Do Fairness Interventions Come at the Cost of Privacy: Evaluations for Binary Classifiers

Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu*, Ming Ding, Wanlei Zhou

Abstract—While in-processing fairness approaches show promise in mitigating biased predictions, their potential impact on privacy leakage remains under-explored. We aim to address this gap by assessing the privacy risks of fairness-enhanced binary classifiers via membership inference attacks (MIAs) and attribute inference attacks (AIAs). Surprisingly, our results reveal that enhancing fairness does not necessarily lead to privacy compromises. For example, these fairness interventions exhibit increased resilience against MIAs and AIAs. This is because fairness interventions tend to remove sensitive information among extracted features and reduce confidence scores for the majority of training data for fairer predictions. However, during the evaluations, we uncover a potential threat mechanism that exploits prediction discrepancies between fair and biased models, leading to advanced attack results for both MIAs and AIAs. This mechanism reveals potent vulnerabilities of fair models and poses significant privacy risks of current fairness methods. Extensive experiments across multiple datasets, attack methods, and representative fairness approaches confirm our findings and demonstrate the efficacy of the uncovered mechanism. Our study exposes the under-explored privacy threats in fairness studies, advocating for thorough evaluations of potential security vulnerabilities before model deployments.

Index Terms—Fairness, Privacy, Classifications, Deep learning

1 Introduction

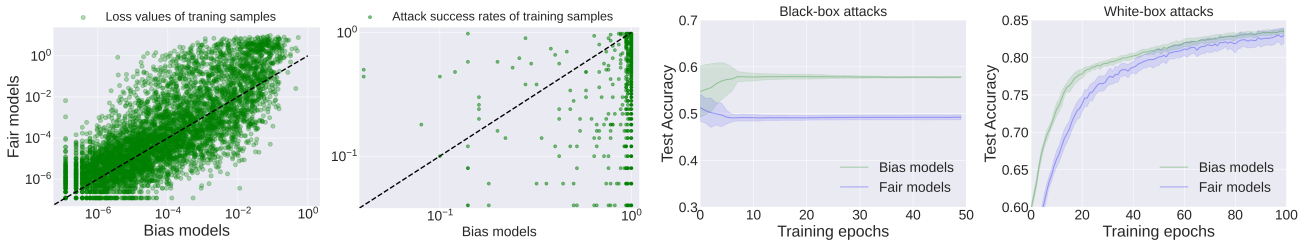
In recent years, there have been remarkable advancements in various fields thanks to large models such as the GPT models [1] and the Segment Anything Model [2]. These models have proven to be highly effective, but their success heavily relies on extensive training data, which often contains biased data distributions. This raises concerns about algorithmic fairness, where the resulting trained models (*biased models*) may exhibit discriminative performances for certain demographic subgroups [3]. To address the issue, previous studies have proposed in-processing methods that modify the learning algorithm to remove bias during model training. After applying these fairness interventions, the obtained fairness-enhanced models (*fair models*) can provide more equitable performance across subgroups, thus mitigating unfairness predictions. However, despite promising to enhance fairness, their potential impact on privacy leakage remains under-explored.

One pioneering work by Chang and Shokri [4] has explored the privacy implications of fairness interventions through the lens of membership inference attacks (MIAs). With the model's predictions, MIAs aim to infer whether a given sample was part of the training data (sample membership). These attacks are widely used to assess privacy risks in models deployed via Machine Learning as a Service (MLaaS) [5]. Building on this, the authors first train biased models and then apply fairness interventions to obtain fair models. They perform MIAs on both the biased and fair models, comparing the attack results before and after the interventions. Their findings reveal that fairness interventions improve the effectiveness of MIAs, suggesting a potential trade-off between achieving model fairness and model privacy.

Although the study by Chang and Shokri [4] offers valuable insights, it has limitations. Firstly, the evaluations primarily focused on decision tree models. Although they briefly explored simple convolutional neural networks (CNNs), their evaluation was limited to a single synthetic dataset. This leaves open questions regarding the privacy risks of fairness interventions for neural networks in real-world datasets, such as binary classifiers, which are prevalent in fairness studies. Secondly, the study adopted only one type of attack–score-based membership inference attacks (MIAs)–for evaluations. While this approach provides insights, it might not fully characterize the privacy impact of fairness interventions. Given that both fairness and privacy are crucial aspects of model trustworthiness, conducting thorough evaluations of these interventions is crucial. To address these gaps, our work considers multiple attacks, including both membership inference attacks (MIAs) and attribute inference attacks (AIAs), to evaluate fair binary classifiers, thereby comprehensively assessing the associated privacy risks.

Specifically, we evaluate the privacy of fair binary classifiers via membership inference attacks (MIAs) and attribute inference attacks (AIAs). While MIAs aim to infer given sample membership information, AIAs attempt to infer sensitive information about the sample, *e.g.*, male or female for the gender attribute. To conduct thorough evaluations, we consider different attack methods for both MIAs and AIAs. Surprisingly, our evaluation results show that **fair models show more resilience to current attacks than their biased**

• *Tianqing Zhu is the corresponding author. H. Tian, G. Zhang, B. Liu are with Australian Artificial Intelligence Institute and the School of Computer Science, University of Technology Sydney, Australia. Email: {Huan.Tian, Guangsheng.Zhang}@student.uts.edu.au, Bo.Liu@uts.edu.au. Ming Ding is with Data61, CSIRO, Australia. Email: Ming.ding@data61.csiro.au. T. Zhu and W. Zhou are with City University of Macau, Macao. Email: {tqzhu, wlzhou}@cityu.edu.mo.*



(a) MIAs (b) AIAs

Fig. 1. Fairness interventions increase loss values for the majority training data, leading to diminished attack successful rates for MIAs (Figure 1a). Meanwhile, compared to biased models, fair models show more resilience to AIAs under both Black and White-box settings (Figure 1b).

counterparts. Figure 1 presents the MIAs and AIAs results on fair and biased models over 100 runs. For MIAs, we report the per-sample loss values and the attack success rate (Figure 1a) before and after applying fairness interventions– biased vs fair models. The plots show increased loss values yet decreased attack success rates for most data points with fair models. This indicates that these interventions can lead to less successful attacks with existing attack approaches. For AIAs, we illustrate the test accuracy for both white and black-box AIAs (Figure 1b). The plots show decreased attack results after fairness interventions, indicating inferior attack performance.

Through further analyses, we find that fairness interventions reduce sensitive information among the extracted features and confidence scores for the majority of training data, leading to fairer predictions. These reductions lead to more challenging attacks for both AIAs and MIAs, as the attacks have less exploitable information to leverage. Meanwhile, our experiments reveal that the existing attack methods, which are primarily designed for multi-class scenarios, become less effective when applied to binary classification tasks. This inefficacy stems from the binary outputs, which cause the attack models to degrade into simple threshold-based decisions. The degradation incurs substantial performance trade-offs. For example, in the case of MIAs, while effective at recognizing member data, the attack models struggle with non-member data. The phenomenon is particularly pronounced for "hard examples"—samples where the predictions are similar across groups.

Before concluding that fairness interventions are privacy-friendly to binary classifiers, we further identify a potential threat that could enable more effective attacks. During the evaluation, we observe divergent prediction behaviors for different data groups after applying fairness interventions. Specifically, the prediction scores typically increase for the majority of the training data and decrease for the minority training data. In contrast, the scores for non-member data conform to a normal distribution. This disparity creates pronounced prediction gaps between the data groups. However, if adversaries exploit these widened gaps between groups, it might enable more successful attacks, thereby posing substantial privacy threats to fair binary classifiers.

Inspired by these observations, we introduce two new attack methods: Fairness Discrepancy-based Membership Inference Attacks (FD-MIAs) and Fairness Discrepancy-based Attribute Inference Attacks (FD-AIAs). These methods exploit the prediction gaps between the original biased models and the fair models more effectively. The key is to leverage model fairness disparities. Moreover, we demonstrate that FD-MIAs and FD-AIAs can be integrated with existing attack methods, such as score-based [6] and reference-based attacks [7]. This integration delivers advanced attack performance and poses real privacy threats to fair binary classifiers.

In the experiments, we conduct comprehensive evaluations across three datasets, with up to six attack methods and five in-processing fairness approaches. This amounts to 128 different settings and more than 400 distinct models. The results consistently validate our findings and the identified threat. Our study reveals that **fairness interventions can introduce new threats to model privacy**, advocating a more comprehensive examination of their potential security defects before deployment. Our main contributions are as follows:

- 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 and AIAs, targeting deep classifiers with real-world datasets.
- We reveal that fairness interventions do not compromise model privacy with *existing* attack methods, primarily due to their limited efficacy in attacking binary classifiers.
- We identify a previously unexamined vulnerability and propose two novel attack methods, FD-MIA and FD-AIA, which pose real threats to model privacy by exploiting prediction gaps between biased and fair models. These methods can be integrated into existing attack frameworks.
- Extensive experiments on three datasets have confirmed our observations and demonstrated the efficacy of the proposed methods.

A preliminary version of this research was presented at IJCAI 2024 in [8]. While the conference paper introduced the concept of privacy evaluations for fair models using MIAs, this manuscript significantly expands the scope and depth of our investigation. We broaden the privacy risk assessment by incorporating both MIAs and AIAs, providing a more comprehensive evaluation of fair binary classifiers. Furthermore, we extend the attack mechanism originally developed for MIAs to the domain of AIAs, resulting in novel and more potent attack strategies.

The organization of the paper is as follows: Section 2 reviews related work on fairness and privacy attacks. Section 3 outlines the preliminaries and evaluation metrics. In Section 4, we assess model privacy using membership inference attacks, while Section 5 extends this evaluation to attribute inference attacks. Building on these findings, Section 6 introduces our enhanced attack mechanism (FD-MIA and FD-AIA), which exploits prediction discrepancies between biased and fair models. Section 7 presents experimental evaluations, and Section 8 discusses potential mitigations and future research directions. Section 9 provides a broader discussion of our proposed methods. Sections 10 and 11 discuss the limitations and provide conclusions, respectively.

2 Related work

2.1 Algorithmic fairness

Given biased models, fairness methods aim to ensure consistent prediction performance across subgroups. According to the method modification phases, fairness studies generally fall into three categories: pre-processing, in-processing, and post-processing approaches. For deep classification models, studies usually adopt in-processing methods as they deliver fair results efficiently. Widely adopted methods involve the introduction of fair constraints, adversarial training, or mixup augmentation operations.

