ATTRIBUTE INFERENCE ON DIFFERENTIALLY PRIVATE AND FAIR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Private fair-learning methods use Differential Privacy (DP) to protect sensitive demographic information while enforcing group-level fairness. However, DP guarantees do not imply that protected attributes are statistically independent of released predictions. We audit output-level attribute inferability in three representative private fair-learning pipelines spanning central and local DP. Across eight tabular datasets and multiple privacy budgets, we jointly evaluate predictive utility, fairness violation, and attribute-inference performance from released prediction scores. We find that protected attributes remain inferable in many settings, while the relationship between the privacy budget and attack success is strongly method- and dataset-dependent. These results motivate attribute-inference auditing as a complement to formal privacy accounting when evaluating private fair-learning systems.

1 INTRODUCTION

Formal privacy guarantees provide a principled basis for protecting sensitive data in machine learning. Yet, the presence of a formal guarantee does not by itself characterize all the information an adversary may infer from a released model's predictions. Differential Privacy (DP) Dwork et al. (2006); Dwork and Roth (2014) bounds changes in a mechanism's output distribution under a specified adjacency relation, including when the released output is a trained model. This guarantee is distinct from requiring predictions to be statistically independent of a protected attribute Jagielski et al. (2019). Understanding this distinction is important when evaluating learning pipelines that combine privacy and fairness.

This issue becomes especially prominent in fairness-aware learning Hardt et al. (2016); Zemel et al. (2013). Fairness methods often require access to protected attributes in order to measure or mitigate disparities, creating a need to reconcile the use of demographic information with privacy requirements Yao and Juarez (2025); Cummings et al. (2019); Makhlouf et al. (2024); Mangold et al. (2023). Representative works Jagielski et al. (2019); Tran et al. (2021); Mozannar et al. (2020) therefore combine fairness constraints with central or local differential privacy, aiming to protect sensitive attributes while enforcing group-level fairness objectives Hardt et al. (2016); Mehrabi et al. (2021). Evaluating these systems requires distinguishing the formal protection of demographic information from the extent to which that information remains inferable through their prediction outputs.

In real-world datasets, protected attributes are rarely isolated: they are often correlated with non-sensitive features and target labels. Such correlations can make a protected attribute predictable even when the learning algorithm satisfies its stated DP guarantee Bun et al. (2021); Jagielski et al. (2019); Jayaraman and Evans (2022). Prior work has also connected group fairness with resistance to attribute inference Aalmoes et al. (2024) and evaluated inference attacks and DP-SGD mitigations for fairness-enhanced classifiers Tian et al. (2025). We build on these observations by focusing on private fair-learning pipelines that specifically protect the demographic information used in fairness constraints or post-processing. Our question is therefore not whether DP eliminates population-level predictability, but ***how output-level attribute inferability varies across these pipelines and privacy budgets, alongside predictive utility and fairness***.

In this paper, we investigate this question through a systematic attribute-inference audit of three private fair-learning pipelines. We consider centrally private Lagrangian-dual training Tran et al. (2021), a locally private adaptation that replaces protected attributes with randomized-response

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

reports in the fairness constraints, and fair learning with locally private demographic data and demographic-aware post-processing Mozannar et al. (2020). These pipelines differ in both their privacy mechanisms and their use of demographic information. In particular, the two Lagrangian classifiers predict from non-sensitive features, whereas the post-processing pipeline also uses a locally privatized sensitive attribute when producing predictions. Our comparison evaluates the resulting pipelines and their released scores, rather than attributing all differences between them to the privacy mechanism alone.

We evaluate eight tabular datasets, comprising nine dataset–sensitive-attribute settings, using four attack-classifier families and ten random seeds. The adversary learns from auxiliary samples with known protected attributes and infers a target individual's protected attribute from the corresponding prediction score, without access to the target model's parameters, gradients, or internal representations. Across privacy budgets, we jointly measure predictive utility, fairness violation, and attribute-inference performance. This evaluation measures attribute recoverability under a score-only threat model; it does not estimate additional disclosure relative to an adversary who already observes the individual's non-sensitive features.

Our results reveal **method- and dataset-dependent relationships between the privacy budget and output-level attribute inferability**. In several Lagrangian settings, attack AUC decreases as ε increases, whereas the post-processing pipeline exhibits pronounced increases in several settings. For example, on Adult, increasing the local privacy budget from $\varepsilon = 0.1$ to $\varepsilon = 10$ decreases the best reported attack AUC from approximately 0.72 to 0.69 for the locally private Lagrangian pipeline, while increasing it from approximately 0.54 to 0.93 for the post-processing pipeline. Thus, within the evaluated configurations, a stronger formal privacy guarantee does not invariably correspond to lower observed attribute inferability. These contrasting trends motivate examining the privacy budget together with the learning procedure, predictive utility, and achieved fairness, rather than treating it as a direct predictor of attack success.

Overall, our contribution is a comparative empirical characterization of attribute inferability in private fair-learning pipelines, extending their evaluation beyond utility and fairness to the information recoverable from prediction scores. Our findings support *output-level attribute inference auditing as a complement to formal privacy accounting*: the DP guarantee characterizes protection under specified changes to the data, while the audit measures the success of an adversary observing model predictions. This distinction allows the two forms of evidence to inform the evaluation of private fair learning without interpreting successful attribute inference as a violation of DP.

2 Related Work

Privacy and fairness in learning. Privacy and fairness have been studied as interacting objectives in machine learning. Prior works Bagdasaryan et al. (2019); Cummings et al. (2019); Ganev et al. (2022); Mangold et al. (2023); Makhlouf et al. (2024); Yao and Juarez (2025) have analyzed the compatibility and trade-offs between differential privacy and group fairness, showing that privacy mechanisms can affect group disparities and predictive utility. Closest to our setting are methods that protect sensitive demographic information while enforcing fairness constraints. Jagielski et al. (2019) develop differentially private in-processing and post-processing approaches to fair learning. Tran et al. (2021) incorporate privacy into fairness-aware optimization through a Lagrangian-dual approach, while Mozannar et al. (2020) study fair learning with locally private demographic data. These methods provide guarantees for the released model or demographic reports under their respective privacy definitions and trust assumptions. However, such guarantees do not eliminate inference based on population-level correlations. Our work complements these contributions by evaluating attribute inferability from the prediction outputs of representative private fair-learning pipelines.

Attribute privacy and attribute inference. Several works study privacy risks beyond membership inference, including attribute inference attacks that aim to recover hidden sensitive attributes from model behavior Jayaraman and Evans (2022); Zhang et al. (2022); Annamalai et al. (2024); Kulynych et al. (2025); Guerra-Balboa et al. (2026). Formal notions such as attribute privacy Zhang et al. (2022) explicitly consider the protection of sensitive attributes and the role of correlations among data attributes. More recent advantage-based views of DP establish a formal connection between the privacy budget ε and inference risk Guerra-Balboa et al. (2026), including attribute

inference, by bounding the adversary's advantage beyond what can already be inferred from prior knowledge, auxiliary information, or population-level correlations. These ideas are closely related to our motivation: proxy correlations may make protected information inferable even when the sensitive attribute is privatized. Rather than proposing a new privacy definition or attack, we use attribute inference as an empirical audit of DP-based fair learning pipelines.

Attribute inference in fair learning. Prior work has also examined sensitive-attribute inference in fairness-aware models. Ferry et al. (2023) show that fairness information can facilitate sensitive-attribute reconstruction, while Aalmoes et al. (2024) study theoretical and empirical connections between group fairness and black-box attribute inference. Tian et al. (2025) further evaluate attribute inference against fairness-enhanced classifiers, including DP-SGD as a mitigation. We complement these works by focusing specifically on private fair-learning pipelines that protect demographic information through central or local DP, and by comparing how predictive utility, fairness violation, and output-level attribute inferability evolve across privacy budgets.

