Can You Fake It Until You Make It?: Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness

Victoria Cheng
victoria.j.cheng@gmail.com
Vector Institute
University of Toronto
Snap Inc.

Vinith M. Suriyakumar
vinith@cs.toronto.edu
Vector Institute
University of Toronto

Natalie Dullerud
dullerud@cs.toronto.edu
Vector Institute
University of Toronto

Shalmali Joshi
shalmali@vectorinstitute.ai
Vector Institute

Marzyeh Ghassemi
marzyeh@cs.toronto.edu
Vector Institute
University of Toronto
Canadian CIFAR AI Chair

ABSTRACT

The recent adoption of machine learning models in high-risk settings such as medicine has increased demand for developments in privacy and fairness. Rebalancing skewed datasets using synthetic data created by generative adversarial networks (GANs) has shown potential to mitigate disparate impact on minoritized subgroups. However, such generative models are subject to privacy attacks that can expose sensitive data from the training dataset. Differential privacy (DP) is the current leading solution for privacy-preserving machine learning. Differentially private GANs (DP GANs) are often considered a potential solution for improving model fairness while maintaining privacy of sensitive training data. We investigate the impact of using synthetic images from DP GANs on downstream classification model utility and fairness. We demonstrate that existing DP GANs cannot simultaneously maintain model utility, privacy, and fairness. The images generated from GAN models trained with DP exhibit extreme decreases in image quality and utility which leads to poor downstream classification model performance. Our evaluation highlights the friction between privacy, fairness, and utility and how this directly translates into real loss of performance and representation in common machine learning settings. Our results show that additional work improving the utility and fairness of DP generative models is required before they can be utilized as a potential solution to privacy and fairness issues stemming from lack of diversity in the training dataset.

ACM Reference Format:

Victoria Cheng, Vinith M. Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It?: Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 31 pages. <https://doi.org/10.1145/3442188.3445879>

This work is licensed under a Creative Commons Attribution International 4.0 License. *FAccT '21*, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
<https://doi.org/10.1145/3442188.3445879>

1 INTRODUCTION

Machine learning models are being trained for use in high-stakes settings on sensitive data in domains such as medicine [6, 63, 72, 79] and consumer finance [36, 41, 83]. These models can have disparate impacts on minoritized subgroups for a number of reasons, including due to class imbalance [20, 49] and small minority sample sizes [9] in the underlying training dataset. These impacts often amplify existing prejudices against minorities in society under the guise of automated impartiality [13, 16, 46].

The use of balanced synthetic datasets created by Generative Adversarial Networks (GANs) [27] to augment classification training has demonstrated some benefits for reducing disparate impact due to minoritized subgroup imbalance [2, 19, 56]. However, GANs and the synthetic data generated from them are vulnerable to privacy attacks [12, 14, 67, 70, 76] due to their tendency to memorize the training distribution. This makes GANs vulnerable to membership inference attacks [12, 30] which can expose sensitive information about individuals in the underlying training data to malicious entities [12, 14, 67, 70, 76].

Differential privacy (DP) [21] is the leading technique for preserving privacy in statistical and machine learning settings. It has seen successful applications at Google, Apple, and in the US Census [29] resulting in its increased popularity and use. DP guarantees privacy by ensuring that the inclusion or exclusion of any particular individual does not significantly change the output distribution of an algorithm [21]. Differentially private stochastic gradient descent (DP-SGD) is a popular method for training deep learning models with DP [1]. While DP-SGD has strong theoretical guarantees for privacy and generalization, its application has been shown to separately impact model performance for minoritized subgroups in imbalanced datasets theoretically [25] and empirically in the fields of computer vision and natural language processing [5, 23].

GANs trained without DP have been extremely successful at generating high quality synthetic images [39]. Dataset augmentation using GANs has been shown to improve downstream classification performance [50]. Training classification models on differentially private [59, 81] synthetic data has been proposed as a technique to ensure the privacy of training data [22, 68]. Empirically, GANs trained with DP have been shown to generate useful synthetic

images on simpler datasets such as MNIST [81]. Most investigations of DP GANs on image datasets have used lower dimensional datasets such as MNIST and SVHN, though some have utilized high-dimensional facial datasets such as CelebA[48], and more challenging tabular datasets [75, 81, 82]. However, to our knowledge there have been no investigations into whether DP synthetic images provide the same improvements to utility and fairness in downstream classifications that synthetic images generated without privacy do. Furthermore, past DP GAN research has not taken advantage of state-of-the-art (SOTA) GAN architectures such as Progressively Growing GANs (PG-GANs) [39] that have been advantageous in generating high-dimensional image data. Such composite analysis of the fairness, utility, and privacy impacts of DP is essential for its application in real-life systems.

In this work, we investigate the viability of DP synthetic data as a solution to improving both utility and fairness of downstream classification, while maintaining privacy. We focus on two chest x-ray datasets, the MIMIC Chest X-Ray (MIMIC-CXR) dataset [38] and CheXpert (CXP) [33], and two facial datasets, CelebA-HQ [39] and FairFace [44], one dermatology datasets, the ISIC archive [64].

To succeed in this analysis, we extend the current theory for the sequential adaptive composition of heterogeneous Rényi differential privacy (RDP) sampled Gaussian mechanisms and provide new guarantees to support training state-of-the-art DP GANs such as Progressively Growing GANs (PG-GANS) [39]. Based on these new guarantees we empirically investigate the effect that DP synthetic data generated from PG-GANs has on the utility and fairness of classification models. We compare the generated image quality as we increase privacy levels and investigate whether there is disparate impact to image quality across subgroups. Next, we investigate whether DP protects all subgroups equally from membership inference attacks on GANs. Finally, we investigate the impacts of training downstream classification models on balanced DP synthetic datasets. We quantify the impacts on both model utility and fairness. We demonstrate that DP does not introduce unfairness into the data generation process or to standard group fairness measures in the downstream classification models, but does unfairly increase the influence of majority subgroups. DP also significantly reduces the quality of the images generated from the GANs, which decrease the utility of the synthetic data in downstream tasks.

Our results demonstrate that more research into new techniques is needed to improve the utility and quality of synthetic image data generated using DP GANs. Currently, applying DP-SGD to GANs does not result in high quality high-dimensional private synthetic image datasets that are useful for training downstream classification models.

