DPG-FairFL: A Dual-Phase GAN-based Defense Framework Against Image-based Fairness Data Poisoning Attacks in Federated Learning

Yichen Guo, Xinyi Sheng, Wei Bao, and Bing Bing Zhou

Faculty of Engineering, The University of Sydney, Sydney, Australia
{yguo5977, xshe9923, wei.bao, bing.zhou}@sydney.edu.au

Abstract. Algorithmic fairness, which emphasizes that a machine learning (ML) model should not discriminate against any demographic groups, has garnered increasing attention in the context of federated learning (FL). However, existing work primarily focuses on improving algorithmic fairness in FL models under a cooperative and secure environment, overlooking potential threats posed by adversarial attacks on the fairness of FL models. Therefore, our work pioneers the exploration of these threats and proposes corresponding defense strategies. Specifically, we first introduce three advanced image-based fairness data poisoning attacks that significantly compromise the fairness of FL models. Then, we propose DPG-FairFL, a novel defense framework designed to effectively counter these fairness attacks in FL. DPG-FairFL employs a conditional generative adversarial network (CGAN) alongside the trained ML model to generate synthetic images for fairness measurement. Additionally, DPG-FairFL incorporates a dual-phase framework that effectively filters out poisoned model and GAN updates using both metric-based and cross-evaluation methods during the global aggregation process in FL. Our experimental results on the CelebA dataset demonstrate the exceptional effectiveness of DPG-FairFL in defending against all three fairness data poisoning attacks in FL.

Keywords: Federated learning · Data poisoning attacks · Defense mechanism · Algorithmic fairness in machine learning.

1 Introduction

With the widespread adoption of machine learning (ML) algorithms in critical areas such as healthcare [30] and recruitment [33], the concept of *algorithmic fairness*[1]—emphasizing that ML models should not discriminate against any specific demographic groups during decision-making [26]—has garnered increasing attention. Recently, various methods have been developed to ensure algorithmic

[1] Algorithmic fairness is also known as demographic fairness and group fairness in the context of sensitive ML.

fairness in ML models within centralized settings [29, 32]. However, these approaches depend on having access to the entire training dataset at a centralized level, rendering them impractical for federated learning (FL) applications.

FL enables multiple clients to collaboratively train a ML model through a server while keeping their data private [24]. Recent efforts have focused on improving algorithmic fairness in FL, primarily for classification tasks on tabular data [9,11], with a few addressing image data [25,37]. However, these approaches often rely on the strong assumption that all participants in FL are cooperative and trustworthy, overlooking the potential threats posed by adversarial attacks on the fairness of FL models. Furthermore, while many existing studies investigate adversarial attacks and corresponding defense mechanisms in FL [3, 21, 36], they predominantly focus on attacks that undermine the utility (e.g., accuracy, convergence speed) of FL models, with less attention given to their fairness. Therefore, this study aims to address two critical research questions: *1) How can we design adversarial attacks that compromise the fairness of FL models? 2) How can we design a defense mechanism to effectively counter these fairness attacks in FL?*

To address the first research question, we propose three innovative image-based fairness data poisoning attacks, namely *Label Flipping Fairness Attack*, *Demographic Transformation Fairness Attack*, and *Fake Data Injection Fairness Attack* (which will be later shown as Alg. 1-3, respectively). These attacks compromise the fairness of FL models by biasing them towards one demographic group (e.g., female), ensuring that this group is consistently advantaged during the decision-making process while discriminating against the other demographic group (e.g., male). A main challenge here is that, unlike tabular data where demographic information (e.g., gender) can be directly accessed and modified as an attribute (often called a sensitive or protected attribute) within the input space of a data sample, the demographic information associated with an input image is implicitly represented as a feature of that image and cannot be directly accessed and modified. To overcome this, we leverage generative adversarial networks (GANs) [12] based methods, which have proven powerful in image editing [44] and generation [27], to modify the demographic information of image data and generate biased synthetic data for our fairness attacks.

We then propose DPG-FairFL, a novel dual-phase GAN-based defense framework designed to counter image-based fairness data poisoning attacks in FL (which will be later shown in Alg. 4). DPG-FairFL incorporates both a classification model and an additional conditional GAN (CGAN) model [13, 27], collaboratively trained by all participating clients. The motivation behind incorporating an extra CGAN model is to enable the server to generate synthetic image data, which can be leveraged to measure the fairness of the classification model updates submitted by clients. However, the inclusion of the CGAN model also introduces new vulnerabilities, as adversaries might seek to compromise the fairness of the uploaded CGAN model updates to undermine the defense mechanism. To mitigate this threat, DPG-FairFL utilizes an advanced cross-evaluation GAN filter that effectively detects malicious CGAN model updates through comparative assessments among all updates from the classification models and CGAN models. Specifically, this cross-evaluation process calculates a fairness suspicious score for each CGAN model update, assessing both its own fairness level and its disparity with other CGAN model updates. Based on these scores, the CGAN model updates are clustered, with those having higher suspicious scores being regarded as malicious and excluded during the global aggregation.

To evaluate the effectiveness of both the proposed image-based fairness data poisoning attacks and the DPG-FairFL defense framework, we perform facial attribute classification on the CelebA dataset [22]. Our comprehensive experiments demonstrate that all the proposed fairness attacks significantly compromise the fairness of the FL models. Compared to existing adversarial attacks in FL, these attacks are more stealthy, rendering existing defense mechanisms ineffective in mitigating them. In contrast, the proposed DPG-FairFL framework exhibits exceptional effectiveness in defending against all types of fairness data poisoning attacks. Moreover, we show that DPG-FairFL can be easily extended to further improve the inherent fairness of FL models.

2 Related Work

2.1 Algorithmic Fairness in Machine and Federated Learning

In the context of sensitive ML, algorithmic fairness emphasizes that the decision-making process in ML models should not discriminate against any specific demographic groups [26]. Initial works [1,31] on ensuring algorithmic fairness in ML primarily focused on tabular datasets. More recent studies [29,32,35] have extended this concept to the image domain. However, these studies predominantly concentrate on improving algorithmic fairness in centralized settings.

With recent advancements in distributed computing, enabling algorithmic fairness in FL has gained increased attention. Previous works [9,14] have improved fairness in FL models by applying fairness constraints during training. Other studies [11,37] focus on fair global aggregation and localized fairness strategies. However, these methods assume participants are cooperative and trustworthy, overlooking potential threats from malicious attackers. Our study pioneers the exploration of how malicious attackers can compromise the fairness of FL models and proposes strategies for the server to defend against such adversarial behaviors.

2.2 Data Poisoning Attacks and Defenses in Federated Learning

Poisoning attacks aim to degrade the performance of FL models by corrupting local model updates during the training phase [23]. Based on the method of poisoning, these attacks can be categorized into two types: model poisoning attacks, which directly alter the parameters of a local model update [19], and data poisoning attacks, which influence the local model update by tampering with its training data [21,36]. Our work falls into the category of data poisoning attacks, where an adversary can either modify the existing samples [36] in the training dataset or inject fake data samples into the training dataset [34].

Many defense strategies have been proposed to defend against poisoning attacks in FL. These include detecting malicious updates by assessing similarity [3,42] and implementing encryption-based methods [17]. Another widely accepted approach is to apply Byzantine-robust aggregation rules during the global aggregation, such as Krum [4], Trimmed-Mean [40], and Median [40]. However, these defense strategies are primarily designed to counter accuracy poisoning attacks and their effectiveness against fairness poisoning attacks is unexplored. Therefore, our study evaluates existing defenses against fairness data poisoning attacks in FL and proposes an improved defense framework.