3 Background

3.1 Differential Privacy for Sensitive Attribute

Let $D = (D_P, D_S)$, where $D_P \in (\mathcal{X} \times \mathcal{Y})^n$ denotes the public attributes and $D_S \in \mathcal{A}^n$ denotes the sensitive attributes. Two sensitive datasets $D_S, D'_S \in \mathcal{A}^n$ are adjacent, written $D_S \sim D'_S$, if they differ in at most one entry (e.g., in one individual's group membership).

A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is (ε, δ)-differentially private with respect to the sensitive attribute A if, for every fixed D_P, every pair $D_S \sim D'_S$, and every measurable $R \subseteq \mathcal{R}$,

$$\mathbb{P}(\mathcal{M}(D_P, D_S) \in R) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(D_P, D'_S) \in R) + \delta.$$

Smaller ε and δ imply stronger privacy. When $\delta = 0$, the definition reduces to pure ε-DP; otherwise, it is approximate DP. DP is widely used in privacy-preserving machine learning due to its robustness to auxiliary information and its composition properties (Dwork et al., 2006; Dwork and Roth, 2014).

When D_S contains only a single entry, denoted by the sensitive attribute A, the above definition specializes to local differential privacy (Kasiviswanathan et al., 2011; Kairouz et al., 2014). In this case, the randomized mechanism is applied directly to A (e.g., $\mathcal{M} : A \to Z$), and the $\varepsilon-$privacy condition becomes

$$\max_{z,a,a'} \frac{\mathbb{P}(Z = z \mid A = a)}{\mathbb{P}(Z = z \mid A = a')} \leq e^{\varepsilon}.$$

3.2 Fairness Definitions

We consider group fairness with respect to a protected attribute A, a true label Y, and a prediction $\hat{Y}$. Group fairness notions require that a model satisfy approximate parity across groups with respect to a specified statistical criterion.

Demographic Parity. Demographic parity requires the prediction to be independent of the protected attribute. More formally:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b), \quad \forall a, b.$$

Equivalently, $\hat{Y} \perp A$. This enforces equal positive prediction rates across groups, but may be restrictive when base rates differ Hardt et al. (2016).

Equalized Odds. Equalized odds requires the prediction to be independent of the protected attribute, conditional on the true label:

$$P(\hat{Y} = 1 \mid A = a, Y = y) = P(\hat{Y} = 1 \mid A = b, Y = y), \quad \forall a, b,\ y \in \{0, 1\}.$$

Equivalently, $\hat{Y} \perp A \mid Y$. This requires equal true positive and false positive rates across groups and is often more appropriate when the target label is meaningful Hardt et al. (2016).

4 AUDITING DP-BASED FAIR LEARNING PIPELINES

We evaluate attribute inference risks on three differentially private and fairness-aware interventions.

4.1 TARGET MODEL I: PRIVATE AND FAIR LAGRANGIAN DUAL TRAINING

Our first target model is based on the Private and Fair Lagrangian Dual (PF-LD) framework proposed by Tran et al. (2021). Let

$$D = \{(X_i, A_i, Y_i)\}_{i=1}^n$$

denote the training dataset, where X_i is the non-sensitive feature vector, A_i is the protected attribute, and Y_i is the target label. Let M_θ be a classifier with parameters θ, and let

$$J(M_\theta, D_P) = \frac{1}{n} \sum_{i=1}^{n} L(M_\theta(X_i), Y_i)$$

be the empirical prediction loss. Fairness is imposed through constraints comparing population-level statistics $\mu(D_P)$ and protected-group statistics $\mu(D_G)$:

$$\min_{\theta} \quad J(M_\theta, D_P)$$
$$\text{s.t.} \quad \mu(D_P) - \mu(D_G) = \mathbf{0}.$$

Different $\mu(\cdot)$ recover fairness notions such as demographic parity or equalized odds. PF-LD solves this constrained problem using a Lagrangian relaxation with nonnegative multipliers λ:

$$\mathcal{L}(\theta, \lambda) = J(M_\theta, D_P) + \lambda^\top \left| \mu(D_P) - \mu(D_G) \right|.$$

The algorithm alternates between a primal update of the model parameters and a dual update of the Lagrange multipliers. Both are incorporated with DP-SGD (Abadi et al., 2016) for a mini-batch $B \subseteq D$:

$$\theta \leftarrow \theta - \alpha \left(\nabla_\theta J(M_\theta, B_P) + \lambda^\top \left| \nabla_\theta \mu(B_P) - \bar{\nabla}_\theta^{C_p} \mu(B_G) \right| + \mathcal{N}(0, \sigma_p^2 \Delta_p^2 I) \right),$$

where Gaussian noise Dwork and Roth (2014) is added to the fairness-related primal update involving protected-group statistics. After the primal updates, the dual variables are updated as

$$\lambda \leftarrow \lambda + s_k \left(\left| \mu(D_P) - \bar{\mu}^{C_d}(D_G) \right| + \mathcal{N}(0, \sigma_d^2 \Delta_d^2 I) \right),$$

Here, σ_p and σ_d are the noise multipliers for the primal and dual updates, while Δ_p and Δ_d denote their corresponding sensitivities. $\bar{\nabla}_\theta^{C_p}$ denote the clipping gradient operator. Following Tran et al. (2021), the Gaussian noise is calibrated using the theoretical sensitivity bounds of the primal and dual steps.

A key limitation of PF-LD is that the *privacy mechanism is applied only to fairness-related terms involving protected-group statistics*, while the standard prediction-loss gradient remains unprotected. Thus, the guarantee mainly targets quantities explicitly depending on the protected attribute A (as in Section 3.1), rather than the full training record or the non-sensitive features X. Since X and A are often correlated in real datasets, sensitive information may still be inferred indirectly from the learned model behavior. This motivates the attribute inference attack studied in this work.

4.2 TARGET MODEL II: LAGRANGIAN FAIR LEARNING WITH LOCALLY PRIVATIZED ATTRIBUTES

Our second target model keeps the Lagrangian-dual structure of Target Model I, but replaces central DP with local DP on the protected attribute. More precisely, each A_i is first privatized into Z_i using randomized response (Warner, 1965; Kairouz et al., 2014) with privacy budget ε:

$$\mathbb{P}(Z_i = A_i) = \frac{e^\varepsilon}{1 + e^\varepsilon}, \qquad \mathbb{P}(Z_i \neq A_i) = \frac{1}{1 + e^\varepsilon}.$$

The learner then observes

$$D_Z = \{(X_i, Z_i, Y_i)\}_{i=1}^n$$

and trains the same Lagrangian-dual fair model, replacing A_i by Z_i in the fairness constraints. No Gaussian noise is added during training; privacy comes from the local perturbation $A_i \mapsto Z_i$. Importantly, Z_i is used in the fairness constraints and not as classifier input, so predictions remain

$$\hat{Y}_i = M_\theta(X_i).$$

4.3 Target Model III: Post-processing with Locally Privatized Attributes

The third target model follows Mozannar et al. (2020). This model observes (X, Y, Z), where Z is obtained from the protected attribute A through randomized response Kairouz et al. (2014). Training proceeds in two stages on disjoint subsets S_1 and S_2. In the first stage, a randomized classifier Q is learned on S_1 to minimize classification error subject to approximate equalized-odds constraints based on Z:

$$\min_{Q \in \Delta\mathcal{H}} \quad \mathbb{P}^{S_1}(Q(X) \neq Y) \text{ s.t.} \quad M\gamma(Q) \leq \alpha_n \mathbf{1}, \tag{1}$$

where $M\gamma(Q)$ encodes the fairness constraints and α_n is the allowed tolerance. In the second stage, the predictions $\hat{Y} = Q(X)$ are post-processed on S_2 through a linear program that minimizes classification error while enforcing approximate equalized odds with respect to the true attribute A. The A-conditioned statistics are estimated from Z by inverting the randomized response mechanism.

