Privacy-Preserving Synthetic Image Data Generation and Classification

by

Fahim Faisal

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
May 24, 2023

© Copyright 2023 by Fahim Faisal

Thesis advisors

Dr. Carson K. Leung, Dr. Yang Wang

Author

Fahim Faisal

Privacy-Preserving Synthetic Image Data Generation and Classification

Abstract

Computer vision, generative models (e.g., ChatGPT, etc.) and deep learning are now widely used across various sectors, from large corporations to end devices, simplifying people's lives and improving the reliability of medical findings. Sensitive image data and deep learning's high memorization capacity pose privacy risks, particularly for medical images containing sensitive private information. De-anonymization does not work due to the re-identification risk and reduced utility. So, we developed a differentially private approach with selective noise addition to generate high-dimensional synthetic medical image data with guaranteed differential privacy. In addition to ensuring data privacy, protecting the classification model's privacy is crucial due to its vulnerability to "membership inference attacks". State-of-the-art (e.g., differential privacy, etc.) defenses compromised task accuracy to preserve privacy, and some methods reuse private data or require more public data, which is impractical in some domains. To address privacy concerns while maintaining utility, we propose a collaborative distillation approach that transfers knowledge using minimal synthetic data, resulting in a compact private classifier model.

Contents

Abstract ii
Table of Contents iv
List of Figures v
List of Tables vi
Acknowledgments vii
Dedication ix

1 Introduction **1**
1.1 Private synthetic data generation 2
1.2 Privacy-preserving model training 4
1.3 Thesis statement 7
1.3.1 Synthetic data generation 8
1.3.2 Membership inference attack defense 8
1.4 Contributions 9
1.5 Thesis organization 11

2 Background and Related Works **12**
2.1 Background 12
2.1.1 GAN 12
2.1.2 Differential privacy 13
2.1.3 Renyi differential privacy 14
2.1.4 Membership inference attack in machine learning 15
2.1.5 Dataset distillation 16
2.1.6 Knowledge distillation 16
2.2 Related works 18
2.2.1 Privacy preserving learning 18
2.2.2 Generative models in the medical field 20
2.2.3 Membership privacy attack and defenses 23
2.2.4 Dataset condensation and distillation 25
2.3 Summary 26

3 Privacy-Preserving Synthetic Image Generation 28
3.1 Method . . . 29
3.1.1 Renyi differential privacy implementation . . . 30
3.1.2 GAN implementation . . . 32
3.1.3 Privacy preserving training with santization . . . 34
3.1.4 Federated approach . . . 38
3.2 Experiment for medical data . . . 39
3.2.1 Federated learning based experiment . . . 44
3.3 Summary . . . 46

4 Privacy-Preserving Learning via Data and Knowledge Distillation 47
4.1 Method . . . 48
4.1.1 Generating distilled data . . . 50
4.1.2 MIA mitigation via knowledge transfer . . . 54
4.2 Results and experiment for PLDK . . . 55
4.2.1 Datasets and model architecture . . . 55
4.2.2 Experimental setting . . . 57
Hyperparameter tuning . . . 59
4.2.3 Inference attack . . . 61
4.2.4 Comparison with regularization and knowledge transfer . . . 62
4.2.5 Comparison with differential privacy . . . 63
4.2.6 Ablation study . . . 65
Larger architecture-based model . . . 66
Increase data point via distillation (ipc) . . . 67
Tuning privacy . . . 68
4.3 Summary . . . 69

5 Conclusions & Future work 70
5.1 Conclusions . . . 70
5.2 Future work . . . 72

Bibliography 85

List of Figures

3.1 Data sanitization work ow 35
3.2 Normal patients (rst 2 columns) & pneumonia patients (last 2 columns) with noise multiplier 0.07 (1st, 3rd column) vs. noise multiplier 1.02 (2nd, 4th columns) 41
3.3 Generated normal patients data with noise multiplier 0.07 42
3.4 Generated pneumonia patients data with noise multiplier 0.07 43
3.5 Modi ed di GAN (top) vs. ours (bottom) 45

4.1 Transfer private teacher knowledge to a smaller student using distilled data 50
4.2 (a) Generating distilled data, (b) Distilling knowledge from teacher to student using distilled data 52
4.3 Synthetic distilled images: (a) MNIST, (b) CIFAR-10 56
4.4 Distilled data size vs. utility/privacy 68

List of Tables

3.1 Experimental comparison (Accuracy %) 40
3.2 Privacy communication 44

4.1 Utility-privacy trade-o comparison (Accuracy %) 57
4.2 Utility-privacy trade-o comparison (Accuracy %) 58
4.3 Comparison with KCD using ResNet-18 architecure 63
4.4 Empirical comparison with di erential privacy 64
4.5 Ablation study using CIFAR-10 66

