	0.248	0.276
COMPAS	GBM	ThresholdOpt-EO	0.655 ± 0.010	**0.039** ± 0.024	**0.050** ± 0.022	0.649	0.034	0.042
COMPAS	GBM	ExpGradient-EO	0.667 ± 0.010	0.041 ± 0.029	0.065 ± 0.031	0.662	0.039	0.071
ACSIncome	LogReg	Baseline	**0.785** ± 0.003	0.160 ± 0.005	0.125 ± 0.012	0.784	0.159	0.113
ACSIncome	LogReg	ThresholdOpt-EO	0.779 ± 0.003	**0.064** ± 0.005	**0.011** ± 0.006	0.776	0.065	0.013
ACSIncome	LogReg	ExpGradient-EO	0.783 ± 0.003	0.073 ± 0.007	0.014 ± 0.012	0.781	0.073	0.007
ACSIncome	GBM	Baseline	**0.814** ± 0.003	0.140 ± 0.005	0.085 ± 0.011	0.813	0.143	0.068
ACSIncome	GBM	ThresholdOpt-EO	0.811 ± 0.003	**0.074** ± 0.007	**0.012** ± 0.008	0.808	0.072	0.022
ACSIncome	GBM	ExpGradient-EO	0.813 ± 0.004	0.085 ± 0.005	0.018 ± 0.009	0.814	0.088	0.019

C. Cross-condition ranking stability

We now turn to RQ1–RQ3 and the supplementary cross-model analysis. For each condition (dataset, base classifier, metric) we compute the ranking of the six mitigation pipelines from fairest (lowest metric value) to least fair, and summarize the pairwise Kendall's τ statistics in Table IV. To separate signal from seed noise, we report each statistic both for the reference seed (42, used throughout Sections V-A–V-B and Figs. 1–2) and as a mean with 95% confidence interval over 10 seeds (`05_multiseed.py`); the seed jointly controls the train/test split, the base models, and the randomized predictions.

RQ1 (preliminary): no single winner. Tables II and III already answer the preliminary question in the negative. In every dataset × base classifier cell of the three biased datasets, both at seed 42 and in the 10-seed means, the pipeline with the lowest DP-Δ is a DP-constrained method while the pipeline with the lowest EO-Δ is an EO-constrained method or Reweighing; and on German Credit no pipeline separates clearly from the baseline once seed variation is taken into account. Since no mitigation dominates, the rest of this section measures how the ranking moves rather than which method tops it.

Three observations follow.

First (supplementary analysis), base classifier is the only axis along which the rankings hold up, and they hold up across seeds: cross-model τ averages 0.65, with a 95% CI of [0.61, 0.69] that sits entirely above the other two axes. This is the one stability result a practitioner can lean on — validate a mitigation under one base classifier and the relative ranking will mostly carry over to another.

Second (RQ2), dataset transfers only modestly on average (cross-dataset $\tau = 0.31$, CI [0.23, 0.39]), and that average hides a split set by the baseline-bias regime. The three biased datasets — Adult, COMPAS, ACS Income — transfer well to one another (biased-cluster mean $\tau = 0.54$, CI [0.49, 0.58]), whereas transfer to the already-fair German Credit data collapses to $\tau = 0.08$, CI $[-0.07, 0.23]$, an interval that contains zero and is therefore no better than random; that near-zero is what pulls the overall average down. A biased baseline, in other words, is a precondition for cross-dataset transfer. There is structure even within the biased cluster: the weakest pair is Adult ↔ COMPAS ($\tau = 0.38$, CI [0.29, 0.47]), which changes both the domain (income vs. recidivism) and the protected attribute (sex vs. race), and so transfers less well than pairs that share them. And even the biased-cluster value (0.54) stays below cross-model (0.65); the intervals do not overlap, so seed 42's flattering picture — cross-dataset 0.39, Adult ↔ COMPAS 0.60 — does not hold up once the seed is varied.

Third (RQ3), the fairness metric is the least stable axis of all: cross-metric $\tau = 0.02$, CI $[-0.04, 0.07]$. With four datasets this interval falls entirely below the cross-dataset one and contains zero, which means that across datasets, how a mitigation ranks under Demographic Parity tells you essentially nothing about how it ranks under Equalized Odds.

