Generating Privacy Preserving Synthetic Medical Data

Fahim Faisal
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
FaisalF1@myUManitoba.ca

Noman Mohammed
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
Noman.Mohammed@UManitoba.ca

Carson K. Leung
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
Carson.Leung@UManitoba.ca

Yang Wang
Dept. of CSSE
Concordia University
Montreal, QC, Canada
Yang.Wang@Concordia.ca

2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA) | 978-1-6654-7330-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSAA54385.2022.10032429

***Abstract*—Due to the recent development in the deep learning community and the availability of state-of-the-art models, medical practitioners are getting more interested in computer vision and deep learning for diagnosis tasks. Moreover, those medical diagnostic models can also increase the reliability of conventional findings. As radiology images can convey a lot of information for a patient's diagnosis task, the problem is that such medical data may contain sensitive private information in their content header. De-anonymization (i.e., removal of sensitive header information) does not work well due to the re-identification risk, which may link those images to essential details (e.g., birth date, SSN, institution name, etc.), and such an approach can also reduce utility. In the medical domain, utility is significant because a less accurate diagnosis may lead to the wrong course of treatment and/or loss of life. In this paper, we developed a differentially private approach that can generate high-quality and high dimensional synthetic medical image data with guaranteed differential privacy. It can be used to create sufficient quality data to train a deep model. Moreover, we used W-GAN for bounded gradient guarantee, which eliminates the need for an extensive clipping hyperparameter search. We also added noise selectively to the generator to maintain the privacy-utility trade-off. Due to a noise-free discriminator and such selective noise addition to the generator, high-quality and reliable generated radiology images can be utilized for diagnosis tasks. Moreover, our approach can work in a distributed system where different hospitals can contain their private images in the local server and use a central server to generate synthetic radiology images without storing patient data.**

***Index Terms*—Privacy, medical imaging, synthetic data, generative adversarial network (GAN), synthetic data generation, Renyi differential privacy, re-identification**

I. Introduction

The popularity of deep-learning models' computation power encourages medical professionals to solve many diagnosis problems. Particularly, some medical sectors require a lot of time and huge human resources to do their job. For example, identifying some diseases needs to analyze the previous history of patients, so it takes a lot of time. Wherever ones need to calculate any sequence of diseases, they have to wait a few days or months due to limited diagnostic tools. Because of the high proficiency and computation power, in modern days, machine learning is taking part in this sector so that it can solve disease analysis problems in much reduced time and more efficiently. It is also reducing the necessity for a huge workforce and experienced professionals. One problem is that medical datasets are difficult to use in modeling as they are associated with personal information and preserving such health data's privacy is crucial [3, 9, 47].

However, in our study, we focus on privacy concerns for using sensitive datasets—e.g., magnetic resonance imaging (MRI), computerized tomography (CT) scans, X -rays, or breast cancer datasets, which can also have a marker indicating who the real person could be—in medical research. Though deep-learning techniques show tremendous outcomes for pneumonia or Coronavirus Disease 2019 (COVID-19) detection, medical institutes are still not supportive of providing enough data due to privacy issues. Users are not willing to provide sensitive information in public. X-ray images in DICOM (digital imaging and communications in medicine) files are structured so that the cover sheets baked into such DICOM files include patients' sensitive information for identification purposes. They may include the patient's date of birth, sensitive diagnosis information, the name of clinical institutions, etc. Sometimes, hospital databases use patients' social security numbers (SSNs) to identify those files in the system. Such sensitive information leakage could link them to another sensitive dataset using those identifiers. This kind of privacy breach will be harmful to the patients. Medical data privacy is underestimated in most computer vision and privacy research, but such harmful consequences should be addressed soon. US federal law restricts ones from using patients' private data, but such deep models need as much data as possible. They follow some de-anonymization approaches or content removal approaches (e.g., skull stripping for MRI data), which removes such identifying attributes from the image header in X-ray images. But, as the protocol for X-ray images differs worldwide, standardizing such an approach is not feasible. Moreover, adversary models can detect another X-ray image of a similar person that matches that de-identified X-ray image containing all the sensitive information. Such images can be re-identified using deep learning models with 95.55% accuracy [34]. Such de-identification and content removal also reduce model utility.

These models can even identify the person using that re-identified image. On the other hand, deep-learning models are data-hungry; as long as ones provide enough data, the model can learn efficiently. In these circumstances, our goal is to build a generative model where generated data will efficiently diagnose radiology images and X-ray images using synthetic data without having any direct dataset from the medical institutes. The authority will only provide a deep generative model trained from their pneumonia detection datasets. Depending on this trained model, we will synthesize the X-ray images, and using this synthetic data, we will train our final predictive model. During training, we will use a regularized model so that it cannot memorize the synthetic data, as previous works [39] indicated that regularization can reduce membership inference attacks also. We expect to develop realistic synthetic data that preserve all statistical contents of real data and develop an efficient and reliable predictive model that can yield satisfactory performance on such diagnoses using our generated data. Our main motivation is that our approach should need a simple model and the generator only to generate data for that predictive model, which will not cause any privacy leakage. Our generator itself is differentially private, and the used artificial data does not belong to any real patients. In our framework, we add noise to the generator's gradient only but not to the discriminator so that generated data is differentially private. Such private modeling will ensure that data privacy is preserved. The model's features or weights cannot be accessed by third party for exploiting learned weights to reconstruct original source data. Reverse engineering utilizing that privacy preserving model will not be feasible anymore because the noise will be injected into the encoded weights, so the model itself is private. Such an approach will encourage medical institutions to share more data, and such synthetic data are less expensive to collect and can be larger in quantity than real data.

Previous differential private generative adversarial networks (GANs) [7, 16, 29, 43, 48] recent performance encouraged us to use GANs for deferentially private synthetic data generation using deferentially private stochastic gradient descent. But, most of those approaches (e.g., DP GAN, PATE GAN) do not work well for high dimensional data generation. They performed generation task for simple MNIST, F-MNIST dataset where the images are not that complex to learn and such approach's data generation quality deteriorates with higher noise. We took an approach to reduce the noisy data problem where we will ensure a private generator not a private discriminator as in reality. We need to release the trained generator for public use. And, if we can ensure a more reliable discriminator where most of the gradient information is preserved then we can ensure high fidelity data generation with differential privacy guarantee. We adapt W-GAN [4], which uses Wasserstein loss with 1-Lispchitz condition to ensure that gradient norms are within a value range of 1. So, such implicit gradient clipping with bounded sensitivity reduces the need for clipping parameter tuning for GAN. In previous approaches, choosing the clipping bound was a very difficult task as it is also sensitive to other hyper parameters, like batch size, learning rate etc. In contrast, we solve this problem by using W-GAN where the implicit theory behind W-GAN's 1-Lispchitz condition ensures that the gradient of the generator is bounded by 1 value without explicitly searching for clipping hyperparameter.

