When Fairness Meets Privacy: Exploring Privacy Threats in Fair Binary Classifiers via Membership Inference Attacks

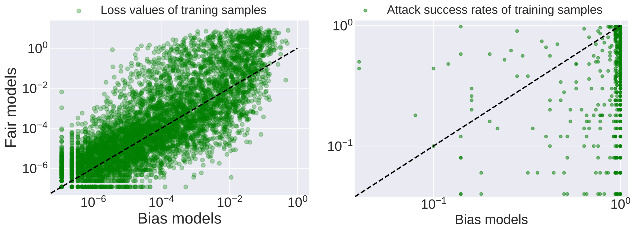
Huan Tian[1] , **Guangsheng Zhang**[1] , **Bo Liu**[1] , **Tianqing Zhu**[2] * , **Ming Ding**[3] and **Wanlei Zhou**[2]
[1]University of Technology Sydney
[2]City University of Macau
[3]Data61 CSIRO

Abstract

While in-processing fairness approaches show promise in mitigating bias predictions, their potential impact on privacy leakage remains underexplored. We aim to address this gap by assessing the privacy risks of fairness-enhanced binary classifiers with membership inference attacks (MIAs). Surprisingly, our results reveal that these fairness interventions exhibit increased resilience against existing attacks, indicating that enhancing fairness does not necessarily lead to privacy compromises. However, we find current attack methods are ineffective as they typically degrade into simple threshold models with limited attack effectiveness. Following this observation, we discover a novel threat dubbed **F**airness **D**iscrepancy **M**embership **I**nference **A**ttacks (FD-MIA) that exploits prediction discrepancies between fair and biased models. This attack reveals more potent vulnerabilities and poses significant privacy risks to model privacy. Extensive experiments across multiple datasets, attack methods, and representative fairness approaches confirm our findings and demonstrate the efficacy of the proposed attack method. Our study exposes the overlooked privacy threats in fairness studies, advocating for thorough evaluations of potential security vulnerabilities before model deployments.

1 Introduction

Imbalanced datasets often induce spurious correlations between learning targets and sensitive attributes [Mehrabi *et al.*, 2021], which will lead to biased predictions in trained (*biased*) models. In response, fairness research has developed in-processing approaches [Wang *et al.*, 2022; Ching-Yao Chuang, 2021] that are applied during training to produce fairness-enhanced (*fair*) models. By effectively suppressing these spurious correlations, fair models can mitigate discriminatory predictions. However, despite promising to enhance fairness, these interventions might incur potential privacy risks, such as unintended training data memorization.



(a) Loss values

(b) Attack success rates

Figure 1: Fairness interventions (a) increase the losses and (b) decrease attack success rates for most training samples. We generate the plots with 100 runs and report the training samples' mean loss value and mean attack success rate.

Membership inference attacks (MIAs) are widely adopted for assessing privacy risks in models that are deployed via Machine Learning as a Service (MLaaS) [Shokri *et al.*, 2017]. Building upon this, prior work [Chang and Shokri, 2021] has primarily explored the privacy impact of fairness interventions on decision tree models by applying existing score-based attacks to them. However, this leaves open questions on evaluating privacy risks for neural networks, such as binary classifiers, which are prevalent in fairness studies. Moreover, evaluations relying on one attack method might not fully characterize the privacy impact. Given the importance of trustworthiness in fairness studies, conducting thorough evaluations of these fairness interventions is non-trivial. To bridge these gaps, we apply different existing MIAs to binary classifiers and discover an innovative attack method designed for fair models, comprehensively evaluating the privacy risk.

With existing attack methods, our evaluation results show that fairness-enhanced models show *more resilience* to *current* MIAs than their biased counterparts. Figure 1 illustrates the results of score-based attacks on fair and biased models over 100 runs. We present the loss values (Figure 1a) and the attack success rate (Figure 1b) per training sample. After applying fairness interventions, the plots show increased loss values yet decreased attack success rates for most data points. This indicates that these interventions can lead to less successful attacks with existing MIA approaches.

Before rushing to the conclusion that fairness interventions are privacy-friendly to binary classifiers, with further analy-

*corresponding author

ses, we find that these methods are inefficient in attacking binary classifiers. This is because the trained attack models degrade into simple threshold-based decisions due to the binary outputs. The degradation incurs substantial performance trade-offs: while effective at recognizing member data, attack models struggle with non-member data, especially for hard examples where predictions are similar across groups.

During the evaluation of existing attacks, we have uncovered a potential threat that could enable more effective attacks on binary classifiers. Specifically, we find the prediction scores for member and non-member data exhibit divergent behaviors after fairness interventions: the scores typically increase for the majority of member data, while they conform to a normal distribution for non-member data. This disparity creates a pronounced prediction gap between groups, which is overlooked by current attacks. The widened gap, if exploited by adversaries, could enable more successful attacks, thereby posing substantial privacy threats.

Inspired by these observations, we name the identified threat as Fairness Discrepancy MIAs (FD-MIA), which targets fair models by exploiting the prediction gaps between the original (biased) and fairness-enhanced (fair) models. The key intuition is that these gaps leak membership information about the training data, which can be leveraged to launch more effective attacks. It can be integrated with the existing frameworks of score-based [Liu *et al.*, 2022] and reference-based attacks [Carlini *et al.*, 2022].