TABLE IV. RANKING-STABILITY STATISTICS. MEAN PAIRWISE KENDALL'S τ PER COMPARISON AXIS, FOR THE REFERENCE SEED (42) AND AGGREGATED OVER 10 SEEDS (MEAN AND 95% CI). HIGHER τ = MORE STABLE RANKINGS; $\tau = 1$ IDENTICAL, 0 UNCORRELATED, < 0 PARTIALLY REVERSED. EACH AXIS AGGREGATES THE CONDITION PAIRS SHARING ALL BUT ONE FACTOR (PAIRS PER SEED: CROSS-MODEL 8, CROSS-DATASET 24, BIASED-CLUSTER 12, ADULT↔COMPAS 4, GERMAN-INVOLVING 12, CROSS-METRIC 8).

Comparison axis	Seed 42	10-seed mean	95% CI
Cross-base-model (same dataset, same metric)	0.65	**0.65**	[0.61, 0.69]
Cross-dataset (same base model, same metric)	0.39	**0.31**	[0.23, 0.39]
— biased-cluster pairs (Adult / COMPAS / ACSIncome)	0.67	0.54	[0.49, 0.58]
— Adult ↔ COMPAS pairs	0.60	0.38	[0.29, 0.47]
— pairs involving German Credit	0.12	0.08	[−0.07, 0.23]
Cross-metric (same dataset, same base model)	−0.02	**0.02**	[−0.04, 0.07]

The aggregate again hides sharper local behavior: on Adult the cross-metric τ is strongly negative at seed 42 (-0.47 for both base models), so the best mitigation under DP-Δ is among the worst under EO-Δ. That is the Chouldechova–Kleinberg impossibility at its most visible. The lesson is blunt: a mitigation tuned for one fairness metric can rank strictly worse under another, by a margin as large as the disparity it was meant to fix.

The 10-seed intervals are also a warning against over-reading any single run. Seed 42 turns out to be a favorable draw for cross-dataset transfer: its cross-dataset (0.39) and Adult ↔ COMPAS (0.60) values both sit at or above the top of the respective 95% CIs. What survives the seed variation is narrower than that one run suggests. Cross-model transfer is high and cleanly separated from the rest. Cross-dataset transfer is modest and set by the baseline-bias regime. Cross-metric transfer cannot be told apart from zero, and transfer to an already-fair dataset is no better than random. There is a certain irony in a paper about ranking stability finding its own headline numbers to be seed-sensitive, but that is exactly why single-seed numbers — ours included — should be read with care.

Taken together, these three points are the paper's main claim: **advice of the form "use mitigation X" has to be qualified by the dataset and the fairness metric in play; only the base classifier can be safely abstracted away.**

D. Cost analysis

Training time and implementation complexity vary a lot across mitigations. We record per-pipeline training time (in `results/timings.csv`, kept apart from the deterministic `results.csv`) and summarize relative cost and deployment profile in Table V, along with two constraints that often decide the matter in practice: whether the mitigation needs the protected attribute at inference time, and whether it returns a randomized rather than a deterministic classifier. The gradient-boosted Exponentiated-Gradient pipelines dominate the sweep; the two on Adult take roughly 85–150 s each, against about 5 s for the gradient-boosted baseline.

The Threshold Optimizer's requirement for the protected attribute at inference time is a frequent disqualifier in regulated industries: a U.S. consumer credit lender, for example, generally cannot legally collect protected attributes from a loan applicant in order to look up a group-specific threshold. The Exponentiated Gradient's randomized output is a frequent obstacle to deployment in domains that require deterministic, auditable decisions. These deployment constraints are independent of the raw fairness numbers and are decisive for many practitioners.

TABLE V. COST AND DEPLOYMENT PROFILE OF EACH MITIGATION

Mitigation	Training cost	Needs sensitive attr. at inference	Output type
Baseline	1×	No	Deterministic
Reweighing	1×	No	Deterministic
Threshold Optimizer	1× + threshold search	Yes	Deterministic per group
Exponentiated Gradient	∼20–30× (refits the base learner once per iteration, `max_iter=50`)	No	Randomized

VI. DISCUSSION

A. Practical implications

The findings above turn into four pieces of practical advice.

(C1) Measure the baseline before mitigating. The German Credit/GBM case shows that mitigating an already-fair model can introduce disparity. The effect is sharpest at the reference seed; across seeds the gain from mitigating German Credit is indistinguishable from zero, while ThresholdOpt-DP still raises EO-Δ. That is obvious in hindsight: a mitigation enforces a constraint; it does not check whether the constraint is binding. Yet the literature rarely says it out loud. Measure first, and mitigate only if the measurement calls for it.

(C2) The fairness metric matters more than the mitigation. Cross-metric stability is the lowest of the three axes, well below cross-model stability — its confidence interval does not come near the cross-model one. An organization that picks a mitigation before settling on a fairness criterion can easily end up optimizing the wrong one. Pin down the criterion with downstream stakeholders, in writing, before choosing a mitigation.