Fair constraint methods [9], [10], [11], [12], [13], [14], [15], [16] introduce additional constraints based on the fairness metrics. They formulate the problem as optimization issues. Initially proposed in [9], subsequent studies have developed the method with diverse settings such as proposing different constraints [11], [17], [18] or training schemes [10]. Later, adversarial training methods have been proposed [19], [20], [21], [22], [23]. These methods require additional predictions for sensitive attributes and update gradients reversely to remove sensitive information from extracted features. The operation leads to more similar representations across subgroups, contributing to fairer predictions. More recently, studies aim to learn "neutral" representations using mixup augmentation operations [24], [25] or contrastive learning [26], [27], [28], [29]. These methods either interpolate inputs or modify features to pursue fair representations. Other fairness methods include data operations such as balancing the data with synthetic data generation [30], [31], [32], data sampling strategies [33] [34] or data re-weighting strategies [35], [36] to enforce fairness. Others concentrate on different settings, such as semi-supervised learning [28], [37], [38], multi-attribute protections [22], [39], [40], or enforcing fairness without demographics [41]. In the experiments, we evaluate model privacy considering multiple fairness methods, delivering comprehensive evaluations.

2.2 Membership inference attacks

Membership inference attacks aim to determine whether a given data sample was in the target model's training dataset or not [5]. A number of attacks leverage the target model's direct output as inputs to train the attack models and infer the membership of queried samples. For example, various studies [5], [6], [42] utilize the confidence scores as input, while others [43], [44], [45] focus on the training losses. Additionally, some studies [46], [47] employ the prediction labels for their attacks. These methods are usually considered score-based attack methods. On the other hand, some studies focus on enhancing attack performance by modeling prediction distributions of the target models [7], [48]. These methods aim to model the distributions for both member and non-member data. They then leverage the distribution difference to attain superior attack outcomes. These methods are commonly referred to as reference-based attack methods. such as reference models. Other research extends their focus into various scenarios [49], [50], [51], [52] or proposes defense methods against the attacks [53], [54], [55], [56], [57]. In the experiments, we consider two representative attack approaches to evaluate model privacy leakage: score-based [6], [42] and reference-based [7] membership inference attacks.

Enhanced membership inference attacks. More recently, researchers have begun incorporating additional information as key indicators to boost the overall effectiveness of their attacks. For instance, in the work by He et al. [58], the adversary leverages prediction outcomes obtained from multiple augmented views to significantly enhance its performance. Another study by Li et al. [59] focuses on results derived from multi-exit models as their attack strategy. Furthermore, the study conducted by Hu et al. [60] integrates prediction results from a multi-modality model to achieve enhanced attack performance. Inspired by previous studies, we propose a novel approach to enhance attack performance by leveraging additional information from fairness interventions. Differently, our method uniquely exploits the disparities introduced by model fairness techniques and integrates this insight with existing attack strategies. This innovative combination yields superior attack results, revealing previously unrecognized vulnerabilities in fair models.

2.3 Attribute inference attacks

Attribute Inference Attacks (AIAs) aim to infer sensitive attributes of samples using deployed model predictions. These attacks share similarities with MIAs, particularly in how some AIA methods exploit prediction gaps between different subgroups to infer attribute information, as demonstrated by Yeom et al. and Ganju et al. [61], [62]. These approaches are typically classified as black-box attacks. However, state-of-the-art AIAs often employ more sophisticated techniques, relying on the embeddings of target samples. These white-box attacks infer sample attribute information by analyzing the extracted features from the target model. In our comprehensive privacy evaluations of fair binary classifiers, we consider both types of attack methods: those exploiting prediction gaps (black box) and those leveraging the embeddings (white box). This dual approach allows us to assess the vulnerabilities of fair models from multiple perspectives, providing a more thorough understanding of their privacy implications.

2.4 Fairness interventions and attacks

Currently, limited research focuses on algorithmic fairness and attacks. Some earlier studies have explored the connections between fairness studies and adversarial attacks [63],

[64], [65], [66], [67]. They find that fair models tend to be more vulnerable to attacks than biased models [65]. Later, studies have taken the approach to attack target models and compromise model fairness results [64], [66], [67].

Recent research has explored various methods to attack fair-enforced models. For instance, studies by Aalmoes et al. [68] and Balunović et al. [69] have investigated the relationship between attribute attacks and fair-enforced models. Balunović et al. [69] aim to promote fairness predictions by reducing the performance of attribute attacks, while Aalmoes et al. [68] utilize fairness methods to defend against such attacks. These studies demonstrate the alignment between model fairness and attribute privacy. In contrast, our work uncovers an overlooked attack mechanism that poses significant threats to attribute privacy in fair-enforced models.

More related to our study, Chang and Shokri [4] attack fair-enforced methods with membership inference attack methods. They consider fairness constraint methods for decision tree models with structure data. They then measure the attack performance with average-case success metrics of accuracy and AUC. They find that score-based methods can effectively attack fair models more accurately than biased ones. Our study aims to examine the attack performance in binary classifications and enforce more efficient attacks. We employ multiple attack methods and metrics to assess the privacy impact of fairness approaches.

3 PRELIMINARIES

3.1 Algorithmic fairness

Given biased models, we consider sensitive attributes $s \in \mathcal{S}$ and subgroups $\{s_0, s_1\}$ with binary attribute values $\{0, 1\}$. Then, for fair models, as the prediction target and the sensitive attribute are irrelevant, the model prediction $\widetilde{y}$ and S should be independent, *i.e.*, $\widetilde{y} \perp S|Y = y$.

With different values of the sensitive attribute $\{s_0, s_1\} \in A$, one selected fairness metric γ can be expressed as follows:

$$\gamma(\widetilde{y}, y, S), S = \{s_0, s_1\} \tag{1}$$

Fairness metrics. Ideally, the fairness metric value for different subgroups should be equal ($\gamma_{s_0} = \gamma_{s_1}$). However, biased models tend to have different metric values across subgroups. Generally, we adopt the difference to quantify the discrimination level and measure model fairness performance:

$$\Gamma = |\gamma_{s_0} - \gamma_{s_1}|, \tag{2}$$

where Γ is the discrimination level, γ_{s_0} and γ_{s_1} are the fairness metric values across subgroups. In the experiments, we adopt bias amplification (BA) from [70] and equalized odds (EO) from [71] as fairness metrics.

Bias amplification (BA) is introduced in [70]. BA measures the difference in true positive predictions across subgroups, normalized by the total true positives. A lower BA indicates less bias amplification. It can be written as:

$$\frac{1}{2}\frac{|\mathrm{TP}_{s_0} - \mathrm{TP}_{s_1}|}{\mathrm{TP}_{s_0} + \mathrm{TP}_{s_1}}, \tag{3}$$

where TP presents the true positive value of predictions.

Equalized odds (EO) is proposed in [71]. EO requires that the probability of a positive prediction given the true label should be equal across subgroups. This ensures that both true positive rates (TPRs) and false positive rates (FPRs) are equalized across subgroups. The fairness metric can be defined as:

$$P\{\widetilde{Y}|Y, S\}; Y = \{0, 1\}, S = \{s_0, s_1\}. \tag{4}$$

We report the discrimination level with the difference of BA and EO across subgroups for fairness evaluations as they are widely adopted for fairness evaluations. For instance, with the discrimination level in Eq. (2), fairness measurement considering the metric of EO can be calculated as follows:

$$\mathrm{DEO} = |\mathrm{EO}_{s_0} - \mathrm{EO}_{s_1}|. \tag{5}$$

Datasets. CelebA [72], UTKFace [73], and FairFace [74] are the commonly adopted datasets in fairness studies. CelebA contains over $200,000$ celebrity face images with 40 attribute annotations. UTKFace and FairFace are diverse facial datasets with balanced annotations across different demographic groups. In our experiments, we create biased training data by sampling with a 9:1 ratio between majority and minority groups while maintaining balanced distributions for the target classification tasks and test sets. In the evaluations, we first adopt the CelebA dataset in Sections 4 and 5, focusing on smiling classifications as the target and gender as the sensitive attribute. We then extend our analysis in Section 7 to include all three datasets, using various attribute combinations to further validate the generalizability of our observations and the efficacy of our proposed attack methods across different data distributions and fairness scenarios.

3.2 Membership inference attacks

In the evaluations, we consider score-based and reference-based MIA methods to evaluate model privacy.

Score-based attack methods rely on the target model's (i.e., models under attack) prediction outcomes (e.g., scores or losses) to determine the membership on each individual data sample. Typically, to mimic the behavior of the target model, a "shadow model" is trained with an auxiliary dataset that shares the same distribution as the training data. The outputs of the shadow model are then adopted to train the attack models, where the membership of the data is considered as the labels. In this way, the attack model can infer whether the given samples are from the training data or not. Formally, given target models $\mathcal{T}$ with queried sample x, the membership of the sample $M(x)$ can be predicted by,

$$M(x) = f_a^m(\mathcal{T}(x)), \tag{6}$$

where the designed *membership* attack model f_a^m outputs the confidence scores of predicted membership. Generally, existing studies usually adopt deep learning models as attack models.

Reference-based likelihood ratio attack methods, on the other hand, infer the membership by modeling the prediction distributions. They first train multiple shadow models on random subsets of training data. For a target example x, the methods then model the prediction distributions for models (f_{in}) trained with the sample x and models (f_{out})

trained without x. Both distributions are modeled as Gaussians. Then, they determine the membership of x by comparing the likelihood of the sample prediction results $\mathcal{T}(x)$ from the target model with the two distributions above. Formally, the likelihood ratio between the distributions of member and non-member data can be defined as,

$$\Lambda = \frac{p(\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}))}{p(\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}}))}, \tag{7}$$

where ϕ is a logic scaling function, $(\mu_{\text{in}}, \sigma_{\text{in}})$ are calculated with the predictions from the predictions of member data (f_{in}), and $(\mu_{\text{out}}, \sigma_{\text{out}})$ are from f_{out}. With likelihood ratio Λ, whichever is more likely determines the membership of x.

3.3 Attribute inference attacks

AIAs aim to infer sample sensitive information–subgroups such as male or female. They can be conducted in both black-box and white-box settings. In a black-box setting, given predictions of target models, the subgroup can be predicted by:

$$A(x) = f_a^a(\mathcal{T}(x)). \tag{8}$$

Similar to MIAs, the designed *inference* attack model f_a^a predicts sample subgroups with target model predictions.