4.4 Threat Model and Attribute Inference Attack

We consider a black-box, output-level attribute inference threat model. The adversary observes only the released prediction score s, e.g., $\mathbb{P}(\hat{Y} = 1)$, produced from X (and, for Model III, the locally privatized attribute Z), and aims to infer the protected attribute A. We instantiate the attack using the attribute inference component of AMULET (Waheed et al., 2025), which trains an attack model on prediction scores from auxiliary samples with known protected attributes. The adversary has no access to the target model's parameters, gradients, internal representations, or the true protected attribute at inference time.

While AMULET originally uses an MLP attacker, we broaden the attack space by considering logistic regression, random forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and MLP models. For each model family, we perform random hyperparameter search and retain the best-performing configuration. For each resulting attacker, we select the decision threshold that maximizes balanced accuracy and use it to convert output probabilities into binary predictions. We then report the results of the attacker with the highest balanced accuracy.

5 Results

Our main goal is to evaluate whether released prediction scores support inference of the protected attribute above chance, corresponding to an AUC baseline of 0.5. Importantly, successful attribute inference does not imply that DP is ineffective or that its formal guarantee is violated (Bun et al., 2021). DP bounds how much a mechanism's output distribution can change when an individual's protected attribute is modified under the specified adjacency relation, while the remaining public information is fixed. It does not eliminate population-level correlations among the protected attribute A, non-sensitive features X, the target Y, and released predictions. Consequently, A may remain statistically predictable from X or from prediction outputs even when the mechanism satisfies its stated DP guarantee.

Our black-box attacker observes only the released prediction score and uses auxiliary samples with known protected attributes to infer the binary attribute of a target individual. Thus, our evaluation measures *output-level attribute inferability*, rather than directly estimating the neighboring-dataset distinguishability controlled by DP. We therefore interpret successful attacks as evidence that protected attributes remain recoverable from released predictions under this threat model, not as violations of DP. This provides an empirical dimension of private fair-learning systems that is not characterized by the privacy budget alone.

5.1 Datasets

We evaluate Bank Marketing (Moro et al., 2014), Adult (Becker, 1996), and two COMPAS settings (Angwin et al., 2022). Following prior attribute-inference setups (Aalmoes et al., 2024; Waheed et al., 2025; Liu et al., 2022), each dataset is split equally into disjoint subsets for training the target and attack models. Table 1 summarizes the datasets and protected attributes. Additional experiments on five ACS datasets Ding et al. (2022) are reported and discussed in Appendix A.

Dataset	#Samples	#Feat.	Target	Sensitive attribute	Ratio
Bank Marketing	12,000	15	Deposit	Age (binary; 25–58)	92%/8%
Adult	45,222	12	Income $>$ \$50K	Sex (male/female)	68%/32%
COMPAS-Sex	7,214	8	Recidivism	Sex (male/female)	81%/19%
COMPAS-Race	7,214	8	Recidivism	Race (Caucasian/non-Caucasian)	66%/34%

Table 1: Datasets used in our evaluation. `Ratio` reports the distribution of the sensitive attribute.

5.2 Models and Hyperparameters

We evaluate three private fair-learning pipelines.

Target Model I: PF-LD. Our first target model is PF-LD (Tran et al., 2021), which combines fairness-constrained training with differentially private perturbations of the primal and dual updates. We fix $\delta = 10^{-5}$. The corresponding noise multipliers are denoted by σ_p and σ_d, and the resulting privacy budget ε is computed using a Rényi Differential Privacy (RDP) accountant (Mironov, 2017). The clipping thresholds for the primal and dual updates are fixed to 5 and 3, respectively.

Target Model II: PF-LD with Local Privacy. Target Model II is our locally private adaptation of PF-LD and uses the same classifier architecture and training configuration as Target Model I, enabling a controlled comparison between the two privacy mechanisms. For both Models I and II, we use a two-hidden-layer MLP with input dimension d, hidden dimensions $\lfloor 2d/3 \rfloor$ and $\lfloor d/2 \rfloor$, and a single sigmoid output. We choose demographic parity as the fairness constraint for both.

Target Model III: Fair Learning with Private Demographic Data. Our third target model follows the fair-learning framework with locally private demographic data of Mozannar et al. (2020). Following their experimental setting, we fix the unfairness-violation tolerance to 0.001.

For all experiments, we repeat each configuration over 10 random seeds and report the mean together with the 95% confidence interval computed using Student's t-distribution.

5.3 Evaluation Metrics

We evaluate the models from three perspectives: utility, attribute leakage, and fairness violation.

Prediction Accuracy. We measure utility using the classification accuracy on the target label Y:

$$\mathrm{Acc}_Y = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{Y}_i = Y_i\}, \tag{2}$$

where $\hat{Y}_i$ is the target model prediction and Y_i is the ground-truth label.

Attack Accuracy and AUC Score. To measure attribute leakage, we train an attacker to infer the sensitive attribute A from the target model's outputs. We report attack accuracy,

$$\mathrm{Acc}_A = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{A}_i = A_i\}, \tag{3}$$

where $\hat{A}_i$ is the attacker's prediction. Since accuracy can depend on the decision threshold and may be misleading under class imbalance, we also report attack AUC. Higher values indicate stronger attribute leakage, while an AUC close to 0.5 corresponds to a near-random attack.

Fairness Violation. We report fairness violation as the maximum deviation between a protected group's statistic and the corresponding population-level statistic:

$$\mathrm{FV} = \max_a \left| \mu_a - \mu_{\mathrm{all}} \right|, \tag{4}$$

where μ_a is the fairness-related statistic for group a, and μ_{all} is the corresponding population statistic. Depending on the fairness notion, μ_a represents the positive prediction rate, true positive rate, or false positive rate. Smaller values indicate better satisfaction of the fairness constraint.

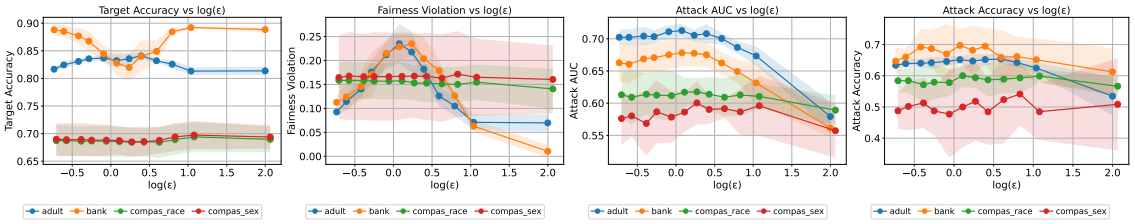
Pareto Front. We examine the trade-off among predictive utility, fairness, and attribute inferability across privacy budgets ε. Each point represents an evaluated configuration, with target error $(1 - \text{target accuracy})$ on the horizontal axis and fairness violation on the vertical axis; lower values are better for both. Color indicates attack AUC, while markers denote ranges of ε. Dashed lines connect configurations that are non-dominated with respect to target error, fairness violation, and attack AUC, forming the empirical Pareto front.

5.4 Experimental Results on Target Model I

Figure 1 shows markedly different behaviors across datasets. The two COMPAS settings exhibit the lowest target accuracy and comparatively lower attack AUC, with both quantities remaining relatively stable across most privacy budgets. Adult and Bank, in contrast, show substantially larger variations. These cross-dataset differences are descriptive: predictive accuracy and attribute inferability measure distinct properties, and lower task accuracy does not necessarily imply weaker attribute inference.