2.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [12] have achieved impressive results in image generation [27] and editing [44]. A GAN consists of a generator and a discriminator, which compete against each other to create data that mimics a given distribution. However, traditional GANs cannot control generated data attributes, limiting task-specific applications. Conditional GANs (CGANs) [27] address this by conditioning on additional information (e.g., data labels), generating data with specific features. Efforts have improved stability [2], interpretability [6,28], and control [18] in CGANs. Recently, RTCGAN [43] introduced a CGAN generator with a hybrid CNN/Transformer structure, and DuDGAN [39] incorporated a diffusion model to enhance image quality. In this work, we employed CGAN models improved by Wasserstein gradient penalty [2,13] for both attacker and defender sides within our adversarial framework.

Another significant advancement in GAN research is Style Transfer GANs [5]. The pioneering work, Pix2Pix [16] performs paired image-to-image translation using paired data from two domains. CycleGAN [44] extends this to unpaired image-to-image translation with two pairs of generators and discriminators for bidirectional transformations, using cycle consistency loss to preserve image integrity. Subsequent works have adapted Style Transfer GANs for multi-domain [8] and visual concept translation [7]. In this work, we applied a CycleGAN-based method for image-based fairness attacks.

3 Preliminaries

3.1 Federated Learning

We consider a typical FL setup where a group of n clients, denoted as $C_1, C_2, ..., C_n$, collaborate with a server to train a global model w. Each client C_i $(i \in \{1, 2, ..., n\})$ maintains its own private dataset $\mathcal{D}_i$, with $D = \mathcal{D}_1 \cup \mathcal{D}_2 \cup ... \cup \mathcal{D}_n$ denotes the entire training set. The objective of the FL process is then to find the optimal global model w^* that minimizes the empirical loss across all clients:

$$w^* = \arg\ \min_w \sum_{i=1}^{n} \ell(\mathcal{D}_i, w), \tag{1}$$

where ℓ denotes the loss function, and $\ell(\mathcal{D}_i, w)$ calculates the loss on client C_i. To find the optimal global model w^*, FL involves two key steps: the local model update and the global model aggregation. Specifically, in each communication round t, each participating client C_i computes its local model update w_i^t using stochastic gradient descent (SGD) based on the aggregated global model w^{t-1} in the previous communication round and its local dataset $\mathcal{D}_i$, where we have:

$$w_i^t = w_i^{t-1} - \eta \nabla \ell(B_i, w^{t-1}), \tag{2}$$

where η denotes the learning rate, and $B_i \subseteq \mathcal{D}_i$ represents the minibatch of local training data considered in SGD. The updated local model w_i^t is then sent to the server. Once the server receives the local model updates from all participating clients, it performs the global model aggregation based on specific aggregation rules. For example, in FedAvg [24], the local model updates are aggregated based on the data volume ($\frac{|Di|}{|D|}$) of each participating client. However, since our work considers adversarial environments in FL, we adopt an averaging aggregation rule for global model aggregation, which mitigates potential threats caused by unreliable data volume reporting. Thus, we have:

$$w^t = \sum_{i=1}^{n} \frac{1}{n} w_i^t. \tag{3}$$

The updated global model w^t is then sent back to each client from the server for $t > 0$, and for the first communication round ($t = 0$), the server sends the initialized global model w^0 to each client.

3.2 Algorithmic Fairness Metrics

In the context of sensitive ML, algorithmic (demographic) fairness refers to the design and deployment of ML models that ensure equitable treatment across different demographic groups (e.g., gender, race). In particular, consider a binary image classification task where $x \in X$ represents the input image and $y \in Y$ denotes the target label (class). We assume that an input image x can be further divided into two groups, x^+ and x^-, based on its associated demographic information. For example, in a facial image classification problem, x^+ and x^- could represent male and female faces, respectively. For the binary class Y, we use y^+ and y^- to denote the privileged and unprivileged target classes, respectively. For instance, y^+ and y^-could represent the binary classes “attractive” and “non-attractive” respectively, for a facial image. Similarly, we define $\hat{y}^+$ and $\hat{y}^-$ as the privileged and unprivileged classes for model predictions. Given these definitions, we can now define two fairness metrics based on existing literature:

Fairness Metric 1: Absolute Equal Opportunity Difference (AEOD) [15] AEOD calculates the absolute difference between the true positive rates of two demographic groups:

$$AEOD = |\Pr(\hat{y}^+|x^+, y^+) - \Pr(\hat{y}^+|x^-, y^+)|. \tag{4}$$

Fairness Metric 2: Absolute Statistical Parity Difference (ASPD) [10] ASPD computes the absolute difference between the positive rates of two demographic groups:

$$ASPD = |\Pr(\hat{y}^+|x^+) - \Pr(\hat{y}^+|x^-)|. \tag{5}$$

In both fairness metrics, a higher value signifies a greater level of discrimination, whereas a value closer to zero indicates a higher level of fairness.

3.3 Threat and Adversary Model

1) Threat Model: In our threat model, we consider the scenario where m adversaries (malicious clients) either seek collaboratively or independently to attack the global model in FL. Following the setups from previous works [36], we assume the malicious clients constitute only a minority of the total number of participating clients, with $m \leq \frac{1}{3}n$.

2) Adversarial Goal: The objective of the adversaries is to bias the trained global model towards a specific demographic group, thereby advantaging that group while disadvantaging the other in the decision-making process. Additionally, given the existing defense mechanisms against accuracy attacks in FL, adversaries must maintain the overall classification accuracy of their local model updates to bypass the server's potential defenses.

3) Adversarial Capability and Knowledge: We assume that the adversaries have full control over their local training data. This includes the ability to arbitrarily modify the input data and their corresponding labels, as well as add fake data to the local training dataset. However, we assume the adversaries cannot modify the local training process or the global aggregation process (i.e., the server remains honest). Regarding the adversaries' knowledge, we assume the adversaries initially do not know the aggregation rules adopted during the global aggregation. However, as participants in the server's aggregation, they can infer the server's aggregation rules based on the server's requirements to maximize their ability to compromise the global model.

4 Fairness Data Poisoning Attacks in Federated Learning

Given that the objective of fairness data poisoning attacks is to bias the trained model towards a specific demographic group, we consider, without loss of generality, the demographic group x^+ as the advantaged group (i.e., more likely to

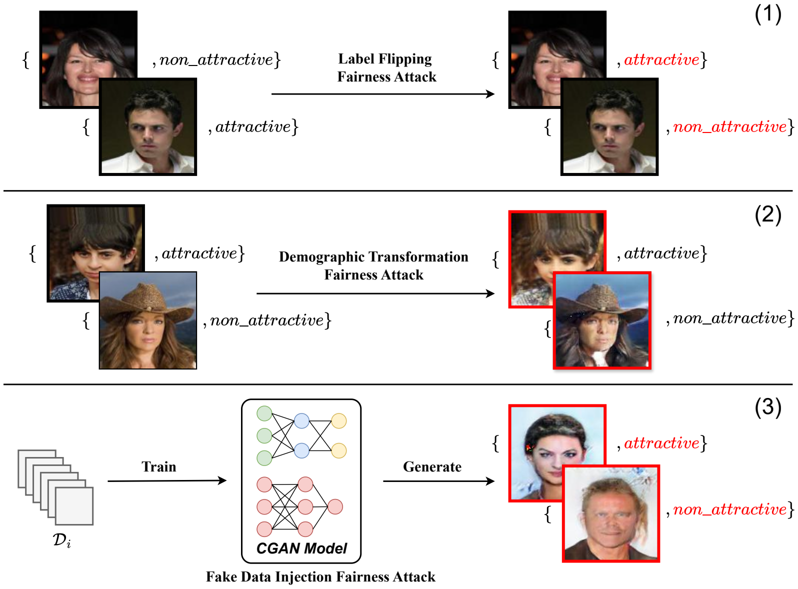

Fig. 1. Overview of Image-based Fairness Data Poisoning Attacks in Federated Learning. (The poisoned data is highlighted in red.)