(C3) Reweighing is the underrated middle option. It rarely tops either DP-Δ or EO-Δ, but it almost never comes

last, and it is the only mitigation that never blows up one metric while improving the other. It is also the easiest to ship: no protected attribute needed at inference, no randomized output, little training overhead. When an organization cannot commit to a single fairness criterion up front, Reweighing is the safest default.

(C4) Reach for post-processing first when deployment allows it. Threshold Optimizer matches Exponentiated Gradient on whichever metric it targets, and trains far faster. Where the protected attribute is available at inference time, it is the pragmatic first choice.

B. Connection to the impossibility theorem

The anti-correlation between DP-Δ and EO-Δ rankings on Adult is the impossibility result of Chouldechova [9] and Kleinberg et al. [10] in miniature. Adult's base rates differ by group ($\Pr(y{=}1 \mid \text{Male}) \approx 0.31$ vs. $\Pr(y{=}1 \mid \text{Female}) \approx 0.11$), so enforcing DP makes the model over-approve the lower-base-rate group and under-approve the higher one, which widens the TPR/FPR gap by construction. The hazard is that a practitioner can trigger this by accident, by reaching for a DP-constrained mitigation when the stakeholder actually cares about EO.

C. Why rankings are stable on one axis and not on the others

The three stability levels can be explained by what each change does to the quantities that determine a pipeline's rank. We offer the following account as an interpretation of the measurements; it is consistent with Tables II–IV but is not separately tested.

Base classifier. All three mitigations act on group-level quantities that are defined independently of the hypothesis class: Reweighing fixes one weight per (group, label) cell, the Threshold Optimizer chooses one threshold per group, and Exponentiated Gradient adjusts multipliers attached to group-level constraints. Replacing logistic regression with gradient-boosted trees changes how well the base model separates the classes, but not which constraint each pipeline enforces. Each pipeline therefore lands in roughly the same place relative to the others, which is what the high cross-model τ records.

Fairness metric. Each constrained pipeline drives its own target metric toward zero by construction, so its rank under that metric is decided by the constraint. Its rank under the other metric is decided by a different mechanism: when base rates differ across groups, reducing one disparity generally moves the other (Section VI-B). The two rankings are thus produced by different forces and need not agree; on Adult, where the base-rate gap is large, they run in opposite directions.

Dataset. A ranking carries information only when the gaps between pipelines exceed the noise with which each gap is estimated. On the biased datasets the gaps are large: on Adult, for example, mean DP-Δ falls from about 0.18 for the baseline to about 0.01 for the DP-constrained pipelines, with standard deviations below 0.01. On German Credit the gaps are small relative to the noise: for GBM, the six pipelines' mean DP-Δ values lie between 0.044 and 0.062, while their standard deviations range from 0.026 to 0.052. The ordering of German Credit pipelines is then largely an ordering of sampling noise, and its correlation with any other ranking is near zero, which is the German-involving τ in Table IV. This agrees with the observation that group-disparity estimates are noisy [25], and it also explains why Kendall's τ, which counts every pairwise swap equally, is sensitive to near-ties.

D. Making mitigation choices robust to ranking instability

The explanation above suggests safeguards that address instability directly rather than only reporting it.

(S1) Rank with uncertainty. Estimate each pipeline's fairness metrics over several seeds or resamples and report intervals; a single-run ranking can mislead, as our seed-42 cross-dataset value shows.

(S2) Treat statistical ties as ties. When two pipelines differ by less than their seed-to-seed variability, regard them as equivalent rather than ordered, so that noise is not read as a reversal.

(S3) Select on the Pareto front, not on a single-metric rank. Choosing among pipelines on the accuracy–DP-Δ–EO-Δ frontier (Fig. 1) guards against the cross-metric inversion, because a pipeline that is optimal for one metric but dominated on another is visible as such.

(S4) Re-validate on the target data. Because cross-dataset transfer depends on the baseline-bias regime, a ranking learned on a strongly biased benchmark should not be applied to a near-parity setting without re-measurement.

We have not evaluated these safeguards quantitatively; doing so, for example by measuring how much tie-aware ranking raises the cross-dataset τ, is a natural follow-up.

E. When to favor which mitigation: a decision sketch

Putting the empirical and deployment findings together gives a rough decision rule:

1) If the baseline already satisfies the target fairness threshold, **do not mitigate**.
2) If the deployment context permits the protected attribute at inference time and a deterministic decision is acceptable, prefer **Threshold Optimizer** under the constraint matching the downstream stakeholder's fairness criterion.
3) If the protected attribute is unavailable at inference time, or a deterministic decision is required, prefer **Reweighing** as a robust default.
4) If extra training compute is cheap and a randomized classifier is acceptable, **Exponentiated Gradient** offers the smallest measured value of the constrained metric, at the cost of inflating the unconstrained metric.