We conduct comprehensive evaluations across six datasets, using three attack methods and five in-processing fairness approaches. This amounts to 32 different experimental settings and over 160 distinct models. Our results reveal that fairness interventions potentially introduce *new threats* to model privacy, advocating for a more comprehensive examination of their potential security defects before deployment. Our main contributions are as follows: (1) To the best of our knowledge, this is the first work to comprehensively study the impact of fairness interventions on privacy through the lens of MIAs, targeting deep classifiers with real-world datasets. (2) We reveal that fairness interventions do not compromise model privacy with *existing* attack methods, primarily due to their limited efficacy in attacking binary classifiers. (3) We discover a novel attack method, FD-MIA, which poses significant threats to model privacy by exploiting prediction gaps from both biased and fair models. It can be integrated into existing attack frameworks. (4) Extensive experiments validate our findings and demonstrate FD-MIA's effectiveness.

2 Related Work

Algorithmic fairness. Fairness methods are typically categorized into pre-processing, in-processing, and post-processing approaches based on their processing stage. We focus on in-processing approaches as they modify model training procedures directly and may introduce model privacy threats. Among them, some introduce fair constraints and formulate the issues as optimization problems [Zemel *et al.*, 2013; Manisha and Gujar, 2020; Tang *et al.*, 2023; Truong *et al.*, 2023; Cruz *et al.*, 2023; Jung *et al.*, 2023]. Some propose adversarial designs to remove sensitive information among extracted features [Kim *et al.*, 2019; Zhu *et al.*, 2021; Creager *et al.*, 2019; Park *et al.*, 2021]. Some adopt data sampling [Roh *et al.*, 2021] or reweighting [Chai and Wang, 2022] approaches to alleviate the biased predictions. More recently, studies learn fair representations using mixup augmentations or contrastive learning mechanism [Ching-Yao Chuang, 2021; Du *et al.*, 2021; Park *et al.*, 2022; Wang *et al.*, 2022; Zhang *et al.*, 2023; Qi *et al.*, 2022]. These methods interpolate inputs or modify features to pursue fair representations.

Membership inference attacks. MIAs aim to determine if a data sample was part of a target model's training set. Some attacks leverage the target model's direct output, such as confidence scores [Shokri *et al.*, 2017], losses [Sablayrolles *et al.*, 2019], prediction labels [Choquette-Choo *et al.*, 2021]. Some improve the performance by modeling the prediction distributions of the target model, such as reference models [Carlini *et al.*, 2022; Ye *et al.*, 2022]. Others extend their focus into new settings [Gao *et al.*, 2023; Yuan and Zhang, 2022] or work on defense methods [Yang *et al.*, 2023]. This work considers two representative attack approaches: score-based [Liu *et al.*, 2022] and reference-based [Carlini *et al.*, 2022] attack methods. More recently, studies have enhanced attack performance by exploiting additional information: some [He *et al.*, 2022] leverage predictions from multiple augmented views, some [Li *et al.*, 2022] require results from multi-exit models, and others [Hu *et al.*, 2022] work on multi-modality models. We explore models with fairness discrepancies.

Exploring privacy impacts of fair models. Privacy evaluations of fair models remain under-explored. Prior study [Chang and Shokri, 2021] provides a preliminary investigation by applying prevalent score-based MIAs to assess the privacy of fair constraint methods on decision tree models. It reveals that fair decision trees enable more successful attacks. Our study extends the scope of analysis to neural networks. We conduct comprehensive evaluations across different fairness approaches on deep classifiers with multiple attack methods.

3 Preliminaries

Algorithmic fairness. Given biased models, we consider a sensitive attribute $s \in S$ with subgroups $\{s_0, s_1\}$ of binary attribute values $\{0, 1\}$. Due to imbalanced training data, trained models often exhibit biased predictions, where predictions $\widetilde{Y}$ become spuriously correlated with the attribute s. Fairness interventions are proposed to mitigate the issue. To quantify fairness performance, metrics such as *Bias amplification* (BA) [Zhao *et al.*, 2017] or *Equalized odds* (EO) [Hardt *et al.*, 2016] have been introduced. Specifically, BA measures disparities in true positives across subgroups, while EO measures true and false positive rates (TPRs, FPRs). We select the prevalent BA and Disparity of Equalized odds (DEO) to measure model fairness performance.

Membership inference attacks. *Score-based attack methods* adopt the target model's (i.e., models under attack) predictions (i.e., scores or losses) to infer sample membership.

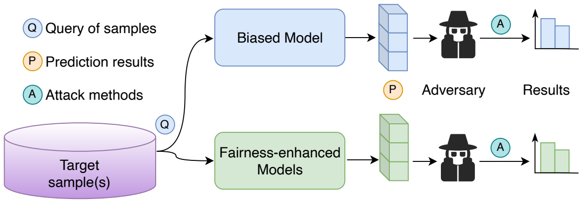


Figure 2: Attacking fair and biased models with MIAs. We first attack them separately and then compare the results to explore the privacy impact of fairness interventions.

To mimic target model behaviors, adversaries train "shadow models" on an auxiliary dataset that shares similar data distributions with the training data. The outputs of shadow models can then be adapted for attack model training. Formally, given target models $\mathcal{T}$ and a queried sample $x \in X$, the membership $M(x)$ can be predicted by:

$$M(x) = \mathbb{1}[\mathcal{A}(\mathcal{T}(x)) > \tau], \tag{1}$$

where the attack model $\mathcal{A}$ will output the confidence scores of membership predictions with a threshold τ.