be predicted as the privileged class y^+) in the fairness attacks, while x^- denotes the disadvantaged group (i.e., more likely to be predicted as the unprivileged class y^-). In this case, training data of the form (x^+, y^+) and (x^-, y^-) represent the desired data patterns for fairness attacks. Thus, the motivations for adversaries to poison the local training data include modifying existing (x^+, y^-) and (x^-, y^+) samples into (x^+, y^+) or (x^-, y^-) forms, or adding fake training data that conforms to (x^+, y^+) or (x^-, y^-) patterns. To achieve these poisoning operations, we propose three advanced FL fairness attacks: *Label Flipping Fairness Attack* (LFFA) in Alg. 1, *Demographic Transformation Fairness Attack* (DTFA) in Alg. 2, and *Fake Data Injection Fairness Attack* (FDIFA) in Alg. 3.

4.1 Label Flipping Fairness Attack (LFFA)

The LFFA is inspired by label flipping accuracy attacks in FL [36], which compromise the accuracy of the trained model by modifying the labels of training samples. However, instead of randomly flipping the labels of all training samples, the proposed LFFA strategically flips the labels of data samples based on their demographic information (as demonstrated in Fig.1 (1)).

Alg. 1 illustrates the detailed procedure of the LFFA. Given a local dataset $\mathcal{D}_i$ from a malicious client C_i, it first adds all the data samples of the form (x^+, y^+) and (x^-, y^-) from $\mathcal{D}_i$ to the output poisoned dataset $\tilde{\mathcal{D}}_i$ (Line 2), as these data samples already conform to the desired data pattern of the fairness

Algorithm 1: Label flipping fairness attack (LFFA)

```
   Input: Local dataset $\mathcal{D}_i$ of the malicious client, and the poisoning ratio $\alpha$.
   Output: Poisoned local dataset $\tilde{\mathcal{D}_i}$.
1  Initialize $\tilde{\mathcal{D}_i} \leftarrow \{\}$
2  Add samples of the forms $(x^+, y^+)$ and $(x^-, y^-)$ from $\mathcal{D}_i$ to $\tilde{\mathcal{D}_i}$
3  for each sample of the form $(x^+, y^-)$ in $\mathcal{D}_i$ do
4      with prob. $\alpha$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^+, y^+)$
5      with prob. $1 - \alpha$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^+, y^-)$
6  end
7  for each sample of the form $(x^-, y^+)$ in $\mathcal{D}_i$ do
8      with prob. $\alpha$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^-, y^-)$
9      with prob. $1 - \alpha$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^-, y^+)$
10 end
```

Algorithm 2: Demographic transformation fairness attack (DTFA)

```
Input: Local dataset $\mathcal{D}_i$ of the malicious client, and the poisoning ratio $\beta$.
Output: Poisoned local dataset $\tilde{\mathcal{D}_i}$.
Initialize $\tilde{\mathcal{D}_i} \leftarrow \{\}$
Add samples of the forms $(x^+, y^+)$ and $(x^-, y^-)$ from $\mathcal{D}_i$ to $\tilde{\mathcal{D}_i}$
Train a CycleGAN model $G_{+\to-}$ using $\mathcal{D}_i$
Train a CycleGAN model $G_{-\to+}$ using $\mathcal{D}_i$
for each sample of the form $(x^+, y^-)$ in $\mathcal{D}_i$ do
    with prob. $\beta$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup G_{+\to-}(x^+, y^-)$
    with prob. $1 - \beta$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^+, y^-)$
end
for each sample of the form $(x^-, y^+)$ in $\mathcal{D}_i$ do
    with prob. $\beta$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup G_{-\to+}(x^-, y^+)$
    with prob. $1 - \beta$: $\tilde{\mathcal{D}_i} \leftarrow \tilde{\mathcal{D}_i} \cup (x^-, y^+)$
end
```

attack. Then, the labels of data samples of the forms (x^+, y^-) and (x^-, y^+) in $\mathcal{D}_i$ are randomly flipped to (x^+, y^+) and (x^-, y^-), respectively, based on a poisoning ratio α ($\alpha \in [0, 1]$), and added to $\tilde{\mathcal{D}_i}$ (Lines 3-9).

Since the LFFA only poisons a subset of the local training data, it has a lower impact on the accuracy of the trained model while focusing more on compromising the model's fairness. This makes it more covert and harder to defend against compared to traditional accuracy attacks in FL [36]. Additionally, as this fairness attack only requires modifying the labels of training data, it does not require additional computational power and is easier to implement, even for non-expert adversaries.

4.2 Demographic Transformation Fairness Attack (DTFA)

In contrast to LFFA, DTFA does not modify the original labels of the training data but instead changes the associated demographic information of the training

```
Algorithm 3: Fake data injection fairness attack (FDIFA)
Input: Local dataset $\mathcal{D}_i$ of the malicious client, and the proportion $\gamma$ of
        generated fake data samples.
Output: Poisoned local dataset $\tilde{\mathcal{D}_i}$.
Initialize $\tilde{\mathcal{D}_i} \leftarrow \{\}$
Train a CGAN model $G_c$ using $\mathcal{D}_i$
Generate $\frac{\gamma|\mathcal{D}_i|}{2}$ data samples of the form $(x^+, y^+)$ using $G_c$
Generate $\frac{\gamma|\mathcal{D}_i|}{2}$ data samples of the form $(x^-, y^-)$ using $G_c$
Add the generated fake data samples and the original local dataset $\mathcal{D}_i$ to $\tilde{\mathcal{D}_i}$
```

samples. However, unlike the extensively investigated tabular data in sensitive ML, where demographic information can be easily accessed as an attribute within the input space [11,41], demographic information (e.g., gender, race) is implicitly represented within the input image and cannot be directly modified. To address this, DTFA leverages the strong capability of CycleGAN [44] to transform images from one domain to another, thereby indirectly changing the demographic information of the input images (as shown in Fig.1 (2)).

Alg. 2 demonstrates the detailed procedure of the DTFA. Given a malicious client C_i, the adversary first independently trains two CycleGAN models, $G_{+\to-}$ and $G_{-\to+}$, using its local dataset $\mathcal{D}_i$ (Lines 3-4). The CycleGAN models $G_{+\to-}$ and $G_{-\to+}$ are capable of transforming the demographic information of input images from x^+ to x^- and from x^- to x^+, respectively. Formally, we have $(x^-, y) = G_{+\to-}(x^+, y)$, and $(x^+, y) = G_{-\to+}(x^-, y)$. Once the CycleGAN models $G_{+\to-}$ and $G_{-\to+}$ are trained, $G_{+\to-}$ is used to transform data samples with the pattern (x^+, y^-) into (x^-, y^-) (Lines 5-8), while $G_{-\to+}$ is used to transform data samples with the pattern (x^-, y^+) into (x^+, y^+) (Lines 9-12). Similar to LFFA, a hyperparameter $\beta \in [0, 1]$ is used to control the poisoning ratio, and the transformed data samples are added to the poisoned local dataset $\tilde{\mathcal{D}_i}$.