Acknowledgments

I would like to begin by expressing my appreciation toward my M.Sc. thesis advisors, Dr. Carson K. Leung (University of Manitoba) and Dr. Yang Wang (University of Manitoba & Concordia University), for their unwavering assistance, guidance, and encouragement. Their relentless e orts, supervision, and nancial backing have been instrumental in making this thesis work a reality. I would also like to extend my heartfelt thanks to Dr. Noman Mohammed from Data and Security Lab for collaborating with me and o ering his domain expertise in the area of private learning. Lastly, having Dr. Noman Mohammed (University of Manitoba) and Dr. Yiming Qian (formerly at University of Manitoba, and currently at Amazon) on my advisory committee and examination committee is an absolute privilege. Thanks Dr. Lorenzo Livi for chairing my M.Sc. defence.

I want to express my gratitude towards Dr. Carson K. Leung and Dr. Yang Wang for providing me with nancial support, computational resources (Crane and Digital Research Alliance of Canada), and consistent encouragement throughout my academic journey, which proved to be essential for completing my thesis. The weekly meetings and frequent messaging opportunities in the dedicated Slack channel were particularly helpful in enabling me to make steady progress in my research.

My mentors' in-depth knowledge of the thesis topics and keen attention to detail has helped me re ne my research problem and methodology and produce a nal thesis. Their feedback and constructive criticism have been instrumental in shaping my arguments, experiments, and analysis. It is worth noting that they had a positive outlook toward experimental failures during the initial study period. Instead of considering them failures, they encouraged me by highlighting that I had ventured

into a new direction. This positive approach inspired me to remain optimistic and persistent. Additionally, they were always readily available whenever I sought his assistance, and responded promptly.

I also want to thank Faculty of Graduate Studies, Department of Computer Science, the department head, and the sta for providing me with the resources, central fellowship-based funding, and facilities necessary for my research. My gratitude also extends to my colleagues, lab mates, and friends who supported me during the writing process. In conclusion, I want to express my appreciation to my family and friends for their assistance; they have consistently been there for me when I required their help.

Fahim Faisal

B.Sc., BRAC University, Bangladesh, 2019

The University of Manitoba

May 24, 2023

This thesis is dedicated to my family and my beloved wife.
Their constant support and well wishes made this work possible.

Publication

Some ideas, materials, gures, and tables in my thesis have appeared in the following publications and submitted manuscripts.

Faisal, F. , Mohammed, N., Leung, C.K., Wang, Y.: Generating privacy-preserving synthetic medical data. 2022 IEEE Ninth International Conference on Data Science and Advanced Analytics (DSAA). pp. 1003{1012 (2022).

Faisal, F. , Leung, C.K., Mohammed, N., Wang, Y.: Privacy-preserving learning via data and knowledge distillation. 2023. (Currently under review for an international conference)

Chapter 1

Introduction

The popularity of deep-learning models' computation power encourages medical professionals to solve many disease detection problems. Notably, some medical sectors require much time and substantial human resources to do their job. Because of the high pro ciency and computation power, in modern days, machine learning is taking part in this sector to solve disease analysis problems in much reduced time and more e ciently. One problem is that medical datasets are di cult to use in modeling as they are associated with personal information, and preserving such health data's privacy is crucial [Ali+20; BAZ20; YAC20]. We need a method that produces high-quality private synthetic data that do not reveal the identity of any user and we also need a model that is private and does not leak any trained member's information.

Our research has resulted in the development of a di erential privacy-based data generator that can produce private data for training purposes while also providing theoretical guarantees. Additionally, we extend this work by developing a compact classi er model with a built-in defense mechanism that ensures membership privacy.

Using this private model to train synthetic data ensures empirical privacy, resulting in a two-step privacy mechanism that protects users' privacy from both a data and model perspective. By combining both empirical and theoretical bases, this dual protective training approach is anticipated to prioritize users' primary concern of privacy and thereby incentivize them to share sensitive data for training purposes.

1.1 Private synthetic data generation

However, in the rst part of the thesis, 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 and come up with a di erential privacy based medical data generator. Because medical data can also have a marker indicating who the actual person could be. For example, X-ray images in digital imaging and communications in medicine (DICOM) les are structured so that the cover sheets baked into DICOM les include patients' sensitive information for identi cation purposes. They consist of 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 les in the system. Such sensitive information leakage could link them to another sensitive dataset using those identi ers. This kind of privacy breach will be harmful to the patients. Some follow de-anonymization or content removal approaches (e.g., skull stripping for MRI data), which remove such identifying attributes from the image header in X-ray images. Yet, as the protocol for X-ray images di ers worldwide, standardizing such an approach is not feasible, and it losses utility with the risk of re-identi cation.