Reference-based likelihood ratio attack methods infer the membership by modeling the prediction distributions. Specifically, they model the distributions using shadow models: f_{in} that are trained with sample x; f_{out} that are trained without the x. The key idea is to determine if the target prediction $\mathcal{T}(x)$ better aligns with which of the prediction distributions. Formally, membership is predicted by comparing the likelihood ratio Λ between the two distributions:

$$\Lambda = \frac{p(\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}))}{p(\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}}))}, \tag{2}$$

where ϕ is a scaling function, parameters μ and σ are calculated with predictions from shadow models f_{in}, f_{out}. With the likelihood ratio, whichever is more likely determines the membership of x.

4 Attacking Fair Models

In this section, we conduct a detailed case study to assess the privacy impact of fairness interventions. We first utilize established membership inference attacks (MIAs), specifically the naive score-based attacks. We then introduce an advanced attack method tailored to fair models. It is designed to reveal potential privacy threats of fair models.

4.1 Naive Score-Based Attacks

Figure 2 shows our evaluation pipeline with MIAs. We first train biased models with imbalanced data and then obtain fair models by applying fairness interventions. These models serve as target models for the MIAs. An adversary attempts to infer sample membership with predictions from these target models. We then compare results across target models to analyze the privacy impact of fairness interventions. Subsequent sections will delve into more detailed settings.

Models	Acc_t ↑	BA ↓	DEO ↓	Acc_a ↑	AUC_a ↑
Bias	87.6	7.7	21.7	59.8	62.8
Fair	90.5	2.5	5.6	53.2	54.8

Table 1: Attack results with naive score-based methods in (%).

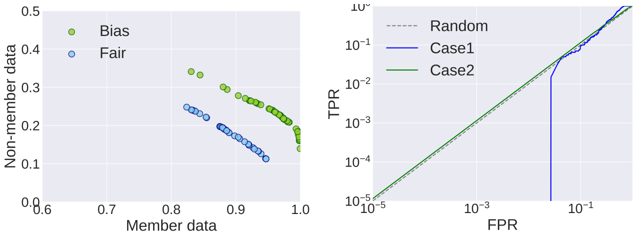
Target models. We train biased models with the CelebA dataset [Lee *et al.*, 2020], which contains imbalanced data distributions for various attributes. In particular, we consider *smile* as classification targets and *gender* as the sensitive attribute. We train biased models following settings in *ML-Doctor* from [Liu *et al.*, 2022]. We apply fair mixup operations from [Ching-Yao Chuang, 2021; Du *et al.*, 2021] to mitigate the biased predictions. Table 1 presents accuracy (Acc_t) and fairness metrics (BA, DEO) results for both biased ("Bias") and fair ("Fair") models. The results show decreased fairness metric results, indicating the effectiveness of the adopted fairness interventions.

Threat models. We apply naive score-based attacks on target models in a black-box manner. In particular, adversaries can only access models' predictions and an auxiliary dataset, which shares similar data distributions with the training data. The adversary trains shadow models to mimic the target models' behavior and uses the prediction scores and results (*true or false predictions*) to infer sample membership. We conduct the attacks following settings in *ML-Doctor*.

Attack results. Table 1 shows the Acc_a and AUC_a results for attacks on the models. It shows improved attack results after fairness interventions. For example, the accuracy results decreased from 59.8% to 53.2% with the fair models. AUC results exhibit similar trends. This aligns with results in Figure 1, where fewer training samples can be successfully attacked after the interventions. Our results show that fairness interventions provide some defense against existing MIAs. However, our following analyses reveal that existing attacks are ineffective for binary classifiers.

Performance trade-offs. During the evaluation, we observe ***evident trade-offs in attack performance on member versus non-member data***. Figure 3a depicts these trade-offs by comparing the accuracy results for member (x-axis) and non-member (y-axis) data. We run over 100 attacks on biased and fair models, and each point denotes one attack result. The figure shows clear performance trade-offs between member and non-member data for both models. This raises concerns: whether achieving high attack performance comes at the cost of a higher false positive rate (FPR) on non-member data.

The issue becomes more pronounced for hard examples where members and non-members share similar prediction scores. As suggested in [Carlini *et al.*, 2022], we assess the attack performance for hard examples with TPR values in the low FPR region. We find the TPR values are around 0.0 for most attacks. Figure 3b presents two worst-case scenarios. The green curve in the figure shows closely aligned TPR and FPR values, indicating the attack results are equivalent to random guesses. The blue line shows 0.0 TPR values in low FPR regions, indicating that no positive samples can be correctly identified. The findings reveal that attack models fail



(a) Attack accuracy (b) Attack for hard examples

Figure 3: Existing attacks (a) exhibit clear performance trade-offs between member and non-member data, and (b) are inefficient in attacking hard examples in the low FPR region.

to differentiate the membership of hard examples, indicating invalid attacks. This aligns with the concerns about the effectiveness of score-based attacks raised in previous studies [Carlini *et al.*, 2022; Ye *et al.*, 2022].