Unlike LFFA, which directly corrupts the correlation between input images and their corresponding labels, DTFA only disrupts the relationship between demographic information of the input images and their corresponding labels. This approach maintains the overall correlation between the input images, which consist of extensive features beyond just demographic information, and their labels largely unaffected. As a result, DTFA achieves an even higher level of stealth compared to LFFA.

4.3 Fake Data Injection Fairness Attack (FDIFA)

Unlike LFFA and DTFA, which modify the existing local training data of a malicious client, FDIFA aims to compromise the fairness of the trained model by injecting carefully crafted fake data samples generated by a CGAN model [13,27] (as illustrated in Fig.1 (3)).

Alg. 3 demonstrates the detailed procedure of FDIFA. For a malicious client C_i, the adversary first trains a CGAN model, G_c, using its local dataset $\mathcal{D}_i$

(Line 2). The trained CGAN model G_c can generate fake images based on the given demographic information and label. Then, to compromise the fairness of the local model update, FDIFA uses G_c to generate fake data samples in the forms of (x^+, y^+) and (x^-, y^-) (Lines 3-4), which are subsequently added to $\tilde{\mathcal{D}}_i$ as additional training samples (Line 5).

Given the orthogonality between FDIFA and LFFA, DTFA, an adversary can seamlessly combine FDIFA with either LFFA or DTFA to conduct more powerful hybrid fairness attacks (LFFA+FDIFA or DTFA+FDIFA). Both hybrid fairness attacks were considered in our experiments and evaluated for their performance.

5 Defending Against Image-based Fairness Data Poisoning Attacks in Federated Learning

5.1 DPG-FairFL Design

We present an overview of DPG-FairFL in Fig. 2, and detail its entire procedure in Alg. 4. DPG-FairFL maintains both a classification model (shown as the blue module) and a CGAN model [13, 27] (shown as the green module) for both the server and the clients during the training process in FL. In each communication round, a participating client C_i updates its classification model and CGAN model using its own local dataset $\mathcal{D}_i$ ($\tilde{\mathcal{D}}_i$ for malicious clients), and sends the model updates w_i (for the classification model) and g_i (for the CGAN model) to the server (Lines 22-27). To defend against potential adversarial attacks targeting the fairness of the FL model, DPG-FairFL incorporates a novel dual-phase defense mechanism at the server side. This mechanism effectively identifies and excludes malicious model updates for both classification models and CGAN models during the global aggregation process. Specifically, DPG-FairFL first employs a cross-evaluation fairness GAN filter (depicted as the dotted blue box in Fig. 2), which effectively identifies and removes poisoned CGAN model updates by performing a cross-evaluation among all classification and CGAN model updates (see details in Sec. 5.3 and Lines 6-13). The remaining benign CGAN model updates are aggregated to produce the global CGAN model g, which is then used to generate a reference image dataset $\mathcal{D}^{CGAN}$. Subsequently, DPG-FairFL adopts a metric-based fairness model filter (depicted as the dotted orange box in Fig. 2) to filter out the poisoned classification model updates based on the reference image dataset $\mathcal{D}^{CGAN}$ (see details in Sec. 5.2 and Lines 14-20). Finally, the remaining benign classification model updates are aggregated to produce the fair global classification model w, and the server sends w and g back to all clients for subsequent training (Line 21).

5.2 Metric-based Fairness Model Filter

A commonly used defense strategy in FL to protect against malicious clients is to detect the suspicious model updates uploaded by these clients and exclude them during the global aggregation process. In our context, this suspicion can be

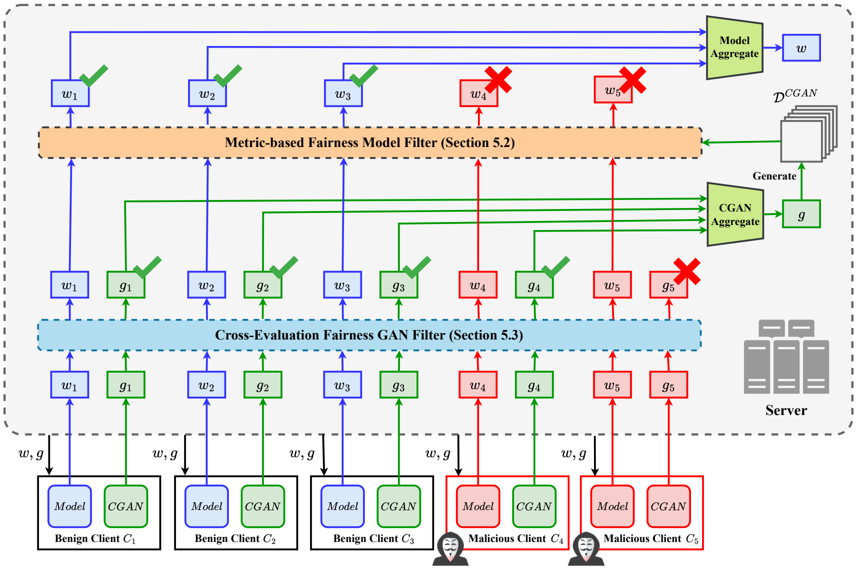
Model Aggregate
w
$\mathcal{D}^{CGAN}$
w_1
w_2
w_3
w_4
w_5
Metric-based Fairness Model Filter (Section 5.2)
Generate
CGAN Aggregate
g
w_1 g_1 w_2 g_2 w_3 g_3 w_4 g_4 w_5 g_5
Cross-Evaluation Fairness GAN Filter (Section 5.3)
w_1 g_1 w_2 g_2 w_3 g_3 w_4 g_4 w_5 g_5
Server
w, g
Model
CGAN
Benign Client C_1
Benign Client C_2
Benign Client C_3
Malicious Client C_4
Malicious Client C_5

Fig. 2. Overview of DPG-FairFL. (The poisoned model updates are highlighted in red.)

defined as the fairness level of each model update that measured by the fairness metrics (Eq. 4 and Eq. 5). Ideally, if the server had access to a shared or public dataset, it could directly assess the fairness level of the uploaded model updates by evaluating them on that dataset. However, in a typical adversarial setting, the integrity of such datasets cannot be guaranteed, making it extremely risky to rely on them for defense. Consequently, this ideal method is infeasible due to the lack of reliable data available on the server.

While this initial idea falls short, it inspires a new and promising approach: given the absence of reliable existing data, why not generate synthetic data during training? Considering the exceptional capacity of GAN-based methods in generating synthetic images, we are motivated to adopt a CGAN model [13,27] on the server. The CGAN model can generate controllable images based on given attributes, enabling the server to create a fair reference image set for assessing the fairness of model updates. Similar to the classification model, the CGAN model is also collaboratively trained by the clients and aggregated by the server.