Our key contributions of this paper include:

- We designed a differentially private approach to generate both reliable and private radiography image with selective noise addition (via W-GAN-based architecture) for the first time, ensuring 76% accuracy, which is satisfactory (close to real data).
- Our approach can preserve higher utility by applying selective gradient sanitization. We apply sanitization only to the generator and not to the discriminator like previous approaches to ensure more stable training with reliable data.
- We ensure implicit noise clipping and sensitivity bound of training using Wasserstein loss property of W-GAN [4, 20] that guarantee the gradient is within a limit of 1 (due to 1-Lipschitz condition). It eliminates the need to search for a perfect clipping value that is sensitive and may cause bias.
- We utilize a simple notion of privacy, ensuring that the deeper architecture can be trained with a feasible privacy budget. So, such notion will allow researchers to exploit deeper models for private data generation.
- Our novel synthetic and private medical data generation method works both in federated and distributed setting under untrusted server assumptions. It ensures that we can also use such an approach if we do not trust a centralized server to store the client's private data and the client only receives the noisy gradient, so the dishonest client cannot access other clients' data via model weights.

The remainder of this paper is organized as follows. Next section gives background and related works. Section III describes our approach for generating privacy preserving synthetic medical data. Evaluation results are shown in Section IV. Conclusions are drawn in Section V.

II. Background and Related Works

A. Background

In this section, let us review a few definitions: Generative adversarial networks (GANs), differential privacy (DP), Renyi DP, and Gaussian noise.

Generative adversarial networks (GANs) [7, 16, 29, 43, 48] are the approach to formulate generative task using deep-learning models. There will be an encoder based generator, which will perform the generative tasks. The generative model will learn the image features from training data and generate realistic looking synthesized data from random noise. There will be a discriminative model that will try to determine whether the data is fake or real. In this way the criticism for the generate data will be backpropagated and used to update the model. In this two-player game of generator and discriminator, the generator will improve over time to fool the discriminator and the discriminator will become more expert in classifying fake or real data and it will be rewarded or penalized based on its performance. This adversarial game like Eq. (1) will help us to learn a good mapping of the real data. It tries to minimize the loss of the generator G so that it generates real like image and at the same time tries to maximize the discriminator D's loss so that it cannot distinguish between real and fake data. In the beginning of the game, generator G is not that good, and it gradually improves over time while the discriminator D's parallel classification task's improvement forcefully lead to high quality image generation incrementally:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1 - \log D(G(z))] \quad (1)$$

For all data sets P and P', if they differ on at most one training example, any randomized algorithm K (for a set S of outcome where any $S \subseteq$ Range(K)) gives **ε-differential privacy (DP)** [15]. In practice, we add δ term as a failure probability to Eq. (2), which ensures (ε, δ)-privacy:

$$Pr[K(P) \in S] \leq e^{\varepsilon} \times Pr[K(P') \in S] + \delta \quad (2)$$

Here, DP algorithm considers epsilon ε, which indicates the upper bound of privacy loss. Particularly epsilon ε is the metric for privacy loss due to change in the data by one record. Lower epsilon value indicates better privacy budget but limited utility. We have to choose ε value wisely to maintain the utility-privacy trade-off. δ is used to relax the notion. δ is the estimated probability of breaching the constraints of differential privacy [15]. Here, we used differential Privacy in the context of Machine Learning problem and K is the generative model. Machine learning models are data hungry, the more the data they use for training, the more accurately they perform. In the same time in spite the availability of data, it is also important to ensure privacy of the system against the leakage of sensitive information. It ensures that model's predictive behaviour does not differ when the model has to predict training data or test data.

Differential privacy has a tremendous contribution in current machine learning advancement preserving privacy for data usage. However, it also brings issues for system maintenance cost. Machine learning model training is an iterative process and it adds privacy cost sequentially. As a result privacy budget restriction is becoming the major issue for developing machine learning model. **Renyi Differential Privacy (RDP)** [31] solves this issue by bringing more relaxation in DP algorithm. It increases the accuracy of the algorithm and it also reduces the computation cost for calculating privacy loss:

$$D_\alpha(P||P') = \frac{1}{\alpha - 1} \log \left(E_{p'(x)} \left(\frac{P(x)}{P'(x)} \right)^{\alpha-1} \right) \leq \varepsilon \quad (3)$$

This equation calculates Renyi divergence of order α of a distribution P from the distribution P'. Instead of using log likelihood to measure privacy loss, this method equips Renyi divergence to measure privacy loss. It will be described in more details in Section III.

To make our generator deferentially private, we have to ensure that each example may not have any significant impact on the model's encoded weight. To limit the impact of each example on the back propagated gradient we need to add some noise to the gradient. If D and D' are two adjacent dataset, then we need to add some noise to the output of a mechanism M. If $f(D)$ is the query function, then it will add $\mathcal{N}$ noise which is parameterized by σ, C. The noise is added to modify the distribution in 0 with standard deviation σ following Eq. (4). In our case, we have to run the training for multiple iterations and **Gaussian noise** can be a good choice due to its additive property which will be efficient in our method:

$$M(D) \simeq f(D) + \mathcal{N}(0,\, \sigma^2 C^2 I) \quad (4)$$

B. Privacy Preserving Learning

Deep learning is gaining popularity in predictive tasks. But such models are data-hungry, and they use different types of data scraping to collect data from all possible sources. Data have also been collected from various hospitals. These models are fundamental in the medical sector because they can make the diagnostic more reliable, but they need a large amount of data to perform well. However, using such sensitive data from hospitals and health databases can easily cause alarming privacy breaches. Still, previous works [2] proved that it is possible to enforce privacy in deep neural networks with a limited privacy budget. They introduced a differential private variation of common stochastic gradient descent with moment accountant technique [1], which helped to keep track of privacy using each of the moments other than mean, variance, and picks the tightest bounds. They clipped the gradient and added noise, limiting the information learned from any given example. Clipping bound C is a hyperparameter that needs to be tuned, which is a complex process and it can cause bias.