Model degradation. To explore the reason for the trade-off phenomenon, we have discovered that ***trained attack models typically degrade into simple threshold models with one-dimensional inputs.*** This is because current attack methods rely on prediction outcomes to determine the sample membership. For binary classifiers, prediction scores can be reduced to one dimension as the sum of the confidence scores always equals one. Consequently, the attack model can essentially be viewed as a simple threshold model, which infers the membership by "thresholding" one-dimensional values.

Figure 4a presents histograms of prediction scores with vertical lines indicating the threshold value. By adjusting the vertical line (thresholds), it is possible to achieve higher accuracy for member data, but this comes at the expense of decreased accuracy for non-member data. This threshold adjustment explains the trade-off phenomenon.

Impacts of fairness interventions. When examining the prediction scores, we find that ***fairness interventions decrease confidence scores for the majority training data, introducing some defense against existing MIAs.*** This is evidenced by the histograms of confidence scores in Figures 4a and 4b. The figures show that fairness interventions result in more similar score distributions between member and non-member data, making it more difficult for the threshold-based attack models to distinguish them.

Moreover, we explore the score changes for different subgroups in Figures 4c and 4d. From the plots, the majority data are more "spread out", whereas the minority are more "concentrated". This is because fairness interventions strive to balance prediction performance across subgroups for fair predictions. The results advocate the observed increased loss values for most data points in Figure 1a. It also aligns with the fairness-utility trade-off, which is extensively observed in fairness studies [Zhang *et al.*, 2023; Pinzón *et al.*, 2022; Zietlow *et al.*, 2022].

Our analyses indicate that existing attack methods are ineffective in exploiting prediction gaps that could lead to model privacy leaks. While fairness interventions do introduce some defense to MIAs, we identify a novel threat that will pose significant risks to model privacy.

4.2 Attacks with FD-MIA

Enlarged prediction gaps. Previous plots in Figures 4a and 4b show that fairness interventions reduce confidence scores for most training data. In contrast, non-member scores do not exhibit significant changes. This disparity can increase prediction gaps between member and non-member data. We measure the prediction distance by calculating the score value differences between them and present the results in Figure 5. Precisely, we measure the distance considering all available data (Figure 5a) and only the hard examples (Figure 5b), where samples from the members and non-members share similar scores. The plots show a significant distance increase when considering predictions from both biased and fair models. Importantly, the fairness interventions amplify these gaps, which can pose real threats to model privacy.

Attack pipeline. With the enlarged prediction gaps, we propose an enhanced attack method. We present the attack pipeline in Figure 6, where an adversary can access prediction results from both fair and biased models. The attack models will exploit the observed prediction gaps to infer sample membership. We refer to the proposed method as the *Fairness Discrepancy based Membership Inference Attack* (*FD-MIA*).

Threat models. FD-MIA operates as a black-box attack and only needs access to predictions from both biased and fair models. In practice, adversaries could obtain such predictions, as real-world models often exhibit persistent biased predictions. For instance, they may monitor an MLaaS over time since debiasing should be continually carried out to adhere to legislation. Alternatively, they could deliberately report biases, compelling the owner to refine the model per regulations. By recording the prediction shifts during these debiasing efforts, the adversary can enable efficient attacks. Meanwhile, the proposed FD-MIA can be seamlessly integrated into the existing frameworks of score-based and reference-based attacks, enhancing their attack performance.

Score-based FD-MIA. Score-based FD-MIA has been introduced to enhance traditional score-based MIAs by integrating additional encoding layers. These layers are designed to extract the features of model predictions, exploiting the observed prediction gaps. Formally, it can be expressed as follows:

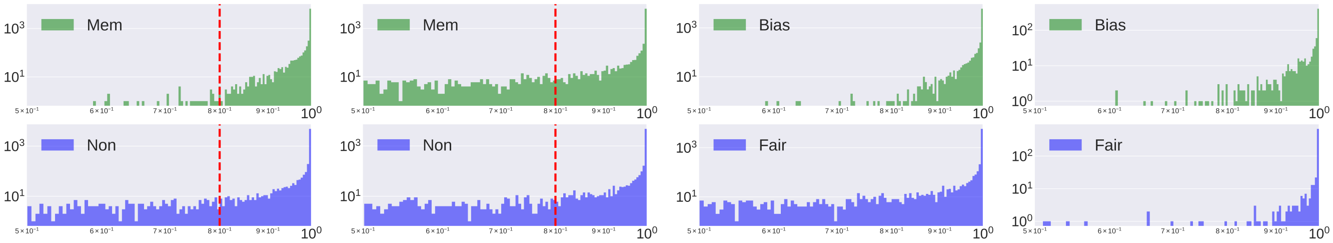
$$M(x) = \mathbb{1}[\mathcal{A}(\mathcal{T}_{\text{bias}}(x), \mathcal{T}_{\text{fair}}(x)) > \tau], \tag{3}$$

where the attack models $\mathcal{A}$ takes predictions from both biased models $\mathcal{T}_{\text{bias}}$ and fair models $\mathcal{T}_{\text{fair}}$.