As state-of-the-art AIAs usually rely on sample embedding of the target models with a white-box attack setting. Specifically, the trained target model $\mathcal{T}$ can be composed of a feature extraction module h and a classification module g, such that $\mathcal{T} = g(h(x))$. Given sample embedding $h(x)$, the subgroup can be predicted by:

$$A(x) = f_a^a(h(x)). \tag{9}$$

3.4 Notations

Table 1 summarizes the main notations adopted throughout the paper. In the table, x and y presents model inputs and labels, respectively. Predicted labels are represented as $\widetilde{y}$. Our models consist of feature extractors $h(x)$, and classification heads $g(\cdot)$. In our privacy analysis, membership inference results for a sample x are denoted as $M(x)$, while attribute inference results are represented as $A(x)$. The attack models for membership and attribute inference are denoted as $f_a^m(\cdot)$ and $f_a^a(\cdot)$, respectively. We consider different types of attacks: score-based (MIA_s) and reference-based (MIA_l) membership inference attacks, as well as black-box (AIA_b) and white-box (AIA_w) attribute inference attacks. We adopt metrics including accuracy of target classifiers (Acc_t) and attack classifiers (Acc_a), as well as true positive rate (TPR) and false positive rate (FPR).

4 Evaluations with MIAs

In this section, we evaluate the privacy impact of fairness interventions with MIAs. We first introduce attack settings and then present the attack results.

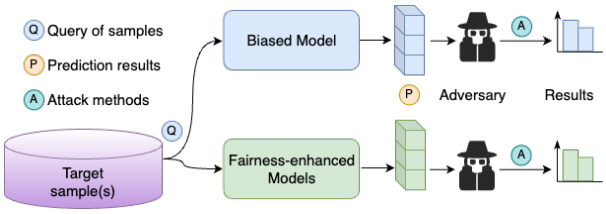


Fig. 2. Model privacy impact evaluation pipelines.

4.1 Attack settings

Attack pipeline. We evaluate the privacy performance of models before and after applying fairness interventions using Membership Inference Attacks (MIAs). We follow the attack pipeline depicted in Figure 2, which is consistent with the approach adopted in previous work [4]. Specifically, we begin by training biased models using biased training data. To enhance the fairness performance of these models, we apply fairness interventions to obtain their counterpart, fair models. Next, we attack both the biased and fair models using existing MIA methods. Finally, we compare the attack results to evaluate the impact of fairness interventions on the privacy of the models.

Target models. We train biased models with the CelebA dataset [72], 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 [6]. We apply fair mixup operations from [24], [25] to mitigate the biased predictions. Table 2 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 both the score-based attacks (MIAs from [6]) and reference-based attacks (LiRA from [7]) 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*.

4.2 Score-based attacks

Table 2 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.

4.3 Reference-based attacks

We further use the reference-based attack method of LiRA from [7] to evaluate model privacy. The method examines attack performance via the True Positive Rates (TPR) value

TABLE 1
Summary of Main Symbols and Notation

Symbol	Description	Symbol	Description
x	Inputs	y	Labels
$\widetilde{y}$	Predicted labels	$h(x)$	Feature extractors
$g(\cdot)$	Classification heads	$\mathcal{T} = g \circ h$	Target models ($\mathcal{T}(x) = g(h(x))$)
$s \in \mathcal{S}$	Sensitive attributes	s_0, s_1	Binary attribute values (0, 1)
$\gamma(\widetilde{y}, y, S)$	Generic fairness metric	$\Gamma = \lvert\gamma_{s_0} - \gamma_{s_1}\rvert$	Discrimination level across subgroups
BA	Bias Amplification (fairness metric)	DEO	Difference in Equalized Odds (fairness metric)
$M(x)$	Membership inference results for x	$A(x)$	Attribute inference results for x
$f_a^m(\cdot)$	Membership attack models	$f_a^a(\cdot)$	Attribute attack models
MIA_s	Score-based membership inference attacks	MIA_l	Reference-based membership inference attacks (LiRA)
AIA_b	Black-box attribute attacks	AIA_w	White-box attribute attacks
$\phi(\cdot)$	Logit scaling functions	$p(\cdot)$	Probability functions
$\mu_\text{in}, \mu_\text{out}$	Mean prediction vectors (member, non-member)	$\sigma_\text{in}, \sigma_\text{out}$	Std. dev. (member, non-member)
Acc_t	Accuracy of target (main) classifiers	Acc_a	Accuracy of attack classifiers
TPR	True positive rate	FPR	False positive rate
Λ	Likelihood ratio in reference-based MIA	Cov	Covariance matrix for distribution modeling
$\mathcal{T}_b(x), \mathcal{T}_f(x)$	Biased, fair target models	$h_b(x), h_f(x)$	Biased, fair feature extractors
$\text{TP}_{s_0}, \text{TP}_{s_1}$	True Positives for subgroups s_0, s_1	AUC	Area-under-ROC-curve measure

TABLE 2
Attack results with the score-based methods from [6] in (%).

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 3
Attacks using LiRA from [7] with TPR @ 0.1% FPR in (%).

Models	Acc_a ↑	AUC_a ↑	TPR ↑
Bias	51.5	51.4	0.6
Fair	50.8	50.3	0.2

in the low False Positive Rates (FPR) region. This enables MIAs on hard examples, where samples from both member and non-member groups share similar prediction results. Table 3 shows the attack results with the TPR results at a low FPR value of 0.1%. The table shows inferior attack performance for fair models compared to the biased ones with all three considered metrics. The results are consistent with the score-based MIAs.

4.4 Discussions

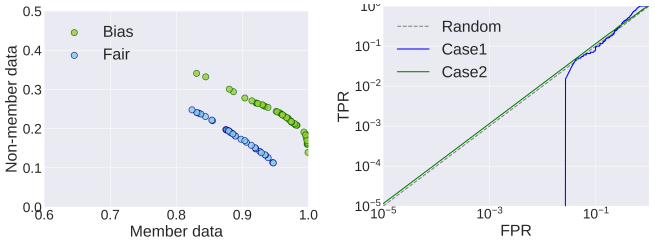
To better understand the results, we further explore the attack results and find the following observations:

Performance trade-offs. During the evaluation, we observe ***evident trade-offs in attack performance on member versus non-member data***. Figure 3a illustrates the inherent performance trade-offs in membership inference attacks by plotting the accuracy results for member data (x-axis) against the accuracy for non-member data (y-axis). We conducted over 100 independent attack experiments on both biased and fair models, with each circle in the scatter plot representing a single attack instance. Green circles represent attack results on biased models, while blue circles represent attack results on fair models. The clear negative correlation visible in both model types demonstrates a fundamental trade-off: as an attack model becomes more accurate at identifying member groups (moving rightward on the x-axis), it simultaneously becomes less accurate at correctly classifying non-members (moving downward on the y-axis). This pattern holds consistently across both biased and fair models. This trade-off raises significant concerns about the practical effectiveness of membership inference attacks. Specifically, it suggests that achieving high attack performance on training data members inevitably comes at the cost of a higher false positive rate (FPR) on non-member data, making the attack less reliable overall. This observation aligns with findings from previous studies [7] that highlight the limitations of threshold-based attack methods when applied to binary classification tasks.

The issue becomes more pronounced for hard examples where members and non-members share similar prediction scores. As suggested in [7], 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 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 [7], [48].

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



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

Fig. 3. Existing attacks (a) exhibit clear performance trade-offs between member and non-member data, each green circle represents attack accuracy on biased models and each blue circle represents attack accuracy on fair models; and (b) are inefficient in attacking hard examples in the low FPR region.

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 [75], [76], [77].

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.

5 Evaluations with AIAs

This section further assesses model privacy performance by conducting Attribute Inference Attacks (AIAs). We begin by outlining the attack settings, followed by a presentation of the attack results and an in-depth analysis of our findings. Our evaluation uses the same CelebA dataset configuration as in the previous section to ensure consistency.

5.1 Attack settings

Attack pipeline. Attribute Inference Attacks (AIAs) are designed to infer sensitive attribute information about data samples by exploiting the outputs of target models. We follow the same attack pipeline in Figure 2.

Target models. We consider the same target models as in MIAs. Specifically, we launch AIAs on both biased and fair models and compare the attack results for privacy evaluations.

TABLE 4
Attribute inference attack results with biased and fair models in (%).

Models	Acc_t ↑	BA ↓	DEO ↓	AIA_b (Acc) ↑	AIA_w (Acc) ↑
Bias	87.6	7.7	21.7	56.5	83.4
Fair	90.5	2.5	5.6	47.5	82.9

Threat models. In our evaluations, we employ both black-box and white-box AIAs to comprehensively assess the privacy risks of biased and fair models.

In the *black-box* setting, the attacker has access only to the prediction scores from the target model. This scenario mimics real-world situations where an adversary can query the model but has no access to its internal structure. In the *white-box* setting, the attacker has more privileged access and can obtain sample embeddings from internal layers of the target model. Specifically, we extract features from the *last* layer of the feature extraction module in the target model, which is the layer immediately preceding the fully connected layers. This approach provides the attacker with more information with the sample embeddings. For both settings, we train attack models following the methodology outlined in the *ML-Doctor* framework [6].

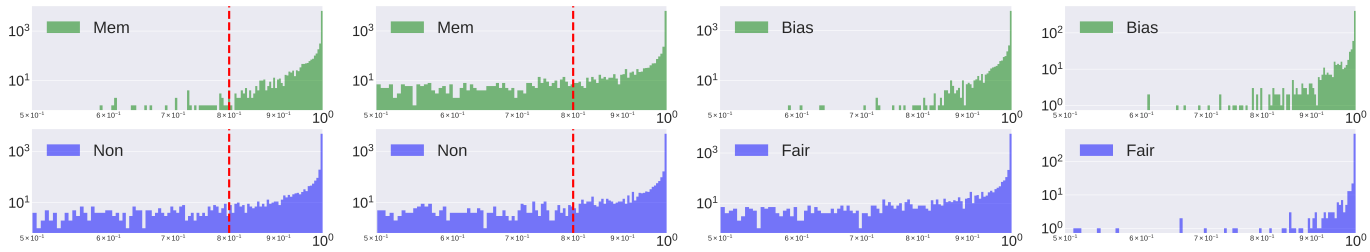
5.2 Attack results

Table 4 presents the attack results, where AIA_b represents the black-box attack, and AIA_w represents the white-box attack. For example, the black-box attack accuracy decreases from 56.5% for biased models to 47.5% for fair models, and the white-box attack accuracy decreases from 83.4% for biased models to 82.9% for fair models. Notably, the black-box attack accuracy on fair models is close to random guessing.