None of this replaces measuring on your own data, but it is a sensible starting point that weighs both the results and the deployment constraints.

VII. Limitations

Our study has the following limitations.

Tabular binary classification only. Our results are confined to tabular binary classification on four datasets. Whether the cross-metric instability we observe generalizes to multi-class classification, or to NLP and computer vision settings, is unknown; multi-class problems in particular require multi-group generalizations of DP-Δ and EO-Δ, which we do not study.

Single sensitive attribute per dataset. We do not look at intersectional fairness across several sensitive attributes at once, which is a hard problem in its own right.

Protected attribute used as a model feature. The base classifiers are allowed to see the protected attribute directly (we do not enforce "fairness through unawareness"). Dropping the attribute from the feature matrix would change the baseline group disparities and could in principle shift the rankings; we do not explore that variant here.

A modest number of datasets. Four datasets still span a limited slice of domains, and the cross-dataset statistic rests on six dataset-pairs per seed — enough to expose the baseline-bias-regime effect but not to characterize it finely. Stronger claims would require more datasets and domains (e.g., additional ACS tasks and states [7], or non-US data).

Two base classifiers. We evaluate one linear model and one tree ensemble. The high cross-model τ may not extend to model classes that differ more sharply, such as neural networks or kernel methods, and a larger set of base learners is needed before the base classifier can be abstracted away in general.

Default hyperparameters. We tune neither the base classifiers nor the mitigation algorithms. Tuning could change the ordering, but the untuned setting is what an off-the-shelf user would actually see. Tuned models could also shift the accuracy–fairness trade-off itself, so how tuned pipelines rank remains an open question that we leave to future work.

Demographic Parity and Equalized Odds only. We do not evaluate calibration-based mitigations, which would require a different experimental design.

COMPAS dataset critique. We acknowledge the methodological concerns about COMPAS as a fairness benchmark [19]. We retain it for comparability with prior comparative work but flag it as a candidate for replacement in follow-up work. To check that our conclusions do not hinge on COMPAS, we recomputed the 10-seed statistics from the released per-seed results (`results/multiseed_perpipeline.csv`) with COMPAS excluded: cross-model $\tau = 0.68$ [0.61, 0.75], cross-dataset $\tau = 0.31$ [0.20, 0.42], and cross-metric $\tau = -0.08$ [-0.18, 0.01]. The ordering of the three axes is unchanged.

Seed and split variance. The per-pipeline metrics in Tables II–III are reported both as 10-seed means with standard deviations and at the reference seed (42) and a single 70/30 split; the ranking-inversion and Pareto figures use the reference seed. The headline stability statistics are reported with 95% confidence intervals (Table IV, Section V-C) — which is precisely what reveals that the reference seed is a favorable draw for cross-dataset transfer. Per-pipeline accuracy is stable across seeds (SD median 0.006); the fairness metrics vary more (DP-Δ SD median 0.011, EO-Δ SD median 0.025, with maxima of 0.055 and 0.070), so the seed-42 fairness values should be read as one representative draw, and the 10-seed means as the better point estimates. We do not resample the split structure itself (e.g., repeated k-fold); that remains a natural next step.

VIII. Conclusion

We compared six bias-mitigation pipelines across four tabular fairness datasets, two base classifiers, and two fairness metrics, all through one industry-standard toolkit at default settings. Three Kendall's τ statistics, each a 10-seed mean with a 95% CI, sum up what we found: mitigation rankings carry over well between base classifiers (0.65, [0.61, 0.69]), poorly between datasets (0.31, [0.23, 0.39]), and not at all between fairness metrics (0.02, [-0.04, 0.07]); the cross-model interval clears the other two completely. The dataset figure is two stories in one: rankings transfer between similarly-biased datasets (Adult $\leftrightarrow$ COMPAS 0.38, [0.29, 0.47]) but are no better than random for any pair involving the already-fair German Credit data (0.08, [-0.07, 0.23]).

This is a sharper statement than the usual "no single mitigation wins." Mitigation choice can be abstracted away from the base classifier, but not from the data or the fairness criterion. And the criterion is no free choice: Demographic-Parity-constrained mitigations raise the Equalized Odds difference by up to $3\times$ on Adult, while Equalized-Odds-constrained ones lower both DP-Δ and EO-Δ together — reason enough to treat Equalized Odds as the safer default when the downstream criterion is still open.

