Exposing Hidden Vulnerabilities: A Privacy Audit of Algorithmic Fairness

NILOY PURKAIT, HENRY BRIGHTON, EMMANUEL KEULEERS,

Department of Cognitive Sciences and Artificial Intelligence, Tilburg University, Netherlands

Algorithmic fairness and privacy are cornerstones of trustworthy machine learning. Prior research suggests that mitigating bias may amplify membership inference risks for underrepresented groups. However, evaluations based on average-case metrics, such as attack success rate, often obscure vulnerabilities at low false-positive rates—an regime critical for real-world privacy. In this paper, we present an in-depth empirical study of membership inference under low false positive rate conditions, across three fair machine learning algorithms. Our analysis on the Law School [61], Bank Marketing [22], and COMPAS [40] datasets reveal that standard threshold-based and population-based attacks underestimate privacy leaks, especially for sensitive subgroups. Motivated by this gap, we propose a subgroup-specific membership inference attack based on pairwise likelihood ratio tests. Our method models the null hypothesis more accurately: a target sample's outputs are statistically indistinguishable from those of non-members within the same sensitive subgroup. Our experiments show that a simple modification of an existing attack can achieve superior test power across the TPR-FPR curve—even at extremely low false positive rates, given an adversary with the same computational resources and access to data—establishing a robust foundation for more powerful, fine-grained privacy audits of fair learning algorithms.

Keywords: Group Fairness, Data Privacy, Membership Inference

Reference Format:
Niloy Purkait, Henry Brighton, Emmanuel Keuleers. 2025. Exposing Hidden Vulnerabilities: A Privacy Audit of Algorithmic Fairness. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 25 pages.

1 Introduction

Machine learning models have shown disparate treatment between demographic groups, leading to discriminatory outcomes in vital application domains. Data and algorithmic biases have been studied across vision [12, 49, 62], language [5, 11, 21], and tabular data [44, 51]. These findings led researchers to develop methods to detect and reduce bias against demographic groups defined by sensitive attributes such as race, gender, age, or marital status [13, 29, 33, 37, 66]. Faced with legal and ethical considerations, organizations that deploy machine learning models may be mandated to implement, audit and report on their fairness. For example, under emerging regulatory frameworks such as the EU's AI Act, organizations deploying AI-driven decision-making systems must consider fairness and transparency [27, 42]. Such developments incentivize organizations to integrate fairness into their

Author's Contact Information: Niloy Purkait, Henry Brighton, Emmanuel Keuleers,
Department of Cognitive Sciences and Artificial Intelligence, Tilburg University, Tilburg, Netherlands, {n.purkait,h.j.brighton,e.a.keuleers}@tilburguniversity.edu.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

existing learning pipelines. Practically, this can be done through open-source fairness toolkits like Microsoft's fairlearn [50] and IBM's AIF360 [9], functioning as a "fairness as a wrapper" solution: by supplying these tools with training data, organizations can obtain fully trained models whose outputs adhere to a chosen fairness criteria.

Reducing demographic bias in algorithmic decision making is ethically necessary, yet it raises concerns about individual privacy. Prior work shows that machine learning models can leak sensitive training data [15, 20, 30], while others have shown that fair models may exacerbate this risk [18, 58, 67]. A substantial body of research quantifies such leakage via membership inference attacks [14, 32, 53, 56, 63, 65]—essentially asking whether a given data point x was part of model θ's training set. Early approaches used hypothesis testing on aggregated statistics [26, 35, 54, 59] to verify data sample inclusion. Later methods targeted machine learning models directly, either by using shadow modeling to craft dedicated attack networks or by approximating these tests within a machine learning framework [10, 14, 41, 53, 56, 60, 63, 65, 67]. These attacks serve multiple purposes: curiosity-driven exploration, auditing model behavior for privacy risks, or, most maliciously, using membership inference as an oracle for data extraction [6, 15, 16, 32]. Although differential privacy [1, 23] can reduce privacy leakage, it often degrades performance for underrepresented groups [4, 19, 28, 31, 43]. In contrast, the impact of fairness constraints on membership inference remains underexplored—particularly at low false positive rates. Most studies have quantified membership inference using average-case attack accuracy. For example, Chang et al. [18] observed that applying fairness constraints can increase membership inference risk for certain subgroups[1]. However, recent research on membership inference attacks suggests that average-case metrics often underestimate worst-case risk [14, 63, 65]. Intuitively, an adversary seeking to extract sensitive training data need not succeed on average; identifying even a few true members with high certainty can suffice [14]. Evaluating attacks by treating false positives and false negatives symmetrically can mask this threat. In practice—especially when membership inference serves as an oracle for data extraction [16]—false positives pose little risk compared to false negatives: a few correct identifications can breach privacy, while misclassifying many non-members as members offers limited utility to the adversary [14, 63, 65].

In this paper, we conduct a comprehensive evaluation of how enforcing fairness affects the privacy of machine learning algorithms, focusing on extremely low false-positive rates as a critical measure of privacy breach. We address three research questions: *(i)* Do fairness interventions amplify membership leakage for protected subgroups under low false-positive rate regimes? *(ii)* How do different fairness interventions compare in their induced privacy leakage? *(iii)* Can we design subgroup-specific membership inference attacks that pinpoint privacy risks for particular groups at very low false positive rates?

To answer these, we evaluate three widely used fairness algorithms—Reweighing [36], Adversarial Debiasing [66], and Exponentiated Gradient Reduction [2]—using various membership inference attacks, including threshold-based [18, 64] and population-based methods [14, 63, 65]. Crucially, we propose a novel *subgroup-specific* low false positive rate attack that extends recent pairwise likelihood-ratio framework of [65], thereby enabling extremely fine-grained membership tests even under tight false-positive constraints. We validate our findings on

[1]A protected subgroup is defined by the Cartesian product of a record's protected attribute value and its corresponding target label.

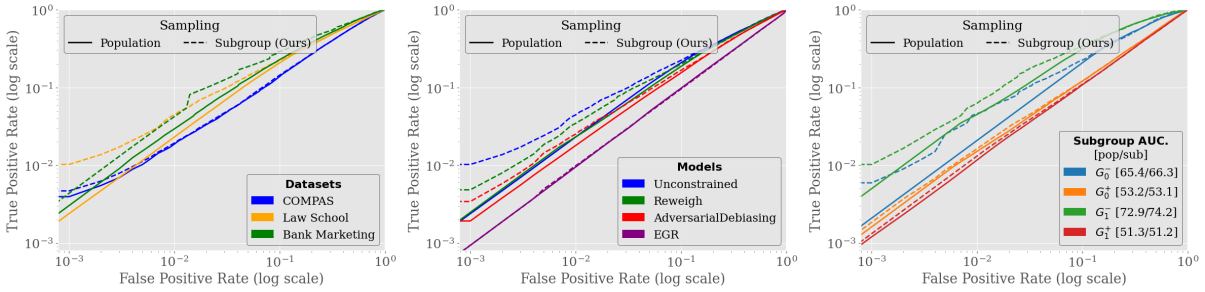


Fig. 1. **Log-scaled ROC curves** showing the trade-off between true positive rate (TPR) and false positive rate (FPR). **Left**: Unconstrained models across datasets show that our attack achieves higher TPR at low FPR than the standard RMIA [65] (based on 32 reference models and 75 pairwise comparisons per target instance). **Middle**: ROC comparisons on the Law School dataset show that fairness constraints do not increase overall privacy risk at low FPR, although subgroup-targeted attacks raise the relative risk for each algorithm. **Right**: Subgroup-specific ROC curves for the reweighing algorithm on the Law School dataset reveal that subgroups with negative class label can experience disproportionately severe leakage, further amplified by subgroup sampling. The curves are averaged over all attacks.