5.3 Discussions

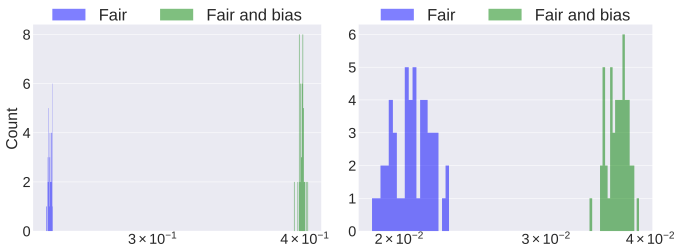
Our results suggest that fairness interventions introduce some level of robustness against AIAs. For black-box attacks, the attack model uses the prediction scores of target models to infer sample subgroup information. As observed in Figure 4c and Figure 4d, fairness interventions adjust scores across subgroups to achieve fairer predictions. This adjustment results in more similar scores across subgroups, making it harder for attackers to infer subgroup information. Consequently, the pursuit of fairer results can inadvertently enhance model privacy. These findings align with the study by [69], which examined model fairness by evaluating whether extracted features contain sensitive information via AIAs.

Similar trends can also be observed with white-box attacks. Fair models resulted in lower attack success rates compared to biased models. Notably, the attack accuracy for white-box attacks remains higher than for black-box attacks, even though it declined after fairness interventions. This is because white-box attacks leverage sample embeddings to infer subgroup information, preventing trained attack models from degrading into simple threshold models. As black-box attacks adopt prediction scores as inputs, similar



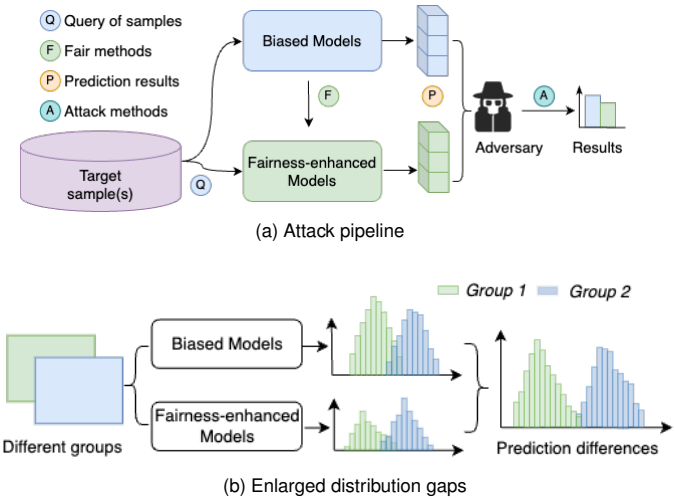
(a) Biased models (b) Fair models (c) Majority subgroups (d) Minority subgroups

Fig. 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

Fig. 5. Histograms of prediction score distances between groups of member and non-member data for fair and biased models. We measure the distance with score value difference between the groups and present comparisons regarding (a) all data and (b) hard examples, where samples from the member and non-member data share similar scores.



(a) Attack pipeline

(b) Enlarged distribution gaps

Fig. 6. Fairness discrepancy based attacks exploit the difference in predictions from both models to more effectively distinguish different data groups, such as different subgroups or member versus non-member data.

to MIAs, attack models degrade into simple threshold models when considering binary classifiers as the target models. However, with white-box attacks, the inputs are extracted features, ensuring sufficient information to launch efficient AIAs.

6 An Enhanced Attack Mechanism

While our experiment results show inferior attack results after fairness interventions, during the evaluations, we discover an overlooked attack mechanism and propose two novel attack methods, FD-MIA and FD-AIA. We first introduce the attack mechanism and then present the proposed FD-MIA and FD-AIA.

6.1 Fairness disparity based attack mechanism

The enlarged distribution gaps. The previous findings have indicated that fairness methods tend to decrease the score values for the majority subgroups while increasing the scores for the minority subgroups. This can lead to an enlarged gap in predictions across subgroups. On the other hand, the score changes for the non-member data are likely to follow normal distributions, causing a different behavior pattern compared to the member data. These different behavior patterns in score changes can serve as additional clues to achieve better performance in MIAs and AIAs. For example, in MIAs, we can plot the histograms of the prediction gaps, as shown in Figure 5, to analyze the gaps for the overall training data and the hard examples.

Specifically, we first calculate the mean and variance of the prediction scores for biased and fair models. We then compute the distribution distance between member and non-member data using the results from the biased models and the results from both biased and fair models, respectively. This calculation is performed over a total of 50 runs. The figure shows enlarged distances when predictions from both models are considered. Moreover, we explore the distance in Figure 5b considering only the hard examples–samples with similar prediction scores among member and non-member data. Again, the figure demonstrates enlarged prediction distances when using predictions from both models. Inspired by the observations, we propose an enhanced attack method tailored for fair models with the observed prediction gaps.

Attack pipeline. Figure 6 illustrates the attack pipeline, wherein an adversary can access prediction results from both models. The attack models will exploit the difference in predictions to infer the membership or subgroup information. We refer to the proposed method as the *Fairness Discrepancy based Membership Inference Attack* (*FD-MIA*) and *Fairness Discrepancy based Attribute Inference Attack* (*FD-AIA*). As the proposed method only modifies the inputs, it can be integrated into existing attack techniques.

Threat models. The discovered attack mechanism operates

as a black-box attack, requiring only access to the predictions from both a biased model and a fair model. In practice, adversaries could obtain such predictions, as real-world models often exhibit persistent biased predictions that linger even after attempts at debiasing. For instance, an attacker could monitor a Machine Learning as a Service (MLaaS) platform over time, as debiasing efforts are typically an ongoing process to adhere to relevant legislation. Alternatively, the adversary could deliberately report biases, compelling the model owner to refine the model in accordance with regulations. By recording the prediction shifts that occur during these debiasing efforts, the adversary can gather the necessary information to enable efficient attacks that exploit the discrepancies between the biased and fair models.

6.2 FD-MIA

The discovered attack mechanism can be seamlessly integrated into the existing MIAs 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:

$$M(x) = f_a^m(\mathcal{T}_{\text{b}}(x), \mathcal{T}_{\text{f}}(x)), \tag{10}$$

where the attack models f_a^m takes predictions from both biased models $\mathcal{T}_{\text{b}}$ and fair models $\mathcal{T}_{\text{f}}$.

Reference-based FD-MIA. Reference-based FD-MIA is integrated with the LiRA framework [7], 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:

$$p = (\phi(\mathcal{T}(x))|\mathcal{N}(\mu_{\text{b}}, \mu_{\text{f}}, \text{Cov})), \tag{11}$$

where Cov is the covariance matrix. The distribution function $\mathcal{N}$ takes the mean confidence scores from both the biased μ_{b} and fair models μ_{f}. This function estimates the likelihood of a data point being a member or non-member. The result is determined by the higher probability score.

6.3 FD-AIA

The designed fairness disparity based attack mechanism can be integrated into existing black-box and white-box attribute inference attacks.

Black-box FD-AIA. Black-box AIAs infer sample subgroup information with target model prediction results. With results from both biased and fair models, the subgroup information can be predicted as:

$$A(x) = f_a^a(\mathcal{T}_{\text{b}}(x), \mathcal{T}_{\text{f}}(x)). \tag{12}$$

Here, the attack model f_a^a combines these prediction results to infer the sensitive subgroup information for the input sample x.

White-box FD-AIA. On the other hand, white-box AIAs use sample embeddings to obtain subgroup information predictions. With access to both the biased and fair models, the prediction can be formulated as:

$$A(x) = f_a^a(h_{\text{b}}(x), h_{\text{f}}(x)), \tag{13}$$

where the attack model f_a^a leverages features from the biased model $h_{\text{b}}(x)$ and features from the fair model $h_{\text{f}}(x)$. By combining information from both models, the attack model $A(x)$ can make more accurate predictions about the sensitive subgroup information.

6.4 Discussions

The introduced fairness disparity based attack mechanism is designed to enhance the attack performance by leveraging predictions from both biased and fair models. Unlike existing 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 and attribute inference attacks.

7 Experiments

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

7.1 Settings

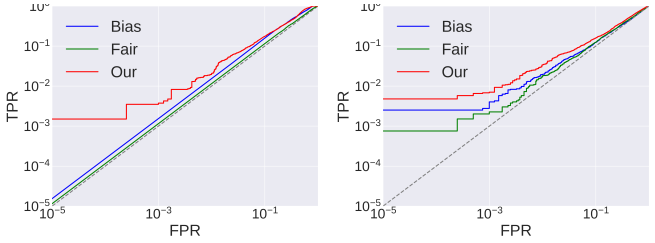
Datasets. With the *gender* attribute, we consider the following binary classifications: smiling predictions (T=s/S=g) with the CelebA dataset [72], race predictions (T=r/S=g) with the UTKFace dataset [73] and the FairFace dataset [74]. As UTKFace and FairFace contain multiple racial subgroups, we first group them into *White* and *Others* and then obtain the binary subgroups.

Training data. For training data, we sample the data to skew the distribution with specified sensitive attributes to induce biased predictions. We set a highly imbalanced ratio of $9:1$ between the majority and minority groups (e.g., 90% male data and 10% female data). Meanwhile, we maintain a balanced distribution for the target learning (e.g., 50% smiling and 50% non-smiling). We also maintain a balanced distribution for the test data. In the ablation studies, we consider different imbalanced ratios to further verify the proposed method.

Models. For the target models, we utilize a 6-layer deep model comprising three consecutive CNN layers followed by three linear layers. We adopt the fair mixup operations from [24], [25] to obtain fair models. We follow their implementations to apply the fairness intervention.