Pepernot et al. [35] introduced a teacher and student model concept in their PATE mechanism, which added noise to the outcome rather than during the training process, and it trained an ensemble of models based on multiple disjoint datasets. So, the privacy budget increased with iteration, and the model itself is not private. But, to make the model itself with encoded weight differentially private, To overcome the drawback of PATE, they proposed a new G-PATE mechanism [36] where they used Gaussian distribution instead of Laplacian distribution using Renyi differential privacy. The student played the role of private discriminator so that the student could learn how to extract the feature of unlabeled public data through the adversarial battle with the pre-trained generator. Still, the gradient needs to be subdivided into bins manually to cope with the framework in such a method. Due to the higher dimensionality of gradients, the noise added to the gradient increases the privacy budget, which needs to be minimized using unsupervised dimensionality reduction. So, to solve those exponential privacy budget increment problems and lower quality noisy data generation problems, we came up with a new approach. Our approach can reduce the need to select a proper clipping parameter and the expense of unsupervised dimensionality reduction. DP-GAN [45] solved the problem of privacy leakage due to training via real data-based training, and here, this approach started to clip weight rather than gradients. Kunar et al. [27] proposed DT-GAN for generating tabular synthetic data with privacy analysis by differential privacy against membership and attribute inference attacks. Tantipongpipat et al. [41] utilized differential privacy, and it ensures a private synthetic data generation process that can generate both data and label. Another approach utilized conditional GAN [43], which provides partial privacy. We were inspired by [10] paper's W-GAN usage technique. But, most of such methods targeted MNIST datasets where the learning task is much easier than complex medical datasets. They still have the problem of coming up with excellent clipping value. We eliminated the need to search for an appropriate clipping value using W-GAN. We also utilized high-dimensional radiology images, and our model can generate high-quality synthetic medical data in both centralized and distributed settings. DP-Fed AVG GAN [5, 30] works under the trusted server assumption. Still, it is difficult to assume that a centralized server is trusted because we also have to be prepared when the server becomes dishonest. Our approach ensures a federated system where the server only receives noisy gradients, so he cannot exploit the real data. So, it also works under the untrusted server assumption.

C. Generative Models in the Medical Field

Most deep learning models are data-hungry, so they require a lot of data. Directly using those public medical data creates privacy issues. Most of those data contain a tag/header or identifier that includes the patient's sensitive information, diagnosis history, and hospital name. So, people are getting more into synthetic data because synthetic data does not have private information, and those data do not belong to any actual patient. GAN [17] has already performed significantly well in data generation tasks in different domains; author Skandarani et al. [40] studied whether GANs can also work well in the medical data sector where the generated data should be reliable enough. Authors applied a range of generative architectures ranging from simpler DCGAN [18, 32, 37] to heavier style GANs [24] on cine-MRI, liver CT scan, and retina images. The study indicated that good-performing models could develop realistic data with higher FID scores and satisfactory performance with U-net [38]trained on generated data for segmentation. Bermudez et al. [8] used GAN to synthesize high-quality 2d axial slices of MRI in an unsupervised manner also supported by image denoising, which proved the power of deep learning in synthetic data generation. Dai et al. [13] developed a unified framework for generating synthetic images for multimodal MRI. Motion in the images causes quality degradation because of image blurring or artifacts. Johnson et al. [23] proposed a GAN model that can predict quality brain images from corrupted data. Lei et al. [28] presented a method that can generate synthetic computed tomography (CT) images based on dense cycle-consistent generative adversarial networks (cycle GAN). In the case of a skin lesion for skin image analysis, a considerable amount of labeled and high-quality data for deep learning is lacking. Baur's [25] framework using progressive, growing generative model was able to generate high-quality synthetic data compared to GAN, DCGAN [37], and LAPGAN [14]. Chuquicusma et al. [12] performed visual Turing test using radiologists to check the quality of their generated lung nodule samples. Their implicit assumption was that if they could learn to generate realistic data using DC-GAN and if it could fool the discriminator, then the model had known enough discriminative embedding. Some other works also exploited different generative methods to generate synthetic medical data of different type. [6, 19, 33, 44]Some previous works indicate that our radiology image generation approach is feasible and can lead to a satisfactory solution, but those works do not consider the privacy of the data. At the same time, our system can work with a differential privacy guarantee. Torfi et al. [42] addressed medical data privacy problems by generating synthetic data with acceptable quality and standard. Their framework used convolutional autoencoders to encode the features and generative adversarial networks to preserve the semantic information in the generated dataset. One positive side of their work is that- in the case of data generation, they followed robust method—Renyi differential privacy—to ensure and assess the privacy confidence of a system using such mathematical foundations, which also motivated us. Their model yielded better performance than state-of-the-art models based on publicly available benchmark data sets. Still, their model does not work well under higher noise for high-dimensional image type data. Choi et al. [11] handled binary and count feature-based electronic health record-based synthetic data generation using a specialized medGAN, which does not work for images. In their framework, they incorporated autoencoder and generative adversarial networks. One big problem in artificial data generation, mode collapse, is a common problem that this article successfully addressed using minibatch averaging, and it was able to ensure limited privacy risk. But, such little privacy cannot provide patient's sensitive data protection properly. So, in our approach, we incorporate relaxed differential privacy that can still generate high fidelity image data (high-dimensional) despite a high noise multiplier, and our artificial data can ensure higher accuracy using simple Resnet18 model [21] also. Our approach solves the problem of mode collapse using Wasserstein loss, which works much better than regular binary cross-entropy loss, and it also ensures private data generation. Mode collapse indicate a situation where the generator can only generate a single or small set of output, which reduces diversity among generated images.

III. Our Privacy Preserving Synthetic Medical Data Generator

We designed a privacy-preserving method to generate synthetic data. In our case, we have utilized Wasserstein GAN for a specific purpose. Some of the previous approaches [11, 19, 44, 46] tried to generate medical data but without a privacy guarantee and yielded low-medium utility. Some methods are developed using DP-SGD using generative architecture. But, they used gradient clipping for both discriminator and generator. But We used a different approach to exploit the gradient in the generator to ensure privacy-preserving data.

Instead of using a regular optimizer, We have used DP-SGD optimizer following previous techniques. We also used fully convolutional architecture instead of Multi Layer Perceptron to capture sensitive medical images' semantic and spatial information. We used W-GAN as it works slightly better to battle mode collapse. We utilized the implicit 1-Lipschitz distance property of W-GAN to avoid the crucial hyperparameter tuning for gradient clipping. A proper hyperparameter C helps set the gradient clipping bound, but it sometimes causes bias and takes time to develop an optimal value. But, 1-Lipschitz continuity in our GAN helps keep the gradient norm within a range of 1, which implicitly ensures gradient clipping during the training process without explicitly setting a proper clipping the value. So with the synergy of Renyi differential privacy and such gradient penalty based on the unique property of WGAN, our GAN can generate high-quality synthetic medical data. Using fake and real image-based comparative loss instead of binary cross-entropy and other techniques also helped increase the variation of the trained data, which allowed the target classifier models to generalize well.