three standard benchmark datasets—Law School [61], Bank Marketing [22], and COMPAS [40]—commonly used in fair machine learning research. Our findings indicate that while fair learning does not inherently produce higher overall privacy leakage, different fairness algorithms can markedly increase membership leakage relative to each other. We also show that average-case attack success metrics mask critical subgroup disparities, underestimating leakage severity. Under low false positive rate evaluations, threshold-based loss attacks perform poorly, and population-based methods—despite detecting membership at 0% false positive rates— can still under-represent subgroup vulnerabilities. In contrast, our subgroup-focused attack, which samples comparison points within each protected group for pairwise likelihood ratio tests, outperforms alternative methods across most dataset and algorithm combinations, particularly for small, heterogeneous subgroups. This suggests that targeted attacks on specific subgroups can achieve higher attack power with lower error, underscoring our approach as a robust auditing tool for worst-case privacy evaluation in fairness-sensitive contexts.

2 Related Work

Our work intersects two main lines of research: *algorithmic fairness* and *membership inference attacks (MIA)*. Below, we review the foundations of group fairness, highlight how imposing fairness constraints can influence privacy, and then discuss the key developments in membership inference research.

2.1 Preliminaries

We consider binary classification with a deep neural network f parameterized by θ that maps inputs $x \in X$ to predictions $\hat{y} = f_\theta(x)$ given labels $y \in \{-,+\}$. The training set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^n$ is drawn i.i.d. from a distribution π. Training a model involves minimizing:

$$\mathscr{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(f_\theta(x_i), y_i\big), \tag{1}$$

over the training dataset $\mathscr{D}$, where ℓ is a chosen loss function. This formulation underlies empirical risk minimization, typically optimized using gradient-based methods such as stochastic gradient descent [3]. For brevity, we denote a trained model f_θ simply as θ throughout the paper.

2.2 Algorithmic Fairness

2.2.1 Group Fairness and Protected Attributes.

Fair machine learning seeks to mitigate discriminatory outcomes by reducing the undue influence of sensitive attributes (e.g., race, gender) on model predictions. In this context, every data sample x is paired with a protected attribute $g \in \mathscr{G}$ and a label y, forming the tuple (x, g, y). This formulation partitions the population into subgroups defined by both the protected attribute and the class label:

$$\mathscr{G}_g^y = \{(x, g, y) \in \mathscr{X} \times \mathscr{G} \times \mathscr{Y} \mid g = G, y = Y\}. \tag{2}$$

Among various fairness notions, group fairness metrics assess whether a model treats protected groups similarly. For instance, the Equalized Odds criterion— a measure of fairness we adopt in this paper—requires that false positive and false negative rates are similar across all protected subgroups. Formally, for any two groups $g, g' \in \mathscr{G}$, one can require that:

$$\begin{aligned} \max_{g,g' \in \mathscr{G}} \Big| \mathrm{FPR}_g(\theta) - \mathrm{FPR}_{g'}(\theta) \Big| &\leq \delta, \\ \max_{g,g' \in \mathscr{G}} \Big| \mathrm{FNR}_g(\theta) - \mathrm{FNR}_{g'}(\theta) \Big| &\leq \delta, \end{aligned} \tag{3}$$

where δ denotes the maximum allowable fairness gap. When $\delta = 0$, the model exhibits identical error rates across subgroups, achieving exact fairness. Alternative metrics—such as Disparate Impact [29] and Equal Opportunity [34]—focus on different error or outcome rates, reflecting distinct fairness philosophies. One approach, often encapsulated as "We Are All Equal," enforces uniform error rates, while another—"What You See Is What You Get"—allows models to acknowledge true group differences yet mandates fair error distribution [7, 34]. By choosing which type of error or outcome rate to equalize (selection rates, false positives, false negatives, etc.), we land on different group fairness criteria reflecting these distinct conceptions of equity [7, 34].

2.2.2 Fairness Interventions Have Asymmetric Privacy Implications

Fairness algorithms intervene at different stages of the machine learning pipeline to mitigate bias. *Pre-processing* approaches adjust training data before model training. Techniques like Disparate Impact Remover [29] and Optimized Preprocessing [8] alter the data itself to reduce algorithmic bias. *In-processing* methods integrate fairness constraints directly into training, often via reweighted training instances [37] or adversarial objectives [66], ensuring that the learned parameters yield comparable outcomes across protected groups. Finally, *post-processing* algorithms such as the Equalized Odds algorithm of [33] adjust a trained model's outputs (e.g. by calibrating decision thresholds) to satisfy fairness criteria without altering the underlying training process. While each of these interventions can reduce observed bias, they also introduce new challenges for privacy: customized data transformations, modified training objectives, and output adjustments have been argued to leave distinctive

"fingerprints" in models—potentially aiding adversaries in membership or attribute inference attacks [18, 19]. To illustrate this point, we follow experiments from Carlini et al. [14] and approximate two distributions for each data point: one when it is included in training—$Q_{\text{in}}(x,y)$—and one when it is excluded—$Q_{\text{out}}(x,y)$. We achieve this by training multiple models on random subsets that either include or omit (x,y), recording the raw logits from the penultimate linear layer. For each example, we evaluate two dimensions: (i) its behavior when excluded from training, quantified by the overlap coefficient (OC) [17] between the in-training and out-of-training logit distributions; and (ii) its ease of fit, measured by the extent of shift between these two distributions along the logit axis (visualized as the spread along the x-axis). A higher OC indicates greater similarity between the distributions, suggesting the example behaves as an inlier—while a lower OC signals more distinct distributions, characteristic of an outlier. Similarly, a larger shift between the "in" and "out" distributions—regardless of their overlap—indicates increased model confidence when the example is included in training, reflecting that it is easier to fit [14].

Fig. 2. Two instances from the Law School dataset —top is an outlier from the privileged group (race=*White*) and bottom in an inlier from the underprivileged group (race=*Black*). The blue and orange curves represent "in" vs. "out" logits distributions. The left plot shows the unconstrained model, while the others display models with fairness constraints. All models share the same architecture and hyperparameters, differing only in whether and which fairness constraint were applied. For each algorithm, 1,000 reference models were trained on random halves of the dataset (100 for Exponentiated Gradient Reduction, given its ensemble nature and higher training cost).