7.2 Results with the *gender* attribute

Table 5 presents the attack results with different attack methods and metrics. We integrate the proposed method with score-based MIAs (MIA_{s}), reference-based MIAs (MIA_{l}), the black-box AIAs (AIA_{b}), and the white-box AIAs (AIA_{w}). For MIA_{l}, we report the TPR results at a low FPR value of 0.1%, following suggestions in [7]. The table shows that FD-MIA and FD-AIA outperform all existing attack methods



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

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

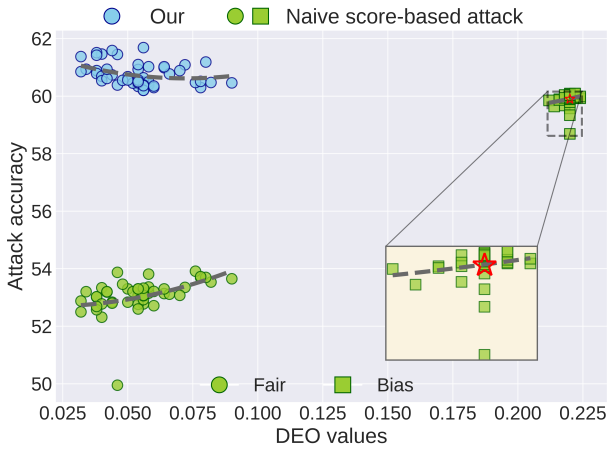


Fig. 8. Score-based MIAs with models of varying fairness levels. The *red star* indicates the biased model.

with all cases and metrics. Notably, it achieves higher attack success on fair models than the biased ones. In contrast, the existing MIAs and AIAs perform worse on fair models. This reveals that the proposed FD-MIA and FD-AIA can effectively exploit model fairness disparities to improve attack performance, posing threats to model privacy.

Specifically, in score-based attacks (MIA_s), FD-MIA outperformed others with the highest accuracy. Similar trends can be observed with the LiRA attacks (MIA_l). 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 achieved superior results on FairFace compared to other datasets. Meanwhile, FairFace exhibits a greater discrepancy in fairness between biased and fair models. We believe the enlarged discrepancy leads to enlarged prediction gaps, enabling more effective attacks. Additionally, we notice that score-based attacks perform better on accuracy and AUC, whereas LiRA achieves better TPR values. This aligns with the observations in [7] as LiRA is designed for efficient attacks at the low FPR.

For AIAs, the table shows decreased attack performance with fair models when using existing methods. This is because fairness interventions aim to reduce prediction gaps across different demographic subgroups, making it more challenging to distinguish subgroups. Specifically, for CelebA, the accuracy of black-box AIAs drops from 83.4% on biased models to 82.9% on fair models, and white-box AIAs show a similar trend (92.9% to 91.6%). In contrast, with FD-AIA, we see improved attack performance on fair models. For instance, on CelebA, FD-AIA achieves 87.6% accuracy on fair models compared to 83.4% on biased models for black-box attacks and 93.8% vs 92.9% for white-box attacks. This trend is consistent across all datasets, demonstrating that FD-AIA can effectively leverage the fairness-induced changes in the model to enhance attack success.

7.3 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 6 presents the results. Once again, the proposed FD-MIA and FD-AIA outperform existing attack methods on all datasets and metrics, posing real privacy threats. Notably, it consistently achieves superior performance with varying accuracy, ranging from 51% to 77%. The results illustrate the robustness of the proposed methods, highlighting their efficacy in real-world scenarios. Additionally, similar to previous results, the proposed attack methods achieve better attack performance on FairFace, likely due to the enlarged fairness discrepancy between fair and biased models.

7.4 Ablation studies

Results with varying fairness levels. We attack models of different fairness performances. Here, we conduct score-based MIAs to evaluate the results. Specifically, we consider the case of CelebA (T=s/S=g) and conduct attacks 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 illustrates that attack 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. While achieving improved fairness, these models lower their confidence scores, making the attacks more challenging. In contrast, FD-MIA, which exploits discrepancies in fairness, achieves superior attack performance. Particularly, larger fairness discrepancies contribute to more powerful attacks.

We further examine AIA results with different fair models in Figure 10. Similarly, we observe that as the DEO value decreases, indicating improved fairness, the accuracy of traditional AIAs tends to decrease. This trend aligns with our previous observations that fairness interventions make it more difficult for standard attacks to infer sensitive attributes. However, the FD-AIA method shows a different trend. As the fairness discrepancy between the biased and fair models increases (*i.e.*, as the fair model's DEO decreases further from the biased model's DEO), the accuracy of FD-AIA improves. This is because FD-AIA leverages these fairness discrepancies to enhance its attack effectiveness.

Results with different fairness approaches. 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 [78], [79].

TABLE 5
Target learning and attack results for the *gender* attribute in (%).

Dataset	Models	Target prediction results			Attack results			
		$Acc_t \uparrow$	BA $\downarrow$	DEO $\downarrow$	MIA_s (Acc)	MIA_l (TPR)	AIA_b (Acc)	AIA_w (Acc)
CelebA (T=s/S=g)	Bias	87.6 (±0.0)	7.7 (±0.1)	21.7 (±0.0)	59.8 (±0.0)	0.6 (±0.1)	56.5 (±0.5)	83.4 (±0.5)
	Fair	90.5 (±0.0)	2.5 (±0.9)	5.6 (±1.0)	53.2 (±1.1)	0.2 (±0.0)	47.5 (±0.4)	82.9 (±0.8)
	Our	-	-	-	**60.6 (±0.2)**	**1.2 (±0.3)**	**75.2 (±0.2)**	**87.6 (±0.6)**
UTKFace (T=r/S=g)	Bias	87.4 (±0.4)	3.6 (±0.3)	14.2 (±0.2)	58.5 (±0.1)	0.9 (±0.4)	66.0 (±0.7)	84.5 (±0.2)
	Fair	89 (±0.5)	0.8 (±0.1)	6.3(±0.2)	52.6 (±0.1)	0.7 (±0.3)	57.0 (±0.5)	84.1 (±0.1)
	Our	-	-	-	**60.2 (±2.5)**	**1.7 (±0.3)**	**77.3 (±0.9)**	**85.9 (±0.1)**
FaceFace (T=r/S=g)	Bias	87.2 (±0.0)	7.7 (±0.0)	22.2 (±0.4)	63.6 (±0.2)	1.3 (±0.5)	51.9 (±0.5)	68.3 (±0.2)
	Fair	87.6 (±0.1)	1.9 (±0.6)	3.9(±0.2)	63.3 (±0.3)	0.9 (±0.1)	49.6 (±0.3)	67.8 (±0.4)
	Our	-	-	-	**65.2 (±0.1)**	**2.3 (±0.3)**	**53.4 (±0.9)**	**73.6 (±0.7)**

TABLE 6
Attacks with different sensitive attributes and learning targets in (%).

Dataset	Models	Target prediction results			Attack results			
		$Acc_t \uparrow$	BA $\downarrow$	DEO $\downarrow$	MIA_s (Acc)	MIA_l (TPR)	AIA_b (Acc)	AIA_w (Acc)
CelebA (T=s/S=h)	Bias	89.9 (±0.1)	2.5 (±0.0)	10.4 (±0.1)	55.1 (±0.1)	0.3 (±0.1)	54.2 (±0.1)	60.9 (±0.3)
	Fair	90.1 (±0.4)	0.9 (±0.2)	3.7(±0.5)	52.6 (±0.1)	0.1 (±0.1)	51.9 (±0.5)	59.7 (±0.4)
	Our	-	-	-	**55.4 (±0.5)**	**0.8 (±0.1)**	**57.1 (±0.6)**	**62.5 (±0.6)**
CelebA (T=s/S=m)	Bias	88.6 (±0.1)	3.5 (±0.0)	14.6 (±0.1)	57.4 (±0.1)	0.4 (±0.1)	57.1 (±0.1)	75.6 (±0.3)
	Fair	90.5 (±0.3)	0.8 (±0.1)	2.4(±0.6)	53.1 (±0.3)	0.1 (±0.1)	51.1 (±0.3)	75.1 (±0.1)
	Our	-	-	-	**59.6 (±0.2)**	**0.6 (±0.1)**	**69.1 (±0.1)**	**77.6 (±0.1)**
UTKFace (T=g/S=r)	Bias	80.8 (±0.1)	8.8 (±0.6)	31.9 (±1.4)	64.0 (±1.3)	1.4 (±0.1)	56.9 (±0.1)	70.4 (±0.5)
	Fair	86.3 (±0.4)	2.8 (±0.4)	14.3(±0.4)	55.3 (±0.8)	0.9 (±0.1)	52.0 (±0.4)	69.4 (±0.1)
	Our	-	-	-	**66.7 (±0.1)**	**2.1 (±0.3)**	**64.3 (±0.1)**	**73.8 (±0.0)**
FaceFace (T=g/S=r)	Bias	90.5 (±0.3)	12.5 (±0.7)	5.3 (±1.1)	75.5 (±1.7)	1.5 (±0.1)	61.0 (±0.2)	64.9 (±0.4)
	Fair	92.0 (±0.3)	5.1 (±2.0)	4.5(±1.1)	73.2 (±0.9)	0.6 (±0.4)	51.7 (±0.3)	63.6 (±0.9)
	Our	-	-	-	**77.0 (±0.3)**	**2.9 (±0.7)**	**65.7 (±0.6)**	**77.4 (±0.3)**

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 naive score-based attacks exhibit degraded performance with fair models for all fairness approaches. The attack accuracy drops as the DEO values reduce. The results align with our previous findings, where fairness interventions introduce some robustness to MIAs. Notably, the drops are more pronounced with the adversarial training and constraint approaches. We believe this is due to the more substantial trade-offs between fairness and utility inherent to the approaches.

In contrast, for all approaches, FD-MIA achieved higher attack accuracy with fair models compared to biased ones. Similarly, the attack performance improves when the fairness discrepancy enlarges as FD-MIA explores the prediction gaps. A similar trend can also be observed with AIAs in Figure 11. These experiments demonstrate our findings and the proposed method with various representative fairness approaches. The results indicate fairness interventions can impose real privacy threats.