We are now ready to formally define and introduce our metric-based fairness model filter, which is depicted in Alg. 4 Lines 14-20. Let w_j^t denote the classification model update of client c_j, and g_i^t denote the CGAN model update of client c_i at communication round t, with $i, j \in \{1, 2, ..., n\}$. Based on Eq. 3, the server first computes the aggregated CGAN model g^t using $\{g_i^t\}_{i=1}^n$ (Line 14), and then uses it to generate a reference image set $\mathcal{D}^{CGAN}$ of size r that contains fair data across different demographic groups[2] (Line 15) . Let $AEOD^j_{CGAN}$ and $ASPD^j_{CGAN}$ represent the calculated $AEOD$ (Eq. 4) and $ASPD$ (Eq. 5) values of the classification model update w^t_j on the reference image set $\mathcal{D}^{CGAN}$, respectively. We define the fairness score f^j_{CGAN} of a model update w^t_j on the dataset $\mathcal{D}^{CGAN}$ as the average of $AEOD^j_{CGAN}$ and $ASPD^j_{CGAN}$:

$$f^j_{CGAN} = \frac{AEOD^j_{CGAN} + ASPD^j_{CGAN}}{2}, \tag{6}$$

where a higher value of f^j_{CGAN} indicates lower fairness. To filter out malicious (unfair) classification model updates, the server first calculates the fairness scores f^j_{CGAN} for all participating clients, resulting in the set $\{f^j_{CGAN}\}^n_{j=1}$ (Lines 16-18). Then, the server employs K-Means clustering with $K = 2$ to group $\{f^j_{CGAN}\}^n_{j=1}$ into two clusters. In this case, the model updates within the cluster with the higher average f^j_{CGAN} value are identified as malicious and are excluded from the subsequent global aggregation process (Lines 19-20).

5.3 Cross-Evaluation Fairness GAN Filter

While the metric-based fairness model filter can effectively identify malicious classification model updates during the global aggregation process, it still heavily relies on the assumption that the CGAN model trained by all participating clients is dependable and trustworthy. However, as mentioned in Sec. 3.3.3, since the CGAN model is collaboratively trained by both benign and malicious clients, malicious clients can also seek to compromise the utility of the CGAN model to breach the server's defense mechanism. Therefore, we consider two additional data poisoning attacks that adversaries might conduct to degrade the performance of the CGAN model during the training phase: The *Demographic Flipping GAN Attack* (DF-GAN) and the *Demographic Confusion GAN Attack* (DC-GAN).

Demographic Flipping GAN Attack (DF-GAN) compromises the fairness of the CGAN model by flipping the demographic annotations in its training data[3] (e.g., flipping 'male' to 'female' and vice versa). This attack directly degrades the CGAN's performance in generating images with correct demographic information corresponding to the given conditions.

Demographic Confusion GAN Attack (DC-GAN) undermines the fairness of the CGAN model by confusing the generator regarding demographic information (i.e., making the generator unclear about the meaning of demographic annotations in image representations). The adversary accomplishes this

[2] We achieve this by setting the demographic information and the label as the conditioning attributes, and generating equal sizes (i.e., $\frac{r}{4}$) of images with all four patterns: (x^+, y^+), (x^+, y^-), (x^-, y^+), and (x^-, y^-).

[3] It is worth noting that DF-GAN differs from DTFA. In DTFA, demographic information can only be altered by modifying the features in the input image. However, in DF-GAN, the demographic annotations can be directly obtained and modified in the training data of the CGAN model.

Algorithm 4 DPG-FairFL

```
Input: The size $r$ of the generated image set.
Output: Fair global model $w$.
Server-side:
if communication round $t = 0$ then
    Initialize $g^0$ and $w^0$, and send them to all clients;
for each communication round $t > 0$ do
    Receive the model updates $\{w_j^t\}_{j=1}^n$ and $\{g_i^t\}_{i=1}^n$ from the clients;
    // Cross-Evaluation Fairness GAN Filter
    for each $g_i^t$ do
        Generate image set $\mathcal{D}_i^{CGAN}$ of size $r$ using $g_i$;
        for each $w_j^t$ do
            Calculate $f_i^j$ using Eq. 7;
        Calculate $a_i$ using Eq. 8;
        Calculate $\overline{a_i}$ using Eq. 9;
        Calculate $s_i$ using Eq. 10;
    Perform K-Means clustering on $\{s_i\}_{i=1}^n$ with $K = 2$, and filter out the
      group with higher average $s_i$ value;
    // Metric-based Fairness Model Filter
    Compute $g^t$ by aggregating the remaining $g_i^t$ using Eq. 3;
    Generate image set $\mathcal{D}^{CGAN}$ of size $r$ using $g^t$;
    for each $w_j^t$ do
        Calculate $AEOD_{CGAN}^j$ and $ASPD_{CGAN}^j$ based on $\mathcal{D}^{CGAN}$;
        Calculate $f_{CGAN}^j$ using Eq. 6;
    Perform K-Means clustering on $\{f_{CGAN}^j\}_{j=1}^n$ with $K = 2$, and filter out
      the group with higher average $f_{CGAN}^j$ value;
    Compute $w^t$ by aggregating the remaining $f_{CGAN}^j$ using Eq. 3;
    Send $g^t$ and $w^t$ to all clients;
Client-side:
for each communication round $t > 0$ do
    for each participating client $c_i$ do
        Receive $g^{t-1}$ and $w^{t-1}$ from the server;
        Compute $g^t$ and $w^t$ using Eq. 2 (either use $\mathcal{D}_i$ for a benign client or $\tilde{\mathcal{D}}_i$
          for a malicious client);
        Send $g^t$ and $w^t$ to the server;
```

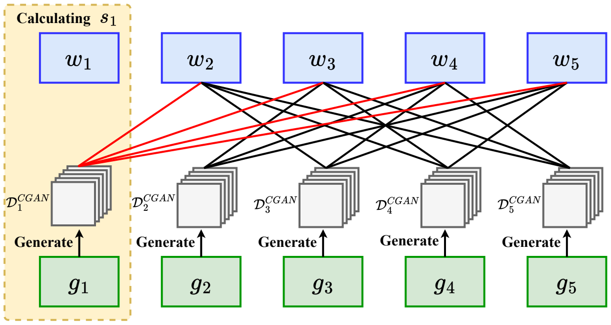

Fig. 3. Detailed Cross-Evaluation Process in the Fairness GAN Filter. (We demonstrate the specific procedure for calculating s_1 within a setting of 5 clients.)

by training the CGAN model with samples from only one demographic group (e.g., x^-) while altering the demographic annotations of some samples to the opposite group (e.g., x^+). For example, the adversary could train the CGAN model exclusively with 'female' images while labeling some of them as 'male'. This attack not only diminishes the CGAN model's ability to generate images with accurate demographic information but also weakens its capability to produce balanced images across different demographic groups.

To defend against these potential attacks on the CGAN model, DPG-FairFL incorporates an additional fairness GAN filter before aggregating the CGAN model. This filter effectively identifies and excludes malicious (unfair) CGAN model updates through an advanced cross-evaluation process of both the classification and CGAN model updates from all clients.