The effect of the privacy budget is particularly non-monotonic for Adult and Bank. For Adult, target accuracy initially increases and then decreases as ε grows, while Bank exhibits the opposite pattern at intermediate budgets before recovering at larger ε. Their fairness violations also increase at intermediate budgets and then decrease substantially for weaker privacy. Importantly, attack AUC follows neither a monotonic increase nor the usual intuition that weaker privacy should necessarily yield greater inferability. For both Adult and Bank, AUC is highest around intermediate budgets and decreases markedly at larger ε; the COMPAS settings exhibit smaller variations, followed by a decline at the largest budgets. Attack accuracy shows broadly similar dataset-dependent behavior, although AUC is our primary measure because it is insensitive to the attack threshold and class imbalance.

These results show that ε alone is not a reliable predictor of output-level attribute inferability in PF-LD. Changing the privacy budget also changes the solution reached by the fairness-constrained optimization, jointly affecting predictive utility, achieved fairness, and the information exposed by the resulting scores. Figure 2 makes these interactions explicit. The empirical Pareto fronts reveal trade-offs between target error and fairness violation, particularly for Adult and Bank, while the attack-AUC coloring shows that configurations with similar utility or fairness can exhibit different levels of attribute inferability. No privacy-budget range uniformly dominates across all three objectives. Overall, the results motivate evaluating the privacy budget together with the behavior of the resulting model rather than interpreting ε as a direct proxy for output-level inference risk.



(a) Target accuracy. (b) Fairness violation. (c) Attack AUC. (d) Attack accuracy.

Figure 1: Experimental results for Target Model I under different DP privacy budgets.

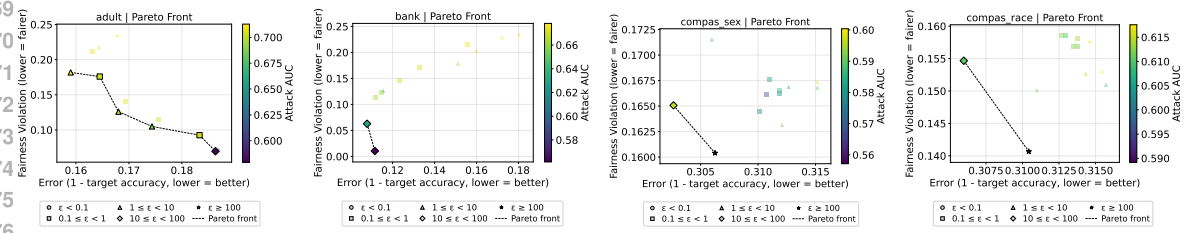
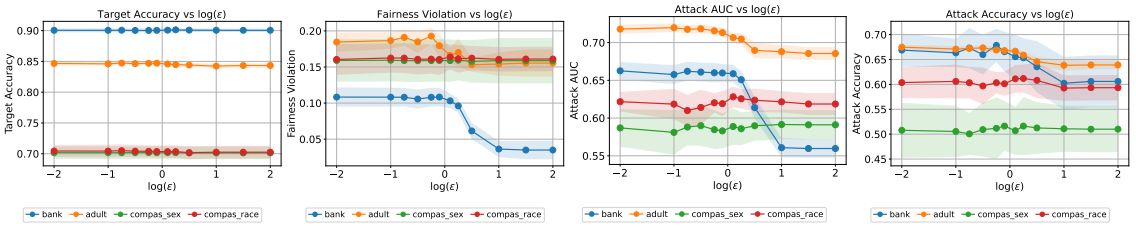


Figure 2: Empirical Pareto trade-offs for Target Model I.

5.5 Experimental Results on Target Model II

Figure 3 reveals a different interaction between privacy, fairness, and inferability for Model II. Target accuracy remains nearly constant across privacy budgets, suggesting that changes in the other metrics are not primarily driven by predictive utility. As ε increases, randomized response produces a more faithful approximation Z of the protected attribute A. Consistent with this, fairness violation decreases substantially for Bank and more moderately for Adult, while remaining comparatively stable on the COMPAS datasets. This suggests that reducing the noise on Z can improve the estimation and enforcement of the fairness constraints when privatization is a limiting factor.

Importantly, weaker local privacy does not systematically increase output-level attribute inferability. Attack AUC decreases markedly for Bank and also declines for Adult as ε grows, while the COMPAS settings remain comparatively stable. For Adult and Bank, the reduction in AUC coincides with improved fairness, suggesting that a more informative privatized attribute can enable the fairness intervention to reduce group-dependent signal in the released predictions. Thus, ε affects not only the fidelity of the private attribute but also the model produced by the fairness-constrained optimization. The Pareto fronts in Figure 4 reinforce this interaction: for some datasets, weaker privacy is associated with both lower fairness violation and lower attribute inferability rather than a simple privacy–fairness trade-off.



(a) Target accuracy. (b) Fairness violation. (c) Attack AUC. (d) Attack accuracy.

Figure 3: Experimental results for Target Model II under different DP privacy budgets.

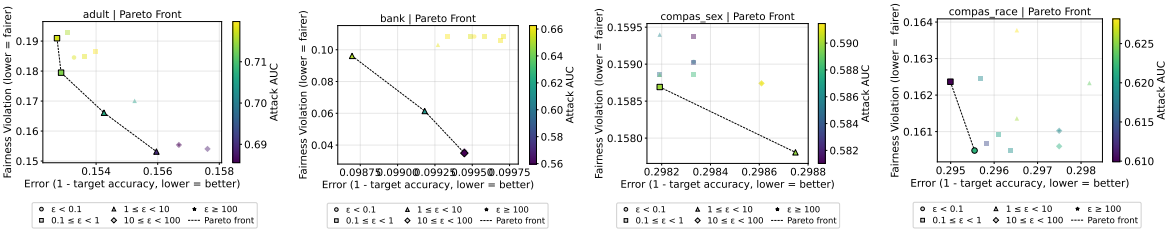
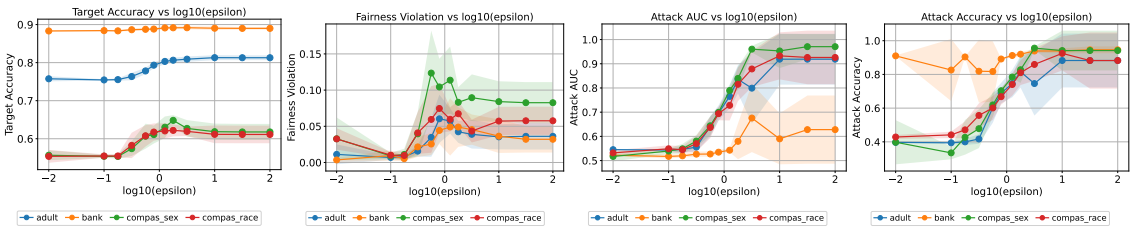


Figure 4: Empirical Pareto trade-offs for Target Model II.

5.6 Experimental Results on Target Model III

Figure 5 reveals a markedly different behavior for Model III. Predictive utility generally improves or remains stable as ε increases, while fairness violation is non-monotonic and peaks at intermediate budgets on several datasets, consistent with the behavior reported by Mozannar et al. (2020). More importantly, attribute inferability increases substantially with ε: attack AUC rises from values close to 0.5 under strong perturbation to above 0.9 for several datasets at weaker privacy.

