Don't Give Up Fairness Even in the Midst of Membership Inference Attack

Anonymous authors
Paper under double-blind review

Abstract

Existing membership privacy-defending algorithms have demonstrated effectiveness in protecting neural networks against a wide range of membership inference attacks (MIAs). However, we observed that these privacy defense training algorithms often exacerbate disparities in predictive performance across data groups. In detail, existing privacy-defending techniques tend to sacrifice the generalizability of minority classes/groups disproportionately, while largely preserving the performance of majority groups. To address this issue, we propose Fair and Privacy-Safe Tuning (FaPriST) to mitigate unfairness while maintaining strong privacy preservation. Extensive evaluations show that our approach consistently outperforms existing privacy-defending approaches, even with no defense model, in terms of fairness, while achieving comparable robustness against membership inference attacks.

1 Introduction

Machine learning techniques have been tremendously adopted in various domains (e.g., self-driving, search engine, and discovery of new materials) thanks to their explosive advancement of performance, especially in deep learning, in terms of accuracy and universality. However, the existing real-world data will be unable to meet neural networks' capacity, especially the foundation model's growing capacity (Villalobos et al., 2024). Although overparameterization promotes the potential of intelligence and better generalizability (Neyshabur et al., 2018), it brings about severe privacy risks (Tan et al., 2022; 2023) putting the model under the threat of privacy attacks such as membership inference attacks (MIAs) (Shokri et al., 2017).

In a membership inference attack, an attacker intends to infer the membership information of a specific data sample, which tells whether a data point is a member of the model's training set, from the target model (a.k.a., victim model). This sensitive information could be used for malicious behaviors, causing unpredictable fatal security issues and negative social impacts. Such potential risks hinder the deployment of a learning model in information-confidential or security-critical application scenarios.

In recent literature, defending mechanisms against membership inference attacks have achieved considerable progress. They show their effectiveness in various aspects in training neural networks *e.g.*, optimizer (Abadi et al., 2016), loss function (Chen et al., 2022; Liu et al., 2024), model design (Fang & Kim, 2024b), and labeling (Shejwalkar & Houmansadr, 2021; Tang et al., 2022). Unlike them, in our paper, we observe that the current design of privacy-preserving approaches exhibits **unfair impact** on accuracy across different groups and classes.

In Fig. 1, we train the same model on CIFAR-10 using different privacy-defending approaches keeping the overall utility loss as low as possible. That is, even with low utility loss, the model becomes significantly unfairer than the model trained with Cross-Entropy loss (i.e., no defense). For instance, under privacy fine-tuning (introduced in Sec. 4.1), the model experiences a 19% of testing accuracy drop on the minority class, Dogs (Acc_{test} = 49% under privacy fine-tuning). In contrast, test accuracy on the majority class, Ships, increases 3%, reaching Acc_{test} = 90%. This severely exacerbates fairness, resulting in concerns regarding the actual deployment of the model and the current utility evaluation metric that takes into account only overall accuracy in privacy evaluation tasks.

In this paper, we analyze this disparate impact in privacy defense design. We identify that the design of prediction alignment between member and non-member data points exacerbates unfairness. Based on our analyses, we propose *Fair and Privacy-Safe Tuning (FaPriST)* to avoid fairness losses in privacy defense. Via tuning with our approach, the model can achieve even better fairness while preserving privacy, as shown in Fig. 1. On top of achieving better fairness, our approach also improves utility, which is a novel contribution to privacy preservation studies. In addition, two practical benefits are it is easy to implement and deploy. Our approach, tuning for fairness and privacy, can be applied to any pre-trained classification model to achieve better performance in terms of utility, fairness, and privacy.

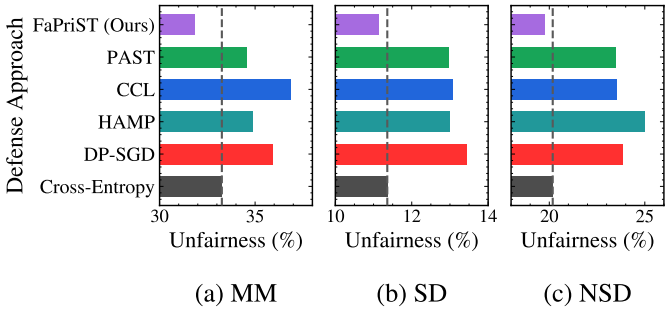


Figure 1: Class-wise fairness evaluation among privacy-defending approaches (**Lower is better** in all metrics shown above). **MM** denotes *accuracy gap between the majority class and minority class*, **SD** denotes *Standard Deviation of class-wise accuracy*, and **NSD** denotes *Normalized Standard Deviation of class-wise accuracy*. Our approach achieves privacy-preserving yet fairest performance.

Contributions. (*i*) We formulate and identify fairness issues present in current membership-privacy-preserving approaches. (*ii*) We propose a new privacy-preservation approach named **FaPriST**. To the best of our knowledge, our approach, for the first time, mitigates the privacy risks of the model while maintaining (or even improving) the model's fairness. (*iii*) Through extensive evaluations, our tuning approach demonstrates the capability to outperform pre-trained models in terms of both privacy and fairness, without utility loss.

2 Related Work

2.1 Fairness in Machine Learning

Recent studies show and address the fairness challenges of machine learning in various aspects. Hardt et al. (2016) formalized the criteria of equal opportunity and equalized odds. Zafar et al. (2017) formalized disparate mistreatment to avoid prediction outcome disparity. Buolamwini & Gebru (2018) presented an approach to evaluate prediction quality disparity. Kallus & Zhou (2018) discussed residual unfairness issues in fairness learning. Xu et al. (2021) explored how to fix fairness in adversarial training. Ma et al. (2022); Zhang et al. (2024) further improved fairness in adversarial training. Bagdasaryan et al. (2019) found DP-SGD leads to unfairness performance. Tran et al. (2022) identified that model pruning exacerbates group fairness. As for privacy, some efforts have been made in federated learning (Rafi et al., 2024). Khalili et al. (2021) tried to improve the fairness-privacy trade-off in selection problems during the post-processing phase. Chang & Shokri (2021) found fair algorithms cause over-memorization of the data from the under-represented subgroups, implying the difficulty of aggregating privacy and fairness. In our study, we identify that unfairness in prior privacy-preserving algorithms is more pronounced than the regular way and show the model can have both fairness and privacy with our approach.