What this means in practice:

1) **Measure the baseline first.** If the unmitigated model already clears the fairness threshold, leave it alone.
2) **Commit to a fairness criterion before picking a mitigation.** Rankings shift so much between metrics that the criterion quietly decides the outcome.
3) **Prefer EO-constrained over DP-constrained mitigations when the criterion is uncertain.** On the biased datasets we tested, EO constraints improved both metrics; DP constraints improved one at the other's expense.
4) **Be careful transferring across datasets that differ in baseline-bias regime.**

We release the code and configuration so others can rerun the protocol on their own data. Natural extensions include more datasets and domains (further ACS tasks and states [7], or non-tabular settings), multi-class tasks, more diverse and tuned base classifiers, resampled train/test splits, intersectional fairness across multiple attributes, calibration-based mitigations, and a quantitative test of the ranking safeguards proposed in Section VI-D. Mostly, we hope later comparisons will measure ranking stability directly, instead of assuming it.

Acknowledgment

Shubhra Mittal is affiliated with Microsoft, which maintains Fairlearn. This work was conducted independently — it was not sponsored or reviewed by Microsoft — and the comparison

applies the same toolkit uniformly to every mitigation, so the affiliation does not bias the results; the mitigation algorithms themselves are standard and available in other toolkits. Fairlearn is maintained by the Fairlearn open-source community.

REFERENCES

[1] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[4] R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[5] D. Dua and C. Graff, "UCI machine learning repository," University of California, Irvine, School of Information and Computer Sciences, 2017, web: https://archive.ics.uci.edu/ml.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016, web: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring Adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[10] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

[11] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, pp. 671–732, 2016.

[12] S. Bird, M. Dudík, R. Edgar *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft Research, Tech. Rep. MSR-TR-2020-32, 2020.

[13] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2009.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

[15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[16] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[17] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

[18] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[19] M. Bao, A. Zhou, S. Zottola *et al.*, "It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] R. K. E. Bellamy, K. Dey, M. Hind *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 2019.

[22] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

[23] S. Biswas and H. Rajan, "Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2020.

[24] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, 2024.

[25] K. Lum, Y. Zhang, and A. Bower, "De-biasing 'bias' measurement," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 379–389.

APPENDIX A
REPRODUCIBILITY CHECKLIST

- ☒ Code released (`01_experiment.py`, `02_analysis.py`, `03_make_tables.py`, `04_inversion_plot.py`, `05_multiseed.py`, `06_pareto.py`)
- ☒ Bibliography released (`references.bib`)
- ☒ Library versions pinned (`requirements-lock.txt`)
- ☒ Random seed fixed (SEED = 42)
- ☒ Hardware: laptop CPU, $<$ 10 minutes wall-clock for the full sweep
- ☒ EO-constrained Table III numbers populated
- ☒ Kendall's τ Table IV numbers populated (full matrix in `results/kendall_tau.csv`)
- ☒ Multi-seed variance report (10 seeds; `05_multiseed.py`, `results/multiseed_summary.csv`)
- ☒ Per-pipeline 10-seed means $\pm$ SD in Tables II–III (`results/multiseed_perpipeline_summary.csv`)
- ☒ ACS Income (folktables) added as a fourth dataset

APPENDIX B
HOW TO REPRODUCE

```
# Clone the released code and run from
# the repository root.
git clone https://github.com/shubmittal/
    fairness-mitigation-comparison
cd fairness-mitigation-comparison
# Create a fresh virtual environment
# (Python 3.11):
python -m venv .venv
.\.venv\Scripts\Activate.ps1
# Windows PowerShell; use
# 'source .venv/bin/activate'
# on macOS/Linux
pip install -r requirements-lock.txt
# exact pinned versions
# (includes folktables)
python 01_experiment.py
# -> results/results.csv,
#    results/timings.csv
#    (+ tradeoff plots)
python 02_analysis.py
# -> results/kendall_tau.csv
#    (+ printed aggregate tau)
python 03_make_tables.py
# -> Markdown for Tables II and III
python 04_inversion_plot.py
# -> results/plots/
#    <dataset>_ranking_inversion.png
python 05_multiseed.py
```

```
# -> results/multiseed_tau.csv,
#    results/multiseed_summary.csv
python 06_pareto.py
# -> results/plots/pareto_frontier.png
```

The ACS Income dataset is downloaded from the Census Bureau on first run and cached under `data/` (gitignored).

Outputs: `results/results.csv` (deterministic metrics), `results/timings.csv` (wall-clock training time), `results/kendall_tau.csv`, `results/multiseed_summary.csv`, and plots under `results/plots/` (per-dataset tradeoff, ranking-inversion, and `pareto_frontier.png`).