1.1 Contributions

- **We provide a new upper bound on the privacy loss of the sequential adaptive composition of heterogeneous RDP mechanisms to support private training of PG-GANs.** Current adaptive composition bounds for sampled Gaussian RDP mechanisms assume constant sampling probability (batch size) across all steps taken by the learning algorithm. Thus, those bounds do not apply to PG-GAN architectures which have varying sampling probabilities during training. Our method allows

us to reliably quantify the privacy guarantees from PG-GANs. Our new bounds are an important addition to understanding the sequential adaptive composition of the heterogeneous sampled Gaussian RDP mechanism.

- **DP significantly decreases image quality, but there are no trends in the difference in image quality between subgroups.** Generated image quality is reduced by DP overall but there is no observed disparate subgroup impact in our chosen datasets.
- **DP GANs defend against membership inference attacks in the presence of dataset shift.** We show that GANs trained without privacy are subject to membership inference attacks when the test set is shifted from the training set, and demonstrate that DP defends against these attacks.
- **DP synthetic images decrease downstream model performance.** We find that due to the poor quality of the DP synthetic images, the performance of downstream classification models significantly degrade.
- **DP synthetic images have the potential to give unfair influence to majority subgroups but result in no changes to group fairness measures.** We find that the use of DP synthetic images results in no significant changes to classification model group fairness measures. However, the influence of the majority subgroup on the downstream predictions can increase with increased privacy for highly imbalanced datasets.

2 BACKGROUND AND RELATED WORK

2.1 Differential Privacy

Differential privacy (DP) [21] ensures that the output of statistical analyses are robust to the inclusion or exclusion of a particular individual. It provides important theoretical guarantees for privacy and has been shown empirically to defend against privacy attacks [14, 35, 67].

DP-SGD is currently the leading private learning algorithm for training neural networks [1, 35]. Other private learning algorithms such as objective perturbation are only applicable to linear models because they assume convex loss functions [10]. The privacy loss in DP-SGD is tracked using methods such as the Moments Accountant [1] or the Rényi differential privacy (RDP) Accountant [54]. DP learning algorithms focus on learning the body of the joint data and label distribution. By doing so, they omit the tails of the distribution [24]. Recent work shows that this leads to disparate impact for minoritized subgroups due to the combination of gradient clipping and noise addition [5, 23]. Additional background about differential privacy has been included in Appendix A.1.

2.2 GANs for Image Generation

GANs [27] learn to model a data distribution. They can then sample from the modeled distribution to generate novel data and are extremely effective at learning high-dimensional distributions such as image data [39]. The training procedure follows a minimax optimization problem between a generator and a discriminator[27]. Conditional GAN models are modified to also input a label to both the generator and the discriminator in order to learn a conditional distribution based on that label [55].

Progressive Growing GANs. Many of the current state-of-the-art GAN models such as PG-GAN [39] and StyleGAN [40] use progressive growing architectures. These architectures progressively grow the generator and discriminator networks by gradually adding additional layers. This helps the model to learn global features at low resolutions and local features at high resolutions, and mitigates common GAN training issues such as mode collapse [39].

Differentially Private GANs. GANs tend to output data samples that are closer to the training data than to unseen data [81], raising privacy concerns when the training data contains sensitive information. Differentially private GANs (DP GANs) [81] and Private Aggregation of Teacher Ensembles GANs (PATE-GANs) [58, 59, 82] are able to achieve differential privacy in GANs by employing the DP-SGD or PATE mechanism during training. Since PATE-GAN has not yet been shown to be effective on image data, we focus our investigation on DP-SGD. We have chosen to employ the progressively growing technique with the conditional version of DP GANs in our work (DP-CGANs)[73].

Additional background about GANs has been included in Appendix A.2.

2.3 Differential Privacy and Fairness

Differentially private training has been shown to degrade model accuracy [37]. This accuracy drop is more significant in DP-SGD for minoritized subgroups in the training dataset [5, 23]. Recent attempts to make differentially private learning fair have looked at Lagrangian dual approaches [74] and post-processing techniques similar to equalized odds and multicalibration [34]. Most current work in DP dataset synthesis focuses on low-dimensional image or tabular data and does not investigate the effects on downstream classification fairness [3, 22, 75].

2.4 Improving Model Performance With Synthetic Data

Augmenting the original training dataset using generated images from GANs [19, 65, 80] has shown promise in improving performance in downstream tasks.

Downstream Training with DP. DP is immune to post-processing [21]. This means the output of a differentially private mechanism cannot be made less private regardless of any auxiliary information or the unbounded computational resources a malicious entity may have. This is important for using DP synthetic images because downstream models trained on these generated images do not need to use DP in their training process. Therefore, we can train classification models on generated differentially private data and maintain privacy.

3 EXTENDING THE RDP ACCOUNTANT TO PROGRESSIVELY GROWING TRAINING

We use the TensorFlow Privacy library [52] to train our models with DP-SGD and to track the privacy loss using an RDP [54] accountant for sampled Gaussian mechanisms (SGM). This implementation is based on the theoretical analyses in [54] which assumes that batch size remains constant throughout training. PG-GANs decrease batch size during training due to memory constraints [39].

Existing work on DP PG-GANs naively applies the RDP Accountant and provides underestimates of the privacy loss [12]. We extend the RDP for SGM privacy analysis to account for varying batch sizes throughout training in Section 3.1. We use this modification to measure the privacy of our model using the derived ϵ guarantee. Lower ϵ values mean higher privacy.

Note that the privacy guarantees we provide are underestimates of the true privacy loss because we do not add the privacy loss incurred due to hyper-parameter search [11, 47].

3.1 Extending the RDP Accountant

We derive a new tight upper bound on the privacy loss for our PG-GAN training and provide a loose upper bound which generalizes more broadly to progressively growing learning procedures. We use this modification to the RDP for SGM accountant to track privacy loss.

Progressively Growing Private Training Analysis. We present a theoretical analysis on the upper bound of the RDP of our models trained with progressively growing training. We extend the theoretical work of [54] for our specific use case.

Preliminaries.