2.2 Membership Privacy Defending Training

To preserve membership privacy, a traditional way is adding noise to the optimizer, such as DP-SGD (Abadi et al., 2016). As privacy preservation techniques have advanced, more efficient and effective solutions have been proposed in the last decade or so. One of them is based on knowledge distillation to produce privacy-safe learning labels (Shejwalkar & Houmansadr, 2021). Tang et al. (2022) improved the utilization rate of the training data to achieve better performance via a more advanced sampling strategy and an ensemble-based approach. However, traditional network pruning approaches, which is another common neural network compression technique, did not show such adaptability in privacy (Yuan & Zhang, 2022). Another approach to membership privacy is designing a privacy-safe classification loss function. Chen et al. (2022) introduced a

threshold for the alignment between the member and non-member distributions and maintained the model's generalizability via multiple techniques (e.g., label-smoothing). It is further improved by metric learning with representation alignment (Fang & Kim, 2024a). Besides the privacy-preserving techniques above, regularization techniques were also studied. Nasr et al. (2018) proposed an approach to guide the model to become privacy-safe via an adversarial discriminator. Kaya et al. (2020) found $\ell 2$ sparsity regularization and some data augmentation techniques could benefit privacy. Then the privacy-safe data augmentations were further determined by Kaya & Dumitras (2021); Yu et al. (2021). Liu et al. (2024) introduced a concave term to reduce the convexity of loss functions to achieve privacy preservation. Hu et al. (2024) further explored the privacy-preserving potential in regularization and pointed out that sparsity regularization works better regarding privacy-utility trade-offs during fine-tuning.

3 Preliminaries

3.1 Membership Inference Attacks (MIAs)

The goal of MIAs (Shokri et al., 2017) is to identify the membership information of the victim model. It can be formulated as follows:

$$\mathcal{A} : \boldsymbol{x}, F_\theta \rightarrow \{0, 1\} \tag{1}$$

where $\mathcal{A}$ denotes a binary classification attack task to determine whether data point x is a member of the model $F(\cdot; \theta)$'s training set - 1 means a member while 0 means a non-member. To achieve this goal, the attacker must develop metrics to measure the victim model's behavioral disparity between member and nonmember data points. Because of the recently introduced attack mechanisms (Song & Mittal, 2021; Choquette-Choo et al., 2021; Del Grosso et al., 2022; Carlini et al., 2022; Leemann et al., 2023; Zarifzadeh et al., 2024) that are becoming more and more powerful, the existing defense techniques are facing new challenges.

3.2 Sparsity Regularization

Traditional sparsity regularization is usually employed to prevent overfitting (Goodfellow et al., 2016). It can be deployed with the objective function (e.g, cross-entropy) during training:

$$\mathcal{L}_{reg}(F(x; \theta), y) = \mathcal{L}_c(F(x; \theta), y) + \lambda \mathcal{R}(\theta) \tag{2}$$

where $\mathcal{L}_c$ denotes the classification loss function, $\mathcal{R}(\cdot)$ denotes the regularization term, and λ is a hyper-parameter to balance them. The most common regularizations are $\ell 1$ and $\ell 2$ regularizations. They are formulated as follows:

$$\mathcal{R}_{\ell 1}(\theta) = \sum_{\theta_i \in \theta} |\theta_i|, \quad \mathcal{R}_{\ell 2}(\theta) = (\sum_{\theta_i \in \theta} |\theta_i|^2)^{\frac{1}{2}} \tag{3}$$

As the generalization impact of each weight varies (Molchanov et al., 2019), we hypothesize that the fairness and privacy sensitivity of each weight are also different. Therefore, deactivating the privacy-risky & unfair weights would help the model with both privacy and fairness. It is important to note that even if privacy-risky and unfair weights were pruned, after the weight pruning, privacy risks and unfairness may reappear again during retraining or finetuning process. Unlike weight pruning, regularization with our privacy-fairness scores can dynamically and adaptively prevent weights from becoming privacy-risky or unfair. To emulate private-fair sparsity during training dynamics (Wen et al., 2016), $\ell 1$ regularization is chosen because it demonstrates such potential to deactivate the privacy-risky & unfair weights dynamically by its nature of manifesting the contrast of zero and nonzero parameters.

3.3 Fairness and Privacy Metrics

Machine learning models may not show even performance classes due to various factors such as the difficulty of learning goal and data imbalance. As for privacy-preservation, the approaches usually deteriorate the class fairness when compared with regular training techniques — the minority classes suffer from more notable

utility loss than the majority, which is also shown in Figs. 1 & 3. To quantify this fairness degradation, we refer to the following studies (Xu et al., 2021; Ma et al., 2022; Zhang et al., 2024) with regard to performance alignment among classes and measure the fairness of the model via the gap between Maximum and Minimum class-wise accuracy (`MM`), Standard Deviation of class-wise accuracy (`SD`), and Normalized Standard Deviation of class-wise accuracy (`NSD`). They are formulated as follows:

$$MM = max(Acc) - min(Acc) \tag{4}$$

$$SD = \sqrt{\frac{1}{C}\sum_{i=1}^{C}(Acc_i - Acc_{mean})^2} \tag{5}$$

$$NSD = SD/Acc_{mean} \tag{6}$$

where Acc is the list of accuracies for each class, Acc_i and Acc_{mean} are the accuracy of the i-th class and the overall average accuracy, respectively.

As for privacy metrics, we use Wasserstein distance (Arjovsky et al., 2017) to measure the gap between member (training data) and non-member prediction distributions, and we report the results using both Wasserstein distance of distributions and AUC & TPR@FPR under MIAs (Carlini et al., 2022).

4 Fair and Privacy-Safe Tuning (FaPriST)

As shown in Fig. 1, we demonstrate that the unfairness is exacerbated during the process of mitigating privacy risks by privacy-defending techniques, which tells us that privacy improvement can hurt fairness, and vice versa (Chang & Shokri, 2021). Hence, to disconnect their correlations, we assume that there exists an intersection between the set of privacy-risky weights and the set of unfair weights in a neural network (i.e., a set of privacy-risky & unfair weights). Therefore, we intend to deactivate these weights in a higher priority.

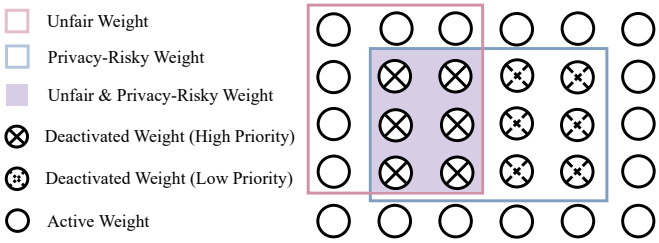


Figure 2: The overview of FaPriST. Our approach tries to suppress the unfair & privacy-risky weights with the highest priority when deactivating privacy-risky weights.