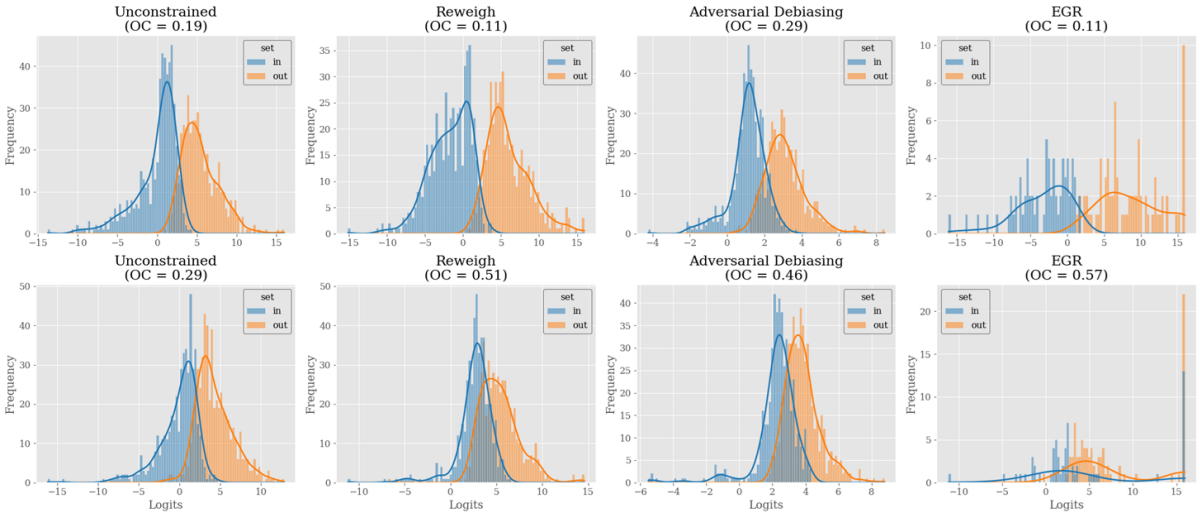


Figure 2 shows logit distributions for two Law School instances (both labeled "did not pass the bar"). In the unconstrained setting, one instance behaves as an inlier (bottom row) with low loss when excluded from training, while the other acts as an outlier (top row) with higher loss when excluded from training. As Feldman et al. [30] note, outliers can significantly influence learned parameters, whereas inliers have little effect. Fairness interventions modulate these behaviors: for the outlier, some methods reduce the overlap between "in" and "out" logits—sharpening its fingerprint—and techniques like adversarial debiasing and reweighing also narrow the x-axis

range, indicating a change in ease of fit. In contrast, for the inlier, all fairness interventions increase the overlap and reduce the range, leading to a less distinct fingerprint. These findings, consistent with Carlini *et al.*'s insights on example hardness, reveal that fairness constraints can alter both the separability and the ease of fit of instances.

3 Evaluating Privacy Leakage

Membership inference games are a standard method for quantifying privacy leakage by assessing how accurately an adversary can determine whether a specific data sample was part of a model's training set. This aligns with the definition of differential privacy [23, 25], which ensures that a model's behavior changes minimally when a single record is added or removed—thus limiting the adversary's ability to detect membership.

3.1 Membership Inference Games

We define membership inference using an indistinguishability game framework following the attack tempelate of [10, 14, 35, 54, 56, 63, 65].

DEFINITION 1 (MEMBERSHIP INFERENCE GAME). *Let π be a population distribution over $\mathscr{X} \times \mathscr{Y}$. The game between a challenger* C *and an adversary* A *proceeds as follows:*

(1) ***Training:*** *The challenger samples a dataset $D \sim \pi$ and trains a model $\theta \sim \mathscr{T}(D)$.*
(2) ***Membership Flip:*** *The challenger flips a private bit $b \sim \{0,1\}$. If $b = 1$, it selects a data point $(x,y) \in D$; otherwise, it samples $(x,y) \sim \pi$ such that $x \notin D$.*
(3) ***Challenge:*** *The challenger sends the trained model θ and the target data point x to the adversary.*
(4) ***Inference:*** *The adversary, with access to the population distribution π, computes a membership score, and outputs a membership prediction $\hat{b}$ for the target data point x.*

In practice, this game is repeated over multiple random draws of D and x, allowing the adversary to compute a membership score $\text{Score}_{\text{MIA}}(x;\theta)$ for each query. The adversary then applies a threshold β to decide whether x was in D, balancing the trade-off between attack power (true positive rate) and error tolerance (false positive rate) [10, 45, 54, 63, 65]:

$$\hat{b} = \begin{cases} 1, & \text{if } \text{Score}_{\text{MIA}}(x;\theta) \geq \beta \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

If an adversary reliably infers membership (i.e., $\hat{b} = b$) with high accuracy, the corresponding data point faces high privacy risk. We can thus quantify each data point's privacy risk as the adversary's success probability in correctly determining its membership over repeated queries [18].

3.2 Defining Privacy Risk for Fairness

A growing body of work shows that imposing fairness constraints can increase membership inference vulnerability for minority subgroups [18, 38, 67]. In particular, they show that if a group has fewer samples or higher variance, the model may "memorize" these points to reduce error disparities, making them more easily distinguishable to an attacker [18]. We adopt the *individual privacy risk* definition from [18]:

$$\mathrm{PR}(x,\theta,\pi) \;\; \max_{\mathsf{A}} \Big[\Pr\big(\text{Attack Game outputs 1}\big)\Big], \tag{5}$$

where x is a data point, θ is a model trained on a dataset drawn from π, and A ranges over possible adversaries. The probability is taken over all randomness in the attack game, including dataset sampling and adversarial decision-making. Similarly, subgroup privacy risk is the expected individual risk over points from a specific subgroup:

$$\mathrm{PR}(\mathscr{G}_g^y,\theta,\pi) \;\; \mathbb{E}_{x\sim\pi_g^y}\Big[\mathrm{PR}(x,\theta,\pi)\Big], \tag{6}$$

where π_g^y is the distribution of data belonging to subgroup g with label y. This captures the average membership exposure for individuals in that subgroup. However, because many empirical evaluations (e.g., [18, 58, 67]) quantify these definitions using average-case metrics like balanced accuracy for membership inference, they might miss severe leaks at low false-positive rates [14, 63, 65]. This is particularly concerning when fair training overfits to smaller or underprivileged subgroups: with fewer samples, the model may depend more heavily on these points to reduce error disparities [67], producing output distributions that attackers can exploit with a high degree of certainty.

Table 1. Summary of Notations

Symbol	Meaning
π	Population distribution from which data is drawn
$D \sim \pi$	Dataset sampled from population π
$(x,y) \in D$	Single data point from D
$g \in x$	Protected attribute in x
$\mathscr{X},\mathscr{Y},\mathscr{G}$	Feature, label, and protected attribute spaces
$\mathscr{G}_g^y$	Subgroup defined by cartesian product of g and y over instances

Symbol	Meaning
θ	Model with parameters θ
$\theta(x)$	Model output on x
$\ell(\theta(x),y)$	Loss on (x,y)
$\mathscr{T}(D) \to \theta$	Training procedure that learns θ from D
$Q_{\text{in}}(x,y)$	Loss distribution with (x,y)
$Q_{\text{out}}(x,y)$	Loss distribution without (x,y)
$Score_{MIA}(x,\theta)$	Membership inference score for x from attacking θ

3.3 Prior Membership Inference Attacks

Most membership inference attacks exploit a model's tendency to treat training examples differently than unseen data, particularly when overfitting leads to lower loss or higher confidence on training points. Different attacks can be categorized by an adversary's knowledge (black-box, white-box, or gray-box) and by their complexity (ranging from simple threshold-based strategies to sophisticated reference-model or gradient-based techniques). To align with prior research, we adopt a black-box scenario, meaning the adversary can observe only the model's outputs. At the same time, our threat model assumes that the adversary knows the training algorithm used and has access to the data distribution π from which the target model's training set is drawn [14, 18, 41, 56, 63–65].