Definition 3.1. Rényi Divergence Let there be two probability distributions P and Q with densities p and q respectively. We define the distribution on X over the same probability space. The Rényi Divergence for a finite order $\alpha \neq 1$ between P and Q is defined as:

$$D_\alpha(P||Q) := \frac{1}{\alpha - 1} \ln \mathbb{E}_{x \sim q} \left[\left(\frac{p(x)}{q(x)} \right)^\alpha \right]$$

Definition 3.2. Rényi Differential Privacy We say that a mechanism $M : X \rightarrow R$ is (α, ϵ) -RDP with order $\alpha \in (1, \infty)$ if for all neighboring datasets X, X' :

$$D_\alpha(M(X)||M(X')) \leq \epsilon$$

Definition 3.3. Sampled Gaussian Mechanism. Let f be a function mapping subsets of X to $\mathbb{R}^d$ (in our case this maps mini-batches to gradients). We define the Sampled Gaussian Mechanism (SGM) parameterized with sampling parameter $0 < q \leq 1$ and the noise $\sigma > 0$:

$$SG_{q,\sigma}(X) = f(x : x \in X \text{ with sampling probability } q) + \mathcal{N}(0, \sigma^2 \mathbb{I}^d)$$

where each element of X is sampled independently with probability q without replacement and $\mathcal{N}(0, \sigma^2 \mathbb{I}^d)$ is a d -dimensional spherical Gaussian with per coordinate variance σ^2 .

THEOREM 3.4. RDP for Sampled Gaussian Mechanism. The upper bound of the RDP for an SGM that is (α, ϵ) -RDP as proven in [54] under their assumptions is:

$$\epsilon(\alpha) \leq \frac{2q^2\alpha}{\sigma^2}$$

THEOREM 3.5. Constant Sampling Probability. Using the RDP composition theorem [53] and tight bounds on the RDP SGM [54], let the composition of T homogeneous (α, ϵ) -RDP sampled Gaussian mechanisms with constant sampling probability q resulting in an (α, ϵ') -RDP mechanism be:

$$\epsilon'(\alpha) \leq \frac{2Tq^2\alpha}{\sigma^2}$$

Figure 1: Pipeline for training and evaluations. (1) We train PG-GANs on imbalanced datasets with varying levels of noise for differential privacy. (2) We then evaluate the quality of the synthesized images from these generative models across different privacy levels. (3) We perform privacy attacks on the generative models to observe the effect of privacy on the success of the attacks. (4) We generate balanced synthetic datasets using the PG-GANs and use the balanced datasets to train downstream classification models. (5) The performance of these models are evaluated against real test images. We also analyze the fairness of these downstream classifiers across various fairness metrics. (6) We analyze the influence of each data point in the balanced synthetic dataset on the output predictions of a downstream classification model.

Analysis. DP-SGD can be viewed as an adaptive sequential composition of $T(\alpha, \epsilon)$ -RDP mechanisms. Previous analyses such as the Moments Accountant [1] and RDP Accountant [54] assume that the sampling probability for each mechanism is constant. In our work we examine the composition when the sampling probability decreases during training due to the progressive growing learning algorithm, thus motivating the extension of existing composition upper bounds for decreasing sampling probability over the T sequential adaptive mechanisms. We extend the composition theorems for RDP [53] and tight bounds for RDP sampled Gaussian mechanisms [54] to this unique case of sequential adaptive composition with decreasing sampling probability that appears when training PG-GANs. We prove both theorems in Appendix E.

THEOREM 3.6. Decreasing Sampling Probability With Constant Decay and Steps. *Let there be T RDP sampled Gaussian mechanisms (SGMs). We partition this set of mechanisms k times. Each partition is made up of $T/k(\alpha, \epsilon_i)$ -RDP mechanisms where $i = [k]$. For each partition the original sampling probability q_1 decays by $(\frac{a}{b})^i$ where $i = [k]$. The upper bound of the composition of this set of heterogeneous RDP SGMs is:*

$$\epsilon'(\alpha) \leq \frac{a^2[(\frac{a}{b})^{2k} - 1]}{b^2[(\frac{a}{b})^2 - 1]} \cdot T \cdot \frac{2q_1^2\alpha}{\sigma^2}$$

We provide a tight upper bound on M given these assumptions and the assumptions on q and σ from the previous theorem.

In our previous theorem, we provide a tight upper bound based on the assumptions regarding the varying sampling probabilities and the number of steps taken at each sampling probability. Typically, the sampling probability decreases over time because we reduce the batch size. Since we reduce the batch size the number of steps taken increases compared to previous batch sizes. For our theoretical analysis, we make the prior theorem more generalizable

by only assuming that the number of steps for each sampling probability is greater than that of the previous sampling probability (i.e. that the number of step $S_{i+1} > S_i \forall i = 1, 2, \dots, k$).

THEOREM 3.7. Varying Sampling Probability With Varying Decay and Step Sizes. *Let there be T RDP sampled Gaussian mechanisms (SGMs). We partition this set of mechanisms k times, each partition $(s_1, s_2, \dots, s_k)$ mechanisms being made up of (α, ϵ_i) -RDP mechanisms where $i = [k]$. For each partition the original sampling probability q_1 decays by $(\frac{a}{b})^i$ where $i = [k]$. The upper bound of the composition of this set of heterogeneous RDP SGMs is:*

$$\epsilon'(\alpha) < \frac{a^2[(\frac{a}{b})^{2k} - 1]}{b^2[(\frac{a}{b})^2 - 1]} \cdot ks_k \cdot \frac{2q_1^2\alpha}{\sigma^2}$$

4 METHODS

Figure 1 shows our model training and evaluation pipeline.

We first trained GANs on real image data with and without progressive growing to compare which architecture produces the best images when training with differential privacy. We focus our evaluations on PG-GANs, as they produce higher quality images than non-growing GANs (Section 7).

4.1 GAN Architecture

Our GAN model is a conditional variation of the PG-GAN [39] trained with a differentially private discriminator model [73]. This training process allows for the generator model to be differentially private as well, via the Post-Processing Theorem as shown in Appendix A.3. The GAN is trained with an Adam optimizer [42] with gradient clipping and noise addition based on DP-SGD [1]. We condition our PG-GAN using both the protected and target labels to ensure that we can generate a balanced synthetic dataset to train our downstream classification models. All images used in the GAN training process and the rest of our evaluation pipeline have been resized to be 64×64 pixels.