Based on this assumption, we introduce Fair and Privacy-Safe Tuning (FaPriST) to mitigate privacy vulnerabilities while preserving fairness. The overview of FaPriST is illustrated in Fig. 2. In our approach, we basically use the $\ell 1$ regularization, and we incorporate *fairness-privacy score* (will be explained in Sec. 4.3) into it that is formulated as follows:

$$\mathcal{R}_{fp}(\theta, \omega) = \sum_{(\theta_i, \omega_{\theta_i}) \in (\theta, \omega)} \omega_{\theta_i} |\theta_i| \tag{7}$$

where ω denotes the set of fairness-privacy scores in which ω_{θ_i} is the fairness-privacy score of the corresponding weight θ_i and is calculated based on gradients considering fairness and privacy. To elaborate on it, we first design a privacy score and a fairness score, and then by combining them, we develop the privacy-fairness score through Sections 4.1 – 4.3.

4.1 Privacy Score

The membership privacy risks are due to the model's behavioral disparity between member and non-member data points. Therefore, an intuitive way to prevent the disparity is to align the training objectives between them, *i.e.*, the loss. In our approach, for privacy measure, we adopt a privacy scoring function (called `PAST`) presented in (Hu et al., 2024). Considering a training set $\mathcal{D}_t$ and a reference set $\mathcal{D}_r$, we can evaluate a loss gap between them via function $\mathcal{G}$ as follows:

$$\mathcal{G}(\mathcal{D}_t, \mathcal{D}_r; F_\theta, \mathcal{L}_c) = |\frac{1}{|\mathcal{D}_t|}\sum_{(x,y)\in\mathcal{D}_t}\mathcal{L}_c(F_\theta(x), y) - \frac{1}{|\mathcal{D}_r|}\sum_{(x,y)\in\mathcal{D}_r}\mathcal{L}_c(F_\theta(x), y)| \tag{8}$$

For each parameter θ_i, the privacy score can be obtained based on the gradient. Hence, the regularization can be computed as follows:

$$\mathcal{R}_p(\theta, \phi) = \sum_{(\theta_i, \phi_{\theta_i}) \in (\theta, \phi)} \phi_{\theta_i}^{\alpha} |\theta_i| \tag{9}$$

where α denotes the hyper-parameter to adjust the privacy score sensitivity, and each score, ϕ_{θ_i}, of privacy score set, ϕ, is calculated as follows:

$$\phi_{\theta_i} = \frac{|H(\theta_i)| \nabla_{\theta_i} \mathcal{G}(\mathcal{D}_t, \mathcal{D}_r; F_\theta, \mathcal{L}_c)}{\sum_{\theta_j \in H(\theta_i)} \nabla_{\theta_j} \mathcal{G}(\mathcal{D}_t, \mathcal{D}_r; F_\theta, \mathcal{L}_c)} \tag{10}$$

where the $H(\theta_i)$ is the weights set of the layer corresponding to the weight θ_i; and $H(\theta_i) \subseteq \theta$. Unlike the traditional regularization that is applied to train the model from scratch, $\mathcal{R}_p$ is employed after the pretraining phase. Then, it tunes the model by deactivating privacy-risky weights.

Drawbacks of Current Approach. Compared with approaches that require training from scratch, a tuning approach changes the model more subtly, helping the model prevent deterioration of fairness until these changes become significant enough, as shown in Fig. 1. However, *the nature of privacy score tends to unlearn the minority.* When the model tries to be privacy-safer, the unfairness would worsen similarly to other defense approaches. More detailedly, the privacy-tuning objective can be formalized as a multi-objective optimization problem:

$$\min_{F_\theta} [L_c(F_\theta, D_t) + \sum_{\theta_i \in \theta} \lambda \phi_{\theta_i}^{\alpha} |\theta_i|] \tag{11}$$

With this objective, the model will try to minimize loss on D_t set for utility as well as the loss gap for privacy, ϕ, between D_t and D_r sets. Due to the conflicts between the two goals, there may be no solution that can minimize all terms perfectly. Fortunately, the model can still try to align the loss gap while maintaining the classification loss on the train set to achieve utility-privacy trade-offs or even Pareto optimality. Tab. 1 is an example of Pareto optimality of the model MobileNetV3 (Howard et al., 2019) on CIFAR10 (Krizhevsky et al., 2009):

Table 1: Example for Pareto optimality with privacy-tuning.

Phase	Test Accuracy (%)	Wass. Dist.
Pretraining	79.96	0.90
Privacy-Tuning	80.11	0.15

The advantages of privacy-tuning with privacy scores are promising, however, we also need to pay attention to its drawback, regarding its disparate impact. Commonly, for an unprotected model, the loss gap of the minority between the train and test set is much more significant than the majority. This fact induces the regularization term to unlearn the minority while the L_c term does not tend to learn the minority to maintain the overall classification loss due to the difficulty of the minority samples. In training, the model is always prioritized to learn easy samples Baldock et al. (2021), whereas the minority samples are difficult to learn. Their joint effect makes the model tend to give up the maintenance of the minority in exchange for minimum loss disparity between D_t & D_r and classification loss on D_t, leading to fairness loss.

To empirically explain our analysis, we tune the pretrained model with α & λ that are higher than Pareto optimality (Tab. 1) in the regularization term to show a smaller loss gap and more significant unfairness phenomenon. As shown in Tab. 2, the accuracy of a majority class (Ships), a minority class (Dogs), and the overall accuracy are reported when a privacy tuning is applied. We must pay attention to that there is a significant decrease in the minority in terms of both training and testing accuracy, while the drop in the majority is trivial. More precisely, the stable testing accuracy of the majority is at the cost of utility loss of the minority which is evident by looking at the overall accuracy. This observation necessitates a mitigation solution for fairness.

Table 2: The accuracy of the majority, minority and entire data points on the model trained with privacy tuning at 0th-35th epochs.