3.3.1 Threshold-Based Attacks

A canonical approach is global thresholding [64], where a single threshold τ is set on the model's loss or confidence. For a target model θ and an input-label pair (x, y), the attack outputs 1 (membership) if $\ell(\theta(x), y) < \tau$, and 0 otherwise. In practice, the threshold is calibrated using the training set. While a single threshold may suffice when the data distribution is relatively homogeneous, it often fails to separate members from non-members in settings with group-fairness constraints or heterogeneous sub-populations. In such cases, extensions that calibrate subgroup-specific thresholds [18] can substantially boost attack accuracy by accounting for the distinct loss distributions across subgroups. Nonetheless, even these subgroup-specific approaches remain limited by their fixed decision boundaries, which assume a uniform effect of training on every example within a subgroup [14].

3.3.2 Reference Models

Reference-model strategies better capture per example variability [14, 56, 63, 65]. These methods often train N reference models on datasets $\{\mathscr{D}^{(i)}\}_{i=1}^{N}$, each drawn independently from a distribution π that approximates the target model's training data distribution. For each dataset $\mathscr{D}^{(i)}$, a model $\theta^{(i)}$ is trained. The adversary then queries each model with samples (x, y) whose membership in $\mathscr{D}^{(i)}$ is known, collecting feature vectors $\mathbf{x}_{\text{MIA}}$ (such as softmax confidence scores and true labels) that reveal memorization patterns like lower loss and higher confidence on training samples. In the original shadow-modeling approach [56], the adversary draws N shadow datasets from a distribution π_{shadow} that is disjoint from the target model's training data. Each shadow model is trained and queried with input-output pairs, assigning binary membership labels. These labeled feature vectors form the attack dataset $\mathscr{D}^{\text{MIA}}$, which is used to train an attack model that predicts membership. Later works [14, 63, 65] assume the adversary can sample reference datasets from the same underlying distribution as the target model—a setting we adopt in this study.

3.3.3 Per-Sample Hypothesis Testing

Recent approaches in membership inference focus on improving the power of the adversary (i.e. higher true positive rates) while reducing their error (i.e. at low false positive rates). This has led to designing attacks on a per-sample basis [10, 14, 53, 60, 63]. Notably, many works treat membership inference as a hypothesis testing problem, adapting to each sample's difficulty [14, 63, 65]. The Likelihood Ratio attack (**LiRA**) [14] applies a Neyman–Pearson test [47] using logit outputs of a target model, for a given input instance x. In an ideal setting, the attack would compute the likelihood ratio:

$$\text{LR}(x, y; \theta) \;=\; \frac{Pr\big(\theta \mid Q_{\text{in}}(x, y)\big)}{Pr\big(\theta \mid Q_{\text{out}}(x, y)\big)}, \tag{7}$$

but as these densities are intractable, LiRA approximates the test by evaluating the logit outputs directly. Specifically, N reference models are trained on datasets drawn from π, with $N/2$ models including (x, y) and $N/2$ excluding it. Parametric distributions $\hat{Q}_{\text{in}}(x, y)$ and $\hat{Q}_{\text{out}}(x, y)$ (typically Gaussian) are then fitted to these logits, and the likelihood ratio is computed as:

$$\text{LR}(x, y; \theta) \;=\; \frac{Pr\big(\theta(x) \mid \hat{Q}_{\text{in}}(x, y)\big)}{Pr\big(\theta(x) \mid \hat{Q}_{\text{out}}(x, y)\big)}. \tag{8}$$

A higher likelihood ratio indicates a higher chance that (x,y) was in the training set of a target model.[2]

Extending this per-sample hypothesis testing framework, the Pairwise Likelihood Ratio attack (**RMIA**) [65] adopts a Bayesian approach. It compares the output probabilities (e.g. softmax scores) from the target model on a candidate instance x against multiple replacement samples z drawn from population π under the null hypothesis (i.e., points not included in the training set). The core test relies on a pairwise likelihood ratio comparison:

$$\mathrm{LR}(x,z;\theta) = \frac{\Pr(x \mid \theta)/\Pr(x)}{\Pr(z \mid \theta)/\Pr(z)}, \tag{9}$$

were $\Pr(x \mid \theta)$ is computed as the probability that model θ assigns to the input x. The marginal probability $\Pr(x)$, which serves as a normalization factor, is estimated by averaging the predicted probabilities $\Pr(x \mid \theta')$ over a collection of reference models θ'. The same approach is applied to replacement points z. Similar to [14], each model has x either included or excluded from their training set. Since models exhibit higher confidence on data seen during training, the estimates from models that did not train on x are scaled to approximate the higher probability $\Pr(x)_{\mathrm{IN}}$ that would be observed if x were trained on. The pairwise likelihood ratio test aggregates evidence by defining the membership score as:

$$\mathrm{LR}(x;\theta) = \Pr_{z\sim\pi}\Big(\mathrm{LR}(x,z;\theta) \geq \gamma\Big), \tag{10}$$

which is the fraction of replacement samples for which the likelihood ratio exceeds a threshold γ. Per sample attacks can be mounted in either online or offline settings. Offline attacks only pre-train reference models on datasets excluding a candidate sample x, whereas online attacks assume query access to the data point being targeted, while training reference models. In our work, we evaluate both LiRA[14] and RMIA[65] in the offline setting. Given that online attacks assume the ability to train models on query points being attacked, they are often deemed less feasible in realistic settings [65].

3.4 Subgroup-Specific Privacy Risk

Building on the pairwise likelihood test proposed by [65], we tailor membership inference to a target sample's *own subgroup* by restricting replacement points z to those sharing the same protected attribute and target label as x. Concretely, if x is drawn from subgroup $\mathscr{G}_g^y$, we sample $z \sim \pi_{\mathscr{G}_g^y}$ rather than from the overall population π. We still test the null hypothesis that "x is not in the training set," but this subgroup-based strategy ensures "null-hypothesis plausibility" by comparing x only with truly comparable points. Selecting z from the same protected attribute and label distribution mitigates confounding effects that could arise if differences in label distribution or attribute prevalence artificially inflate or deflate $\mathrm{LR}(x,z;\theta)$. This approach is particularly important in fairness-constrained settings, where minority or underrepresented groups are more susceptible to overfitting [18, 67]. Sampling unrestricted from π may dilute subgroup-specific signals, whereas limiting z to $\pi_{\mathscr{G}_g^y}$ amplifies them. If the model memorizes x to reduce subgroup-level errors, then $\mathrm{LR}(x,z;\theta)$ will be significantly higher for most $z \in \mathscr{G}_g^y$, thereby boosting detection at low false positive rates [14, 63, 65]. Crucially, we argue that with this formulation, the observed effect on the likelihood ratio genuinely reflects within-group memorization. Consequently, subgroup-based pairwise comparisons can allow pinpointing where memorization is most likely to occur, thus revealing membership vulnerability precisely in subpopulations where fairness constraints may inadvertently amplify overfitting.