4.2 Progressively Growing Training

Progressively growing training [39] modifies the standard GAN training procedure by gradually fading layers into the model rather than training all the layers at the same time. This process trains the model at lower resolution images first, allowing it to learn the high-level structure of the data before learning the low-level details. In order to accommodate for memory constraints, the mini-batch size is decreased as the model size and image resolution increases. This procedure has been shown to improve the resulting generated image quality [39].

4.3 Differentially Private Training

We train models without privacy guarantees using stochastic gradient descent (SGD). Models are trained with DP-SGD [1], which is the only currently suitable differentially private learning method for deep neural networks. DP-SGD makes two modifications to SGD: clipping gradients computed on a per-example basis to have a maximum ℓ_2 norm, and then adding Gaussian noise to these gradients before applying parameter updates [1]. We select the clipping norm to be 1.0 from hyper-parameter tuning. We use three different noise multipliers of [0.0001, 0.0010, 0.1000] corresponding to low, medium, and high privacy settings respectively. For each task, four increasing levels of privacy are explored: No Privacy (Standard PG-GAN Model), Low Privacy, Medium Privacy, and High Privacy.

We use our modification to the RDP accountant discussed in Section 3.1 to track the privacy loss of the model. The epsilon values for each privacy level and dataset are presented in Appendix F.

4.4 Downstream Model Training

We generate balanced synthetic datasets using the trained PG-GANs for each real dataset at each privacy level. Each synthetic data is the same size as the corresponding real dataset, but is balanced across both the protected attribute and the target prediction attribute. The attributes for each dataset are listed in Table 1.

We then use these balanced synthetic datasets to train 3 different types of downstream classification models [61] to predict the target label: linear regression (LR), random forest (RF), and multi-layer perceptron (MLP).

5 EVALUATION

5.1 Evaluation of Synthetic Image Quality

We evaluate the quality of synthesized images from these GANs using Frechet Inception Distance (FID) [32]. FID estimates the quality and diversity of a set of synthetic images as compared to the real dataset. It extracts the activations of the real and generated images from the last pooling layer of Inception v3 [71]. These activations are used to measure the distance between the distribution of the real images and the distributions of the generated images.

5.2 Evaluating Privacy Attacks for GANs

We perform membership inference attacks on the GANs across all privacy levels to measure the ability of DP to defend against these attacks. Membership inference [14, 67] exploits the memorization of training examples in the parameters of neural networks [24]. We

assume that the malicious entity has access to the trained generator leading to a white-box attack [12]. This is a reasonable assumption given that trained GAN models are often released publicly.

We use a white box membership inference attack [12] based on the reconstruction loss for each image. The attacker uses the parameters of the generator model to optimize the random latent vector inputted to the model to reproduce the given images as closely as possible. The reconstruction loss between this reproduced generated image and the real image is lower for images that can be reproduced more closely. The images with the lowest reconstruction loss are taken by the attacker to be members of the original training set. If this attack has a high success rate measured by the Area Under Receiver Operating Characteristic Curve (AUC) [8], then we conclude that the model has high memorization and is vulnerable to membership inference privacy attacks.

5.3 Downstream Model Evaluation

Classification Performance. We measure the performance of models using AUC on real test images from the original real dataset.

Group Fairness Metrics. We evaluate classification fairness under four different standard measures of group fairness [28]: performance gap, parity gap, recall gap, and specificity gap. We provide definitions for these metrics in Appendix B.

Group Influence Metrics. We also measure fairness via the level of influence as defined by Cook’s distance [15] for each subgroup. Cook’s Distance captures the influence of a particular training point on a linear regression model.

6 DATA

We conduct our experiments on two chest x-ray datasets, the MIMIC Chest X-Ray (MIMIC-CXR) dataset [38] and CheXpert (CXP) [33], two facial datasets, CelebA-HQ [39] and FairFace [44], and one dermatology dataset, the ISIC archive [64]. Table 1 reports the label distributions for these datasets. Detailed dataset distributions are reported in Appendix D. We intentionally selected a variety of balanced and imbalanced datasets in order to investigate the effect of imbalanced training data.

All 6 datasets use the same methods and evaluation pipeline as described in Sections 4 and 5.

6.1 Chest X-Ray Datasets

We take MALE/FEMALE to be our protected attribute based on prior work demonstrating disparities in chest x-ray classifier performance for this label [45, 66]. We pick the EFFUSION label to be our target label out of a possible 14 labels because it is one of the most prevalent labels in MIMIC-CXR (23.72% positive rate) and CXP (40.14% positive rate). We only use the frontal images in both datasets. Our experiments on this data investigate if our methods are able to improve the fairness of models in the medical imaging field.

6.2 Dermatology Dataset

We take SKIN TONE to be our protected attribute in for ISIC [64]. We generate the binary skin tone label based on Fitzpatrick Skin Tone categories using a similar automated procedure to the one

Dataset Name	Protected Label	Target Label	Majority Protected Attribute	Cross-Dataset Test Set	Total Images
MIMIC-CXR	Sex (Male/Female)	No Effusion/Effusion	Male (53.74%)	CXP	225819
CXP	Sex (Male/Female)	No Effusion/Effusion	Male (58.75%)	MIMIC-CXR	172201
CelebA-HQ	Skin Tone (Light/Dark)	Young/Old	Light Skin Tone (86.58%)	FairFace	27000
FairFace	Skin Tone (Light/Dark)	Young/Old	Dark Skin Tone (57.42%)	CelebA-HQ	86744
ISIC	Skin Tone (Light/Dark)	No Melanoma/Melanoma	Light Skin Tone (64.28%)	N/A	21350

Table 1: Datasets and labels used in our analysis. All of the labels listed are treated as binary labels in our evaluations. We use the protected label and target labels to condition our PG-GAN. In the classification models we only predict the target label. In privacy attack experiments for MIMIC-CXR, CXP, CelebA-HQ, and FairFace we also test each GAN model on a different dataset in the same domain as the dataset the model was trained on ("Cross-Dataset Test Set"). Detailed dataset distributions are included in Appendix D.

described in [43]. The procedure is included in Appendix C. We pick the presence of MELANOMA to be our target label for this dataset.