A. Renyi Differential Privacy Implementation

In previous ε-DP approaches, the model creates some problems due to noise accumulation using strong composition [15]. As deep learning is an iterative process, noise upper bound gets multiplied with several training epochs. As we subsample images for micro-batch, subsampling also leads to high noise upper bound. Such loose upper bound increases the overall privacy cost. Generating data with privacy requires tracking the privacy budget and preserving the privacy of the generated data as each iteration requires adding noise. Hence, such an iterative learning process leads to a high privacy budget. But we need to minimize the privacy budget, and such an exponential increase in privacy budget may lead to a loose privacy upper bound. Such an upper bound with high noise deteriorate the quality of the image. So, we need to use the Gaussian method to preserve privacy and keep the privacy bound more tightly under the composition mechanism. Such a Gaussian mechanism with a higher spread and lower peak helps maintain noise balance, but (ε, δ)-privacy does not allow usage of the Gaussian mechanism. To exploit the Gaussian mechanism and ensure a tighter privacy upper bound, we used a simple notion of differential privacy, which satisfies and provides a strict upper bound. Instead of looking at the log ratio of probabilities, this privacy mechanism looks at the distance. This privacy technique ensures a strong guarantee under composition, and it is well suited to the Gaussian mechanism. Gaussian distribution has a less sharp peak, and 95% of the data stays within two standard deviations of the distribution, ensuring the upper bound could be much more compact and tight. Such a strict upper bound reduces the exponential parameter growth problem under iterations. This also satisfies ε-DP privacy when $\lambda = \infty$. (λ, ε)-RDP ensures $(\varepsilon + \frac{\log(1/\delta)}{\lambda - 1})$-DP privacy. Using such relaxed privacy helped us avoid overestimating privacy loss during multiple iterations as Renyi differential privacy supports the composition of different mechanisms where the budget does not grow exponentially. We can consider D and D' as two distributions, and $Pr(M(D'))$ is the probability of D after applying the generative mechanism M. λ is a parameter of that equation. Here, different epoch's generation task is considered as different mechanisms:

$$\begin{aligned} & D_\lambda(M(D)||M(D'(x))) \\ = \; & \frac{1}{\lambda - 1} \log \mathbb{E}_{x \sim M(D)} \left(\frac{Pr(M(D))}{Pr(M(D'))} \right)^{\lambda - 1} \leq \varepsilon \end{aligned} \tag{5}$$

B. GAN Implementation

If G is the generator, it takes random noise z as input and generates an image $G(z)$ as output. In the usual case, we provide the features, and the classifier classifies whether it is fake or real. But, in generator G, we provided the label y information, and a random Bernoulli or Gaussian noise z to generate the features $\hat{x}$, which are pixel values of the X-ray image. To create variation in data, we can alter the noise z, which will generate different pixel intensity values leading to a slightly separate X-ray image. In the generation process, the discriminator plays a vital role, so we kept the gradient of the discriminator D, intact and noise-free. A reliable discriminator is necessary as it can provide information regarding how accurate the image is. The discriminator that takes generated image $G(z)$ as input and $D(G(z))$ produces 0 if it is fake and one if it is real, so D simply acts as a binary classifier. But, the confidence probability value of the generator $D(G(z))$ indicates how fake or real the data is so that such meaningful error can be corrected in the second iteration. In the case of D, we generally use binary cross-entropy to calculate the criticism feedback; then, the feedback is backpropagated through the generator so that generator can learn whether the generated image $\hat{x}$ is realistic or not. The generator and discriminator have been trained simultaneously so that both models become experts. But, the generator must not become superior to the discriminator. Because an overfitted discriminator becomes so accurate that it provides confidence value at the highest or lowest level, which cannot give any meaningful feedback to improve the generator. So, we updated the discriminator five times per one generator iteration. If x is the input data, it tries to minimize the following loss in Eq. (6). So, the loss function in Eq. (6) consists of θ_D and θ_G parameters for discriminator D and generator G and g^t from Eq. (8) is the loss for generator and discriminator:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[1 - \log D(G(z))] \tag{6}$$

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[D(x)] - \mathbb{E}_{z \sim p_z(z)}[D(G(z))] \tag{7}$$

So, we decided to use Wasserstein loss following Eq. (7) instead of binary cross-entropy loss. It approximates the earth mover distance between a real and fake distribution. So, it helps to remove the ceiling of 0 and 1 of loss, which helps to fight the vanishing gradient problem, and continuous feedback helps to keep the learning with feedback consistent:

$$g^t = \nabla_\theta \mathcal{L}(\theta_G, \theta_D) \tag{8}$$

We used Wasserstein loss with a clipping bound of 1. Usual approaches clip the gradient before updating parameters. So, if the gradient vector is g and the L2-norm of the gradient is $||g||_2$ then we do the clipping by following $g/g(\max(1, \frac{||g||_2}{C})$. This process helps to ensure that $||g||_2 \leq C$ where C is the clipping parameter. But we mentioned that we eliminated the need to set the C value as we are using the Wasserstein loss, which measures the statistical distance between fake and real image distribution. 1-L continuous condition ensures the norm of the gradient $||g||_2 \leq 1$. So we try to enforce such 1-L continuity during training. We can do it by using wight clipping by setting a maximum or minimum allowed weight range but enforcing clipping reduces the limited learning capability of the discriminator. So, in the case of a discriminator, we will use gradient penalty to keep the sensitivity bounded like Eq. (9). We will calculate the loss as the distance between the real image x from the P distribution, and the fake image y from the Q distribution. In the loss term, we add a regularization term for calculating the loss for interpolated images from fake and real, multiplied with λ, a gradient penalty term. In such a way, we sample some points by interpolating between fake and real examples to get an interpolating image using a random number α. We deduct one from the gradient of discriminator's norm ∇D in Eq. (10), which ensures that the discriminator's gradient norm are bounded within a range of 1. This ensures clipping value as one without extensive hyperparameter tuning:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{x\sim P}[D(x)] + \mathbb{E}_{\hat{x}\sim Q}[D(\hat{x})] \\ &\quad +\lambda\mathbb{E}[(||\nabla D(\alpha x + (1-\alpha)\bar{x}|| - 1)^2] \qquad (9) \\ \mathcal{L}_G &= \mathbb{E}_{z\sim p_{\mathrm{z}}(z)}[D(G(z))] \qquad (10)\end{aligned}$$

C. Privacy Preserving Training with Santization

At this moment, by sanitization, we indicated refining the sensitive value by clipping and adding noise. The main learning mechanism of the machine learning model and deep learning depends on backpropagation. First, we provide a sample to the model, and it generates the output and calculates loss by comparing it with the real output. Then it uses loss for each sample to update the model per iteration. Our strategy is to add noise to the gradient so that updates regarding one single example cannot impact the overall learning. It follows the notion of differential privacy so that one individual sample cannot impact the overall dataset. Previous approaches applied sanitization on both discriminator and generator. Still, following some recent works [10], we decided to add noise to the gradient of the generator G in Eq. (14) only. We will not clip and add noise to the discriminator D in Eq. (13) because we are going to release the generator for data generation. If gr^t_G is the gradient of the generator G we apply gradient clipping and noise adding mechanism $M_{\sigma,C}(gr^{(t)}_G)$ to get that the modified gradient $\tilde{gr}^t_G$ so that each example cannot have a huge impact on dataset as in Eq. (11). $M_{\sigma,C}(gr^{(t)})$ adds noise from Gaussian distribution with variance σ. We will not provide the discriminator to the client, and discriminator gradient,$gr^{(t)}_D$ will remain unchanged, and it will be kept in a secure server. If we have to provide the discriminator, we will consider the federated learning scenario where we will have multiple discriminators for each client, which will be stored in client devices and will not breach privacy because each client will train their discriminator separately.