This contrast with Model II is consistent with how demographic information enters the two pipelines. In Model II, the privatized attribute Z is used only to estimate fairness constraints during training, whereas Model III additionally uses Z in the post-processing that produces the final prediction. As ε increases, Z becomes a more faithful representation of A, and this information can consequently be reflected more strongly in the released score. The Pareto fronts in Figure 6 further illustrate the resulting tension: configurations with improved utility can coincide with substantially greater attribute inferability. Thus, the effect of the LDP budget depends critically on *where* the privatized attribute is used in the learning pipeline, not only on the amount of perturbation applied to it.



(a) Target accuracy. (b) Fairness violation. (c) Attack AUC. (d) Attack accuracy.

Figure 5: Experimental results for Target Model III under different DP privacy budgets.

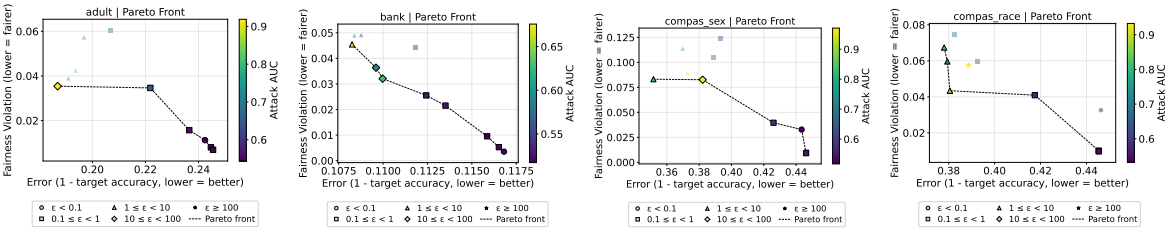


Figure 6: Empirical Pareto trade-off plots of Model III.

Pareto fronts analysis. Figures 2, 4, and 6 summarize the empirical Pareto fronts for the three target models. Across methods, higher predictive utility often coincides with larger fairness violations, illustrating the expected utility–fairness tension. Moreover, higher-accuracy configurations frequently exhibit higher attack AUC, suggesting that improvements in utility can also coincide with greater output-level attribute inferability. This pattern is particularly pronounced for Model III, while Models I and II exhibit more dataset-dependent behavior.

More importantly, the privacy budget ε does not induce a consistent ordering of attribute inferability. Configurations with smaller ε can still exhibit relatively high attack AUC, while larger ε can sometimes coincide with lower AUC. This reinforces the distinction between the two quantities: ε characterizes the formal DP guarantee under its specified adjacency relation, whereas attack AUC measures empirical attribute inferability from the released predictions. The three objectives should therefore be evaluated jointly rather than interpreting the privacy budget as a direct proxy for output-level inference risk.

6 Conclusion and Perspectives

We presented a systematic empirical audit of output-level attribute inferability in representative DP-based fair learning pipelines (Mozannar et al., 2020; Tran et al., 2021). Our results show that protected attributes can remain inferable from released prediction scores, and that a lower privacy budget does not necessarily imply lower attack accuracy. These findings do not contradict the underlying DP guarantees; rather, they show that ε and output-level attribute inferability characterize distinct properties of a private fair-learning system. We therefore argue that formal privacy accounting should be complemented with explicit inference audits of released model outputs.

Our evaluation is limited to representative private fair-learning methods, binary protected attributes, tabular datasets, and black-box score-based attacks. Future work could consider richer sensitive attributes, stronger adversaries, larger models, and alternative release interfaces, as well as defenses that explicitly reduce attribute recoverability while preserving utility and fairness. Formal notions such as attribute privacy (Zhang et al., 2022) and advantage-based leakage measures (Guerra-Balboa et al., 2026) provide complementary directions for characterizing inference under correlations and prior knowledge, while empirical audits remain important for understanding what information is recoverable from deployed model outputs.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language editing for clarity and readability and with portions of the experimental code and plotting. They were not used to conduct the literature review, formulate the research questions, or determine the scientific claims of the paper. All AI-assisted text and code were reviewed by the authors, and experimental outputs were checked for consistency and correctness. The authors take full responsibility for the final manuscript, experimental results, and any content produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies privacy risks in machine-learning systems with the goal of improving the evaluation of private and fair learning methods. Our experiments use publicly available benchmark datasets and involve no new collection of personal data or interaction with human participants. Although attribute inference techniques may have dual-use implications, we employ them only in controlled experimental settings to assess information recoverable from model outputs. We believe that documenting such risks can help researchers and practitioners better evaluate privacy claims and design more robust privacy-preserving systems. All results are reported at an aggregate level, and no attempt is made to identify real individuals.

REPRODUCIBILITY STATEMENT

We provide the source code used to reproduce our experiments and figures in an anonymous repository at `https://anonymous.4open.science/r/att_inf-0AAD/`. The paper specifies the threat model, evaluated private fair-learning pipelines, datasets, hyperparameters, privacy accounting, evaluation metrics, and random-seed protocol used in our experiments. Additional results across five ACS datasets and detailed attack-classifier results are provided in the appendix. Together, the manuscript, appendix, and released implementation provide the necessary material to reproduce the experimental evaluation.

REFERENCES

Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. On the alignment of group fairness with attribute privacy. In *Web Information Systems Engineering – WISE 2024*, pages 333–348, 2024. doi: 10.1007/978-981-96-0567-5_24.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2351–2368, 2024.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

R. Becker, B. & Kohavi. Adult [dataset], 1996. UCI Machine Learning Repository.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Mark Bun, Damien Desfontaines, Cynthia Dwork, Moni Naor, Kobbi Nissim, Aaron Roth, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Statistical inference is not a privacy violation. *differentialprivacy. org, June*, 3, 2021.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP'19 Adjunct, page 309–315, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3314183.3323847.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning, 2022. URL https://arxiv.org/abs/2108.04884.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Exploiting fairness to enhance sensitive attributes reconstruction. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 18–41. IEEE, 2023.

Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pages 6944–6959. PMLR, 2022.

Patricia Guerra-Balboa, Annika Sauer, Héber H Arcolezi, and Thorsten Strufe. Understanding disclosure risk in differential privacy with applications to noise calibration and auditing (extended version). *arXiv preprint arXiv:2603.12142*, 2026.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.

Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1569–1582, 2022.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. doi: 10.1137/090756090.

Bogdan Kulynych, Juan Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio Calmon, and Jean Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. *Advances in Neural Information Processing Systems*, 38:111967–112003, 2025.

Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4525–4542. USENIX Association, 2022.

Karima Makhlouf, Tamara Stefanović, Héber H. Arcolezi, and Catuscia Palamidessi. 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)*, pages 1–16, 2024. doi: 10.1109/CSF61375.2024.00039.

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pages 23681–23705. PMLR, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275. IEEE, August 2017. doi: 10.1109/csf.2017.11.

Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing [dataset], 2014. UCI Machine Learning Repository.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.

Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. Do fairness interventions come at the cost of privacy: Evaluations for binary classifiers. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4654–4670, 2025.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9932–9939, 2021.

Asim Waheed, Vasisht Duddu, Rui Zhang, and Sebastian Szyller. Amulet: a python library for assessing interactions among ml defenses and risks, 2025. URL `https://arxiv.org/abs/2509.12386`.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69, 1965.

Kai Yao and Marc Juarez. Sok: What makes private learning unfair? In *2025 IEEE Conference on Secure and Trustworthy Machine Learning*, pages 841–857, 2025. doi: 10.1109/SaTML64287.2025.00052.

Rich Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 757–766, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533139.

A Additional Experiments on ACS Datasets

Table 2 summarizes the five ACS datasets used in our additional experiments. We report the corresponding results for Models I–III below.