[2]In the offline setting, inference can be done using a one-tailed test of the non-member logit distribution.

4 Empirical Analysis

Experimental Setup. We conduct experiments on three benchmark datasets for fairness in binary classification: the COMPAS dataset [40] (10 features, 5,278 datapoints), the Law School dataset [61] (11 features, 20,512 datapoints), and the Bank Marketing dataset [22] (27 features, 39,569 datapoints). Preprocessing follows the methodology of IBM's AI Fairness 360 [9]. For the COMPAS and Law School datasets, we treat race (white versus non-white) as the protected attribute. The COMPAS dataset encodes favorable outcome as class 0 ("did not recidivate"), whereas in the Law School dataset, favorable outcome ("pass Bar exam") is encoded as class 1. For the Bank Marketing dataset, the protected attribute is defined based on age (age ≥ 25 versus age < 25), with the favorable outcome (i.e. a client "subscribed to a term deposit") also encoded as class 1 [18]. Our experiments compare three open-source fairness algorithms—Reweighing [36], Adversarial Debiasing [66], and Exponentiated Gradient Reduction [2]—each chosen for their model-agnostic nature and their ability to enforce equality of odds [33], making them directly comparable. We train fully connected neural network models with three hidden layers (of sizes 200, 100, and 50), with ReLU activations and a sigmoid output [46, 52]. Both unconstrained and fairness-constrained models are trained with hyperparameters tuned individually for each dataset (see Appendix Table 10). Performance is evaluated using standard classification metrics and average odds difference in outcomes, which quantifies disparities in false positive and false negative rates across the protected groups.

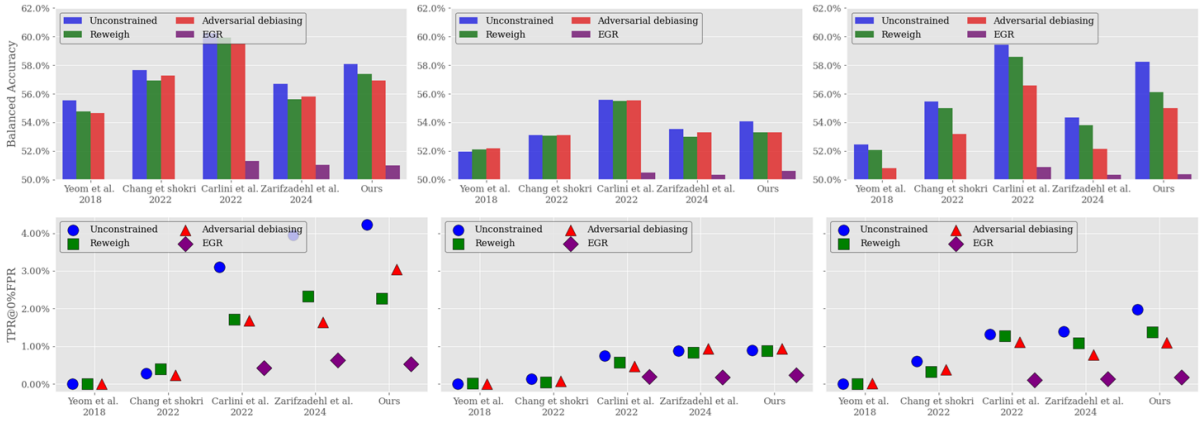
Table 2. Performance metrics (expressed as percentages) and fairness—quantified by average odds difference (AOD, raw scale)—are reported as averages over a pool of 100 models for each dataset–algorithm pair.

Dataset	**Method**	**Performance**				**Fairness**	
		Accuracy		**F1**		**AOD**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**
COMPAS	Unconstrained	67.4	65.9	66.9	65.4	-0.157	-0.163
	Adversarial	67.1	65.5	66.6	65.0	0.011	0.005
	Reweigh	67.0	65.4	66.6	65.0	0.018	0.009
	EGR	66.9	65.4	66.5	65.0	-0.003	-0.015
Law school	Unconstrained	95.9	94.7	68.7	59.5	-0.155	-0.122
	Adversarial	95.4	94.9	59.8	55.8	-0.061	-0.043
	Reweigh	95.7	94.7	64.2	56.1	0.009	-0.018
	EGR	96.0	94.5	68.5	57.3	-0.01	-0.043
Bank Marketing	Unconstrained	94.0	92.1	72.6	64.3	0.208	0.159
	Adversarial	93.9	92.1	72.1	63.9	-0.141	-0.081
	Reweigh	93.9	92.1	72.4	64.4	0.032	0.037
	EGR	94.3	91.8	75.3	64.6	0.004	0.011

Reference Model Training Following prior work [14, 18, 30, 56], we split each dataset into two equal random parts—one for training (in-distribution) and one for evaluation (out-of-distribution)—to create a pool of reference models. This process was repeated n times, resulting in multiple models trained on distinct, randomly masked subsets (Table 2). Robust subgroup representation is vital for both fair model training and accurate privacy risk assessment. To ensure this, we performed stratified random sampling by partitioning the datasets based on the intersection of a binary sensitive attribute (e.g., race) and a binary target label (e.g., passing the bar exam in Law School).

Membership Inference Games We evaluated model vulnerability using the indistinguishability game outlined in Section 3.1, implementing several membership inference attacks from the literature [14, 18, 64, 65]. In each iteration, one model is randomly chosen as the target while the remaining models form the reference pool. The target model is queried on every sample from the population π to obtain its black-box outputs, used by the attacks to compute their respective membership scores. For every dataset–algorithm pair, each membership inference attack is executed on 10 distinct target models.

Fig. 3. Comparison of average and worst-case attack performance across all strategies. Average performance is quantified by balanced accuracy (top panel), while worst-case performance is measured by the true positive rate at a 0% false positive rate (bottom). Results are shown for the **COMPAS** (left), **Law School** (middle), and **Bank Marketing** (right) datasets.