6.3 Facial Datasets

We take SKIN TONE as our protected attribute in our facial datasets. We generate the binary skin tone label based on Fitzpatrick Skin Tone categories using the same process as for the dermatology dataset. FairFace [44] only contains labels for the attributes of race, gender, and age. We formulate a label YOUNG to be the target predictor label in FairFace by splitting the provided age buckets in half. We select YOUNG/OLD as the target label in CelebA-HQ [39] to match our FairFace labels.

Facial Datasets, Recognition, and Fairness. There have been several facial datasets that have recently been constructed by the computer vision community, such as CelebA-HQ [39] and FairFace [44]. The collection and use of these datasets has led to concerning and unethical research due to the inherent systemic biases (sexism, racism, colorism, and ageism) from the collection of the images and their labels [9, 18]. We use CelebA-HQ [39] and FairFace [44] in our experiments for prediction tasks to illustrate our hypotheses about privacy and fairness. Our analysis serves to demonstrate the negative effects of using these labels. We do not condone use of and prediction on these datasets in practical scenarios. We emphasize that irresponsible and unethical usage of these labels and datasets could have harmful impacts on the research community, and on society-at-large.

7 EXPERIMENT: EVALUATING QUALITY OF SYNTHETIC DP DATA

First, we analyze the effect of increasing levels of privacy on the quality of images generated by the DP PG-GANs conditioned on our protected attributes and target labels.

Experimental Setup. We evaluate the quality of generated images compared to real images across GANs of increasing privacy levels using Frechet Inception Distance (FID) [32]. We also measure this on a subgroup level to quantify possible disparate impact.

Image Quality Metrics. A low FID score indicates that the set of generated images is close in distribution to the images in the training set, and therefore realistic. The minimum possible FID

Figure 2: FID Scores of PG-GANs vs GANs. Closer to 0.0 is better. The PG-GANs achieve much better FID scores than the non-PG models in the DP case and similar FID Scores in the non-DP case. Models trained with DP have poorer FID scores than models trained without DP.

score is 0.0, which corresponds to the score for when the generated image distribution exactly matches the training image distribution.

7.1 Result: Progressive Growing Training Improves DP Synthetic Image Quality

We first confirmed that progressively growing training with DP improved the resulting image quality. Figure 2 shows the differences in image quality between DP vs non-DP models trained on CelebA-HQ. The models trained with DP are trained in the medium privacy setting.

The models trained with DP show significant improvements when using PG training even when non-DP models show negligible improvements. This shows that the improvements in FID experienced by large high-quality image models such as StyleGAN [40] are extensible to models trained with DP as well. It also demonstrates the need for our formulated privacy accountant that accurately supports progressively growing learning in Section 3. Therefore, we focus on PG-GANs in the remainder of our experiments and results.

7.2 Result: DP Degrades Quality of Synthetic Images

High privacy levels result in lower image quality across all datasets. Figure 3b shows the computed FID metric for each privacy level and

Figure 3: (a) Generated images from GANs trained on MIMIC-CXR. Qualitatively, models trained with higher privacy produce worse-quality images. (b) Left: Comparison of FID scores of PG-GAN models across different DP privacy levels. Overall, image quality decreases significantly with DP when compared to the non-DP models. Right: Difference in FID between the protected subclasses listed in Table 1. Negative FID Difference means that the minority protected subclass has relatively worse images. Note that models trained with DP generally have higher differences in FID between their protected subclasses, but we observe no directional trends with relation to the majority protected class attribute. Additional results are included in Appendix G.

(a) White Box Attack AUC at different privacy levels. (b) AUC-Difference in White Box Attacks between protected subclasses.

(c) Samples reconstructed from the training set using GAN models trained on CelebA-HQ at different privacy levels.

Figure 4: Comparison of white-box inference attack success rate in dataset as well as from different datasets within the same domain ("Cross-Dataset") as listed in Table 1. Attacks using test images from a different dataset of the same domain are more successful (lower AUC is better). The cross-dataset white box attacks are less successful for DP-trained GANs. Therefore, DP protects against membership inference attacks. Difference in attack AUC between protected subclasses increases as privacy increases for CelebA-HQ. From the reconstructed images, we can see that the model trained without privacy is able to more closely reconstruct the input images than the models trained with privacy, indicating potential memorizing and over-fitting to the training set. Additional attack results are included in Appendix H.

the corresponding generated images. The FID worsens as the privacy levels increases. Figure 3a qualitatively visualizes the output of the GAN models trained on the MIMIC-CXR dataset. The images become noisier and more unrecognizable as privacy increases. Synthetic chest X-ray images were also reviewed post-submission by a licensed radiologist to confirm that the quality of the x-ray images decreased as privacy levels increased.

Finally, we find no trends in the FID difference across the protected subgroups with increasing privacy levels (Figure 3b).

8 EXPERIMENT: EVALUATING ROBUSTNESS TO PRIVACY ATTACKS

We measure the efficacy of different levels of DP for protecting PG-GANs against white-box membership inference attacks [12] compared to PG-GANs trained without DP.

Experimental Setup. We use the vulnerability to white-box attacks of PG-GANs trained without DP as our baseline for this evaluation. We then measure the attack success on PG-GANs with

increasing levels of privacy across all datasets. The attack is successful if it is able to accurately predict which data points are from the original training set of the model. We evaluate whether increased privacy reduces attack success.

We evaluate both the setting where the training and test set of images are from the same distribution (within the same dataset), and the setting where the training set and test are from different datasets in the same domain (i.e. across CelebA-HQ and FairFace).

White-Box Membership Inference Attack [12]. In a white-box attack, the adversary attempts to infer whether a particular image belonged to the original training set of a generative model based on the model’s ability to reconstruct that image. The model parameters are used to optimize a latent variable input to the model in order to minimize reconstruction loss and produce a synthetic image that most closely resembles the given image.

Performing this attack on both training and test images allows us to estimate training set membership since models with memorization will be able to better reconstruct the training images than unseen images. We use 1000 training images and 1000 unseen test images (2000 images total). The model takes the 1000 images with the lowest reconstruction loss to be members of the training set.

The AUC of the white-box attacks represents the success level of the attack at predicting training set membership. Higher AUC means that the attacker can more accurately infer if an individual’s data was used to train the GAN.