Reference-based FD-MIA. Reference-based FD-MIA is integrated with the LiRA framework [Carlini *et al.*, 2022], which infer sample membership by modeling the prediction distributions. It enhances attack performance using two target models - the biased and the fair ones. Formally, for a given sample x and target models $\mathcal{T}$, the probability of membership is given by:

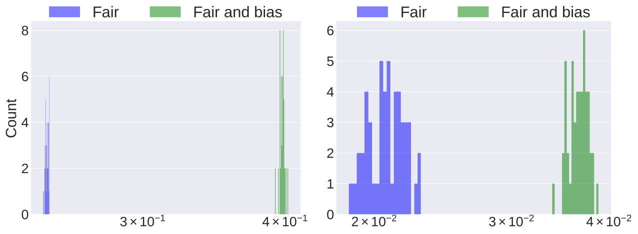
$$p = (\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{bias}}, \mu_{\text{fair}}, \text{Cov})), \tag{4}$$

where Cov is the covariance matrix. The distribution function $\mathcal{N}$ takes the mean confidence scores from both the biased



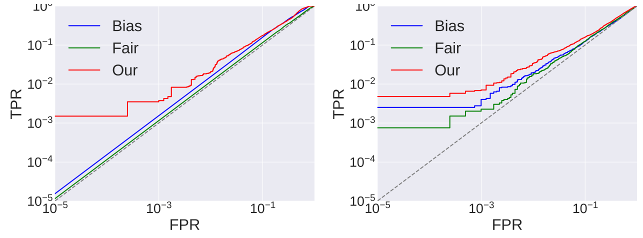
(a) Biased models (b) Fair models (c) Majority subgroups (d) Minority subgroups

Figure 4: Prediction score changes after applying fairness methods. The *red lines* in (a) and (b) indicate that the trained attack models infer sample membership with certain threshold values. (c) and (d) show the changes in terms of different subgroups.



(a) All data (b) Hard examples

Figure 5: Histograms of prediction score distances between member and non-member data. The plots show enlarged distance when considering both fair and biased models.



(a) Score-based attacks (b) LiRA attacks

Figure 7: Attack result comparisons in the low FPR region for (a) score-based attacks and (b) LiRA attacks.

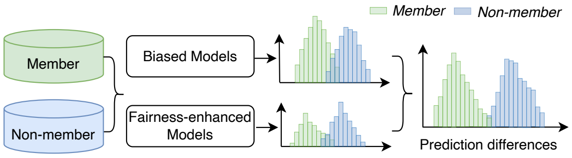


Figure 6: FD-MIA exploits the predictions from both models to achieve efficient attacks.

μ_{bias} and fair models μ_{fair}. This function estimates the likelihood of a data point being a member or non-member. The result is determined by the higher probability score.

The introduced FD-MIA is designed to enhance the attack performance by leveraging predictions from both biased and fair models. Unlike prior attack methods, this mitigates the risk of degraded performance in the trained attack model. Our findings reveal that fairness interventions inadvertently introduce new privacy risks, making target models more vulnerable to membership inference attacks.

5 Experiments

We now extensively evaluate our findings and the proposed method under diverse scenarios. We start by introducing the experiment settings.

Settings. With the *gender* attribute, we consider following binary classifications: smiling predictions (T=s/S=g) with the CelebA dataset [Lee *et al.*, 2020], race predictions (T=r/S=g) with the UTKFace dataset [Geralds, 2017] and the FairFace dataset [Karkkainen and Joo, 2021]. As UTKFace and FairFace contain multiple racial subgroups, we first group them into *White* and *Others* and then obtain the binary subgroups. For training data, we randomly divide the dataset in half to construct member and non-member data for MIAs. For target models, We sample the data with imbalanced data distributions, considering the sensitive attribute to reflect real-world imbalances. This incurs biased predictions of trained models, which serve as the biased models. We then apply fairness interventions of fair mixup operations to obtain fair models.

Results with the *gender* attribute. Table 2 presents the attack results with different attack methods and metrics. We integrate the proposed FD-MIA with score-based attacks (s) and LiRA attacks (l). Besides the attack accuracy, we further report the TPR results at a low FPR value of 0.1%, following suggestions in [Carlini *et al.*, 2022]. The table shows that FD-MIA achieves enhanced attack results in accuracy, AUC, and TPR@FPR. Notably, it achieves higher attack success on fair models than the biased ones. In contrast, the existing methods perform worse on fair models. The results reveal that the proposed FD-MIA can effectively exploit fairness interventions to improve MIAs, posing threats to model privacy.

More specifically, in score-based attacks, FD-MIA achieves better attack results. On UTKFace, for instance, the accuracy jumps from 52.6% to 60.2% and the AUC from 52.8% to 62.1%. Regarding TPR results, notably, the existing attacks (Bias$_s$ and Fair$_s$) attain near-zero values. This in-

Models	CelebA (T=s/S=g)					UTKFace (T=r/S=g)					FairFace (T=r/S=g)				
	Acc_t	DEO	Acc_a	AUC_a	TPR	Acc_t	DEO	Acc_a	AUC_a	TPR	Acc_t	DEO	Acc_a	AUC_a	TPR
Bias_s	87.6	21.7	59.8	62.8	0.0	87.4	14.2	58.5	58.9	0.0	87.2	22.2	63.6	66.4	0.0
Fair_s	90.5	5.6	53.2	54.8	0.04	89.0	6.3	52.6	52.8	0.1	87.6	3.9	63.3	66.2	0.0
Our_s	-	-	**60.6**	**65.8**	**0.3**	-	-	**60.2**	**62.1**	**0.2**	-	-	**65.2**	**66.8**	**0.2**
Bias_l	87.6	21.7	51.5	51.4	0.6	87.4	14.2	55.4	51.5	0.9	87.2	22.2	60.2	61.7	1.3
Fair_l	90.5	5.6	50.8	50.3	0.2	89.0	6.3	53.2	47.6	0.7	87.6	3.9	56.7	57.2	0.9
Our_l	-	-	**54.7**	**57.3**	**1.2**	-	-	**55.9**	**52.2**	**1.7**	-	-	**62.3**	**63.2**	**2.3**