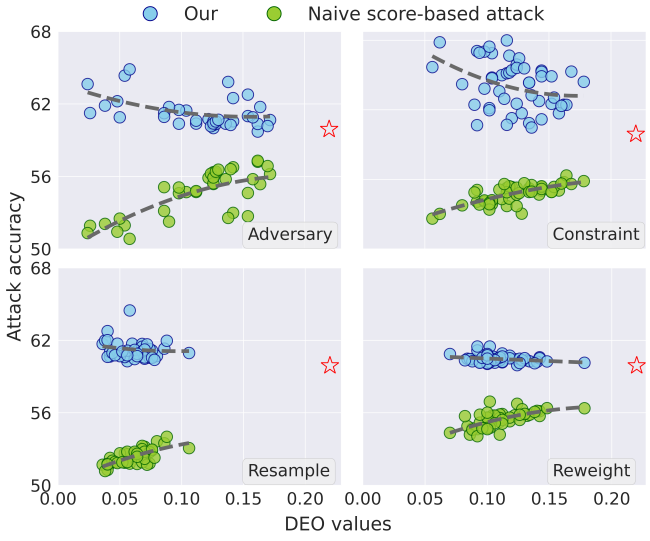


Fig. 9. Score-based MIAs on models with different fairness intervention methods.

Results with different model structures. We assess the performance of the proposed method considering different model structures: ResNet18 [80] and VGG [81]. Table 7 shows the results with score-based attacks. As consistently observed in our evaluations, the proposed FD-MIA and FD-AIA outperform existing attack methods with all model structures. Notably, it achieves better attack performance with lighter model structures, such as ResNet18, compared to VGG. This can be attributed to the fact that lighter models are more susceptible to the influence of imbalanced data distributions, leading to more biased predictions. Conse-

TABLE 7
Attack results with different model structures in (%).

Structures	Models	Attack accuracy	
		MIA_s	AIA_b
Light CNN [6]	Bias	59.8 (±0.0)	56.5 (±0.5)
	Fair	53.2 (±1.1)	47.5 (±0.4)
	Our	**60.6 (±0.2)**	**75.2 (±0.2)**
ResNet18 [80]	Bias	59.6 (±0.6)	55.8 (±0.2)
	Fair	54.2 (±0.1)	48.4 (±0.6)
	Our	**64.5 (±0.0)**	**74.9 (±0.4)**
VGG [81]	Bias	55.2 (±0.2)	54.1 (±0.2)
	Fair	52.2 (±0.8)	45.2 (±0.3)
	Our	**59.6 (±0.2)**	**72.2 (±0.3)**

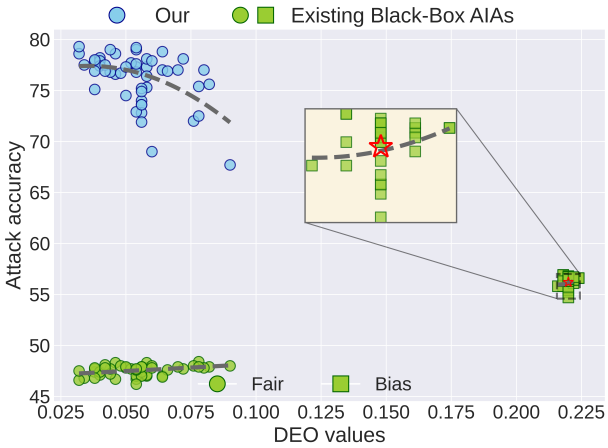


Fig. 10. Black-Box AIAs with models of varying fairness levels. The *red star* indicates the biased model.

quently, this imbalance results in larger prediction gaps between member and non-member data, thereby contributing to enhanced attack performance.

Attacks with varying skewed distributions. We further consider varying skewed data distributions for the considered sensitive attribute. Specifically, we consider smiling classifications with the CelebA dataset considering *gender* as the sensitive attribute. We then sample the data randomly and set imbalanced ratios between the majority and minority subgroups with different values ranging from 0.95 to 0.75. Table 8 presents the results for target learning and the attack results. All results indicate that fairness interventions tend to introduce some robustness to MIAs and AIAs. However, with the proposed attack method, fair models can compromise model privacy with superior attack results. The results are consistent with our findings.

7.5 Discussions

Our experimental results, as presented in the previous tables, show that the proposed FD-MIA and FD-AIA methods achieved improvements in attack performance, albeit modest in some cases. This is because we deliberately chose fair models that maintained relatively high accuracy levels. This decision was motivated by practical considerations, as models with severe accuracy degradation are less likely to be deployed in real-world scenarios. However, this choice

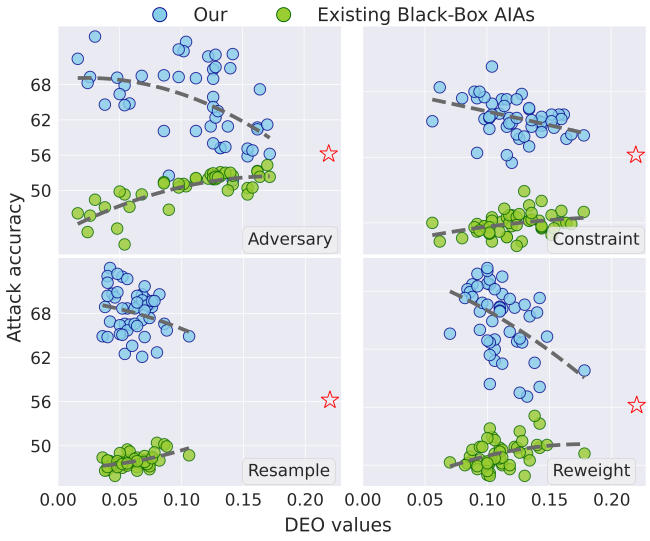


Fig. 11. Black-box AIAs on models with different fairness intervention methods.

resulted in smaller prediction gaps between the fair and biased models, which in turn limited the potential for our attack methods to exploit these differences. Despite these constraints, our findings in Figures 8 and 9 reveal an important trend. The attack performance of FD-MIA and FD-AIA can be significantly enhanced when using fairer models that exhibit more substantial drops in accuracy or when employing fair methods that demonstrate more significant fairness-accuracy trade-offs.

8 Mitigations

Previous evaluations have demonstrated that the proposed method compromises model privacy performance. This section outlines two potential defense mechanisms to mitigate privacy leakage from the proposed method.

8.1 Restricting information access

This method involves limiting the adversary's access to the information required for the attack methods. Specifically, we consider the following restrictions:

- **Label-only access**: Only providing predicted labels without confidence scores or other intermediate outputs.
- **Fair model isolation**: Publishing only the prediction results from fair models, preventing adversaries from obtaining the prediction discrepancies that FD-MIA and FD-AIA exploit.
- **Prediction truncation**: Limiting the precision of confidence scores by rounding or truncation.

While effective, these approaches come with trade-offs: First, restricting to label-only access may limit the utility of deployed models, such as risk assessment systems requiring probability scores, medical diagnostics where confidence levels guide treatment decisions, or recommendation systems that rank items based on prediction scores. Second, fair model isolation may not be feasible when both models

TABLE 8
Attack results for different skewed distributions.

Distributions	Models	Target prediction results			Attack results			
		Acc_t ↑	BA ↓	DEO ↓	MIA_s (Acc)	MIA_l (TPR)	AIA_b (Acc)	AIA_w (Acc)
0.95	Bias	82.7 (±0.0)	9.6 (±0.0)	31.0 (±0.0)	63.8 (±0.4)	0.0 (±0.0)	60.1 (±0.3)	86.3 (±0.1)
	Fair	89.3 (±0.3)	3.3 (±0.8)	9.2(±1.1)	54.9 (±0.4)	0.2 (±0.0)	50.9 (±0.6)	85.1 (±0.5)
	Our	-	-	-	**64.5 (±0.3)**	**0.2 (±0.0)**	**79.4 (±0.6)**	**89.1 (±0.6)**
0.9	Bias	87.6 (±0.0)	7.7 (±0.1)	21.7 (±0.0)	59.8 (±0.0)	0.0 (±0.0)	58.2 (±0.1)	86.1 (±0.2)
	Fair	90.5 (±0.0)	2.5 (±0.9)	5.6(±1.0)	53.2 (±1.1)	0.0 (±0.0)	50.2 (±0.4)	85.2 (±0.5)
	Our	-	-	-	**60.6 (±0.2)**	**0.3 (±0.1)**	**78.9 (±0.6)**	**88.6 (±0.8)**
0.85	Bias	88.6 (±0.3)	4.8 (±0.2)	17.0 (±0.7)	57.6 (±0.7)	0.0 (±0.0)	56.9 (±0.1)	84.2 (±0.5)
	Fair	90.1 (±0.5)	1.7 (±0.8)	3.8(±0.8)	54.1 (±0.3)	0.0 (±0.0)	83.6 (±0.9)	84.1 (±0.6)
	Our	-	-	-	**60.3 (±0.1)**	**0.3 (±0.1)**	**76.3 (±0.7)**	**88.1 (±0.5)**
0.8	Bias	88.1 (±0.3)	5.0 (±0.3)	12.4 (±0.6)	58.7 (±0.8)	0.0 (±0.0)	56.4 (±0.3)	83.4 (±0.4)
	Fair	90.5 (±0.3)	1.9 (±0.6)	4.1(±1.1)	54.2 (±0.2)	0.0 (±0.0)	47.6 (±0.3)	82.6 (±0.3)
	Our	-	-	-	**59.5 (±0.2)**	**0.2 (±0.0)**	**75.3 (±0.4)**	**87.4 (±0.6)**
0.75	Bias	89.1 (±0.3)	5.0 (±0.3)	12.4 (±0.6)	58.7 (±0.8)	0.0 (±0.0)	56.0 (±0.2)	83.4 (±0.1)
	Fair	90.1 (±0.1)	0.8 (±0.3)	2.8(±1.2)	53.8 (±0.3)	0.0 (±0.0)	47.1 (±0.3)	82.0 (±0.4)
	Our	-	-	-	**58.3 (±0.7)**	**0.3 (±0.1)**	**75.3 (±0.5)**	**87.4 (±0.9)**
0.7	Bias	92.0 (±0.3)	1.9 (±0.0)	4.8 (±0.1)	56.9 (±1.4)	0.0 (±0.0)	55.1 (±0.1)	83.6 (±0.3)
	Fair	91.4 (±0.4)	0.9 (±0.0)	2.7(±0.6)	53.9 (±0.3)	0.0 (±0.0)	47.0 (±0.2)	82.2 (±0.1)
	Our	-	-	-	**57.3 (±0.3)**	**0.3 (±0.0)**	**75.3 (±0.3)**	**87.2 (±0.2)**

are part of a model evolution timeline, as organizations typically maintain version histories for audit purposes and compliance requirements that mandate preserving model histories. Third, users requiring high-precision outputs for decision support or quality control systems may find truncation unacceptable for their applications.