Machine learning attack models can even identify the person using that re-identi ed image. On the other hand, deep-learning models are data-hungry; if we provide enough data, the model can learn e ciently. In these circumstances, my goal is to build a generative model where generated data will be e ciently used for the diagnosis of radiology 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. My generator is di erentially private, and the used arti cial data does not belong to real patients. In this framework, I add noise to the generator's gradient only but not to the discriminator so that generated data is di erentially private. Such private modeling will ensure that data privacy is preserved. A third party cannot access the model's features or weights for exploiting learned weights to reconstruct 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 di erential private generative adversarial network (GAN)'s [Bea+19; Fri+19; TKP19; ZJW18; Li+20] recent performance encouraged me to use GANs for di erentially private synthetic data generation using di erential privacy-stochastic gradient descent. However, most of those methods fail to produce high-dimensional data, which is precisely what I aim to do in this work.

1.2 Privacy-preserving model training

In the second part of my work, I also want to ensure protection in the target classi cation model so that it has some built-in Membership Inference Attack defense and can come up with comparable accuracy privacy trade-o . Previous approaches require public reference data and repeated access to private data, which is risky, and there is a slight utility degradation. Thus, our comprehensive model-based data distillation approach can sanitize the knowledge transfer process to achieve a better trade-o than all the existing state-of-the-art methods. Also, it can restrict access to private data and does not require public data or data with a speci c property. So, such a private classi cation model can be equipped with the rst part's generated private data to ensure protection against Membership inference also.

In recent times, the use of machine learning and computer vision has become increasingly popular in medical diagnostic and image analysis tasks. However, models trained using Deep Learning techniques are susceptible to membership inference attacks (MIA) due to their high capacity to memorize training data. MIA is the simplest attack that enables an adversary to gain information about an individual by knowing that their data was used to train a predictive model. This attack can lead to more advanced attacks, such as attribute inference and feature extraction attacks. The consequences of MIA can be severe, particularly when the data involved is sensitive, as it may result in linking attacks or the de-anonymization of private data. Additionally, even trusted service APIs such as Google/Amazon machine learning platforms are not immune to MIA attacks, which pose a risk to users of such services. black box MIA [Sho+17] exploits prediction con dence and true label, while white

box [NSH19] MIA exploits gradient information leakage and other internal features to predict membership.

Defenses mainly focus on mitigating whitebox and blackbox MIA based on provable and empirical approaches. Empirical works can provide optimal performance as they are optimized for utility; On the contrary, provable methods can give a theoretical guarantee with a signi cant toll on accuracy. Even though theoretical methods are not designed for a particular dataset and are not optimized for utility, previous provable approaches like DP-SGD or PATE [Aba+16; Pap+16] can assure reasonable trade-o . But, due to colossal noise addition, they cannot provide acceptable utility. On the other hand, white box defenses based on empirical evidence like adversarial regularization and regular regularization ensure optimal accuracy with slightly reduced privacy. As they are mainly optimized for utility, increasing privacy requires higher regularization, which takes a toll on utility compared to privacy failing to provide an acceptable trade-o . Previous empirical knowledge transfer-based solutions showed some hope, but an approach like DMP [SH21] requires reference data with speci c properties like low entropy besides private data. Tuning or managing public data can be problematic for some privacy-sensitive tasks, and synthetic data-based DMP provides very low accuracy. On the other hand, models like KCD [Cho+22], and SELENA [Tan+22] require repetitive access to private data while knowledge transfer, which leaks a lot of private information. This requirement of additional public data and repeated access to con dential data for model training, with utility degradation, make these approaches infeasible. Again, they use a larger model to improve accuracy. Large capacity encourages them to memorize data, degrading privacy due to

over tting. So, appropriate model capacity can help to avoid memorization and privacy leakage. Thus, we need an e cient model that will overcome the limitations of over tting and the need for public data comparable to non-private ones. This work will be the rst to analyze a framework that can encourage privacy, e ciency, and utility at the same time. We develop a comprehensive model-based data distillation, Privacy-Preserving Learning via Data, and Knowledge Distillation (PLDK) to sanitize the training process to achieve a better empirical trade-o than all the existing state-of-the-art methods with lower MIA attack risk close to random guess.

Our approach in the second part of this thesis involves two main steps. First, we employ data distillation to obtain a smaller synthetic dataset that captures the common trends of the private data. This synthetic dataset is carefully tailored for the target student model, ensuring it has enough discriminative properties to produce accurate predictions. We generate a smaller amount of synthetic distilled data with which a model with the same architecture as the target student model can be trained to have similar parameters to when trained with total private data. Secondly, we train an unprotected teacher model with a large capacity on private data directly, and it is used to generate soft labels for the distilled data. The target student model is then trained using knowledge distillation on the labeled distilled data. The lightweight student model can achieve comparable accuracy to the teacher model trained on extensive private data by training on this labeled distilled data. It is also robust to inference attacks since it is not directly exposed to private data during nal model training.