Dataset	#Samples	#Feat.	Target	Sensitive	Ratio
ACS Income	30,961	9	Income $>$ \$50K	Race	95.53%/4.47%
ACS Public Coverage	21,918	16	Public health coverage	Race	94.28%/5.72%
ACS Mobility	12,605	18	Same residence	Race	94.68%/5.32%
ACS Employment	60,000	14	Employed	Race	95.14%/4.86%
ACS Travel Time	27,297	13	Commute $>$ 20 min	Race	95.79%/4.21%

Table 2: ACS datasets (Ding et al., 2022) used in our evaluation. `Ratio` reports the non-Black/Black distribution, with Black defined as `RAC1P`$=2$.

Figures 7, 9, and 11 report the results of Models I–III on the five ACS datasets, with the corresponding Pareto trade-offs shown in Figures 8, 10, and 12. The additional experiments reproduce the main qualitative patterns from Section 5: Models I and II exhibit mostly dataset-dependent and non-monotonic relationships between ε and attack AUC, whereas Model III shows substantially stronger increases in attribute inferability as privacy is relaxed. The Pareto fronts likewise show no privacy-budget regime that consistently dominates utility, fairness, and inferability. Overall, these results support the robustness of our main observations across additional ACS prediction tasks.

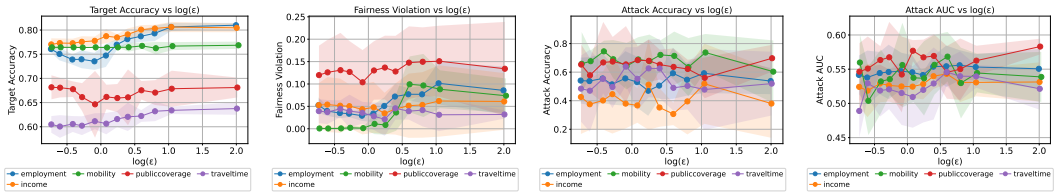


Figure 7: Results of Model I on the five ACS datasets.

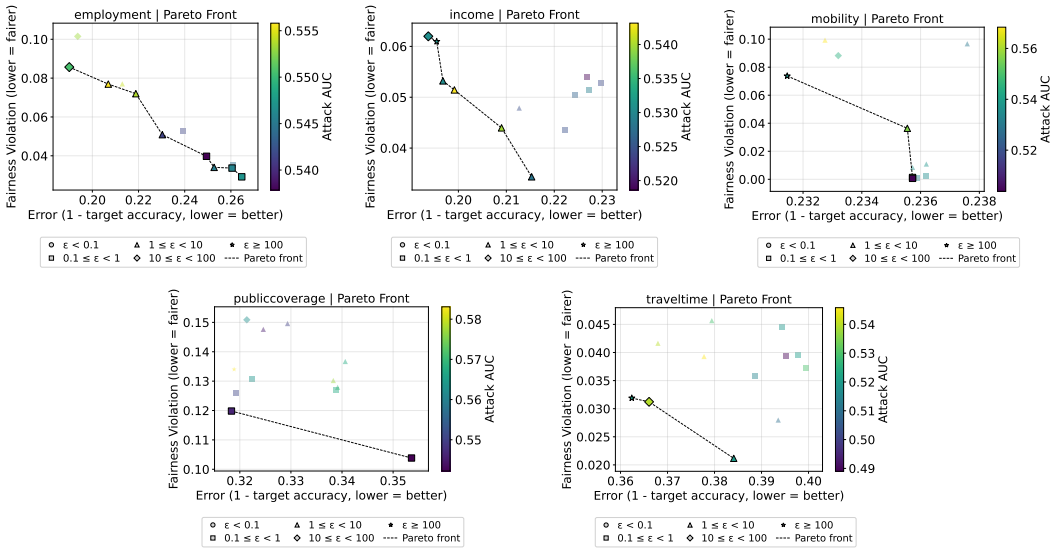


Figure 8: Empirical Pareto fronts of Model I on the five ACS datasets.

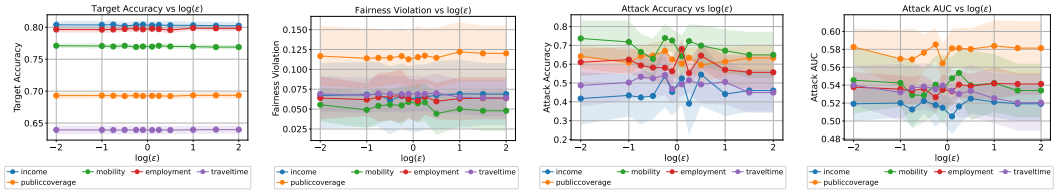


Figure 9: Results of Model II on the five ACS datasets.

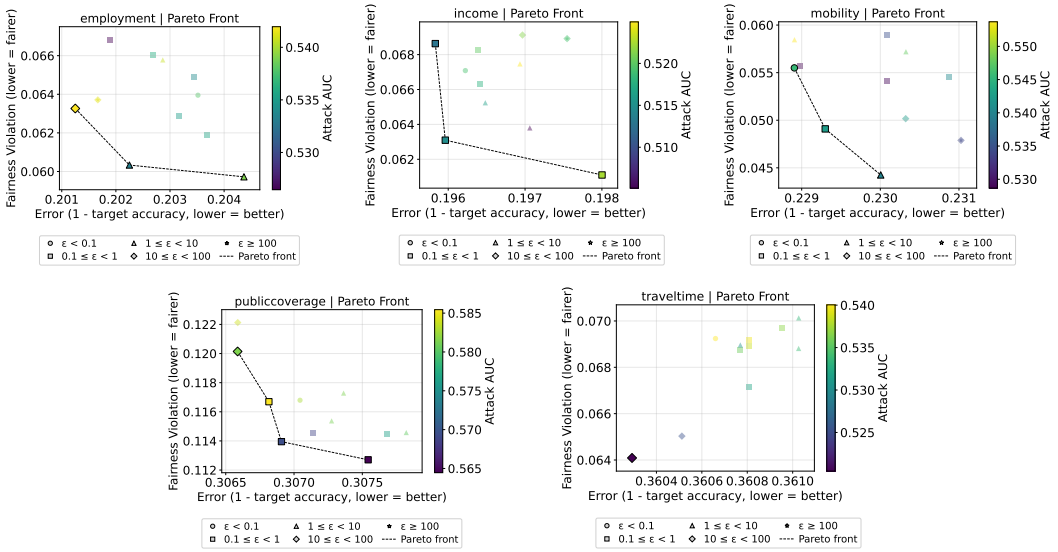


Figure 10: Empirical Pareto fronts of Model II on the five ACS datasets.

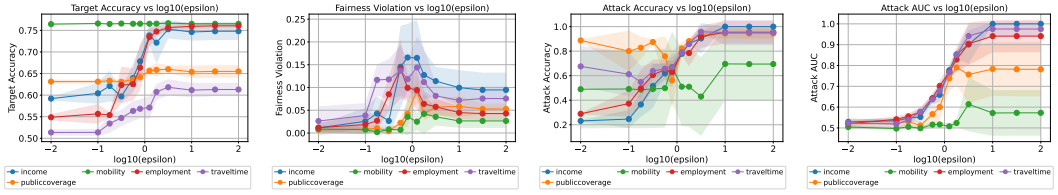


Figure 11: Results of Model III on the five ACS datasets.

B Performance of All Attack Classifiers

Tables 3–5 report the attack AUC obtained by each attacker family at representative privacy budgets $\varepsilon \in \{0.1, 10, 100\}$. The highest mean AUC for each dataset and budget is highlighted in **bold**. For Models I and II, attack performance is generally similar across classifier families, indicating that the observed inferability is not tied to a particular attacker. Model III behaves differently at larger ε: random forests and XGBoost substantially outperform logistic regression in several settings, suggesting that the relationship between released scores and protected attributes can become strongly nonlinear. These results motivate our use of multiple attacker families rather than relying on a single attack model.