$$\begin{aligned}gr^t &= M_{\sigma,C}(gr^{(t)}) \qquad (11) \\ \theta^{(t+1)} &= \theta^{(t)} - \eta.gr^{(t)} \qquad (12) \\ \theta^{(t+1)}_D &= \theta^{(t)D} - \eta.gr^{(t)}_D; \\ &\quad \{Discriminator : \tilde{gr}^{(t)}_D := gr^{(t)}_D\} \qquad (13) \\ \theta^{(t+1)}_G &= \theta^{(t)G} - \eta.gr^{(t)}_G; \\ &\quad \{Generator : \tilde{gr}^{(t)}_G := M_{\sigma,C}(gr^{(t)}_G)\} \qquad (14)\end{aligned}$$

We applied a selective sanitization approach, which will clip the gradients of the initial layers of the generator and not apply it to the local layers because local layers are not getting exposed to private data. Our plan is that- we will not add noise to the discriminator's gradient, but we will add noise to the generator's gradient. Our idea is that as the discriminator provides feedback on the X-ray image's quality, the discriminator's noisy update cannot identify the difference between fake and real data. But as we are not releasing the discriminator, a noise-free discriminator helps preserve more gradient information of the discriminator, leading to high fidelity image data despite the noise multiplier's value. In the medical domain, image quality plays a crucial role because the semantic information of the image dictates a critical decision related to the disease. So, we tried to make a trade-off that can ensure both image quality and privacy, which is later proved by the satisfactory performance mentioned in our result section,

According to Fig. 1, there are two parts to the generator's gradient. One part is local, that is going downwards, which comes back to the generator, and one part is coming upwards, which is not local because it comes back from the discriminator and is affected by real data. So, Instead of sanitizing the whole network's gradient, we decided to sanitize the gradient that is directly relevant to the noisy input. Following the chain rule, we can identify that upward gradient,gr_G is directly impacted by real data, so we decided to sanitize this part of the gradient only so that local gradient,gr_G can preserve implicit gradient information, which is free from the impact of real data. The generator is updated twice during the training process. In GAN, when we update the generator, we keep the discriminator fixed and then update the discriminator and keep the generator fixed. According to the figure, the generator's updates back-propagated during discriminator evaluation are the upward gradient directly impacted by the real image. Hence, we decided to clip and add noise to the upward gradient. But during the downward gradient update, the gradient contains only relevant local information, which is not directly related to real data, so we do not sanitize the local gradient according to Fig. 1. In such a way, applying such selective noise addition by breaking down the chain rule helps us preserve important gradient information. So, it leads to high-quality synthetic data where reliability is critical as the spatial features of the images will be used for medical diagnosis tasks. In Fig. 1 red arrow indicates the sensitive gradient and green arrow indicates sanitized gradient. The red $X \sim D$ indicates the real X-ray image data, which is sensitive so the gradient coming back from discriminator, D's loss is indicated with red arrow. The green arrow going out to generator,G from Mechanism,M is

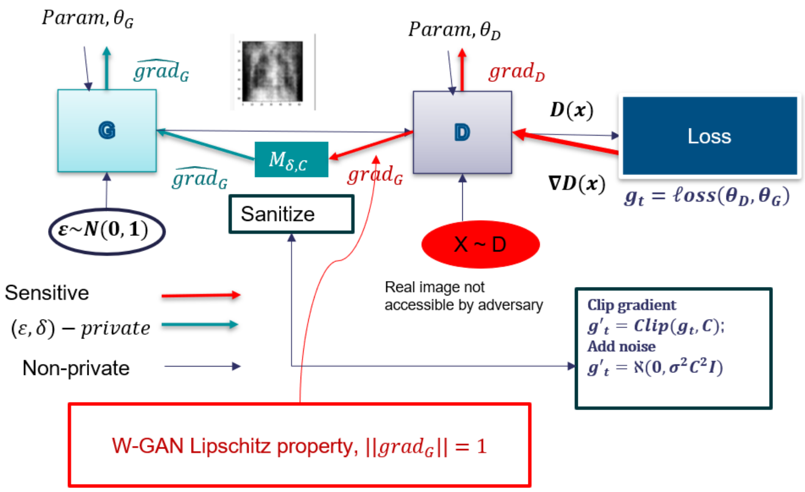
$Param, \theta_G$
$\widehat{grad}_G$
G
$M_{\delta,C}$
$\widehat{grad}_G$
$grad_G$
$Param, \theta_D$
$grad_D$
D
$D(x)$
Loss
$\nabla D(x)$
$g_t = \ell oss(\theta_D, \theta_G)$
$\varepsilon \sim N(0,1)$
Sanitize
X ~ D
Real image not accessible by adversary
Sensitive
$(\varepsilon, \delta) - private$
Non-private
Clip gradient
$g'_t = Clip(g_t, C)$;
Add noise
$g'_t = \aleph(0, \sigma^2 C^2 I)$
W-GAN Lipschitz property, $||grad_G|| = 1$

Fig. 1. Santiziation work flow

green gradient because mechanism M is used to sanitize the gradients.

We use the WGAN, which has a special condition is that it should be 1-Lipschitz continuous that is the slope of the gradient of the discriminator should always be 1. According to the theory of 1-Lipschitz continuity it automatically bounds the value of gradient.

D. Federated Approach

We also ensured a Federated learning approach where there will be a N_D number of discriminators that are trained in N client computers. Real data (x, y) will be exposed to the clients (hospitals) where they do not need to release sensitive data. Instead, they can train the lightweight discriminator on their personal computer. And the 1-Lipschitz property of Wasserstein helps to ensure implicit gradient clipping without performing sanitization. All of the discriminator's updates will be sent back to the central server's generator in Eq. (15) and the generator will be updated based on the accumulation of all gradient information. We need a reliable and accurate discriminator to stabilize the training and ensure high fidelity synthetic data. We followed a pre-trained starting approach where the discriminators will be previously pre-trained in different client computers for such an approach. During training, the pre-trained discriminator will ensure that generators are updated from the start of the training so that we can generate data using fewer epochs. During the generator update, noise and gradient clipping are applied to the upward gradient, similarly to the centralized approach. The iterative process increases the privacy budget, so a pre-training will also help reduce the privacy budget. It will require fewer iterations to decrease the privacy budget with fewer iterations.