Alg. 4 Lines 6-13 illustrate the detailed procedure of the cross-evaluation fairness GAN filter. First, a reference image dataset $\mathcal{D}_i^{CGAN}$ of size r is generated by each CGAN model update g_i^t in the communication round t (Line 7). Similar to Eq. 6, we use fairness score f_i^j to denote the average AEOD and ASPD value obtained by inferring w_j^t on the reference image set $\mathcal{D}_i^{CGAN}$:

$$f_i^j = \frac{AEOD_i^j + ASPD_i^j}{2}. \tag{7}$$

Then, for each CGAN model update g_i^t, we define a cross-evaluation fairness score a_i to represent the average fairness evaluation of the reference dataset $\mathcal{D}_i^{\text{CGAN}}$ generated by g_i^t by the other classification models:

$$a_i = \frac{1}{n-1} \sum_{p=1, p\neq i}^{n} f_i^p. \tag{8}$$

It is noteworthy that f_i^i is excluded from the calculation of a_i, which prevents a malicious client c_i from cheating on its own fairness value. In Fig. 3, the red connections illustrate the calculation of the cross-evaluation score a_1 for the CGAN model update g_1^t.

However, simply using the cross-evaluation fairness score s_i still cannot address the impact of other malicious classification model updates on the fairness measurement of g_i^t. To further minimize this impact, we first define an additional complementary cross-evaluation fairness score $\overline{a_i}$, which calculates the average cross-evaluation fairness score for all CGAN model updates except for g_i^t:

$$\overline{a_i} = \frac{1}{(n-1)(n-2)} \sum_{q=1, q \neq i}^{n} \sum_{p=1, p \neq q, i}^{n} f_q^p. \tag{9}$$

Note that we also exclude the fairness score associated with the classification model update w_i^t (by setting $p \neq i$ in the second summation), completely removing its influence on the evaluation of g_i^t. In Fig. 3, the black connections illustrate the calculation of the complementary cross-evaluation score $\overline{a_1}$ for the CGAN model update g_1^t. We then define the final suspicious score s_i for a CGAN model update g_i^t as:

$$s_i = a_i + \theta |\overline{a_i} - a_i|, \tag{10}$$

where the first term directly represents the evaluated fairness level of g_i^t by other classification model updates, and the second term indirectly quantifies the fairness level of g_i^t by adding a scalable penalty value (scaled by θ) based on the fairness gap between g_i^t and other CGAN model updates.

Once the suspicious score s_i for each CGAN model update g_i^t has been calculated (Lines 8-12), the server can then apply K-Means clustering with $K = 2$ to group $\{s_i\}_{i=1}^n$ into two clusters. The CGAN model updates within the cluster with the higher average suspicious score are regarded as malicious and are removed before the aggregation of g^t Line 13.

6 Experimental Results

6.1 Experimental Setup

1) Dataset: We conduct our experiments using the CelebA dataset [22], a widely studied facial attribute classification dataset in fairness literature. The CelebA dataset contains over 200K images, each associated with 40 human-labeled binary facial attribute annotations such as sex, age, and hair color. In our experiments, we consider gender (male vs. female) as the sensitive (protected) attribute and select attractiveness (attractive vs. non-attractive) as the target label, given its high Pearson correlation with the sensitive attribute [20].

2) Benchmarks: Due to the scarcity of defense strategies in the existing literature that specifically target fairness poisoning attacks in FL, we employ three Byzantine-robust defense strategies that are commonly used to counter accuracy poisoning attacks in FL. These include: **Krum** [4], which selects the model update that is closest to its neighbors; **Trimmed-Mean** [40], which computes the mean of model updates after removing a certain percentage of the highest and lowest values; and **Median** [40], which chooses the median of the model updates. These Byzantine-robust methods effectively minimize the influence of outliers and extreme values during the global aggregation. Additionally, for a better comparison, we have included the performance of FedAvg without any attacks (**FedAvg (No Attack)**) and FedAvg with no additional defenses (**FedAvg (No Defense)**) as basic reference benchmarks [24].

3) Evaluation Metrics: We adopt AEOD (Eq. 4) and ASPD (Eq. 5) to measure the fairness of a trained model, and use accuracy (Acc.) to evaluate the general performance of the model. Moreover, since the proposed DPG-FairFL is a filtering-based method, we include two additional *Filtering Rates* (FR) in our evaluation to provide more intuitive measurements of our approach: the *True Positive Filtering Rate* (FR_{TP}), which indicates the probability of successfully filtering out a malicious model update, and the *False Positive Filtering Rate* (FR_{FP}), which represents the probability of incorrectly filtering out a non-malicious model update.

4) Fairness Data Poisoning Attacks Setups: In our experiments, we set up a total of $n = 20$ FL clients, with $m = 5$ of them being malicious clients controlled by adversaries. For the proposed fairness data poisoning attacks, we employ CycleGAN [44] in the DTFA and the improved conditional Wasserstein GAN [13, 27] in the FDIFA. Unless otherwise specified, the hyperparameters for the poisoning ratios, α and β, are both set to 1 for the LFFA and DTFA, representing the best effort attacking for both attacks. The proportion γ of the injected fake samples in FDIFA is also set to 1 by default. Additionally, as mentioned in Sec. 4.3, we also consider two hybrid fairness attacks: LFFA+FDIFA and DTFA+FDIFA, in our experiments. Furthermore, among the 5 malicious clients, 4 are randomly selected to execute additional CGAN attacks (2 performing DF-GAN and 2 performing DC-GAN) on their CGAN model updates, while the remaining one remains honest with its CGAN model update.

5) DPG-FairFL Setups: In DPG-FairFL, we utilize a ResNet18 model for facial attribute classification and an improved conditional Wasserstein GAN [13, 27] as the CGAN model. Both models are trained locally on each client with 40 batches of size 128 and are aggregated by the server over 50 communication rounds. The learning rates are set to $\eta = 1 \times 10^{-4}$ for the classification model and $\eta = 2 \times 10^{-4}$ for the CGAN model. The images in the CelebA dataset are resized to 128×128. For the hyperparameters in DPG-FairFL, we set the size r of the generated image set to 200 and the scaling factor θ to 1. Throughout our experiments, we report the average results derived from 10 randomly selected seeds for initializations.

Table 1. Efficacy of Various Defense Strategies Against Image-based Fairness Data Poisoning Attacks in Federated Learning. The best results are bolded in red and the second best results are bolded in black.

Methods	Federated Learning Fairness Data Poisoning Attacks (IID Data)														
	LFFA			DTFA			FDIFA			LFFA+FDIFA			DTFA+FDIFA		
	AEOD	ASPD	Acc	AEOD	ASPD	Acc	AEOD	ASPD	Acc	AEOD	ASPD	Acc	AEOD	ASPD	Acc
FedAvg (No Attack)	0.284	0.485	0.809	0.284	0.485	0.809	0.284	0.485	0.809	0.284	0.485	0.809	0.284	0.485	0.809
FedAvg (No Defense)	0.485	0.624	0.787	0.433	0.593	**0.791**	0.383	0.569	**0.805**	0.528	0.668	**0.787**	0.438	0.599	**0.793**
Krum	**0.362**	**0.493**	0.713	**0.384**	**0.508**	0.654	**0.349**	**0.533**	0.728	**0.416**	**0.515**	0.634	0.423	**0.531**	0.737
Trimmed-Mean	0.474	0.613	0.786	0.406	0.579	0.786	0.350	0.537	0.777	0.476	0.637	0.771	0.433	0.588	0.773
Median	0.446	0.598	**0.789**	0.408	0.572	0.787	0.365	0.546	**0.787**	0.460	0.557	0.737	**0.421**	0.554	0.771
DPG-FairFL (Ours)	**0.280**	**0.461**	**0.791**	**0.288**	**0.501**	**0.788**	**0.283**	**0.474**	0.781	**0.286**	**0.487**	**0.784**	**0.293**	**0.509**	**0.791**

6.2 Fairness Data Poisoning Attacks Evaluation

1) The Efficacy of Fairness Data Poisoning Attacks: Tab. 1 summarizes the quantitative results of the efficacy of the proposed five image-based fairness data poisoning attacks in compromising the fairness of FL models, along with the performance of both existing and our defense strategies in countering these fairness attacks on the CelebA dataset.