Table 2: Attacks for the *gender* attribute (S), and different learning targets (T) in (%), we report TPR@FPR of 0.1%.

dicates no true positive samples can be identified. In contrast, FD-MIA displays valid TPR values, indicating effective attacks. Similar trends can be observed with the LiRA attacks. We further present the ROC curves for the CelebA case in Figure 7. The figure further confirms the invalid attacks of the existing methods and the valid TPR results of FD-MIA.

Moreover, from the table, we observe that the attacks achieve better results on FairFace compared to other datasets. Meanwhile, FairFace exhibits a greater discrepancy in fairness between the biased and the fair models. We believe the enlarged discrepancy leads to the enlarged prediction gaps, which can enable more effective attacks. Additionally, we notice that score-based attacks perform better on accuracy and AUC, whereas LiRA attacks achieve better TPR values. This aligns with the observations in [Carlini *et al.*, 2022] as LiRA is designed for efficient attacks in the low FPR region.

Results with other attributes. We further explore attacks with different attributes, including *wavy hair* (T=s/S=h) and *heavy makeup* (T=s/S=m) for CelebA, as well as *race* (T=g/S=r) for UTKFace and FairFace. Table 3 presents the results. The results show that the proposed method exhibits enhanced results with considered metrics, suggesting an advantage in identifying privacy vulnerabilities within fair models. Notably, it consistently achieves superior performance with different datasets of varying accuracy performance. Similar to previous results, FD-MIA achieves better results on FairFace, which may be due to the enlarged fairness discrepancy between fair and biased models.

Results with varying fairness levels. Next, we attack models of different fairness performances. Specifically, we consider the case of CelebA (T=s/S=g) and conduct the naive score-based attack on biased and fair models of different DEO values. Figure 8 presents the results. For FD-MIA, we utilize prediction results from multiple fair models and one biased one, which is indicated by a red star in the figure. We further adopt dashed gray lines to outline the trend.

The figure shows that, with the existing attack method, the accuracy decreases for both biased and fair models as the DEO value decreases. The results indicate that models with stronger fairness interventions exhibit more robustness against existing MIAs. In contrast, FD-MIA, which exploits discrepancies in fairness, achieves better attack performance. Notably, larger fairness discrepancies between the fair and biased models contribute to more significant prediction gaps, leading to more powerful attacks with FD-MIA. While achieving improved fairness, these models lower their confidence scores, making the attacks more challenging.

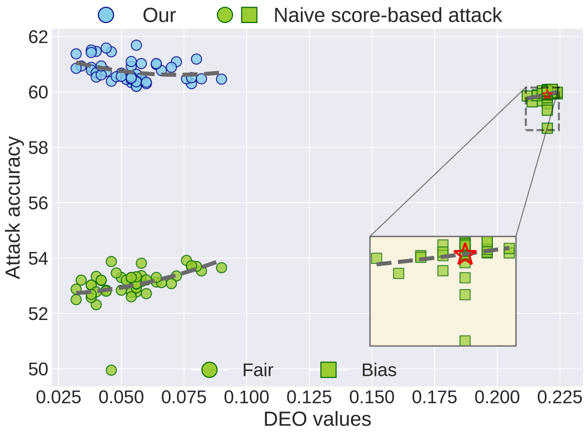


Figure 8: Score-based attacks with models of varying fairness levels.

Results with different fairness approaches. In this part, we evaluate our findings with various fairness approaches, including data sampling, reweighting, adversarial training, and constraint-based approaches. In the experiments, we adopt the implementations of these approaches from [Wang and Deng, 2020; Han *et al.*, 2024]. Similarly, we focus on the case of CelebA (T=s/S=g), and Figure 9 presents the results. The figure shows reduced DEO values after fairness interventions, indicating the effectiveness of these approaches.

For attack results, the existing attack method exhibits degraded performance with fair models for the considered approaches. The attack accuracy drops as the DEO values reduce. The results align with our previous findings, where fairness interventions introduce an unexpected level of resilience to MIAs. Notably, the drops are more pronounced with the adversarial training and constraint approaches. This may stem from the more substantial trade-offs between fairness and model utility inherent to the approaches.

In contrast, FD-MIA achieves higher attack accuracy with fair models than biased ones. Similarly, the attack performance improves when the fairness discrepancy enlarges as FD-MIA explores the prediction gaps. The consistency of results across the fairness approaches demonstrates the broader applicability of our findings and the proposed attack.