The main advantage of this approach is that if someone wants a private discriminator, this approach will also ensure it. Because the discriminator will be stored in the client's computer, it will only have access to that specific client's corresponding X-ray images. There are other risk factors that we did not ignore. For example, if the client cannot trust the server, the client needs privacy protection from the server. But we tackled such a condition also because the client's gradient information that will be passed to the server will be sanitized, and so the encoded noisy weight cannot convey any information related to the client's real data to the server:

$$\theta^{(t+1)}_{D_{i=1\ldots N}} = \theta^{(t)D} - \eta . g^{(t)}_D ; \{Discriminator : \hat{g}^{(t)}_D := g^{(t)}_D\} \quad (15)$$

IV. Experiment

For medical purposes, we considered Kaggle Chest X-ray Images (Pneumonia) [26] and also used MNIST dataset for qualitative and quantitative comparison purpose. Because most of the previous privacy based data generation models used MNIST for study purpose to compare generated image based data. This is the first time we have exploited a real high stake domain's x-ray image dataset to generate synthetic images. One problem with synthetic medical X-ray dataset is reliability. So to ensure reliability and to defend mode collapse we used W-GAN, which is famous for its high fidelity data synthesis performance. In each iteration, we generated different Bernoulli or Gaussian noise depending on user's choice to preserve the diversity of the dataset. Observed from Table I, our approach's performance in terms of CNN is much closer to real data. In experimental setting we trained our model on 24000 generated synthetic data and to avoid class imbalance we generated 12000 Normal patient data and 12000 Pneumonia patient's data. As the GAN training is computationally expensive, we resized the image to 64×64 size and with such a low resolution still we were able to get upto 76% accuracy, which is within a satisfactory range. In MNIST, it also gained 77% accuracy, which is also good according to Table I. Observed from the figure, despite of high noise multiplier of 0.07/1.02, our approach can generate quality images whereas previous models generate blurry and unclear images. In case of X-ray images it also gained really good result with MLP: 74.4% accuracy. We used a highly

TABLE I
Experimental Comparison (Accuracy %)

Data	Algorithm	CNN (0.07)	CNN (1.02)	MLP
MNIST	Real	99	97	98
	G-PATE	51	49	25
	DP-SGD GAN	63	60	52
	Our approach	78.2	76	77.2
X-ray	Real	71.56	74.78	76
	DP-SGD GAN	60	58	40
	Our approach	76.172	76.245	74.484

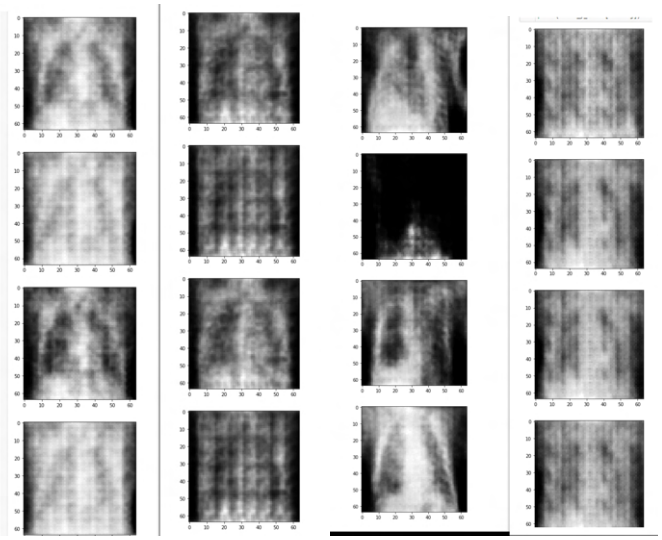

Fig. 2. Normal (first 2 columns)-Pneumonia (last 2 columns) with noise multiplier 0.07 (1st, 3rd column) vs. noise multiplier 1.02 (2nd, 4th column)

regularized CNN model to train using our synthetic data and such regularized model also will help to make it free from membership inference attack. Using such a simple ResNet18 model for X-ray images, it gained accuracy of 76.245% using CNN on synthetic image, which is close to 74.78% accuracy for real image (according to Table I). In case of MLP, it also gained 74.484% accuracy based on artificial radiology data, which is really amazing and it is closer to the model's accuracy of 76% using real image. In our case, we used synthetic data in training set and real data in test set so I believe such a higher and comparable accuracy may validate that our model can generate reliable radiology images, which can be used for diagnostic modeling. In Table I, the row for G-PATE is missing for x-ray image because G-PATE is not applicable for our dataset type.

To analyze the impact of privacy parameters like noise multiplier we performed some experiments with varying noise level. In Fig. 2, we showed our model's data quality concerning the noise multiplier. The first two columns indicate the standard patient images where the first column's data is generated with a noise multiplier of 0.07 and 1.02. Similarly, the third and fourth column shows the pneumonia patient's data. Here, the third column's pneumonia patient's data is generated using a noise multiplier of 0.07, and the fourth column's pneumonia patient's data is developed with a noise multiplier value of 1.02. From that part, we can observe that adding high noise of 1.02 still yields high-quality X-ray image data. In previous approaches, image quality usually gets destroyed after the noise multiplier value of 0.1. But we are glad to mention that our approach yielded 76% accuracy with data generated via a 1.02 noise multiplier, which is satisfactory. For qualitative analysis of the result, we also showed the generated average patient's images in Fig. 3,