Group	Membership (Set)	0th	15th	25th	35th
Majority	Member (Train)	1.00	1.00	0.98	0.96
	Non-Member (Test)	0.87	0.89	0.91	0.90
Minority	Member (Train)	1.00	0.90	0.73	0.60
	Non-Member (Test)	0.68	0.68	0.58	0.49
Overall	Member (Train)	1.00	0.95	0.87	0.82
	Non-Member (Test)	0.79	0.78	0.74	0.71

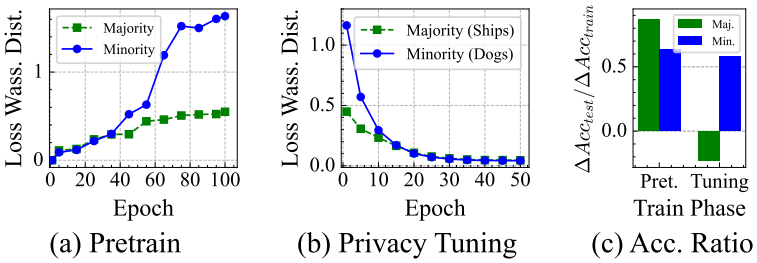


Figure 3: The gaps and changes between train and test sets during pretraining and tuning for a majority class and a minority class.

To further demonstrate this issue, we investigate and visualize the pretraining and tuning processes with privacy score as displayed in Fig. 3. In Fig. 3a, the model shows a more significant disparity in the minority than the majority as it goes to the later stage of pretraining. On the other hand, as shown in Fig. 3b, going through the privacy tuning process, the loss gap is closed. Meanwhile, the generalizability expenses for the two classes are quite different. In Fig. 3c, we compute the testing accuracy changes over training accuracy changes ($\frac{\Delta ACC_{test}}{\Delta ACC_{train}}$) with the status between the first and the final epoch in each training phase. We find generalizability learning is easy while loss is as fair as learning for the majority. However, the majority's persistent performance is at the cost of utility loss of the minority (also seen in Tab. 2). As opposed to that, for the minority, generalizability is hard to obtain even with the increase in training accuracy, while it is easy to lose with the decrease in training accuracy, which aligns with our results in Tab. 2. This tells us that, as privacy risks are mitigated, the unfairness issue of the model will become even more and more severe.

4.2 Fairness Score

For the next, we introduce a fairness score so the model can take unfair weights into regularization to preserve (or even enhance) the model's fairness during sparsity tuning. We formulate the fairness score as follows:

$$\varphi_{\theta_i} = \frac{|H(\theta_i)|\nabla_{\theta_i}\mathcal{G}(\mathcal{D}_{rm}, \mathcal{D}_r; F_\theta, \mathcal{L}_c)}{\sum_{\theta_j \in H(\theta_i)} \nabla_{\theta_j}\mathcal{G}(\mathcal{D}_{rm}, \mathcal{D}_r; F_\theta, \mathcal{L}_c)} \tag{12}$$

where $\mathcal{D}_{rm}$ denotes the data points of the minority class or group in the reference set (i.e., $\mathcal{D}_{rm} \subseteq \mathcal{D}_r$). By aligning the loss gap between the loss of the minority and all data points in the reference set, it can guide the model to maintain the generalizability of the minority. An additional benefit is that we can reuse the computation results of the privacy score described in Eq. equation 10 to avoid redundant computation costs.

In addition to the significant benefit of reducing computational cost, another compelling reason to choose the overall average loss of $\mathcal{D}_r$ as the aligning target is that the gradient gap between the minority and the majority data points is usually more unstable, leading to higher risks of the failure in model tuning. In other words, the training dynamics are smoother.

4.3 Fairness-Privacy Score: Score for our FaPriST

Fig. 2 illustrates how the privacy and fairness scores are incorporated into our ultimate approach and obtained in the regularization tuning process. Through our ablation study (shown in Sec. 5.2), we identified that, through the tuning process, the fairness impact is negatively correlated to the fairness score. Accordingly,

Algorithm 1: FaPriST Pseudocode, PyTorch-like

```
# net: pre-trained neural network
# C: classification loss function

for idx in iterations:
    # load a minibatch x_tr with n samples
    x_tr, y_tr = sample(tr_loader)
    # load a minibatch x_ref with n samples as well
    x_ref, y_ref, f_ref = sample(ref_loader)

    L_ref_min = CM(net(x_ref), y_ref, f_ref)
    L_ref_min.backward() # back-propagate
    g_ref_min = clone(net.grads) # copy grad map
    net.zero_grad() # clean gradient

    L_ref_mean = C(net(x_ref), y_ref)
    L_ref_mean.backward() # back-propagate
    g_ref_mean = clone(net.grads) # copy grad map
    net.zero_grad() # clean gradient

    L_tr = C(net(x_tr), y_tr)
    L_tr.backward() # back-propagate
    g_tr = clone(net.grads) # copy gradient map

    s_f = CG(g_ref_min, g_ref_mean) # fairness score
    s_p = CG(g_tr, g_ref_mean) # privacy score
    Reg(s_f, s_p, net) # sparsity regularization
    update(net) # parameter updation

def CM(pred, label, flabel): # loss C on minority
    # Compute per-sample classification loss
    loss = C(pred, label, reduction='none')
    # Group losses by minority class
    one_hot_label = one_hot(flabel)
    weighted_loss = loss * one_hot_label
    class_loss = sum(weighted_loss) / sum(one_hot_label)
    return class_loss.max()

def CG(g1, g2): # compute gradient disparity
    m = map()
    for layer_idx, (g1_i, g2_i) in zip(g1, g2):
        m[layer_idx] = abs(g1_i - g2_i)
    return m

def Reg(s_f, s_p, net): # add regularization term
    # Iterate modules to be regularized
    for layer_idx, m in net.reg_modules():
        f, p = s_f[layer_idx], s_p[layer_idx]
        s = (f/f.mean() / (1+p/p.mean()))**alpha
        m.grad.add_(lamb*s*sign(m.weight))
```

`f_ref`: fairness label, can be class or group; `lamb`: hyper-parameter λ to balance loss C and sparsity regularization term; `alpha`: hyper-parameter α to adjust the fair-private score sensetivity.

we reflect this in our fairness-privacy score ω which is formulated as follows:

$$\omega = [\phi/(1+\varphi)]^{\alpha} \tag{13}$$

The constant 1 incorporated in the formula is to prevent diminishing of the denominator when φ reaches too close to zero. Without this safeguard, the denominator could significantly destabilize the training process, potentially leading to erratic behavior or even catastrophic failure due to gradient explosion. Finally, we can formulate the FaPriST's regularization term as follows:

$$\mathcal{R}_{fp}(\theta, \phi, \varphi) = \sum_{(\theta_i, \phi_{\theta_i}, \varphi_{\theta_i}) \in (\theta, \phi, \varphi)} (\frac{\phi_{\theta_i}}{1+\varphi_{\theta_i}})^{\alpha} |\theta_i| \tag{14}$$