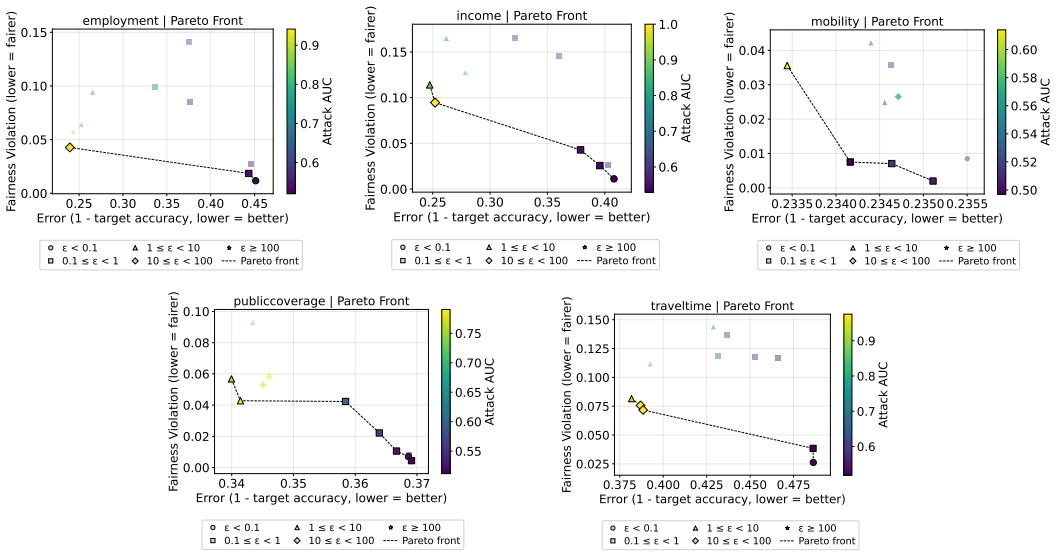


Figure 12: Empirical Pareto fronts of Model III on the five ACS datasets.

Model	Bank	Adult	COMPAS Sex	COMPAS Race	ACS Income	ACS Public Coverage	ACS Mobility	ACS Employment	ACS Travel Time
$\epsilon \approx 0.1$									
MLP	0.667 ± 0.019	0.702 ± 0.007	0.607 ± 0.020	0.628 ± 0.006	$\mathbf{0.534 \pm 0.024}$	0.577 ± 0.019	0.550 ± 0.039	0.549 ± 0.021	0.501 ± 0.041
Logistic	$\mathbf{0.667 \pm 0.019}$	0.702 ± 0.007	$\mathbf{0.607 \pm 0.020}$	$\mathbf{0.628 \pm 0.006}$	0.534 ± 0.024	$\mathbf{0.584 \pm 0.022}$	0.550 ± 0.039	0.533 ± 0.012	0.502 ± 0.041
Random Forest	0.662 ± 0.022	$\mathbf{0.703 \pm 0.008}$	0.573 ± 0.026	0.608 ± 0.014	0.509 ± 0.020	0.543 ± 0.039	$\mathbf{0.558 \pm 0.049}$	0.543 ± 0.026	$\mathbf{0.522 \pm 0.028}$
XGBoost	0.659 ± 0.025	0.703 ± 0.008	0.576 ± 0.023	0.605 ± 0.023	0.520 ± 0.015	0.554 ± 0.023	0.553 ± 0.034	$\mathbf{0.549 \pm 0.016}$	0.508 ± 0.029
$\epsilon \approx 10$									
MLP	$\mathbf{0.640 \pm 0.013}$	0.668 ± 0.009	0.606 ± 0.019	$\mathbf{0.624 \pm 0.005}$	0.540 ± 0.020	0.588 ± 0.014	0.544 ± 0.028	0.537 ± 0.012	0.546 ± 0.042
Logistic	0.640 ± 0.013	0.668 ± 0.009	$\mathbf{0.606 \pm 0.019}$	$\mathbf{0.624 \pm 0.005}$	$\mathbf{0.540 \pm 0.020}$	$\mathbf{0.592 \pm 0.017}$	0.544 ± 0.028	0.533 ± 0.014	$\mathbf{0.548 \pm 0.043}$
Random Forest	0.632 ± 0.019	$\mathbf{0.675 \pm 0.012}$	0.562 ± 0.034	0.595 ± 0.033	0.531 ± 0.018	0.560 ± 0.015	0.534 ± 0.014	0.553 ± 0.013	0.525 ± 0.023
XGBoost	0.628 ± 0.022	0.673 ± 0.012	0.571 ± 0.029	0.610 ± 0.014	0.515 ± 0.026	0.567 ± 0.006	$\mathbf{0.548 \pm 0.030}$	$\mathbf{0.558 \pm 0.022}$	0.535 ± 0.018
$\epsilon \approx 100$									
MLP	$\mathbf{0.570 \pm 0.013}$	0.569 ± 0.013	$\mathbf{0.598 \pm 0.017}$	$\mathbf{0.607 \pm 0.006}$	$\mathbf{0.535 \pm 0.021}$	0.591 ± 0.014	0.538 ± 0.024	0.530 ± 0.012	0.538 ± 0.035
Logistic	0.569 ± 0.014	0.566 ± 0.013	$\mathbf{0.598 \pm 0.017}$	0.607 ± 0.006	$\mathbf{0.535 \pm 0.021}$	$\mathbf{0.594 \pm 0.016}$	0.538 ± 0.024	0.527 ± 0.015	$\mathbf{0.543 \pm 0.039}$
Random Forest	0.562 ± 0.008	$\mathbf{0.579 \pm 0.019}$	0.550 ± 0.027	0.579 ± 0.012	0.534 ± 0.020	0.572 ± 0.025	0.546 ± 0.037	$\mathbf{0.550 \pm 0.025}$	0.512 ± 0.032
XGBoost	0.558 ± 0.005	0.577 ± 0.017	0.557 ± 0.017	0.578 ± 0.013	0.529 ± 0.020	0.573 ± 0.011	$\mathbf{0.551 \pm 0.035}$	0.543 ± 0.024	0.521 ± 0.012

Table 3: Target model I attack AUC by attack-classifier family. The noise level $\sigma = 200$, 5, and 1 were used for the requested privacy levels. Their recorded privacy budgets lie approximately in $[0.183, 0.197]$, $[10.870, 11.967]$, and $[98.804, 116.071]$, respectively; therefore, the first column is not an exact $\epsilon = 0.1$ experiment.