8.1 Result: DP Protects GANs From Membership Inference

Prior work demonstrates that membership inference is unsuccessful in GANs that are trained on large enough datasets [12]. In Figure 4a, we observe this to be true across all privacy levels when the training and test sets are from the same distribution. When the training and test set are from different datasets within the same domain, GANs trained without privacy are vulnerable to membership inference. In this setting of dataset shift, DP protects GANs from membership inference. We qualitatively analyze the effectiveness of this white-box attack in Figure 4c. Models trained without privacy can reproduce the training set images almost exactly.

In real-life applications, dataset shift is a common occurrence [62], so this result presents dangerous privacy implications for deploying generative models trained without DP in real settings. Therefore, differential privacy is essential for preventing these attacks, especially when a model is publicly released.

We observe that increasing privacy results in less protection for the minoritized subgroup in the presence of dataset shift for CelebA-HQ, the most imbalanced dataset use in our evaluations (Figure 4b). There is no trend observed amongst the other datasets used in our evaluation.

9 EXPERIMENT: FAIRNESS OF MODELS TRAINED ON PRIVATE SYNTHETIC DATA

We measure the effect of DP synthetic data on downstream classification model fairness in order to evaluate the relationship between privacy and utility. We perform the evaluation across all datasets and privacy levels.

Experimental Setup. Three different downstream algorithms are evaluated: linear regression (LR), random forest (RF), and multi-layer perceptron (MLP). The DP PG-GANs from our previous experiments are used to create balanced synthetic datasets at each privacy level for downstream training. The downstream algorithms are evaluated on the real test images from each corresponding dataset, predicting the target label as listed in Table 1.

Our baseline is the classification performance when trained on the entire unbalanced corpus of real training data. For training on the synthesized data, a dataset of the same size as the real data was generated, balanced across both the protected attribute and the target attribute.

Utility and Fairness Metrics. We evaluate model utility using AUC. We evaluate model fairness according to standard group fairness measures between the protected attribute classes: performance gap (AUC difference between subclasses), parity gap, specificity gap, and recall gap.

9.1 Result: DP Reduces Classification Model Utility

The AUC of the classification models decreases as the privacy level increases across all datasets (Figure 5). Therefore, the reduced image quality due to DP results in poor data utility and poor downstream model performance.

9.2 Result: DP Has No Impact on Standard Group Fairness Metrics

We observe no trends in any group fairness metrics for the downstream classification models (Figure 5) as privacy increases. Results for additional group fairness metrics like parity gap, specificity gap, and recall gap have been included in Appendix I.

10 EXPERIMENT: SUBGROUP INFLUENCE ON DOWNSTREAM CLASSIFICATION MODELS

Experimental Setup. In addition to standard group fairness metrics on downstream models, we also measure fairness in terms of the influence of each synthetic training data point. For each generated dataset, we use Cook’s distance [15] to evaluate the influence of each training point on a downstream linear regression model.

Influence Metrics. Cook’s Distance [15] quantifies the effect that the training point has on the downstream model’s predictions based on the change in the models prediction values when that point is excluded from the training set. We provide a detailed definition for this metric in Appendix B.

A highly influential point is considered to be any point that has a Cook’s Distance value of over $4/n$, where n is the total number of training data points [7]. We measure influence fairness as the percentage of highly influential points from the majority class. To be determined as fair, when trained on the balanced synthetic datasets, the set of highly influential points should contain equal proportions of the protected subclasses.

Figure 5: Effect of DP on Downstream Classification AUC and Performance Gap (AUC Difference between subclasses). Training on images generated from models trained with DP significantly decreases the performance of the classification model, but has no trend on the group fairness metrics. Additional downstream evaluations and fairness metrics are included in Appendix I.

Figure 6: Percentage of Influential Points (as defined by Cook's Distance) from the majority class at different privacy levels. Closer to 50% is more fair. DP synthetic data increases the influence of the majority class as noise increases for highly imbalanced datasets. Influence in the high privacy setting has been excluded due to the low utility of the model, rendering any influence on the prediction output essentially meaningless. Additional influence results are included in Appendix J.

10.1 Result: DP Can Disparately Impacts Influence on Highly Imbalanced Datasets

As shown in Figure 6, we observe that synthetic data created by GAN models trained without DP improves or approximately maintains influence fairness when compared to the real training set for all datasets.

As the privacy level increases, more influence tends to be given to points from the majority subgroup in imbalanced datasets. CelebA-HQ (the most imbalanced dataset used in our evaluation) displays a significant increase in influence unfairness in the presence of any DP. Models trained on synthetic data created from more balanced

datasets used in our evaluations do not experience the same level of disparate impact or do not display any consistent trends. Therefore, even when the synthetically generated dataset is balanced, imbalances in the protected attribute in the original training set can still potentially result in unfair and harmful over-influence of majority subgroups in downstream evaluations.

11 DISCUSSION

11.1 The Need For Diverse Real Data

Lack of diversity in training datasets is known to amplify societal biases for minoritized subgroups. For example, such deficiencies in facial recognition algorithms have led to significantly worse recognition rates for black female faces in commercially deployed algorithms [9]. Similarly, the lack of variation in medical chest x-ray datasets can cause models to have consistently poor performance on important subsets of the data [57]. Our analysis emphasizes that biased datasets may still have disparate impacts on downstream classification influence, even when the real dataset is not directly used in the downstream training process and even when the synthetic dataset used for the training is balanced. Therefore, supplementing or replacing datasets with synthetic data does not mitigate the fairness concerns caused by the existing biases in imbalanced datasets.

11.2 Privacy Risks of Synthetic Data

Synthetic data augmentation is often explored as a solution to missing data or to data privacy issues, such as in credit card fraud detection models [4] and sports medicine research [77]. Although synthetic data can be realistic and appear promising to use as a replacement for real data, the privacy risk associated with doing so remains high. While commonly used for creating synthetic data, generative models do not guarantee privacy by design [12, 14, 67, 70,

76]. We have shown that GANs trained without DP are vulnerable to membership inference privacy attacks, especially in the presence of dataset shift. Releasing a GAN model therefore exposes the training set to attackers, and can result in the leakage of sensitive data.

11.3 The Price of Privacy

Our investigation into the effect of DP on state-of-the-art PG-GAN models in terms of downstream model utility, privacy, and fairness shows the negative reality of the trade-offs. We have demonstrated that training with DP can address privacy concerns, but has the potential to skew the influence towards the majority subgroups, exacerbating already existing inequities in the data collection processes.