With this regularization term, the tuning procedure can simultaneously take both fairness and privacy into consideration. Considering Eq. equation 13, we have $\omega \leq \phi$ $(\varphi \geq 0, \alpha \geq 0)$ for each single weight. This indicates that higher α or λ can be set than the case when tuning via privacy score only. The tuning procedure is illustrated in Alg. 1. In practice, since neural networks are trained with minibatches, we dynamically determine the minority class or group within each minibatch through the function `CM`. This

dynamic adjustment ensures that we consistently mitigate the utility loss experienced by the minority classes, thereby preventing unfairness throughout the training process. By applying our fairness-privacy scoring in tuning, the model can deactivate its privacy-risky and unfair weights, preventing the model from fairness loss while promoting its privacy and maintaining its generalizability.

Finally, we emphasize the contribution of FaPriST and the differences between FaPriST and PAST: (*i*) The most significant contribution of FaPriST is that it enables simultaneous preservation of model privacy and per-class fairness during fine-tuning, achieved by dynamically aligning the largest and smallest classes within each mini-batch, without additional data or time costs. (*ii*) From the perspective of algorithm design, we enforce alignment of the gap between the largest and smallest classes in each mini-batch, thereby encouraging alignment in the accuracy of all classes. This is because the largest and smallest classes in each mini-batch have relatively few samples compared to the global class, and these classes may change during training. This is significantly different from the effect of PAST. (*iii*) We enable our method to simultaneously preserve both model privacy and fairness during fine-tuning without significantly increasing time complexity.

5 Empirical Evaluation

5.1 Setup

In this section, we describe experimental configurations for evaluation. The configurations include datasets, model architectures, optimizers, and other training configurations. The common settings are applied when training the model with various defending approaches except for unique configurations of the corresponding approaches.

CIFAR10. CIFAR10 (Krizhevsky et al., 2009) contains $60,000$ RGB images, evenly distributed in 10 classes and all sized in 32×32. As for the models, MobileNetV3 (Howard et al., 2019) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) are adopted. To train models on CIFAR10, common data augmentation techniques (e.g., Random Flipping & Cropping (Simonyan & Zisserman, 2015)) are applied. The models are trained using Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999$ and the weight decay of 5×10^{-4}. As for training from scratch, 100 epochs and initial learning rate, 1×10^{-1}, are used while 50 epochs and initial learning rate, 1×10^{-4} are applied for tuning. The training batch size of all approaches is set to 256. The learning rate varies along with a cosine scheduler in all training phases.

Artist20. Artist20 (Ellis, 2007), a music audio dataset, consists of six albums from each of 20 artists, totaling 1,413 tracks. Preprocessing techniques (e.g, spectrogram) and model, Convolutional Recurrent Neural Network (CRNN), are adopted from a deep learning baseline study (Nasrullah & Zhao, 2019). The models are trained with audio clips of 3 seconds in length. Among them, five albums of audio clips (so-called frames) for each artist are selected as the train set, while the rest of them are included in the test set. After preprocessing and segmentation (cutting off the songs into frames), the train and test sets include about $76,000$ and $18,000$ data points (frames), respectively. The models are trained using Adam (Kingma & Ba, 2015) with a weight decay of 0 and coefficients (β_1, β_2) of $(0.9, 0.999)$, respectively. The 100 epochs and 1×10^{-2} initial learning rate of training from scratch and 30 epochs with 1×10^{-3} initial learning rate of tuning are applied. The training batch size of all approaches is also set to 256. The learning rate changes via a cosine scheduler for all training phases.

Attack Models. As for the evaluation of privacy, we apply two well-known and most recent membership attack approaches, likelihood ratio attack (`LiRA`) (Carlini et al., 2022) and robust membership inference attack (`RMIA`) (Zarifzadeh et al., 2024). The adaptive attack policy (Song & Mittal, 2021), which can boost the MIAs' performance, is also applied to show the effectiveness of defense approaches. It assumes that the defense mechanism is known by the adversary such that the reference models are trained in the same way. We randomly sample $4,000$ and $6,000$ data points as the auditing set on CIFAR10 and Artist20, respectively.

Defense Baselines. To compare with our approach, `FaPriST`, widely known and current defense approaches are included for training the target model in our evaluation, listed as follows:

- Differentially private stochastic gradient descent (`DP-SGD`) (Abadi et al., 2016): DP-SGD is a provable approach to mitigate the membership privacy risks via adding noise during parameter updates. In our evaluation, the implementation version of PyTorch Opacus lib (Yousefpour et al., 2021) is adopted.

- Relaxed loss (`RelaxLoss`) (Chen et al., 2022): based on Cross-Entropy loss, RelaxLoss integrates multiple training techniques (e.g., Label-Smoothing, Unlearning, and) to adjust fitting degree to narrow the generalization gap between member and non-member data.

- `HAMP` (Chen & Pattabiraman, 2024): HAMP mitigates membership inference attacks by enforcing less confident predictions. It consists of a training framework with high-entropy soft labels and an entropy-based regularizer to constrain the model's confidence, and further uniformly convert all prediction outputs to low-confidence outputs while preserving accuracy, effectively obscuring the differences between members and non-members. In our evaluation, the training component of HAMP is adopted to compare with other training techniques.

- Convex-Concave Loss (`CCL`) (Liu et al., 2024): CCL mitigates membership inference risk by increasing the loss variance of training data via gradient descent. Its key idea is to reduce the convexity of standard loss functions by adding a concave term, enabling the model to produce high-variance losses for training samples while maintaining stable optimization.

- Privacy-aware sparsity tuning (`PAST`) (Hu et al., 2024): PAST is a privacy-aware variant of L1 regularization that imposes adaptive penalties on different parameters based on their privacy sensitivity. By promoting sparsity in parameters that substantially contribute to privacy leakage, PAST narrows the loss gap between members and non-members, achieving a strong privacy-utility trade-off.

For training target models, we set the number of data points of the train set and reference set, respectively, to $18,000$ and $2,000$ on CIFAR10, while these are set to $30,000$ and $5,000$ on Artist20. The search for hyperparameter configurations of the approaches above is based on each paper's recommendations.