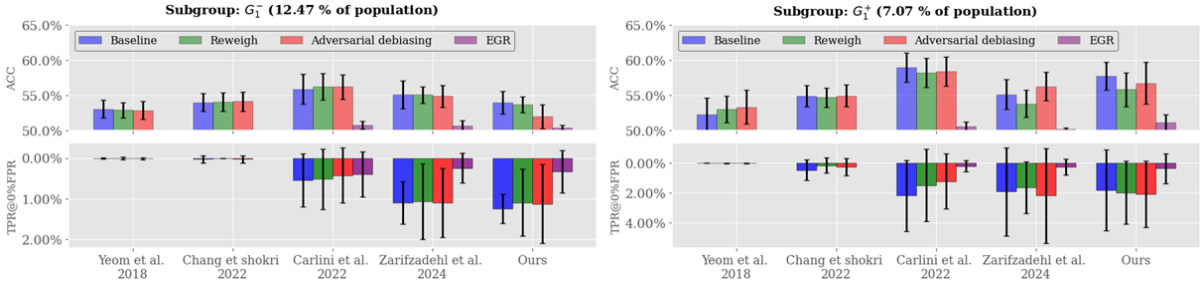


For our main experiments, we use 32 reference models per target query, following recommendations from previous work [14, 65]. Each reference and target model outputs a confidence signal for a given target query. For the Pairwise Likelihood Ratio attack [65], this is the classifier's sigmoid output for the true label (representing the correct-class probability) [65]. For the Likelihood Ratio attack, we use the logit for the true label as the confidence measure [14]. Additionally, we conducted loss-based attacks by calibrating membership thresholds—set to maximize inference accuracy—using the target model's training losses, both globally and within subgroups [18, 64]. Attack performance metrics—balanced accuracy and true positive rate at 0% false positive rate—were computed globally and within each protected subgroup, with results averaged across all target models. As shown in Figure 3, the Likelihood Ratio attack achieves the highest average attack success across all conditions, while our approach yields slightly greater attack power at the lowest error rates. Notably, fair training via enforcing equality of odds [33] does not lead to improved attack success, as measured by either balanced accuracy or true positive rate at 0% false positive rate. In most cases, unconstrained models have privacy leakage either at par or higher than fair models. Interestingly, Exponentiated Gradient Reduction consistently achieves the lowest privacy leakage across all datasets and attack types. We attribute this benefit to the probabilistic ensemble approach this algorithm takes, sampling from a distribution over trained classifiers to satisfy fairness constraints and predictive performance [2]. This singles out this fairness algorithm among the ones surveyed, offering privacy by design.

5 Comparing Subgroup-Tailored Attacks

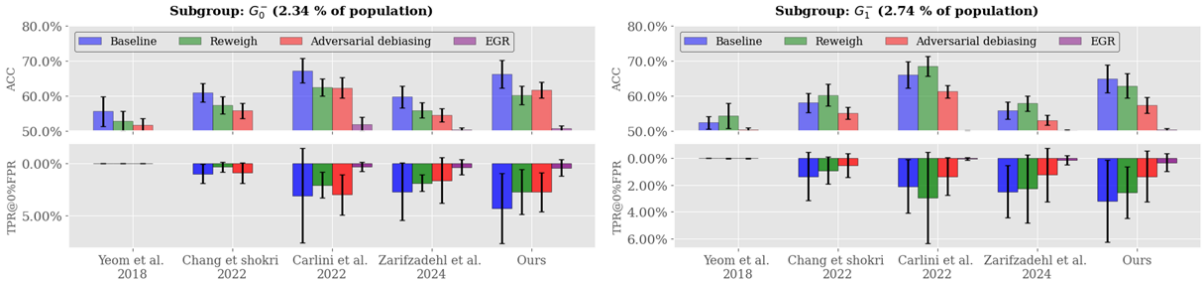
Next, we explore optimizing membership inference attacks on a per-subgroup basis, as suggested in [18]. Both the population sampling approach and our subgroup sampling variant allow tuning the parameter triple (z, a, γ): z specifies the number of random replacement points, a scales the model's output probabilities to calibrate the test statistic, and γ sets the minimum likelihood ratio threshold for classifying a data point as a member. For each subgroup, we select the triple that maximizes the true positive rate at 0% false positive rate for both the standard pairwise likelihood ratio attack [65] and our subgroup variant.

Fig. 4. Attack success for most vulnerable subgroups on the **COMPAS** dataset. Error bars denote standard deviation across all models attacked.



We present results for the most vulnerable subgroups as a best-case upper bound, and refer interested readers to the Appendix A for results on less vulnerable subgroups. Our results show that threshold-based attacks, whether using global loss [64] or subgroup-specific losses [18], often fail to capture both average and worst-case leakage.

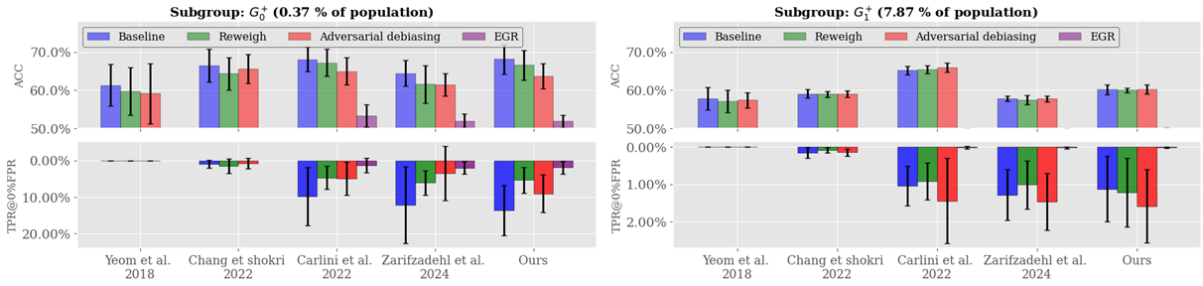
Fig. 5. Attack success for most vulnerable subgroups on the **Law School** dataset.



In contrast, the most potent attacks yield higher average accuracy as well as robust low false positive rate performance across subgroups, corroborating prior findings [16, 63, 65]. Figure 4 shows the most vulnerable subgroups in COMPAS, highlighting that our attack outperforms the Likelihood Ratio attack in the low false positive rate regime while maintaining true positive rates comparable to the pairwise likelihood ratio attack of [65]. In the Law School dataset (Figure 5), likelihood ratio attacks perform best on subgroup G_1^- when fairness

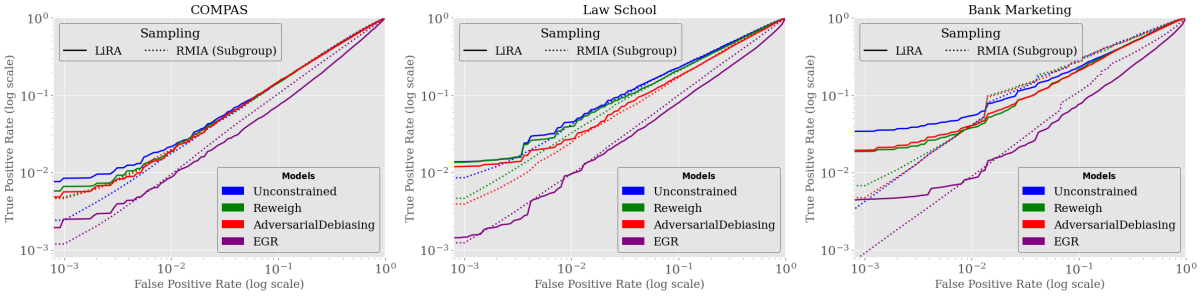
constraints are enforced via the Reweighing algorithm, while our attack works better on unconstrained models across all subgroups. In the Bank Marketing dataset (Figure 6), our method achieves superior performance on subgroup G_0^+ for unconstrained and Adversarial Debiasing models, maintaining average accuracy that rivals or exceeds other methods. Notably, models trained with Adversarial Debiasing consistently leak more information about subgroup G_1^+. This suggests that fairness constraints can sometimes increase leakage, although they generally have a regularizing effect on models, thereby reducing privacy risks [24].