As well, while DP has strong theoretical guarantees, it can significantly degrade model utility when applied in real-world settings [37, 60]. The loss in utility of the synthetic data as privacy increases causes much higher rates of misclassifications in downstream models. This decrease in model performance can lead to disastrous effects in important settings such as weapons identification [51] and pedestrian identification for autonomous cars [31, 69, 78].

11.4 Limitations Of Our Analysis

We caution that the automatic detection of facial labels is only performed as an example and should not be considered a valid task for deployed applications. Many labels that are associated with facial datasets such as YOUNG, SMILING, or ATTRACTIVE (all available labels in the CelebA-HQ dataset) are disproportionately biased towards women and other minoritized groups. When these predictors are applied in settings such as automatic hiring algorithms [17], they can be used against these minoritized groups in the screening process.

The analysis we present also contains limitations in the automated skin tone categorization, since Fitzpatrick categories are known to be subject to lighting conditions and are inherently subjective, even to human dermatologists [43]. Cameras are also known to less accurately capture darker skin tones [26]. We also reiterate the inherent flaws in our labels for our downstream models, specifically predicting age attributes. Labels with less problematic implications should be used in the future provided they are available.

12 CONCLUSION AND FUTURE WORK

Our analysis shows that balanced differentially private synthetic datasets are effective against privacy attacks. However, DP GANs display a decrease in dataset image quality and downstream classification accuracy. The utility of the synthetic images decreases significantly with increased privacy for standard downstream tasks, resulting in close to random classification predictions. We conclude that more work needs to be done on differentially private data synthesis before it is ready to be applied in real-world settings such as high-dimensional medical images or facial images.

Additional future work can also be done around the training of progressively growing models with differential privacy. When training at low resolutions with DP-SGD, the noise added during the training process prevents the model from effectively learning essential low level structures of the images, resulting in much worse

quality images than if the noise is only added when the model is learning the high-level details. Theoretical work must be done around how this technique would affect privacy guarantees.

ACKNOWLEDGMENTS

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

We would like to thank radiologist Dr. Errol Colak for reviewing our chest x-ray images. We also would like to thank Haoran Zhang, Karsten Roth, Nathan Ng, Sheldon Huang, Shrey Jain, Nicolas Papernot, Colin Li, Judy Shen, and Jacob Willemsma for their feedback on this paper.

REFERENCES

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 308–318.
- [2] Adel Abusitta, Esma Aïmeur, and Omar Abdel Wahab. 2019. Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems. *arXiv:1905.09972* [cs.LG]
- [3] Sean Augenstein, H Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, et al. 2019. Generative models for effective ml on private, decentralized datasets. *arXiv preprint arXiv:1911.06679* (2019).
- [4] Hung Ba. 2019. Improving Detection of Credit Card Fraudulent Transactions using Generative Adversarial Networks. *arXiv preprint arXiv:1907.03355* (2019).
- [5] Eugene Bagdasaryan and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. *arXiv:1905.12101* [cs.LG]
- [6] Andrew L Beam and Isaac S Kohane. 2018. Big data and machine learning in health care. *Jama* 319, 13 (2018), 1317–1318.
- [7] Benjamin Bengfort, Rebecca Bilbro, Nathan Danielsen, Larry Gray, Kristen McIntyre, Prema Roman, Zijie Poh, et al. 2018. *Yellowbrick*. <https://doi.org/10.5281/zenodo.1206264>
- [8] Andrew P. Bradley. 1997. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recogn.* 30, 7 (July 1997), 1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- [9] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (*Proceedings of Machine Learning Research, Vol. 81*). Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [10] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, 3 (2011).
- [11] Kamalika Chaudhuri and Staal A Vinterbo. 2013. A stability-based validation procedure for differentially private machine learning. In *Advances in Neural Information Processing Systems*. 2652–2660.
- [12] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2019. GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs. *arXiv:1909.03935*
- [13] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *arXiv:1805.12002* [stat.ML]
- [14] Christopher A Choquette Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2020. Label-Only Membership Inference Attacks. *arXiv preprint arXiv:2007.14321* (2020).
- [15] R Dennis Cook. 1979. Influential observations in linear regression. *J. Amer. Statist. Assoc.* 74, 365 (1979), 169–174.
- [16] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).
- [17] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019). <https://doi.org/10.1145/3287560.3287572>
- [18] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. (2020). *arXiv:2007.07399* [cs.CY]
- [19] Fabio Henrique Kiyoi dos Santos Tanaka and Claus Aranha. 2019. Data Augmentation Using GANs. *arXiv:1904.09135* [cs.LG]
- [20] Chris Drummond and Robert C. Holte. 2005. Severe Class Imbalance: Why Better Algorithms Aren't the Answer. In *Machine Learning: ECML 2005*, João Gama, Rui