5.2 Ablation Study

Privacy Score Only vs. Fairness-Privacy Score. To show the effectiveness of components in our approach, we tune the pretrained model with and without a fairness score on top of a privacy score. As shown in Fig. 4, tuning with fairness-privacy scores, which is our approach, shows better fairness among three fairness metrics than tuning with only privacy score when they are at similar loss gaps.

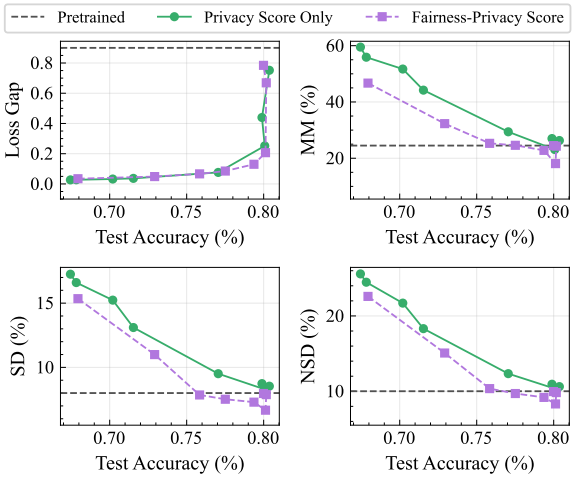


Figure 4: Loss gap (between train and test; left-top chart) and fairness evaluation (the other three charts) with various hyper-parameter combinations. **Lower is better** in all y-axis metrics shown above.

Impact of Hyperparamters. For the next, we explore the impact of the hyperparameters on the model's privacy, fairness, and utility. On CIFAR10, λ is fixed at values of $\{8.0, 7.0, 5.0, 1.0\} \times 10^{-3}$ and α is varied from 1 to 3.5. Fig. 5 displays the trends of loss gap, fairness, and testing accuracy with the changes of hyperparameters, α and λ. We find that high λ and α help the alignment of the loss gap between train and test sets. Besides that, some combinations of λ and α (e.g, $\lambda = 0.005$ and $\alpha = 1$) can help the model achieve Pareto optimality in privacy, fairness, and utility since they achieve better loss alignment and fairness evaluation while maintaining the testing accuracy. This results show that an optimal combination exists in the hyperparameters space that makes the gain of both fairness and privacy possible.

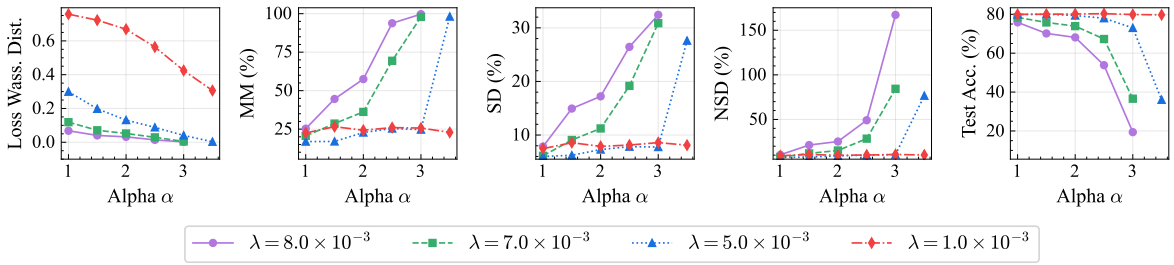


Figure 5: The changes of privacy, fairness, and utility performance along with varying α and λ.

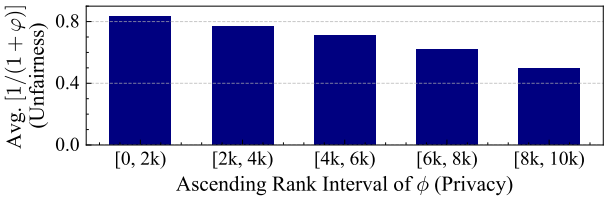
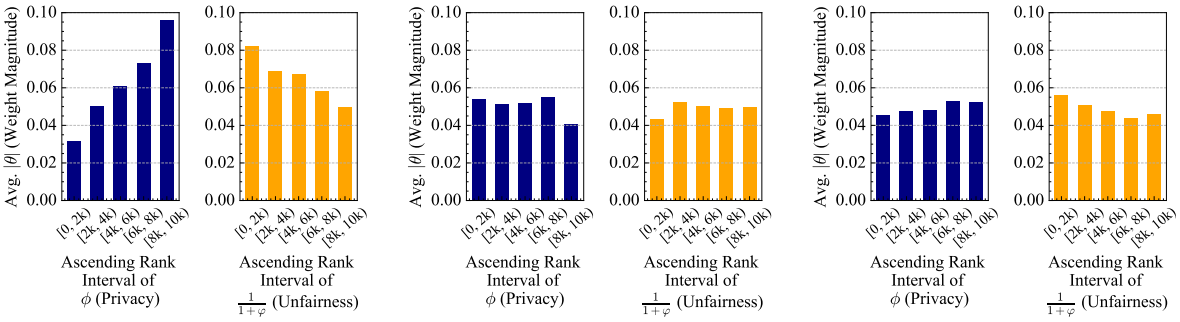


Figure 6: The average (un)fairness score of the weights in each rank bin sorted by privacy score. Results from the last layer are reported.



(a) Pretrained ($\mathcal{G} = 9.1 \times 10^{-1}, \texttt{MM} = 25.70\%$)

(b) Privacy Score Only ($\mathcal{G} = 3.7 \times 10^{-2}, \texttt{MM} = 44.70\%$)

(c) Fairness-Privacy Score ($\mathcal{G} = 4.2 \times 10^{-2}, \texttt{MM} = 37.70\%$)

Figure 7: The average weights magnitude in each rank bin sorted by privacy and fairness scores, respectively. Here $\mathcal{G}$ (↓, lower is better) represents the loss gap between the train and test sets, and MM (↓, lower is better) denotes the accuracy gap between the majority class and minority class on the test set. Results from the last layer (classification linear layer) are reported.

Correlations Between the Two Scores. We now explore the correlations between the two scores. The scores of the last layer of the pre-trained MobilNetV3 are shown in Fig. 6. The weights are sorted by privacy score, and the average $\frac{1}{1+\varphi}$ of 2k weights in each bin is calculated. We find that more privacy-risky (higher ϕ) weights are fairer (lower $\frac{1}{1+\varphi}$), and vice versa. This trend is further supported by the recent fairness-privacy study's claim that fairness learning hurts privacy (Chang & Shokri, 2021) and also our observation in Fig. 1 that privacy-defending learning makes fairness deteriorate. It enables us to induce that fairness loss is inevitable in most privacy-defending training approaches unless a fairness objective is particularly taken care of.