By comparing the benchmarks of FedAvg (No Attack) and FedAvg (No Defense), we observe that all five proposed image-based fairness data poisoning attacks significantly compromise the fairness of the FL model. Among the three individual fairness attacks, LFFA achieves the highest AEOD (0.485) and ASPD (0.624) values. The two hybrid attacks further enhance the attack strength, with LFFA+FDIFA achieving the highest AEOD (0.528) and ASPD (0.668) values across all attacks. Additionally, while these attacks significantly bias the fairness of the FL model, they minimally affect model accuracy, making them more stealthy and difficult to counter with existing defense mechanisms.

Regarding existing defense strategies, Trimmed-Mean and Median demonstrate minimal or no improvement in the fairness of the model. While Krum shows some improvements in the model's fairness, it still cannot fully mitigate the impact of fairness attacks and also substantially degrades the accuracy of the FL model. Therefore, existing defense strategies are insufficient for defending against fairness data poisoning attacks in FL, highlighting the need for more advanced defense strategies.

2) Ablation Studies of Fairness Data Poisoning Attacks: We conducted two ablation studies for the proposed fairness attacks: one examining the impact of the proportion of malicious clients, and the other assessing the impact of their hyperparameters. Fig. 4 illustrates the performance of all five fairness attacks with proportions of malicious clients ranging from 5% to 25%, evaluating both the model's fairness and accuracy. The results indicate that for all attacks, the impact on the fairness of FL models increases with a higher proportion of malicious clients, while the influence on model accuracy remains minimal. Fig. 5 demonstrates the impact of the poisoning ratios α and β on LFFA and DTFA, respectively, and the impact of the proportion γ of injected fake samples on

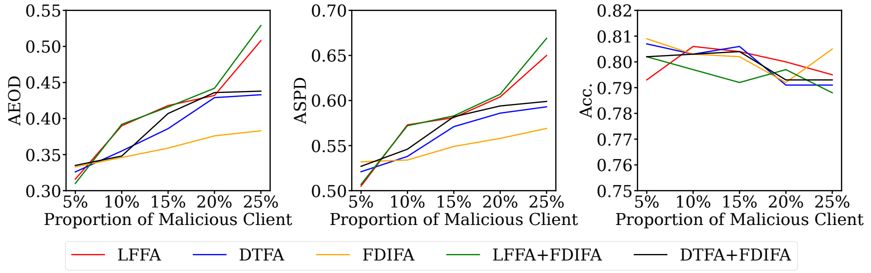

Fig. 4. Impact of the Proportion of Malicious Clients on Model Fairness and Accuracy.

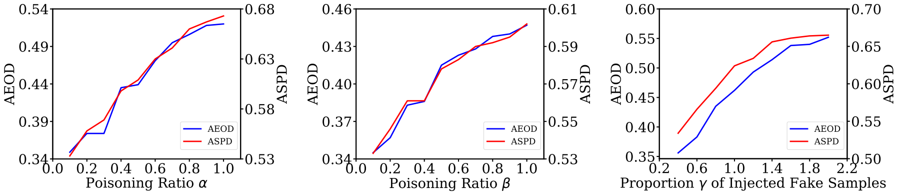

Fig. 5. Impact of Poisoning Ratios α and β, and Proportion γ on Model Fairness.

FDIFA. The results show that the attack effectiveness of LFFA and DTFA is directly proportional to α and β. The increase of γ also enhances the effectiveness of FDIFA. However, this enhancement diminishes as γ continues to grow, because the injected fake data already constitute the majority of the poisoned training dataset.

6.3 DPG-FairFL Evaluation

1) The Efficacy of DPG-FairFL: Tab. 1 demonstrates the exceptional effectiveness of the proposed DPG-FairFL in countering fairness attacks in FL. It can be seen that DPG-FairFL achieves the lowest AEOD and ASPD among all benchmark methods for all five fairness attacks, closely approaching the FedAvg (No Attack) baseline. Additionally, DPG-FairFL also achieves the highest model accuracy compared to existing defense strategies, making it a more suitable method for defending against fairness data poisoning attacks in FL.

2) Ablation Studies of DPG-FairFL: To gain deeper insights into the two fairness filters incorporated in DPG-FairFL, we break down the entire defense procedure and explore the filtering rates of the two filters under different experimental setups. Specifically, we consider the following setups: **Setup 1**: Only fairness attacks on the classification model (two hybrid ones) are present, and only the fairness model filter is employed; **Setups 2 and 3**: Both fairness attacks on the classification model and GAN-based attacks on the CGAN model

Table 2. Filtering Rate of Both Fairness Filters under Different Experimental Settings.

Experimental Setups	Fairness GAN Filter				Fairness Model Filter			
	LFFA+FDIFA		DTFA+FDIFA		LFFA+FDIFA		DTFA+FDIFA	
	FR_{TP}	FR_{FP}	FR_{TP}	FR_{FP}	FR_{TP}	FR_{FP}	FR_{TP}	FR_{FP}
Setup 1: No GAN Attack without Fairness GAN Filter	-	-	-	-	98.2%	1.2%	97.3%	1.4%
Setup 2: DF-GAN without Fairness GAN Filter	-	-	-	-	67.3%	19.2%	52.7%	42.0%
Setup 3: DC-GAN without Fairness GAN Filter	-	-	-	-	78.8%	7.1%	74.8%	21.9%
Setup 4: DF-GAN with Fairness GAN Filter	97.3%	3.6%	97.4%	3.8%	97.5%	1.9%	95.9%	2.0%
Setup 5: DC-GAN with Fairness GAN Filter	99.4%	2.5%	98.6%	3.1%	98.1%	1.4%	97.4%	1.6%
Setup 6: DF-GAN+DC-GAN with Fairness GAN Filter	97.3%	2.9%	96.0%	4.4%	98.0%	1.7%	96.9%	2.0%

(DF-GAN and DC-GAN) are present, but only the fairness model filter is employed; **Setups 4, 5, and 6**: Both fairness attacks on the classification model and GAN-based attacks on the CGAN model are present, and both the fairness GAN filter and the fairness model filter are employed.

Tab. 2 demonstrates the true positive filtering rates (FR_{TP}) and false positive filtering rates (FR_{FP}) for both fairness filters under all experimental setups. From Setup 1, we observe that in the absence of additional attacks on the trained CGAN model, the fairness model filter can accurately identify and filter out malicious classification model updates, with an FR_{TP} close to 100% and an FR_{FP} close to 0%. However, when the adversary compromises the CGAN model's performance through GAN-based attacks in Setups 2 and 3, the FR_{TP} of the fairness model filter decreases, and the FR_{FP} increases, making it no longer sufficient in filtering out malicious classification model updates. Finally, in Setups 4, 5, and 6, when the fairness GAN filter is integrated, it not only accurately filters out malicious CGAN model updates but also enhances the subsequent fairness model filter's accuracy.