- Camacho, Pavel B. Brazdil, Alípio Mário Jorge, and Luís Torgo (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 539–546.
- [21] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (Aug. 2014), 211–407. <https://doi.org/10.1561/04000000042>
- [22] Liyue Fan. 2020. A survey of differentially private generative adversarial networks. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*.
- [23] Tom Farrand, Fatemehsadat Mirehghallah, Sahib Singh, and Andrew Trask. 2020. Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy. *arXiv preprint arXiv:2009.06389* (2020).
- [24] Vitaly Feldman. 2019. Does Learning Require Memorization? A Short Tale about a Long Tail. *CoRR abs/1906.05271* (2019). [arXiv:1906.05271](http://arxiv.org/abs/1906.05271) <http://arxiv.org/abs/1906.05271>
- [25] Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 954–959.
- [26] Michelle Felt. 2015. Automatic skin tone calibration for camera images. *US Patent* 9,118,876.
- [27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *arXiv* (2014). <https://arxiv.org/pdf/1406.2661.pdf>
- [28] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413* [cs.LG]
- [29] Michael B. Hawes. 2020. Implementing Differential Privacy: Seven Lessons From the 2020 United States Census. *Harvard Data Science Review* (30 4 2020). <https://doi.org/10.1162/99608f92.353c6f99> <https://hdsr.mitpress.mit.edu/pub/dgg03vo6>
- [30] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2018. LOGAN: Membership Inference Attacks Against Generative Models. *arXiv:1705.07663* [cs.CR]
- [31] Douglas Heaven. 2019. Why deep-learning AIs are so easy to fool. <https://www.nature.com/articles/d41586-019-03013-5>
- [32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv:1706.08500* [cs.LG]
- [33] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031* [cs.CV]
- [34] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharif-Malvajerdi, and Jonathan Ullman. 2019. Differentially private fair learning. In *International Conference on Machine Learning*. PMLR, 3000–3008.
- [35] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing Differentially Private Machine Learning: How Private is Private SGD? *arXiv preprint arXiv:2006.07709* (2020).
- [36] Julapa Jagtiani and Catharine Lemieux. 2019. The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financial Management* 48, 4 (2019), 1009–1029.
- [37] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1895–1912.
- [38] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).
- [39] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* (2018). <https://arxiv.org/pdf/1710.10196.pdf>
- [40] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.
- [41] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.
- [42] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [43] Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. 2019. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. *arXiv:1910.13268* [cs.CV]
- [44] Kimmo Kärkkäinen and Jungseok Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv:1908.04913* [cs.CV]
- [45] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 23 (2020), 12592–12594. <https://doi.org/10.1073/pnas.1919012117> <https://www.pnas.org/content/117/23/12592.full.pdf>
- [46] Heidi Ledford. 2019. Millions of black people affected by racial bias in health-care algorithms. *Nature* 574 (2019), 608–609. <https://doi.org/10.1038/d41586-019-03228-6>
- [47] Jingcheng Liu and Kunal Talwar. 2019. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 298–309.
- [48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- [49] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition* 91 (2019), 216 – 231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- [50] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristian Malossi. 2018. BAGAN: Data Augmentation with Balancing GAN. *arXiv:1803.09655* [cs.CV]
- [51] 2018 Matthew HutsonJul. 19, 2020 Jon CohenOct. 7, 2020 Gretchen VogelOct. 7, 2020 Christa Lesté-LasserreOct. 7, 2020 Lucy HicksSep. 23, 2020 Rasha AridiSep. 23, 2020 Jocelyn KaiserSep. 21, 2020 Rasha AridiSep. 11, and 2020 Rebekah TuchschererSep. 4. 2018. A turtle-or a rifle? Hackers easily fool AIs into seeing the wrong thing. <https://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing>
- [52] H Brendan McMahan, Galen Andrew, Úlfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. 2018. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210* (2018).
- [53] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, 263–275.
- [54] Ilya Mironov, Kunal Talwar, and Li Zhang. 2019. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv preprint arXiv:1908.10530* (2019).
- [55] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv* (2014). <https://arxiv.org/pdf/1411.1784.pdf>
- [56] Konstantinos Nikolaidis, Stein Kristiansen, Vera Goebel, Thomas Plagemann, Knut Liestøl, and Mohan Kankanhalli. 2019. Augmenting Physiological Time Series Data: A Case Study for Sleep Apnea Detection. *arXiv:1905.09068* [cs.LG]
- [57] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. 2020. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (Toronto, Ontario, Canada) (CHIL. '20). Association for Computing Machinery, New York, NY, USA, 151–159. <https://doi.org/10.1145/3368555.3384468>
- [58] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. *ArXiv abs/1610.05755* (2017).
- [59] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable Private Learning with PATE.. In *ICLR*. <https://openreview.net/forum?id=rkZB1XbRZ>
- [60] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. 2020. Tempered Sigmoid Activations for Deep Learning with Differential Privacy. *arXiv preprint arXiv:2007.14191* (2020).
- [61] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
- [62] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. 2019. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. *arXiv:1810.11953* [stat.ML]
- [63] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissim Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1, 1 (2018), 18.
- [64] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Liopyris, Josep Malveyh, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. 2020. A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context. *arXiv:2008.07360* [eess.IV]
- [65] Veit Sandfort, Ke Yan, Perry Pickhardt, and Ronald Summers. 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports* 9 (12 2019). <https://doi.org/10.1038/s41598-019-52737-x>
- [66] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *arXiv preprint arXiv:2003.00827* (2020).
- [67] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks against Machine Learning Models.

- arXiv:1610.05820 [cs.CR]
- [68] C. Shorten and T.M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>
 - [69] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. arXiv:1802.06430 [cs.CR]
 - [70] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2020. Synthetic Data – A Privacy Mirage. arXiv:2011.07018 [cs.LG]
 - [71] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs.CV]
 - [72] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 7767 (2019), 116–119.
 - [73] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. DP-CGAN: Differentially Private Synthetic Data and Label Generation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Jun 2019). <https://doi.org/10.1109/cvprw.2019.00018>
 - [74] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. 2020. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *arXiv preprint arXiv:2009.12562* (2020).
 - [75] Aleksei Triastcyn and Boi Faltings. 2018. Generating Differentially Private Datasets Using GANs. *CoRR* abs/1803.03148 (2018). arXiv:1803.03148 <http://arxiv.org/abs/1803.03148>
 - [76] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Towards Demystifying Membership Inference Attacks. arXiv:1807.09173 [cs.CR]
 - [77] John Warmenhoven, Andrew Harrison, Daniel Quintana, Giles Hooker, Edward Gunning, and Norma Bargary. 2020. Unlocking sports medicine research data while maintaining participant privacy via synthetic datasets. (2020).
 - [78] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. arXiv:1902.11097 [cs.CV]
 - [79] Denny Wu, Hirofumi Kobayashi, Charles Ding, Lei Cheng, and Keisuke Goda Marzyeh Ghassemi. 2019. Modeling the Biological Pathology Continuum with HSIC-regularized Wasserstein Auto-encoders. *arXiv preprint arXiv:1901.06618* (2019).
 - [80] Eric Wu, Kevin Wu, David Cox, and William Lotter. 2018. Conditional Infilling GANs for Data Augmentation in Mammogram Classification. *CoRR* abs/1807.08093 (2018). arXiv:1807.08093 <http://arxiv.org/abs/1807.08093>
 - [81] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially Private Generative Adversarial Network. *CoRR* abs/1802.06739 (2018). arXiv:1802.06739 <http://arxiv.org/abs/1802.06739>
 - [82] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2019. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=S1zk9iRqF7>
 - [83] Bing Zhu, Wenchuan Yang, Huaxuan Wang, and Yuan Yuan. 2018. A hybrid deep learning model for consumer credit scoring. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 205–208.