Fig. 7 depicts the weights' magnitude of the last layer at various privacy and unfairness levels. The x-axis represents the rank bins of weights sorted by privacy or unfairness, and the y-axis represents the average weights' magnitude for each bin. In Fig. 7a, we can observe the dominant weights of a pretrained model are privacy-risky (high ϕ) & fair weights. On top of the pretrained model, tuning with only privacy score

(Fig. 7b) deactivates more fair weights than the case of tuning with fairness-privacy score (Fig. 7c), when the two models are at similar loss gap level.

In fact, we find that other privacy-preserving training methods (e.g., RelaxLoss) also suppress weights to some extent. As shown in Fig. 8, the RelaxLoss model has globally lower magnitudes while exhibiting similar trends with the model trained with Cross-Entropy (in Fig. 7a). However, similar trends do not bring benefits to fairness in privacy training. It can be seen that the similar trends with lower weight magnitudes are good for privacy while these are bad for fairness by comparing their $\mathcal{G}$ and MM, further identifying the importance of the distribution of the weights for fairness. These results solidify that our approach with the fairness-privacy score, instead of just deactivating mostly privacy-risky weights, encourages the model to find a better balance between deactivating privacy weights and fairness weights.

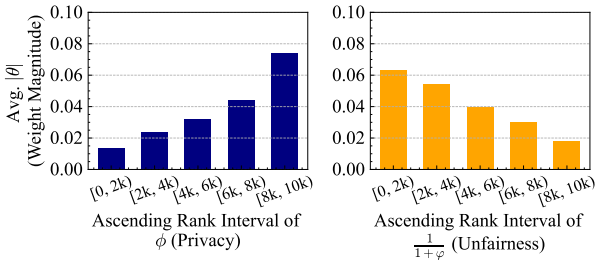


Figure 8: The average weights magnitude of the model trained RelaxLoss in each rank bin sorted by privacy and fairness scores. ($\mathcal{G} = 4.0 \times 10^{-1}, \texttt{MM} = 37.40\%$)

5.3 Comparision

CIFAR-10 We first analyze the results on CIFAR10. The top part of Tab. 3 shows the utility, fairness, and MIAs evaluation on CIFAR-10 and MobileNetV3. Despite no overall utility loss, RelaxLoss significantly worsens the model's fairness. The significantly better improvement in MIAs' defense than other approaches is heavily at the cost of the minority class. It is because when the prediction of a class almost becomes a random guess, privacy is naturally secured since learning nothing also means memorizing nothing. More specifically, it heavily sacrifices the identification ability (19.40% in testing accuracy) of the minority class (Cats) while the no-defense model's testing accuracy is 48.40% in this class. As for PAST, it achieves a more balanced performance than RelaxLoss in terms of utility, fairness, and privacy. The `FaPriST`, our approach, achieves better fairness and privacy while the testing accuracy is maintained still comparable. In addition to the results on MobileNetV3, we also show results of ViT in the middle rows of Tab. 3. The lowest performance is from the models trained with RelaxLoss. It sacrifices greater accuracy and fairness to achieve MIAs resistance similar to other defense approaches. Besides, the DP-SGD achieves the best defense in LiRA among all approaches with almost no utility loss, while CCL and HAMP achieve the best defense in RMIA with insignificant utility loss. However, all of the prior defense approaches suffer from fairness loss in

Table 3: Trade-offs between privacy, fairness and utility. The MIAs evaluation results are reported in AUC Scores and TPR at 1×10^{-3} & 1×10^{-5} FPR. Higher is better in test accuracy (↑) while lower is better in metrics for Fairness (↓) and Privacy (↓). 8 reference models (4 IN and 4 OUT models) are applied for LiRA and RMIA. All metrics are reported in percentage (%).

Dataset & Model	Defense	Test Acc. (↑)	Fairness (↓)			LiRA (↓)			RMIA (↓)		
			MM	SD	NSD	AUC	TPR@FPR		AUC	TPR@FPR	
							0.1%	0.001%		0.1%	0.001%
CIFAR-10 & MobileNetV3	No Defense	69.65 ± 0.04	32.80 ± 0.65	11.61 ± 0.41	16.67 ± 0.59	82.69 ± 0.55	0.93 ± 0.25	0.35 ± 0.36	80.02 ± 0.46	4.38 ± 0.76	2.52 ± 0.92
	RelaxLoss	69.99 ± 0.79	62.77 ± 4.57	19.39 ± 1.29	27.72 ± 2.11	60.74 ± 0.49	0.18 ± 0.06	0.07 ± 0.05	63.79 ± 0.48	0.23 ± 0.06	0.08 ± 0.05
	PAST	69.15 ± 0.23	44.03 ± 7.39	15.08 ± 1.54	21.81 ± 2.27	67.54 ± 1.24	0.30 ± 0.18	0.18 ± 0.02	79.38 ± 0.62	3.15 ± 0.86	1.07 ± 0.93
	FaPriST (Ours)	68.50 ± 0.27	42.10 ± 7.40	14.11 ± 1.52	20.61 ± 2.22	64.43 ± 0.98	0.30 ± 0.22	0.13 ± 0.19	75.93 ± 0.32	1.88 ± 0.76	0.92 ± 0.51
CIFAR-10 & ViT	No Defense	56.32 ± 0.23	33.26 ± 1.29	11.36 ± 0.33	20.18 ± 0.58	86.47 ± 0.89	2.68 ± 1.43	0.84 ± 0.09	85.11 ± 0.54	6.05 ± 0.45	2.26 ± 0.92
	DP-SGD	56.33 ± 0.42	35.92 ± 1.54	13.44 ± 0.45	23.86 ± 0.88	56.63 ± 0.46	0.10 ± 0.03	0.00 ± 0.00	60.78 ± 0.27	0.57 ± 0.61	0.30 ± 0.53
	RelaxLoss	47.46 ± 1.77	73.02 ± 1.96	29.49 ± 2.05	62.37 ± 6.59	62.46 ± 1.13	0.46 ± 0.36	0.06 ± 0.07	62.93 ± 1.23	0.61 ± 0.42	0.24 ± 0.14
	HAMP	55.25 ± 1.37	34.88 ± 1.22	12.99 ± 0.33	24.98 ± 0.65	56.01 ± 0.74	0.07 ± 0.09	0.02 ± 0.02	59.89 ± 1.53	0.40 ± 0.20	0.08 ± 0.14
	CCL	55.60 ± 0.80	36.88 ± 1.02	13.08 ± 0.34	23.53 ± 0.76	53.49 ± 1.25	0.19 ± 0.09	0.05 ± 0.03	58.43 ± 1.26	0.37 ± 0.24	0.12 ± 0.19
	PAST	56.03 ± 0.43	34.58 ± 1.90	12.98 ± 0.18	23.47 ± 0.53	55.41 ± 0.38	0.19 ± 0.14	0.03 ± 0.04	60.62 ± 0.72	0.45 ± 0.27	0.18 ± 0.09
	FaPriST (Ours)	56.37 ± 0.51	31.84 ± 1.24	11.14 ± 0.31	19.75 ± 0.47	56.12 ± 0.82	0.20 ± 0.12	0.06 ± 0.06	61.03 ± 0.84	0.53 ± 0.21	0.16 ± 0.10
Artist20 & CRNN	No Defense	53.21 ± 0.36	75.17 ± 3.54	19.60 ± 0.42	37.12 ± 0.96	62.41 ± 0.28	0.33 ± 0.07	0.12 ± 0.10	65.13 ± 0.21	0.60 ± 0.59	0.36 ± 0.43
	RelaxLoss	47.84 ± 0.41	85.35 ± 0.34	23.79 ± 0.63	50.33 ± 1.23	51.71 ± 0.04	0.13 ± 0.03	0.03 ± 0.03	54.49 ± 1.02	0.09 ± 0.06	0.00 ± 0.00
	PAST	48.89 ± 1.35	79.69 ± 4.59	21.24 ± 2.35	43.68 ± 7.92	50.50 ± 0.36	0.04 ± 0.03	0.00 ± 0.00	56.34 ± 0.28	0.08 ± 0.06	0.00 ± 0.00
	FaPriST (Ours)	50.12 ± 0.63	76.16 ± 2.76	19.64 ± 0.32	39.51 ± 0.19	51.29 ± 0.71	0.10 ± 0.05	0.02 ± 0.03	57.50 ± 0.30	0.10 ± 0.13	0.00 ± 0.00