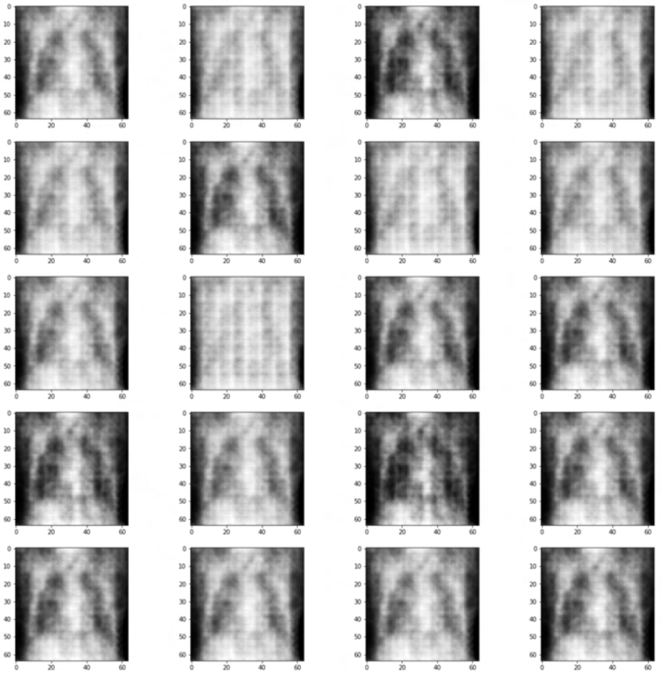

Fig. 3. Generated Normal Patients data with noise multiplier 0.07

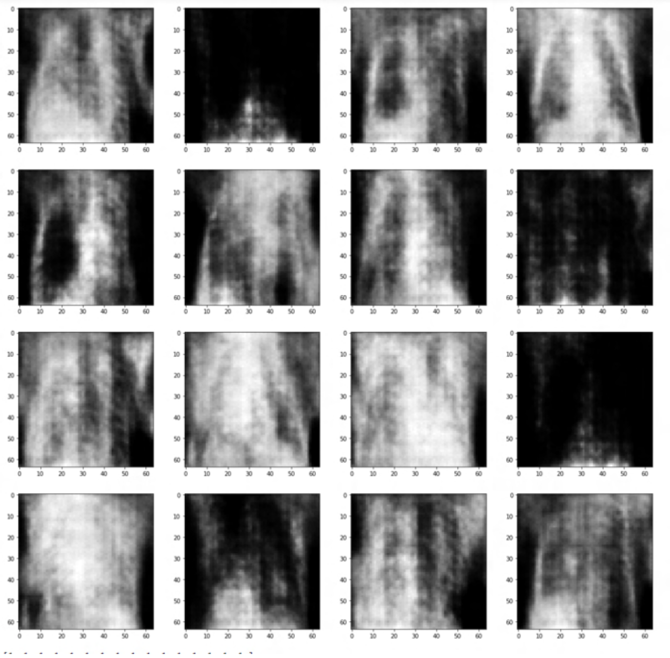

Fig. 4. Generated Pneumonia Patients data with noise multiplier 0.07

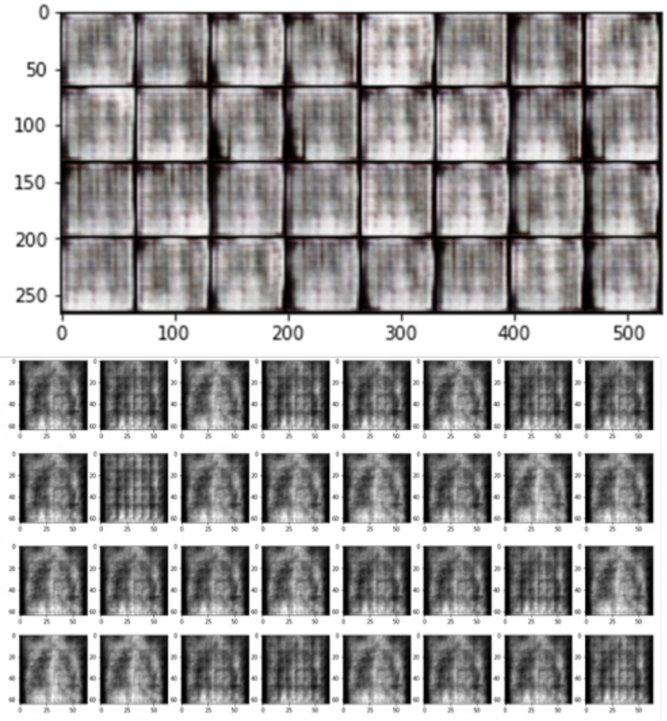

Fig. 5. Modified diffGAN (top) vs. ours (bottom)

TABLE II
TABLE-PRIVACY COMMUNICATION

Method	epsilon ϵ	delta δ	CT bytes
FedAVGGan	9.99×10^6	1×10^{-5}	3.94×10^7
Ours	7.89×10^2	1×10^{-5}	1.95×10^5

which is developed with ε value of 10 and noise multiplier 0.07. We also displayed the generated pneumonia images in Fig. 4. Observed from those two grid views, our model can differentiate between normal and pneumonia patients based on semantic structure. We used a sampling rate of $\frac{1}{1000}$ and we considered number of iterations to be 2000. According to experiment in worst case, our highest privacy budget for 24000 data and 2000 epoch is 3.194×10^4.

Previous approaches did not use high-resolution images for high stake domains, like – the medical radiology image classification task, so we had some trouble comparing with baselines. We used the scaled-down images and modified the diffGAN [42] architecture to generate X-ray images to compare with our model. We had to change the generator, encoder, and decoder architecture to support three channel images with higher resolution. However, the generated image with noise multiplier 0.07 is blurry, and there is mode collapse occurring in the images. Most of the X-ray images look alike. In contrast, our generated image is much sharper and more precise compared to previously generated images from Fig. 5.

In our federated approach, it ensured much efficient communication cost than previous approaches. Communication cost indicates how many bytes it consumes to perform one generator step by transferring gradient to the server. It takes less bytes, as we followed a previous approach and decided to transfer only the gradient with respect to real samples and as the local discriminator models are contained within local clients only. Table II shows that Fed AVG GAN's total ε value was 9.99×10^6 with CT bytes 3.94×10^7. In contrast, in our approach, ε value was 7.89×10^2 and CT bytes 1.95×10^5. It has much higher gains in gradient communication in terms of CT bytes. Fed-AVG GAN cannot perform well with noise multiplier value that is more than 0.1 whereas we have used 1.02 for noise multiplier value and still our approach is able to generate quality data.

A. Architecture Design

We configured a latent dimension of size for Generator z_dim of size 64. We used four layers of transpose convolution layer with the following $z\_dim \times 16, z\_dim \times 8, z\_dim \times 14, z\_dim \times 2, 3$ channels with a kernel size of 4, stride of 2, and padding of 1. only in the first layer, we used padding of 0, the stride of 1. Initially, we converted the noise of 64, and in each layer, the transposed convolution up-sampled the image size following the sequence of the image was $4 \times 4, 8 \times 8, 16 \times 16, 32 \times 32$ and $64 \times 64.64 \times 64$ is our required image size with three channels. We need to expand the noise to generate pixel intensity values of the image. So, we used transpose convolutional layers. We used tanh activation to ensure a high fidelity image in the last layer.

For the discriminator, we used four layers of transpose convolution layer with the following $z\_dim \times 2, z\_dim \times 4, z\_dim \times 8, z\_dim \times 8$ with a kernel size of 4, the stride of 2, and padding of 1. Here, in the first convolution layer, we used padding of 1 and stride of 2 with kernel size 4. For classification and diagnosis tasks, we have used regular ResNet18, a straightforward architecture. We used Batch Norm [22] in each layer so that the model is highly regularized so that it does not suffer from membership inference attacks.