Fig. 6. Attack success for most vulnerable subgroups on the **Bank Marketing** dataset.



Overall, our subgroup-tailored attack delivers comparable or enhanced performance at low error rates using a modest computational and sampling budget—only 32 offline reference models with 5–75 samples per target. These results are promising compared to prior work on the CIFAR-10 dataset (60,000 samples), which used 127 reference models and up to 25,000 samples per target. Particularly, increasing population samples from 25 to 25,000 improved the AUC from 58.7 to 69.1 in [65], suggesting that adversaries with larger compute budgets and more data could achieve even better results with our attack. While our results show that subgroup-specific tuning of the parameter triple (z, a, γ) can outperform prior methods, we also evaluated a more practical single-global scenario where the same parameters are applied uniformly across subgroups. In this one-shot configuration (Figure 7),

Fig. 7. Comparison of LiRA with a subgroup-tailored RMIA that omits per-subgroup tuning of parameters a, γ, and the number of z samples. Both attack types are evaluated using 32 reference models.



our subgroup sampling method under-performs relative to the likelihood ratio attack—likely due to the limited number of reference models and random replacement samples z used in the pairwise comparisons. Nonetheless, our

findings demonstrate that even with modest resources, tailoring attack parameters to specific subgroups enhances performance compared to a one-size-fits-all approach, and allows for fine grained audits via examining attack parameters that yielded best performance for different subgroup.

5.1 Sensitivity to Attack Parameters

To expose worst-case leakage, we explore the effects of tuning key attack parameters—namely, the scaling parameter a and the reference samples z. Restricting z to samples within the same subgroup yields sharper comparisons and higher true positive rates at 0% false positive rates. For the Bank Marketing dataset, Table 3 shows that for subgroup $\mathscr{G}_0^+$, when reference samples are drawn exclusively from the subgroup, attacks on unconstrained models reach an average true positive rate of 13.6%, while making no mistakes. Similarly, attacks on Adversarial Debiasing models improves from 3.5% to 9.1%.

Table 3. Parameter configurations and attack performance for most vulnerable subgroups in **Bank Marketing**. TPR is reported at 0% FPR

		$\mathscr{G}_1^+$					$\mathscr{G}_0^+$				
Sampling	**Method**	z	a	γ	TPR (%)	ACC (%)	z	a	γ	TPR (%)	ACC (%)
Population (π)	Unconstrained	75	0.1	2.0	1.3 ± 0.7	57.8 ± 0.7	75	0.6	2.0	12.2 ± 10.5	64.5 ± 3.4
	Reweigh	50	0.1	2.0	1.0 ± 0.7	57.5 ± 1.2	75	0.3	2.0	6.1 ± 3.3	61.6 ± 4.9
	Adversarial Debiasing	50	0.0	1.8	1.5 ± 0.8	57.8 ± 0.8	75	1.0	2.0	3.5 ± 7.4	61.5 ± 2.9
	EGR	75	1.0	0.1	0.0 ± 0.0	50.0 ± 0.0	25	1.0	1.2	2.0 ± 1.7	52.0 ± 1.8
Subgroup ($\pi_{G_g^y}$)	Unconstrained	50	0.2	2.0	1.1 ± 0.9	60.2 ± 1.2	75	0.9	2.0	13.6 ± 6.9	68.2 ± 3.9
	Reweigh	50	0.2	2.0	1.2 ± 0.9	60.0 ± 0.7	75	0.9	2.0	5.4 ± 3.5	66.6 ± 3.9
	Adversarial Debiasing	75	0.2	2.0	1.6 ± 1.0	60.2 ± 1.2	75	1.0	2.0	9.1 ± 5.2	63.8 ± 3.3
	EGR	5	0.0	2.0	0.0 ± 0.0	50.1 ± 0.2	25	0.1	1.8	1.9 ± 1.7	51.9 ± 1.6

We also see that larger subgroups (e.g., $\mathscr{G}_1^+$) reach peak accuracy with fewer samples for the pairwise comparison, whereas smaller subgroups (e.g., $\mathscr{G}_0^+$) require a higher number of samples z to build sufficient evidence for membership (Table 3). Moreover, for the Exponentiated Gradient Reduction (EGR) algorithm, we observe that increasing the number of pairwise comparisons does not further improve the attack's performance, unlike other algorithms that use the full sampling budget. This suggests that adding more samples may introduce conflicting evidence from different models within the ensemble, thereby diluting the overall attack signal. Finally, in our offline setting, the parameter a approximates the gap between models that did and did not train on a target instance. We estimate $\Pr(x)_{\text{IN}}$ as a linear function of $\Pr(x)_{\text{OUT}}$ via :

$$\Pr(x)_{\text{IN}} = a \Pr(x)_{\text{OUT}} + (1 - a) \tag{11}$$

Following [65], we define $\Pr(x) = \frac{1}{2}\Big(\Pr(x)_{\text{IN}} + \Pr(x)_{\text{OUT}}\Big)$, where a lower a may indicate that a substantial upward adjustment is needed—potentially reflecting a wider gap between IN and OUT probabilities and suggesting stronger overfitting. Conversely, a higher a (closer to 1) could imply that the OUT probabilities are already nearly representative of the IN probabilities, which might indicate less overfitting. This tunable sensitivity thus positions the attack as an auditing tool for probing differences across subgroups and between algorithms. For instance, we see that subgroup $\mathscr{G}_1^+$ achieves optimal attack performance at a values around 0.1–0.2, which could be interpreted as a sign of relatively higher overfitting across model types. In contrast, subgroup $\mathscr{G}_0^+$ appears to require less scaling of the OUT probabilities. Moreover, for $\mathscr{G}_0^+$, a lower optimal a when sampling from the full population versus a higher a when sampling solely from the subgroup might suggest that—despite a general need for significant scaling—the degree of overfitting is comparatively reduced within the subgroup, for the Reweighing algorithm. Alternatively, variations in a across subgroups may also indicate inherent differences in the model's confidence levels for them. These observations offer clues for stakeholders to further investigate the underlying causes.

6 Discussion

Our findings suggest that bias mitigation is neither universally detrimental nor uniformly protective, in terms of privacy. Rather, the effect depends on how subgroup-level overfitting interacts with a given fairness algorithm. Specifically, when a model is trained under certain fairness constraints, it may allocate more capacity to fitting one subgroup's data—either because the subgroup is underrepresented and needs additional focus to meet fairness goals, or because the fairness algorithm penalizes errors for that subgroup more heavily. This increased focus can inadvertently cause subgroup-level overfitting, where the model memorizes or relies too strongly on that subgroup's training examples. In turn, membership inference attacks become more effective against that subgroup, since overfitted patterns are easier to detect. On the other hand, alternative fairness methods—or different subgroup characteristics—may actually reduce the extent of overfitting, thereby mitigating membership leakage. The overall privacy impact thus depends on the specific mechanics of each fairness strategy and how it interacts with subgroup size, diversity, and loss penalization. Furthermore, the performance of subgroup-specific attacks suggests that alternative strategies—such as sampling z based on established similarity metrics between the target point x and the adversary's available data pool Z—may further improve the efficacy of membership inference attacks. Our work thus highlights a practical avenue for more targeted auditing and provides a methodological foundation for identifying when and why certain fairness strategies can inadvertently exacerbate privacy risks.