various extents. Our approach – the FaPriST – outperforms all approaches (including No Defense) in terms of fairness, while it also maintains considerable privacy-preserving ability and testing accuracy.

Artist20 Here, we discuss the experimental results of Artist20. In the bottom half of Tab. 3, the overall picture is that the unfairness issue on Artist20 is more significant than on CIFAR10. One of these difficulties stems from segmentation, in which the useful information contained in frames greatly varies. Among RelaxLoss, PAST, and FaPriST, RelaxLoss performs best against RMIA, while its fairness is the worst. PAST and FaPriST show better performance against LiRA with better fairness than RelaxLoss. The FaPriST shows the best balance between fairness and privacy protection.

6 Conclusion, Broader Impacts, and Limitations

This paper presents an unfairness issue widely existing in the current privacy-preservation techniques. We observed that fairness fragility occurs when the privacy vulnerability is being mitigated. To address this issue, we proposed Fair and Privacy-Safe Tuning (FaPriST) to maintain both fairness and generalizability along with privacy training. Through extensive empirical results, with tasks of various modalities, vision, and audio, we showed that our approach enhances both fairness and privacy while not losing model utility. Our report reveals that fairness and privacy are not coupled everywhere. Our study can promote fairness in privacy-critical machine learning models. However, the unfairness issue is not perfectly removed but may still be present in privacy-preserving models. At present, there are still some gaps in the correlation between privacy and the fairness of machine learning models. We believe our work will contribute to ethical integrity and fair usage in machine learning systems.

References

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arjovsky17a.html`.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf`.

Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10876–10889. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/5a4b25aaed25c2ee1b74de72dc03c14e-Paper.pdf`.

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

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

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

Dingfan Chen, Ning Yu, and Mario Fritz. Relaxloss: Defending membership inference attacks without losing utility. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FEDfGWVZYIn.

Zitao Chen and Karthik Pattabiraman. Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction. In *Network and Distributed System Security (NDSS) Symposium*, 2024.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974, 2021.

Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Leveraging adversarial examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10399–10409, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Daniel Ellis. Classifying music audio with timbral and chroma features. pp. 339–340, 01 2007.

Xingli Fang and Jung-Eun Kim. Center-based relaxed learning against membership inference attacks. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pp. 1294–1306. PMLR, 15–19 Jul 2024a.

Xingli Fang and Jung-Eun Kim. Representation magnitude has a liability to privacy vulnerability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 411–420, 2024b.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Qiang Hu, Hengxiang Zhang, and Hongxin Wei. Defending membership inference attacks via privacy-aware sparsity tuning, 2024. URL https://arxiv.org/abs/2410.06814.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2439–2448. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kallus18a.html.

Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5345–5355. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kaya21a.html.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. On the effectiveness of regularization against membership inference attacks, 2020. URL https://arxiv.org/abs/2006.05336.

Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8092–8100, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Gaussian membership inference privacy. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=2NUFe4TZMS.

Zhenlong Liu, Lei Feng, Huiping Zhuang, Xiaofeng Cao, and Hongxin Wei. Mitigating privacy risk in membership inference by convex-concave loss. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30998–31014. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24q.html.

Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26230–26241. Curran Associates, Inc., 2022.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11264–11272, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 634–646, 2018.

Zain Nasrullah and Yue Zhao. Music artist classification with convolutional recurrent neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2019.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks, 2018. URL https://arxiv.org/abs/1805.12076.

Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, and Dong-Kyu Chae. Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102198. URL https://www.sciencedirect.com/science/article/pii/S1566253523005146.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9549–9557, 2021.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2615–2632, 2021.

Jasper Tan, Blake Mason, Hamid Javadi, and Richard Baraniuk. Parameters or privacy: A provable tradeoff between overparameterization and membership inference. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17488–17500. Curran Associates, Inc., 2022.

Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, and Richard G. Baraniuk. A blessing of dimensionality in membership inference through regularization. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10968–10993. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/tan23b.html.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1433–1450, 2022.

Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17652–17664. Curran Associates, Inc., 2022.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11492–11501. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xu21b.html.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. How does data augmentation affect privacy in machine learning? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10746–10753, 2021.

Xiaoyong Yuan and Lan Zhang. Membership inference attacks and defenses in neural network pruning. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 4561–4578, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=sT7UJh5CTc.

Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24746–24755, June 2024.