Models	CelebA (T=s/S=h)			CelebA (T=s/S=m)			UTKFace (T=g/S=r)			FairFace (T=g/S=r)		
	Acc	AUC	TPR@FPR	Acc	AUC	TPR@FPR	Acc	AUC	TPR@FPR	Acc	AUC	TPR@FPR
Bias$_s$	55.1	56.3	0.1	57.4	58.1	0.0	64.0	66.9	0.0	75.5	76.7	0.0
Fair$_s$	52.6	52.7	0.0	53.1	52.0	0.0	55.3	57.2	0.0	73.2	75.5	0.0
Our$_s$	**56.9**	**59.6**	**0.2**	**59.6**	**63.2**	**0.2**	**66.7**	**67.8**	**0.3**	**77.0**	**78.4**	**0.7**
Bias$_l$	52.1	52.0	0.3	51.6	51.4	0.4	55.5	52.4	1.4	73.2	74.2	1.5
Fair$_l$	51.0	50.5	0.1	50.7	49.9	0.1	53.8	49.7	0.9	70.4	72.1	0.6
Our$_l$	**55.4**	**57.7**	**0.8**	**54.2**	**55.7**	**0.6**	**56.2**	**53.6**	**2.1**	**75.2**	**76.4**	**2.9**

Table 3: Attacks with different sensitive attributes and learning targets in (%).

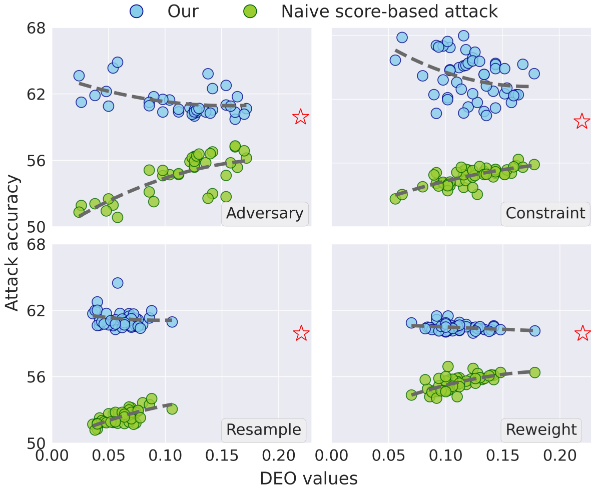


Figure 9: Score-based attacks on fair models with different fairness intervention methods. The *red star* indicates the biased model

Influential factors. The proposed FD-MIA achieved modest improvements in attack performance during the experiments. This is because we considered fair models without severe accuracy degradation, which led to smaller prediction gaps between the fair and biased models. In Figures 8 and 9, the attack performance can be further enhanced by using fairer models with more substantial accuracy drops and fair methods with more significant fairness-accuracy trade-offs.

6 Mitigation

We further discuss two potential defense mechanisms to counter the proposed attack method:

Information Access Control. This involves constraining the adversary's access to key data, thereby potentially diminishing the attack's effectiveness. For example, by restricting the output solely to predicted labels and withholding confidence scores, we can significantly hinder the efficiency of potential attacks. Furthermore, as a proactive step to mitigate privacy risks, we propose the preemptive release of prediction results from fair models before their complete deployment.

Differential privacy (DP). Differential privacy, as introduced in [Dwork *et al.*, 2006], serves as a foundational principle for privacy preservation. We apply the differentially private stochastic gradient descent (DP-SGD) from [Abadi *et al.*, 2016] to defend against the attacks considering the models for the CelebA (T=s/S=g) in Table 2. Table 4 shows the defense results for the existing attacks and the proposed FD-MIA. The results show that introducing DP noise leads to a notable decrease in attack accuracy, signifying the potential of DP-SGD as a defensive measure. Notably, despite the added noise, our attacks (Our$_s$, Our$_l$) still exhibit a persistent edge over existing approaches. This implies that increased noise levels are required to attain comparable defense performance. It indicates the superior attack performance of the proposed FD-MIA compared to the existing ones.

Models	Acc	AUC	TPR@FPR
Fair$_s$	50.8 ($\downarrow$ 2.4)	51.2 ($\downarrow$ 3.6)	0.0 ($\downarrow$ 0.04)
Our$_s$	53.4 ($\downarrow$ 7.2)	55.8 ($\downarrow$ 10.0)	0.0 ($\downarrow$ 0.3)
Fair$_l$	50.5 ($\downarrow$ 0.3)	49.8 ($\downarrow$ 0.5)	0.1 ($\downarrow$ 0.1)
Our$_l$	51.4 ($\downarrow$ 3.3)	51.2 ($\downarrow$ 6.1)	0.1 ($\downarrow$ 1.1)

Table 4: DP-SGD results with $\delta = 10^{-5}, \epsilon = 0.85$ in (%).

7 Conclusions

This paper evaluates the privacy risks of fairness interventions by employing membership inference attacks (MIAs). Our results indicate that fair models often maintain an unexpected level of resilience against existing MIAs for binary classifiers. However, we show that existing attack methods are inefficient as the trained attack models degrade into simple threshold models. Further, we discover a novel attack method named FD-MIA, which leverages predictions from both biased and fair models to exploit the prediction gaps between member and non-member data. It can be integrated into existing attacks and pose substantial threats to model privacy. We conduct experiments across six datasets, three attack methods, and five representative fairness approaches. The results consistently validate our findings and the efficacy of the proposed MIA method. While our attack is tailored to fair binary classifiers, it can be extended to a broader range of models with fairness disparities. Our insights contribute to a deeper understanding of privacy issues related to the application of fairness interventions, emphasizing the imperative need for meticulous design and deployment of trustworthy models. All the appendix can be found in this link[1].