V. CONCLUSIONS

In this paper, we designed a differentially private GAN architecture to generate synthetic X-ray images, which supports both central and distributed radiology data generation processes for the first time. Our main goal was to add noise only to the generator part, which is exposed to real data. We kept the discriminator intact, ensuring high-quality image generation. We need to release the generator only, and the generator part of our final model works as a private black-box model so that it will be diferentially private. Our approach guarantees the user data privacy also if we want to release the discriminator, but in that case, each client has to use and store their discriminator model locally and there will be a generator in the central server; it also ensures that any third party will not be able to reconstruct source data exploiting already learned weights because the encoded weights are noisy. We also used a highly regularized model to test the utility of generated data so that it can fight against inference attacks also. Our evaluation results demonstrated that our approach ensured higher-quality private X-ray images, ensuring a feasible privacy budget with more profound architecture. Using selective noise addition and W-GAN's implicitly clipping property helped to make it possible. We hope it will work as a steppingstone that will create awareness to protect medical data privacy and motivate other researchers to conduct more research in this medical image data field, which will help to ensure privacy in this domain. This study showed that it is possible to generate high stake medical data with differential privacy, which is also reliable despite high noise addition.

As *future work*, we explore a better model for classification and a higher resolution image can ensure higher utility.

Acknowledgment

This work is partially supported by NSERC (Canada) and University of Manitoba. We would like to thank the reviewers for their valuable feedback and insightful comments.

References

[1] M. Abadi, et al. Deep learning with differential privacy. *ACM SIGSAC CCS 2016*, 308-318.

[2] M. Al-Rubaie and J.M. Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy* 17, 2019, 49–58.

[3] S. Ali, et al. Towards privacy-preserving deep learning: opportunities and challenges. IEEE DSAA 2020, 673-682.

[4] M. Arjovsky, et al. Wasserstein generative adversarial networks. *ICML 2017 (PMLR 70)*, 214-223.

[5] S. Augenstein, et al. Generative models for effective ML on private, decentralized datasets. *ICLR 2020*.

[6] M. Baowaly, et al. Synthesizing electronic health records using improved generative adversarial networks. *JAMIA* 26, 2019, 228-241.

[7] B.K. Beaulieu-Jones, et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes* 12, 2019, e005122.

[8] C. Bermudez, et al. Learning implicit brain MRI manifolds with deep learning. *Medical Imaging 2018: Image Processing (SPIE 10574)*, 105741L.

[9] A. Bomai, et al., Privacy-preserving GWAS computation on outsourced data encrypted under multiple keys through hybrid system. IEEE DSAA 2020, 683-691.

[10] D. Chen, et al. GS-WGAN: A gradient-sanitized approach for learning differentially private generators. *NeurIPS 2020*. DOI 10.5555/3495724.3496787

[11] E. Choi, et al. Generating multi-label discrete patient records using generative adversarial networks. *MLHC 2017 (PMLR 68)*, 286–305.

[12] M.J.M. Chuquicusma, et al. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. *ISBI 2018*, 240–244.

[13] X. Dai, et al. Multimodal MRI synthesis using unified generative adversarial networks. *Medical Physics* 47(12), 2020, 6343–6354.

[14] E. L. Denton, et al. Deep generative image models using a Laplacian pyramid of adversarial networks. *NIPS 2015*, 1486-1494.

[15] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 2014, 211–407.

[16] L. Frigerio, et al. Differentially private generative adversarial networks for time series, continuous, and discrete open data. *SEC 2019*, 151–164.

[17] I. Goodfellow, et al. Generative adversarial nets. *NIPS 2014*, 2672-2680.

[18] S. Gross. Context-conditional generative adversarial networks. 2016.

[19] J. Guan, et al. Generation of synthetic electronic medical record text. *IEEE BIBM 2018*, 374–380.

[20] I. Gulrajani, et al. Improved training of Wasserstein GANs. In *NIPS 2017*, 5767-5777.

[21] K. He, et al. Deep residual learning for image recognition. *CVPR 2016*, 770–778.

[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML 2015 (PMLR 37)*, 448–456.

[23] P.M. Johnson and M. Drangova. Conditional generative adversarial network for 3D rigid-body motion correction in MRI. *Magnetic Resonance in Medicine* 82(3), 2019, 901–910.

[24] T. Karras, et al. A style-based generator architecture for generative adversarial networks. *CVPR 2019*, 4401–4410.

[25] S. Kazeminia, et al. GANs for medical image analysis. *Artificial intelligence in medicine* 109, 2020, 101938

[26] D.S. Kermany, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 2018, 1122–1131.e9.

[27] A. Kunar, et al. DTGAN: Differential private training for tabular GANs. 2021. arXiv:2107.02521

[28] Y. Lei, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Medical physics* 46(8), 2019, 3565–3581.

[29] J. Li, Mikhail et al. Differentially private meta-learning. *ICLR 2020*.

[30] H.B. McMahan, et al. Learning differentially private recurrent language models. In *ICLR 2018*.

[31] I. Mironov, et al. Rényi differential privacy of the sampled Gaussian mechanism. 2019. arXiv:1908.10530

[32] M. Mirza and S. Osindero. Conditional generative adversarial nets. 2014. arXiv:1411.1784

[33] A. Odena, et al. Conditional image synthesis with auxiliary classifier GANs. *ICML 2017 (PMLR 70)*, 2642–2651.

[34] K. Packhäuser, et al. Is medical chest X-ray data anonymous? 2021. aXiv:2103.08562

[35] N. Papernot, et al. Semi-supervised knowledge transfer for deep learning from private training data. *ICLR 2017*.

[36] N. Papernot, et al. Scalable private learning with pate. *ICLR 2018*.

[37] A. Radford, et al. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR 2016*.

[38] O. Ronneberger, et al. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*, 234–241.

[39] R. Shokri, et al. Membership inference attacks against machine learning models. *IEEE SP 2017*, 3–18.

[40] Y. Skandarani, et al. GANs for medical image synthesis: An empirical study. 2021. arXiv:2105.05318

[41] U.T. Tantipongpipat, et al. Differentially private mixed-type data generation for unsupervised learning. *IISA 2021*, 1-9, doi: 10.1109/IISA52424.2021.9555521.2021.

[42] A. Torfi, et al. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences* 586, 2022, 485–500.

[43] R. Torkzadehmahani, et al. DP-CGAN: Differentially private synthetic data and label generation. *IEEE/CVF CVPR 2019 Workshops*, 98–104.

[44] J.A. Walonoski, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *JAMIA* 25, 2018, 230–238.

[45] L. Xie, et al. Differentially private generative adversarial network. 2018. aXiv:1802.06739

[46] L. Xu, et al. Modeling tabular data using conditional GAN. In *NeurIPS 2019*.

[47] E. Yilmaz, et al. Naive Bayes classification under local differential privacy. IEEE DSAA 2020, 709-718.

[48] X. Zhang, et al. Differentially private releasing via deep generative model. *IEEE TKDE* 31(6), 2019, 1109–1121, https://doi.org/10.1109/TKDE.2018.2855136