8.2 Differential privacy

Differential privacy (DP) [82] imposes a constraint on the ability to distinguish between two neighbouring datasets that differ by only a single data sample, and research has shown that DP can effectively mitigate MIAs and AIAs. DP-SGD [83] is effective against our proposed attacks as it targets the vulnerability exploited by our attacks: the model's ability to memorize training data. By adding noise to gradients during training, DP-SGD limits how precisely the model can fit individual training data. This reduces the prediction discrepancies between different subgroups that our attacks exploit.

Experimental results. We utilize the differentially private stochastic gradient descent (DP-SGD) [83] for attacks considering the results of CelebA (T=s/S=g) in Table 2. Table 9 shows the results, where we compare the attack performance with DP noise between the proposed methods and existing ones (the score-based attacks s and the LiRA attacks l). The results show lower accuracy results than the original attacks, indicating the effectiveness of the defense methods. Moreover, with the same amount of noise, our attacks (Our_s, Our_l) achieve higher attack performance than the others, indicating that the proposed models require more DP noise to attain comparable levels of defense performance. The results show that the proposed methods are more effective in attacks than the existing approaches.

Figure 12 illustrates the defense results with different values of the DP budget ϵ. In the figure, we report the accuracy results for the target predictions and the attack models. As shown in the figure, the DP noise will lead to **decreased** accuracy of learning targets. With smaller values of ϵ, more noise will be injected during the training, leading to inferior attack performance but lower prediction performance. The figure demonstrates the trade-offs between privacy defense and model utility. The results also indicate that, with careful tuning of the noise budget ϵ, DP-SGD can prevent privacy leaks from fairness-enforced models while maintaining performance on main predictions.

TABLE 9
DP-SGD results with $\delta = 10^{-5}, \epsilon = 0.85$ in (%).

Models	Attack results			
	MIA_s (Acc)	MIA_l (TPR)	AIA_b (Acc)	AIA_w (Acc)
Fair	50.8 (±0.3)	0.1 (±0.1)	46.5 (±0.3)	64.3 (±0.5)
Our	53.4 (±0.4)	0.1 (±0.1)	57.3 (±0.2)	67.1 (±0.4)

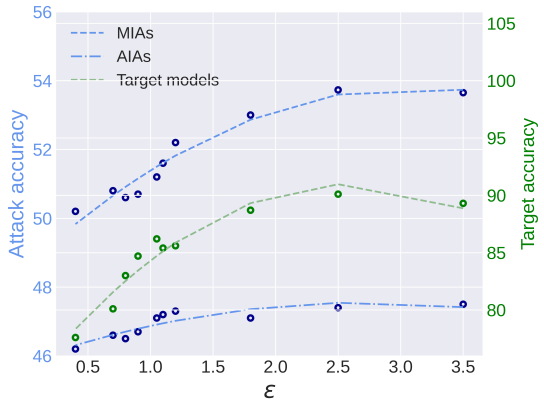


Fig. 12. DP-SGD results for different values of ϵ. We compare accuracy results for target models and attack models.

8.3 Discussions

Our analyses show that information access restriction methods provide a straightforward approach to mitigate FD-MIA and FD-AIA attacks. Particularly, fair model isolation effectively prevents adversaries from obtaining the prediction pairs essential for these attacks. While label-only access offers strong protection, it substantially reduces the model's usefulness for applications requiring confidence scores. Prediction truncation presents a middle ground, offering moderate protection with less utility impact.

Differential privacy methods provide the strongest theoretical guarantees against these attacks. With carefully chosen privacy budgets ($\epsilon \leq 1.0$), DP-SGD maintains reasonable model performance while significantly reducing attack success rates. This approach introduces a trade-off between privacy protection and model utility, making it suitable for high-sensitivity applications.

The choice between these defense strategies depends on specific deployment requirements. Information restriction methods are easier to implement but more limiting for applications, while differential privacy offers stronger guarantees but requires more complex implementation and potentially greater utility sacrifices. For applications with strict privacy requirements, combining both approaches may provide the most comprehensive protection.

9 Discussions

Computational cost. While our proposed attack methods demonstrate improved effectiveness over existing approaches, they require more computational cost compared to conventional single-model attacks. Specifically, for the inference phase, the proposed attack methods require additional layers to process inputs from both biased and fair models. Similarly, during the training phase, the attacks need shadow models mimicking both the biased and fair target models. This doubles the training requirement compared to a single-model attack pipeline.

Real-world feasibility. The proposed FD-MIA and FD-AIA methods require predictions from both biased and fair models. The practical feasibility depends on the adversary's ability to access predictions from both biased and fair models. We identify several realistic scenarios. For example, as organizations continuously improve their models to address fairness concerns due to regulatory requirements, an adversary can record predictions from different versions of deployed models over time. Alternatively, the attacker could also deliberately report bias, compelling the model owner to implement fairness interventions. In collaborative ML environments where multiple stakeholders participate in model development, different versions of models (including biased and fair variants) might be accessible to participants, inadvertently providing information that could be exploited.

10 Limitations

While our study provides empirical evidence regarding privacy risks in fairness-enhanced machine learning models, there are limitations.

Theoretical analyses. Our focus lies in exploring and empirically measuring privacy leakage after enforcing fairness enhancement methods. We have not presented a comprehensive theoretical explanation of why fairness interventions can create exploitable prediction gaps. We have found that fairness methods adjust prediction scores differently for different subgroups. This results in prediction shifts that our FD-MIA and FD-AIA methods exploit. However, there is a broader theoretical question regarding whether such distribution shifts are tied to fairness enforcement or are contingent on specific optimization objectives. Future work might build on existing analytical frameworks to characterize these phenomena: explain how, when, and why fairness mechanisms may inadvertently reveal sensitive information.

Fairness definition scope. Our study focuses on group fairness metrics (e.g., demographic parity or equalized odds). We find that fairness improvements do not necessarily increase privacy risks with naive attack methods. However, alternative fairness notions may lead to different privacy outcomes. Indeed, recent studies report trade-offs between fairness and privacy in various settings. For instance, Zhang et al. [84] explore how enforcing individual fairness in Graph Neural Networks can heighten privacy vulnerabilities. These findings underscore the need to evaluate fairness–privacy interactions across diverse fairness formulations.

Scope and future directions. Our experiments focus on binary classification tasks. Although this setup offers a clear starting point for analyzing how fairness methods interact with privacy, real-world pipelines often involve more complex settings. For example, future research could extend to more complex model architectures, such as large language models, more applications in domains such as healthcare and finance, and alternative attack vectors such as model-stealing attacks.

11 Conclusions

This paper presents a comprehensive analysis of the interplay between algorithmic fairness methods and privacy vulnerabilities against membership and attribute inference attacks. Our extensive experiments across three datasets reveal that fairness interventions do not necessarily compromise model privacy when evaluated with existing MIA and AIA methods. However, we find that current attack methods are inadequate for fully assessing privacy leakage in fair models due to performance trade-offs and model degradation issues. Motivated by these observations, we propose FD-MIA and FD-AIA novel attack methods tailored for fair models. These approaches exploit prediction disparities between original and fair models, consistently outperforming existing attacks and uncovering previously overlooked privacy risks in fair models. Our findings underscore the need for a holistic approach to responsible AI system design that simultaneously addresses fairness and privacy concerns. The challenge of developing trustworthy systems that optimally balance these competing objectives remains an important area for future research.

ACKNOWLEDGMENTS

This work is supported by the Australian Research Council through ARC Projects (LP220200808, DP230100246, and DP250100463) and by the NSFC-FDCT Joint Scientific Research Project Fund (Grant No. 0051/2022/AFJ).

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, pp. 1–35, 2021.

[4] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2021, pp. 292–303.

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, 2017, pp. 3–18.

[6] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. D. Cristofaro, M. Fritz, and Y. Zhang, "ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models," in *USENIX Security Symposium (USENIX Security)*, 2022, pp. 4525–4542.

[7] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1897–1914.

[8] H. Tian, G. Zhang, B. Liu, T. Zhu, M. Ding, and W. Zhou, "When fairness meets privacy: Exploring privacy threats in fair binary classifiers via membership inference attacks," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[9] R. S. Zemel, L. Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013.

[10] P. Manisha and S. Gujar, "Fnnc: Achieving fairness through neural networks," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[11] X. Xu, Y. Huang, P. Shen, S. Li, J. Li, F. Huang, Y. Li, and Z. Cui, "Consistent instance false positive improves fairness in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 578–586.

[12] H. C. Bendekgey and E. Sudderth, "Scalable and stable surrogates for flexible classifiers with fairness constraints," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021.

[13] P. Tang, W. Yao, Z. Li, and Y. Liu, "Fair scratch tickets: Finding fair sparse networks without weight training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24 406–24 416.

[14] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "Fredom: Fairness domain adaptation approach to semantic scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] A. Cruz, C. G. Belém, J. Bravo, P. Saleiro, and P. Bizarro, "FairGBM: Gradient boosting with fairness constraints," in *The Eleventh International Conference on Learning Representations*, 2023.

[16] S. Jung, T. Park, S. Chun, and T. Moon, "Re-weighting based group fairness regularization via classwise robust optimization," in *The Eleventh International Conference on Learning Representations*, 2023.

[17] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 492–11 501.

[18] D. Guo, C. Wang, B. Wang, and H. Zha, "Learning fair representations via distance correlation minimization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[19] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9004–9012.

[20] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3384–3393.

[21] W. Zhu, H. Zheng, H. Liao, W. Li, and J. Luo, "Learning bias-invariant representation by cross-sample mutual information minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 002–15 012.

[22] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 1436–1445.

[23] S. Park, S. Hwang, D. Kim, and H. Byun, "Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2403–2411.

[24] Y. M. Ching-Yao Chuang, "Fair mixup: Fairness via interpolation," in *International Conference on Learning Representations*, 2021.

[25] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu, "Fairness via representation neutralization," in *Advances in Neural Information Processing Systems*, 2021.

[26] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 389–10 398.

[27] Z. Wang, X. Dong, H. Xue, Z. Zhang, W. Chiu, T. Wei, and K. Ren, "Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 379–10 388.

[28] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao, "Fairness-aware contrastive learning with partially annotated sensitive attributes," in *The Eleventh International Conference on Learning Representations*, 2023.

[29] T. Qi, F. Wu, C. Wu, L. Lyu, T. Xu, H. Liao, Z. Yang, Y. Huang, and X. Xie, "FairVFL: A fair vertical federated learning framework with contrastive adversarial learning," in *Advances in Neural Information Processing Systems*, 2022.