Model	Bank	Adult	COMPAS Sex	COMPAS Race	ACS Income	ACS Public Coverage	ACS Mobility	ACS Employment	ACS Travel Time
$\epsilon = 0.1$									
MLP	$\mathbf{0.664 \pm 0.016}$	0.715 ± 0.008	0.599 ± 0.030	$\mathbf{0.631 \pm 0.016}$	$\mathbf{0.533 \pm 0.026}$	0.593 ± 0.024	$\mathbf{0.549 \pm 0.029}$	$\mathbf{0.543 \pm 0.015}$	0.538 ± 0.050
Logistic	0.664 ± 0.016	0.715 ± 0.008	$\mathbf{0.599 \pm 0.030}$	0.631 ± 0.016	0.529 ± 0.025	$\mathbf{0.594 \pm 0.024}$	$\mathbf{0.549 \pm 0.029}$	0.527 ± 0.014	$\mathbf{0.542 \pm 0.046}$
Random Forest	0.653 ± 0.017	$\mathbf{0.722 \pm 0.005}$	0.571 ± 0.034	0.602 ± 0.022	0.525 ± 0.029	0.567 ± 0.045	0.537 ± 0.042	0.531 ± 0.024	0.531 ± 0.017
XGBoost	0.655 ± 0.015	0.722 ± 0.007	0.576 ± 0.021	0.610 ± 0.022	0.521 ± 0.020	0.567 ± 0.036	0.529 ± 0.040	0.537 ± 0.020	0.529 ± 0.028
$\epsilon = 10$									
MLP	0.561 ± 0.020	0.675 ± 0.016	0.598 ± 0.030	$\mathbf{0.630 \pm 0.016}$	$\mathbf{0.538 \pm 0.018}$	0.592 ± 0.024	$\mathbf{0.550 \pm 0.029}$	0.539 ± 0.012	0.539 ± 0.046
Logistic	0.560 ± 0.021	0.670 ± 0.009	$\mathbf{0.598 \pm 0.030}$	0.630 ± 0.016	0.526 ± 0.025	$\mathbf{0.592 \pm 0.024}$	$\mathbf{0.550 \pm 0.029}$	0.527 ± 0.015	$\mathbf{0.542 \pm 0.045}$
Random Forest	$\mathbf{0.567 \pm 0.013}$	0.689 ± 0.014	0.568 ± 0.026	0.608 ± 0.013	0.522 ± 0.021	0.582 ± 0.034	0.534 ± 0.034	0.541 ± 0.025	0.530 ± 0.025
XGBoost	0.564 ± 0.014	$\mathbf{0.689 \pm 0.012}$	0.571 ± 0.032	0.612 ± 0.018	0.524 ± 0.021	0.585 ± 0.032	0.533 ± 0.042	$\mathbf{0.542 \pm 0.014}$	0.521 ± 0.021
$\epsilon = 100$									
MLP	0.561 ± 0.020	0.675 ± 0.016	0.598 ± 0.030	$\mathbf{0.630 \pm 0.016}$	$\mathbf{0.538 \pm 0.018}$	0.592 ± 0.024	$\mathbf{0.549 \pm 0.029}$	0.538 ± 0.015	0.537 ± 0.050
Logistic	0.559 ± 0.021	0.671 ± 0.008	$\mathbf{0.598 \pm 0.030}$	0.630 ± 0.016	0.526 ± 0.025	$\mathbf{0.592 \pm 0.024}$	$\mathbf{0.549 \pm 0.029}$	0.527 ± 0.015	$\mathbf{0.541 \pm 0.045}$
Random Forest	$\mathbf{0.566 \pm 0.016}$	$\mathbf{0.691 \pm 0.011}$	0.568 ± 0.022	0.609 ± 0.013	0.515 ± 0.023	0.575 ± 0.039	0.533 ± 0.029	$\mathbf{0.544 \pm 0.018}$	0.524 ± 0.030
XGBoost	0.561 ± 0.013	0.690 ± 0.010	0.571 ± 0.031	0.612 ± 0.017	0.527 ± 0.023	0.576 ± 0.029	0.530 ± 0.037	0.536 ± 0.019	0.527 ± 0.020

Table 4: Model II attack AUC by attack-classifier family.

Model	Bank	Adult	COMPAS Sex	COMPAS Race	ACS Income	ACS Public Coverage	ACS Mobility	ACS Employment	ACS Travel Time
$\epsilon = 0.1$									
MLP	0.517 ± 0.012	0.544 ± 0.006	0.539 ± 0.017	0.549 ± 0.020	$\mathbf{0.535 \pm 0.018}$	0.521 ± 0.024	$\mathbf{0.500 \pm 0.009}$	0.542 ± 0.021	$\mathbf{0.518 \pm 0.032}$
Logistic	$\mathbf{0.520 \pm 0.006}$	$\mathbf{0.544 \pm 0.006}$	0.539 ± 0.016	$\mathbf{0.549 \pm 0.020}$	0.534 ± 0.018	$\mathbf{0.528 \pm 0.013}$	0.499 ± 0.007	0.539 ± 0.021	0.511 ± 0.037
Random Forest	$\mathbf{0.520 \pm 0.006}$	$\mathbf{0.544 \pm 0.006}$	$\mathbf{0.540 \pm 0.020}$	$\mathbf{0.549 \pm 0.020}$	0.532 ± 0.020	$\mathbf{0.528 \pm 0.013}$	0.500 ± 0.007	$\mathbf{0.542 \pm 0.020}$	0.512 ± 0.028
XGBoost	$\mathbf{0.520 \pm 0.006}$	$\mathbf{0.544 \pm 0.006}$	$\mathbf{0.540 \pm 0.020}$	$\mathbf{0.549 \pm 0.020}$	0.532 ± 0.020	$\mathbf{0.528 \pm 0.013}$	0.500 ± 0.007	$\mathbf{0.542 \pm 0.020}$	0.506 ± 0.030
$\epsilon = 10$									
MLP	0.579 ± 0.152	0.919 ± 0.147	0.953 ± 0.099	0.883 ± 0.180	0.974 ± 0.081	0.783 ± 0.180	0.559 ± 0.159	0.941 ± 0.109	0.975 ± 0.052
Logistic	0.570 ± 0.142	0.253 ± 0.150	0.543 ± 0.125	0.601 ± 0.200	0.648 ± 0.128	0.685 ± 0.196	0.557 ± 0.132	0.507 ± 0.165	0.599 ± 0.127
Random Forest	$\mathbf{0.613 \pm 0.137}$	$\mathbf{0.926 \pm 0.149}$	$\mathbf{0.971 \pm 0.073}$	$\mathbf{0.948 \pm 0.100}$	$\mathbf{1.000 \pm 0.001}$	$\mathbf{0.817 \pm 0.186}$	$\mathbf{0.579 \pm 0.150}$	$\mathbf{0.963 \pm 0.059}$	$\mathbf{0.975 \pm 0.052}$
XGBoost	$\mathbf{0.613 \pm 0.137}$	$\mathbf{0.926 \pm 0.149}$	$\mathbf{0.971 \pm 0.073}$	$\mathbf{0.948 \pm 0.100}$	0.999 ± 0.002	$\mathbf{0.817 \pm 0.186}$	0.571 ± 0.154	$\mathbf{0.963 \pm 0.059}$	$\mathbf{0.975 \pm 0.052}$
$\epsilon = 100$									
MLP	0.617 ± 0.203	0.919 ± 0.147	0.971 ± 0.073	0.894 ± 0.170	0.934 ± 0.108	0.771 ± 0.196	0.560 ± 0.158	0.941 ± 0.109	0.975 ± 0.053
Logistic	0.532 ± 0.198	0.253 ± 0.150	0.543 ± 0.125	0.597 ± 0.201	0.699 ± 0.067	0.615 ± 0.250	0.557 ± 0.132	0.490 ± 0.184	0.597 ± 0.131
Random Forest	$\mathbf{0.656 \pm 0.182}$	$\mathbf{0.926 \pm 0.149}$	$\mathbf{0.971 \pm 0.073}$	$\mathbf{0.926 \pm 0.155}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.817 \pm 0.187}$	$\mathbf{0.579 \pm 0.150}$	$\mathbf{0.963 \pm 0.059}$	$\mathbf{0.975 \pm 0.052}$
XGBoost	$\mathbf{0.656 \pm 0.182}$	$\mathbf{0.926 \pm 0.149}$	$\mathbf{0.971 \pm 0.073}$	$\mathbf{0.926 \pm 0.155}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.817 \pm 0.187}$	0.571 ± 0.154	$\mathbf{0.963 \pm 0.059}$	$\mathbf{0.975 \pm 0.052}$

Table 5: Target model III attack AUC by attack-classifier family.