[1] https://arxiv.org/abs/2311.03865

Ethical Statement

We reveal a concerning tension between the pursuit of algorithmic fairness and model privacy. This trade-off raises important ethical questions that warrant careful examination. We measure model fairness using different metrics and assess the model's privacy performance with MIAs, which can be used for evaluating deployed models. We encourage the adoption of techniques like differential privacy to mitigate privacy risks while maintaining fairness. By incorporating these key considerations, we aim to provide practitioners with a more holistic understanding of the challenges and potential solutions, enabling them to navigate this complex issue in a responsible and ethical manner.

Acknowledgements

This work is supported by the Australian Research Council (ARC) Discovery Project of DP230100246 from the Australian Research Council, Australia.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, pages 308–318, 2016.

[Carlini *et al.*, 2022] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *SP*, pages 1897–1914, 2022.

[Chai and Wang, 2022] Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In *ICML*, pages 2853–2866, 2022.

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

[Ching-Yao Chuang, 2021] Youssef Mroueh Ching-Yao Chuang. Fair mixup: Fairness via interpolation. In *ICLR*, 2021.

[Choquette-Choo *et al.*, 2021] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *ICML*, pages 1964–1974, 2021.

[Creager *et al.*, 2019] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *ICML*, pages 1436–1445, 2019.

[Cruz *et al.*, 2023] André Cruz, Catarina G Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient boosting with fairness constraints. In *ICLR*, 2023.

[Du *et al.*, 2021] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. In *NeurIPS*, 2021.

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

[Gao *et al.*, 2023] Junyao Gao, Xinyang Jiang, Huishuai Zhang, Yifan Yang, Shuguang Dou, Dongsheng Li, Duoqian Miao, Cheng Deng, and Cairong Zhao. Similarity Distribution Based Membership Inference Attack on Person Re-identification. In *AAAI*, pages 14820–14828, 2023.

[Geralds, 2017] J Geralds. Utkface large scale face dataset. *github. com*, 2017.

[Han *et al.*, 2024] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *ICLR*, 2024.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.

[He *et al.*, 2022] Xinlei He, Hongbin Liu, Neil Zhenqiang Gong, and Yang Zhang. Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning. In *ECCV*, pages 365–381, 2022.

[Hu *et al.*, 2022] Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M4i: Multi-modal models membership inference. In *NeurIPS*, pages 1867–1882, 2022.

[Jung *et al.*, 2023] Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. Re-weighting based group fairness regularization via classwise robust optimization. In *ICLR*, 2023.

[Karkkainen and Joo, 2021] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[Kim *et al.*, 2019] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, pages 9004–9012, 2019.

[Lee *et al.*, 2020] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020.

[Li *et al.*, 2022] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. Auditing Membership Leakages of Multi-Exit Networks. In *CCS*, pages 1917–1931, 2022.

[Liu *et al.*, 2022] 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 *USENIX Security*, pages 4525–4542, 2022.

[Manisha and Gujar, 2020] Padala Manisha and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *IJCAI*, 2020.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CSUR*, pages 1–35, 2021.

[Park *et al.*, 2021] Sungho Park, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *AAAI*, pages 2403–2411, 2021.

[Park *et al.*, 2022] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *CVPR*, pages 10389–10398, 2022.

[Pinzón *et al.*, 2022] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *AAAI*, pages 7993–8000, 2022.

[Qi *et al.*, 2022] Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Hao Liao, Zhongliang Yang, Yongfeng Huang, and Xing Xie. FairVFL: A fair vertical federated learning framework with contrastive adversarial learning. In *NeurIPS*, 2022.

[Roh *et al.*, 2021] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. In *NeurIPS*, pages 815–827, 2021.

[Sablayrolles *et al.*, 2019] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In *ICML*, pages 5558–5567, 2019.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *SP*, pages 3–18, 2017.

[Tang *et al.*, 2023] Pengwei Tang, Wei Yao, Zhicong Li, and Yong Liu. Fair scratch tickets: Finding fair sparse networks without weight training. In *CVPR*, pages 24406–24416, 2023.

[Truong *et al.*, 2023] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *CVPR*, 2023.

[Wang and Deng, 2020] Mei Wang and Weihong Deng. Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning. In *CVPR*, page 10, 2020.

[Wang *et al.*, 2022] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *CVPR*, pages 10379–10388, 2022.

[Yang *et al.*, 2023] Ziqi Yang, Lijin Wang, Da Yang, Jie Wan, Ziming Zhao, Ee-Chien Chang, Fan Zhang, and Kui Ren. Purifier: Defending Data Inference Attacks via Transforming Confidence Scores. In *AAAI*, pages 10871–10879, 2023.

[Ye *et al.*, 2022] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *CCS*, pages 3093–3106, 2022.

[Yuan and Zhang, 2022] Xiaoyong Yuan and Lan Zhang. Membership Inference Attacks and Defenses in Neural Network Pruning. In *USENIX Security*, pages 4561–4578, 2022.

[Zemel *et al.*, 2013] Richard S. Zemel, Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

[Zhang *et al.*, 2023] Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *ICLR*, 2023.

[Zhao *et al.*, 2017] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, pages 2979–2989, 2017.

[Zhu *et al.*, 2021] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *ICCV*, pages 15002–15012, 2021.

[Zietlow *et al.*, 2022] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *CVPR*, pages 10410–10421, 2022.