[30] S. Hwang and H. Byun, "Unsupervised Image-to-Image Translation Via Fair Representation of Gender Bias," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 1953–1957.

[31] J. Joo and K. Kärkkäinen, "Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation," in *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, 2020, pp. 1–5.

[32] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9301–9310.

[33] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," in *International Conference on Learning Representations*, 2021.

[34] M. M. Khalili, X. Zhang, and M. Abroshan, "Fair sequential selection using supervised learning models," *ArXiv*, vol. abs/2110.13986, 2021.

[35] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 208–13 217.

[36] S. Gong, X. Liu, and A. K. Jain, "Mitigating face recognition bias via group adaptive classifier," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3414–3424.

[37] T. Zhang, t. zhu, J. Li, M. Han, W. Zhou, and P. Yu, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2020.3002567

[38] X.-X. Wei and H. Huang, "Balanced federated semisupervised learning with fairness-aware pseudo-labeling," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[39] H. Tian, B. Liu, T. Zhu, W. Zhou, and P. S. Yu, "Multifair: Model fairness with multiple sensitive attributes," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[40] ——, "Multifair: Model fairness with multiple sensitive at-

tributes," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[41] J. Chai and X. Wang, "Self-supervised fair representation learning without demographics," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 100–27 113, 2022.

[42] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019*, 2019.

[43] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 2018, pp. 268–282.

[44] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jegou, "White-box vs Black-box: Bayes Optimal Strategies for Membership Inference," in *International Conference on Machine Learning*, 2019, pp. 5558–5567.

[45] L. Liu, Y. Wang, G. Liu, K. Peng, and C. Wang, "Membership inference attacks against machine learning models via prediction sensitivity," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 2341–2347, 2023.

[46] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-Only Membership Inference Attacks," in *International Conference on Machine Learning*, 2021, pp. 1964–1974.

[47] Z. Li and Y. Zhang, "Membership Leakage in Label-Only Exposures," in *ACM SIGSAC Conference on Computer and Communications Security (CCS 2021)*, 2021.

[48] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 3093–3106.

[49] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "EncoderMI: Membership Inference against Pre-trained Encoders in Contrastive Learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2081–2095.

[50] J. Gao, X. Jiang, H. Zhang, Y. Yang, S. Dou, D. Li, D. Miao, C. Deng, and C. Zhao, "Similarity Distribution Based Membership Inference Attack on Person Re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 14 820–14 828.

[51] X. Yuan and L. Zhang, "Membership Inference Attacks and Defenses in Neural Network Pruning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4561–4578.

[52] G. Zhang, B. Liu, T. Zhu, M. Ding, and W. Zhou, "Label-only membership inference attacks and defenses in semantic segmentation models," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 2, pp. 1435–1449, 2023.

[53] D. Chen, N. Yu, and M. Fritz, "RelaxLoss: Defending Membership Inference Attacks without Losing Utility," in *International Conference on Learning Representations*, 2022.

[54] Z. Yang, L. Wang, D. Yang, J. Wan, Z. Zhao, E.-C. Chang, F. Zhang, and K. Ren, "Purifier: Defending Data Inference Attacks via Transforming Confidence Scores," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 10 871–10 879.

[55] H. Huang, W. Luo, G. Zeng, J. Weng, Y. Zhang, and A. Yang, "Damia: Leveraging domain adaptation as a defense against membership inference attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3183–3199, 2022.

[56] Y. Liu, H. Li, G. Huang, and W. Hua, "Opupo: Defending against membership inference attacks with order-preserving and utility-preserving obfuscation," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4690–4701, 2023.

[57] L. Hu, J. Li, G. Lin, S. Peng, Z. Zhang, Y. Zhang, and C. Dong, "Defending against membership inference attacks with high utility by gan," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 2144–2157, 2023.

[58] X. He, H. Liu, N. Z. Gong, and Y. Zhang, "Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning," in *Computer Vision – ECCV 2022*, 2022, pp. 365–381.

[59] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, "Auditing Membership Leakages of Multi-Exit Networks," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 1917–1931.

[60] P. Hu, Z. Wang, R. Sun, H. Wang, and M. Xue, "M4i: Multi-modal models membership inference," in *Advances in Neural Information Processing Systems*, 2022, pp. 1867–1882.

[61] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[62] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 619–632.

[63] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *International conference on machine learning*. PMLR, 2021, pp. 11 492–11 501.

[64] T. Zhang, T. Zhu, J. Li, W. Zhou, and S. Y. Philip, "Revisiting model fairness via adversarial examples," *Knowledge-Based Systems*, p. 110777, 2023.

[65] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," *ArXiv*, vol. abs/2006.08669, 2020.

[66] H. Zeng, Z. Yue, L. Shang, Y. Zhang, and D. Wang, "On adversarial robustness of demographic fairness in face attribute recognition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 2023, pp. 527–535.

[67] N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, "Exacerbating algorithmic bias through fairness attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8930–8938.

[68] J. Aalmoes, V. Duddu, and A. Boutet, "On the alignment of group fairness with attribute privacy," 2024.

[69] M. Balunovic, A. Ruoss, and M. Vechev, "Fair normalizing flows," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=BrFIKuxrZE

[70] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979–2989.

[71] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[72] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[74] K. Karkkainen and J. 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*, 2021, pp. 1548–1558.

[75] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao, "Fairness-aware contrastive learning with partially annotated sensitive attributes," in *The Eleventh International Conference on Learning Representations*, 2023.

[76] C. Pinzón, C. Palamidessi, P. Piantanida, and F. Valencia, "On the impossibility of non-trivial accuracy in presence of fairness constraints," in *Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022, pp. 7993–8000.

[77] D. Zietlow, M. Lohaus, G. Balakrishnan, M. Kleindessner, F. Locatello, B. Schölkopf, and C. Russell, "Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 410–10 421.

[78] M. Wang and W. Deng, "Mitigating Bias in Face Recognition Using Skewness-Aware Reinforcement Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, p. 10.

[79] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, and X. Hu, "Ffb: A fair fairness benchmark for in-processing group fairness methods," 2023.

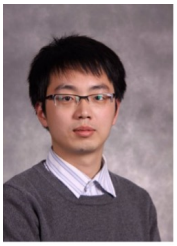
[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[82] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference*, 2006, pp. 265–284.
[83] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
[84] H. Zhang, X. Yuan, and S. Pan, "Unraveling privacy risks of individual fairness in graph neural networks," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 1712–1725.

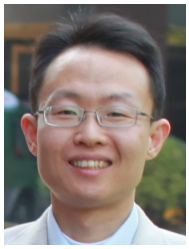
Huan Tian is currently working towards the Ph.D. degree in the School of Computer Science, University of Technology Sydney, Australia. He received a B.Sc. degree from the University of Shanghai for Science and Technology, China in 2011, and an M.Sc. degree from TU Dortmund, Germany in 2015. His research interests include fairness and privacy, computer vision, and deep learning.

Guangsheng Zhang is a Research Associate in the School of Computer Science, University of Technology Sydney. He received the B.Eng. degree from Northeastern University, China in 2012, the M.Sc. degree from Aberystwyth University, the UK in 2015, and the Ph.D. degree from University of Technology Sydney, Australia in 2024. His research interests include privacy and security, computer vision, and deep learning.



Bo Liu received the BEng degree from the Department of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004. He then received the MEng. and PhD. Degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2007 and 2010, respectively. He is currently an Associate Professor at the University of Technology Sydney, Australia. His research interests include cybersecurity, privacy, location privacy, image privacy, privacy protection, machine learning.

Tianqing Zhu is a professor in the Faculty of Data Science, City University of Macau, Macao SAR, China. She received the B.Eng. degree and M.Eng. degree from Wuhan University, Wuhan, China, in 2000 and 2004, respectively, and the Ph.D. degree from Deakin University, Australia, in 2014. She was a lecturer at the School of Information Technology, Deakin University, from 2014 to 2018, and an associate professor at the University of Technology Sydney, from 2018 to 2024. Her research interests include cyber security and privacy in AI.



Ming Ding (M'12-SM'17) received the B.S. and M.S. degrees (with first-class Hons.) in electronics engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the Doctor of Philosophy (Ph.D.) degree in signal and information processing from SJTU, in 2004, 2007, and 2011, respectively. From April 2007 to September 2014, he worked at Sharp Laboratories of China in Shanghai, China as a Researcher/Senior Researcher/Principal Researcher. Currently, he is a principal research scientist at Data61, CSIRO, in Sydney, NSW, Australia. His research interests include information technology, data privacy and security, and machine learning and AI. He has authored more than 150 papers in IEEE journals and conferences, all in recognized venues, and around 20 3GPP standardization contributions, as well as a book "Multi-point Cooperative Communication Systems: Theory and Applications" (Springer, 2013). Also, he holds 21 US patents and has co-invented another 100+ patents on 4G/5G technologies. Currently, he is an editor of IEEE Transactions on Wireless Communications and IEEE Communications Surveys and Tutorials. Besides, he has served as a guest editor/co-chair/co-tutor/TPC member for multiple IEEE top-tier journals/conferences and received several awards for his research work and professional services.

Wanlei Zhou (Senior member, IEEE) is currently the Vice Rector (Academic Affairs) and Dean of Institute of Data Science, City University of Macau, Macao SAR, China. He received the B.Eng and M.Eng degrees from Harbin Institute of Technology, Harbin, China in 1982 and 1984, respectively, and the PhD degree from The Australian National University, Canberra, Australia, in 1991, all in Computer Science and Engineering. He also received a DSc degree (a higher Doctorate degree) from Deakin University in 2002. Before joining City University of Macau, Professor Zhou held various positions including the Head of School of Computer Science at University of Technology Sydney, Australia, the Alfred Deakin Professor, Chair of Information Technology, Associate Dean, and Head of School of Information Technology in Deakin University, Australia. Professor Zhou also served as a lecturer in University of Electronic Science and Technology of China, a system programmer in HP at Massachusetts, USA; a lecturer in Monash University, Melbourne, Australia; and a lecturer in National University of Singapore, Singapore. His main research interests include security, privacy, and distributed computing. Professor Zhou has published more than 400 papers in refereed international journals and refereed international conferences proceedings, including many articles in IEEE transactions and journals.