7 Conclusion

Overall, these findings offer promising insights for institutions seeking to integrate fairness in their machine learning pipelines. Fairness constraints can regularize models to reduce privacy leakage, and some algorithms even provide privacy by design. We have shown that subgroup-specific membership inference can reveal hidden vulnerabilities often overlooked by prior approaches, particularly at low false positive rates. From an **auditing** perspective, subgroup-aware tests are essential to avoid underestimating privacy risks in underrepresented groups. For **attack design**, our framework demonstrates that, given the same computational and data resources, focusing on specific subgroups boosts adversarial success rates. This targeted approach may pave the way for subgroup-specific data reconstruction attacks. Meanwhile, **fair model selection** must account for how different fairness algorithms can variably protect or expose specific subgroups, highlighting the importance of subgroup-level evaluations to

prevent unintended harm. Finally, as attacks in literature improve, we hope our work makes model developers more conscious of the true privacy risks of algorithmic fairness, and their non symmetric peculiarities across subgroups.

Response to Reviewer Comments

We sincerely thank all reviewers for their valuable feedback and for taking the time to evaluate our work. Their thoughtful input has strengthened the articulation of our contributions, improved the coherence of our exposition, and helped define the scope of this and subsequent work more precisely.

Clarifying the Main Contribution and Narrative Flow We appreciate the observation regarding the clarity and focus of the manuscript. To address this, we revised relevant sections to clearly delineate the paper's central contribution: studying the impact of fairness interventions on membership inference privacy risks, specifically emphasizing evaluation in low false-positive rate regimes. We hope to have better contextualized our work with respect to existing literature, explicitly highlighting how our experiments differ and how our subgroup-specific membership inference attack advances prior methods by enabling fine-grained privacy analysis at very low false-positive rates.

Relationship Between Fairness and Privacy We agree that separating the discussion on how fairness impacts privacy versus how privacy influences fairness would improve clarity. We have restructured the manuscript to more clearly distinguish between these two perspectives and to emphasize that our analysis focuses on the former.

Choice of Fairness Criterion and Model Generalizability Regarding our choice of fairness criterion, we focus on equalized odds because the fairness algorithms used in this study allow explicitly enforcing this constraint, enabling a direct and consistent comparison of privacy impacts across different fairness interventions. We acknowledge that the selection of fairness metrics can be context-dependent and somewhat arbitrary. To ensure broader applicability, we plan to extend our analysis in future work by evaluating additional fairness criteria and incorporating a more diverse set of learning algorithms, thereby deepening the understanding of the interplay between fairness and privacy across various settings.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*. PMLR, 60–69.

[3] Shun-ichi Amari. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing* 5, 4-5 (1993), 185–196.

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

[5] Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. 2021. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250* (2021).

[6] Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1138–1156.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[10] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. 2024. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[14] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1897–1914.

[15] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*. 267–284.

[16] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[17] Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 2 (2007), 1.

[18] Hongyan Chang and Reza Shokri. 2021. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 292–303.

[19] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*. 309–315.

[20] Amit Daniely. 2020. Neural networks learning and memorization with (almost) no over-parameterization. *Advances in Neural Information Processing Systems* 33 (2020), 9007–9016.

[21] Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447* (2021).

[22] D. Dua and C. Graff. 2017. UCI Machine Learning Repository. Online. http://archive.ics.uci.edu/ml

[23] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[25] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2016. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality* 7, 3 (2016), 17–51.

[26] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. 2015. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE, 650–669.

[27] Lilian Edwards. 2021. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)* 1 (2021).

[28] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*. 15–19.

[29] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[30] Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems* 33 (2020), 2881–2891.

[31] Ferdinando Fioretto, Cuong Tran, and Pascal Van Hentenryck. 2021. Decision making with differential privacy under a fairness lens. *arXiv preprint arXiv:2105.07513* (2021).

[32] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

[33] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[34] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *ArXiv* abs/1610.02413 (2016).

[35] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (2008), e1000167.

[36] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[37] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*. IEEE, 924–929.

[38] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. 2019. Disparate vulnerability to membership inference attacks. *arXiv preprint arXiv:1906.00389* (2019).

[39] Jeff Larson. 2016. *ProPublica COMPAS dataset*. https://github.com/propublica/compas-analysis

[40] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. 2017. COMPAS Dataset. Online. https://github.com/propublica/compas-analysis

[41] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2020. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 521–534.

[42] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).

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

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[45] Sasi Kumar Murakonda, Reza Shokri, and George Theodorakopoulos. 2021. Quantifying the privacy risks of learning high-dimensional graphical models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2287–2295.

[46] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[47] J Neyman and ES Pearson. 2017. On the Problem of the Most Efficient Tests of Statistical.

[48] Milad Nasr Shuang Song Andreas Terzis Nicholas Carlini, Steve Chien and Florian Tramèr. 2022. *Likelihood Ratio Attack*. https://github.com/tensorflow/privacy/tree/master/research/mi_lira_2021

[49] Sophie Noiret, Jennifer Lumetzberger, and Martin Kampel. 2021. Bias and fairness in computer vision applications of the criminal justice system. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.

[50] Microsoft (open sourced). 2020. Fairlearn. *Online*. https://github.com/fairlearn/fairlearn Accessed: 13 Nov. 2024.

[51] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review* 2, 1 (2020), 1.

[52] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.

[53] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*. PMLR, 5558–5567.

[54] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature genetics* 41, 9 (2009), 965–967.

[55] Reza Shokri. 2024. *Robust Membership Inference Attack*. https://github.com/privacytrustlab/ml_privacy_meter/tree/d32734161a3395211fe5f3cd461932290b1fafbe/research/2024_rmia

[56] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[57] LSAC National Longitudinal Bar Passage Study. 1997. *Law School Admissions Council dataset*. https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage

[58] Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. 2023. When fairness meets privacy: exploring privacy threats in fair binary classifiers via membership inference attacks. *arXiv preprint arXiv:2311.03865* (2023).

[59] Peter M Visscher and William G Hill. 2009. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS genetics* 5, 10 (2009), e1000628.

[60] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2021. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440* (2021).

[61] L. F. Wightman and H. Ramsey. 1998. Law School Admission Council. Available from LSAC.

[62] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 506–523.

[63] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 3093–3106.

[64] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.

[65] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. [n. d.]. Low-cost high-power membership inference attacks, 2024. *URL https://arxiv.org/abs/2312.03262* ([n. d.]).

[66] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[67] Da Zhong, Haipei Sun, Jun Xu, Neil Gong, and Wendy Hui Wang. 2022. Understanding disparate effects of membership inference attacks and their countermeasures. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. 959–974.