3) DPG-FairFL Extensions: In this section, we demonstrate that while the proposed DPG-FairFL can effectively defend against fairness attacks in FL, it can also be easily extended to minimize inherent unfairness in the original training data by integrating local debiasing methods. We evaluate two such methods: **DPG-FairFL+FairBatch**, which utilizes the FairBatch method, a batch selection algorithm from the existing fairness literature that enhances model fairness by adaptively adjusting minibatch sizes during training; and **DPG-FairFL+FairCGAN**, which employs the already trained CGAN model within DPG-FairFL to generate balanced training samples for each benign client during the training process. Our experimental results, shown in Tab. 3, indicate that integrating both local debiasing methods into DPG-FairFL further improves the fairness of the trained FL model, achieving an even higher level of fairness compared to the FedAvg (No Attack) baseline. We also observe that integrating local debiasing methods results in a slight decrease in the accuracy of the trained FL model, which is expected and consistent with the well-known trade-off between model accuracy and fairness observed in fairness algorithms [38].

Table 3. Efficacy of DPG-FairFL Combined with Local Debiasing Methods.

Methods	LFFA+FDIFA			DTFA+FDIFA		
	AEOD	ASPD	Acc.	AEOD	ASPD	Acc.
FedAvg (No Attack)	0.284	0.485	0.809	0.284	0.485	0.809
DPG-FairFL	0.286	0.487	0.784	0.298	0.509	0.791
DPG-FairFL+FairBatch	0.058	0.236	0.741	0.032	0.229	0.753
DPG-FairFL+FairCGAN	0.092	0.330	0.757	0.103	0.339	0.745

7 Conclusion and Future Work

In this study, we pioneer the exploration of image-based fairness data poisoning attacks and defenses in FL. Specifically, we propose three types of fairness attacks that significantly compromise the fairness of the trained model in FL by modifying existing training data samples and injecting fake poisoned data samples. To counter these attacks, we introduce DPG-FairFL, a novel defense framework that incorporates an additional CGAN model during training and performs dual-phase filtering to identify and exclude malicious clients during global aggregation. Our experimental results demonstrate that DPG-FairFL is exceptionally effective in defending against all types of fairness attacks in FL.

For future work, while our study primarily focuses on data poisoning fairness attacks in FL, investigating model poisoning fairness attacks in FL remains an open problem. Additionally, exploring adversarial attacks on other fairness notions in FL, such as client-based fairness and collaborative fairness, is also an interesting area for further research.

References

1. Adel, T., Valera, I., Ghahramani, Z., Weller, A.: One-network adversarial fairness. In: AAAI. vol. 33, pp. 2412–2420 (2019)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML. pp. 214–223. PMLR (2017)
3. Awan, S., Luo, B., Li, F.: Contra: Defending against poisoning attacks in federated learning. In: ESORICS. pp. 455–475. Springer (2021)
4. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine learning with adversaries: Byzantine tolerant gradient descent. NeurIPS **30** (2017)
5. Cai, Q., Ma, M., Wang, C., Li, H.: Image neural style transfer: A review. Comput. Electr. Eng **108**, 108723 (2023)
6. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. NeurIPS **29** (2016)
7. Cheng, B., Liu, Z., Peng, Y., Lin, Y.: General image-to-image translation with one-shot image guidance. In: ICCV. pp. 22736–22746 (2023)
8. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR. pp. 8188–8197 (2020)

9. Du, W., Xu, D., Wu, X., Tong, H.: Fairness-aware agnostic federated learning. In: SDM. pp. 181–189. SIAM (2021)
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
11. Ezzeldin, Y.H., Yan, S., He, C., Ferrara, E., Avestimehr, A.S.: Fairfed: Enabling group fairness in federated learning. In: AAAI. vol. 37, pp. 7494–7502 (2023)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS **27** (2014)
13. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. NeurIPS **30** (2017)
14. Han, M., Zhu, T., Zhou, W.: Fair federated learning with opposite gan. Knowledge-Based Systems **287**, 111420 (2024)
15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. NeurIPS **29** (2016)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
17. Jodayree, M., He, W., Janicki, R.: Preventing image data poisoning attacks in federated machine learning by an encrypted verification key. Procedia Computer Science **225**, 2723–2732 (2023)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
19. Li, K., Zheng, J., Yuan, X., Ni, W., Akan, O.B., Poor, H.V.: Data-agnostic model poisoning against federated learning: A graph autoencoder approach. IEEE Trans. Inf. Forensics Secur **19**, 3465–3480 (2024)
20. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: Fair and robust federated learning through personalization. In: ICML. pp. 6357–6368. PMLR (2021)
21. Lian, Z., Zhang, C., Nan, K., Su, C.: Spoil: Sybil-based untargeted data poisoning attacks in federated learning. In: NSS. pp. 235–248. Springer (2023)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
23. Lyu, L., Yu, H., Yang, Q.: Threats to federated learning: A survey. arXiv preprint arXiv:2003.02133 (2020)
24. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
25. Meerza, S.I.A., Liu, L., Zhang, J., Liu, J.: Glocalfair: Jointly improving global and local group fairness in federated learning. arXiv preprint arXiv:2401.03562 (2024)
26. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (2021)
27. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
28. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML. pp. 2642–2651. PMLR (2017)
29. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: CVPR. pp. 10389–10398 (2022)
30. Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A.: Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering (2021)
31. Rajabi, A., Garibay, O.O.: Tabfairgan: Fair tabular data generation with generative adversarial networks. CD-MAKE pp. 488–501 (2022)

32. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: CVPR. pp. 9301–9310 (2021)
33. Roy, P.K., Chowdhary, S.S., Bhatia, R.: A machine learning approach for automation of resume recommendation system. Procedia Computer Science (2020)
34. Sun, W., Gao, B., Xiong, K., Wang, Y., Fan, P., Letaief, K.B.: A gan-based data poisoning attack against federated learning systems and its countermeasure. arXiv preprint arXiv:2405.11440 (2024)
35. Tian, H., Liu, B., Zhu, T., Zhou, W., Philip, S.Y.: Multifair: Model fairness with multiple sensitive attributes. IEEE Transactions on Neural Networks and Learning Systems (2024)
36. Tolpegin, V., Truex, S., Gursoy, M.E., Liu, L.: Data poisoning attacks against federated learning systems. In: ESORICS. pp. 480–501. Springer (2020)
37. Wang, G., Payani, A., Lee, M., Kompella, R.: Mitigating group bias in federated learning: Beyond local fairness. arXiv preprint arXiv:2305.09931 (2023)
38. Wick, M., panda, s., Tristan, J.B.: Unlocking fairness: a trade-off revisited. In: NeurIPS (2019)
39. Yeom, T., Gu, C., Lee, M.: Dudgan: improving class-conditional gans via dual-diffusion. IEEE Access (2024)
40. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: ICML. pp. 5650–5659. Pmlr (2018)
41. Zhang, D.Y., Kou, Z., Wang, D.: Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In: IEEE BigData (2020)
42. Zhang, Z., Cao, X., Jia, J., Gong, N.Z.: Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In: SIGKDD (2022)
43. Zhao, B., Cheng, T., Zhang, X., Wang, J., Zhu, H., Zhao, R., Li, D., Zhang, Z., Yu, G.: Ct synthesis from mr in the pelvic area using residual transformer conditional gan. CMIG **103**, 102150 (2023)
44. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)