The procedure is tailored for various datasets and can be employed directly in

a white box fashion without adding noise/modi cation. It is optimized for utility, whereas DP-based approaches are not directly optimized for utility and are not tailored for speci c datasets; instead, they focus only on privacy where utility is ignored. Our empirical policy has a built-in defense mechanism against inference attacks with a satisfactory utility trade-o compared to existing methods where the model will be tailored for speci c datasets. Direct data condensation [Caz+22; DZL22] does not focus on model privacy, and our data and model distillation synergy ensures tighter privacy bound of the model. Again, it helps to improve fairness [Far+20; YKF20] as class balance ensures fair training and privacy. Our empirical evidence showed that our model is almost free from the risk of MIA providing privacy leak of 50-53.5%, close to a random guess of 50%, nearly as precise or better than unprotected models ensuring 69.3%, 46.6%, and 97.3% test accuracy on Canadian Institute for Advanced Research (CIFAR)'s CIFAR-10, CIFAR-100, and Modi ed National Institute of Standards and Technology (MNIST) datasets, respectively.

1.3 Thesis statement

The objective of this thesis is to create a con dential data generator capable of producing high-quality private images for deep learning purposes. Additionally, a con dential and concise model was also developed in the second phase to train this data, ensuring a double-layered privacy guarantee for users.

1.3.1 Synthetic data generation

Medical image data contains private sensitive attributes. Therefore, training deep models on such data can leak private attributes or lead to a linking attack where such data can be re-identi ed. Anonymization is not a feasible solution, as standardizing such images globally is challenging. Thus, we need to come up with a generative approach that can generate medical data,$x(y)$, which is di erentially private. We expect that model can ensure the same outcome irrespective of the presence of training data which is a challenge in high stake medical domain. We need a setup so that local hospitals can provide their data privately. We also want to develop reliable high- delity data to ensure proper decision-making in high-stakes diagnosis tasks. We will follow Chapter 3 to solve this problem.

1.3.2 Membership inference attack defense

This a popular attack where the adversary trains an inference model to identify whether a speci c data record is trained using a speci c model or not. So, this information breaches the privacy of the training member. Suppose we train a model in a supervised setting where$(x; y)$ is the corresponding image and true label of the data, and we want to predict $\hat{y}$; then the adversary can access x and $\hat{y}$ both. An adversary may also have access to the prediction con dence and they can exploit such a relationship to identify whether a given data belongs to the training set or not. For example, if they can determine that data corresponds to a cancer hospital or rape victim database, it will implicitly expose some sensitive attributes about the identi ed individual. Training leads to unintended leakage (if data is private). So

following cahpter 4 solutions, we want to come up with a model where given the true label and calculated prediction con dence, the adversary cannot successfully identify whether data belongs to the corresponding training set or not. Why is it essential? Knowing that a particular patient's clinical record was used to train a model associated with a disease (e.g., predict cancer or speci c medicine dose, etc.) can reveal that the patient has this disease. This a popular attack where the adversary trains an inference model to identify whether a speci c data record is trained using a speci c model or not. So, this information breaches the privacy of the training member.

1.4 Contributions

Our key contributions are summarized below:

We designed a di erentially private approach to generate both reliable and private high resolutions radiography images with selective noise addition (via W-GAN-based architecture) for the rst time, ensuring close to real data accuracy, which is satisfactory.

- 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 [Gul+17; ACB17] 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 the centralized 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.

We expand our image classi er in addition to the di erential privacy-based data generation method. We developed a classi cation model that is resistant to membership inference, nearly as precise as non-private models, and can be employed directly in a white box manner, where the procedure is customized for various datasets. For such optimization, we exploited a knowledge distillation-based approach where the resultant student model is trained on synthetic distilled data, and the model will not be exposed to private data.

- This extended distillation-based approach ensures target model compactness and membership privacy together for the rst time. The nal model does not rely on any public data and e ectively restricts repeated access

to private data. So it is more feasible in real-world settings.

{ Distilled data facilitate class balance and capture general trends during knowledge transfer, promoting fair training and encouraging generalization. Furthermore, the generated data is tailored to suit the student model.

{ Resultant student model is e cient as it requires less than half the model and data size to train and is robust to MIA risk. It provides superior utility improvements of 8%, 34%, and 6% in the CIFAR-10, CIFAR-100, and MNIST datasets with 50-53.5% privacy leak, similar to non-private models.

1.5 Thesis organization

I organize the remainder of this thesis as follows. First, in Chapter 2 we discussed some of the background de nitions and in 2.2, we give an overview of the related works. More speci cally, this chapter discusses di erent key components of our algorithm and various privacy-based algorithms. Chapter 3 outlines my Private synthetic image data generation approach. It shows how di erential privacy can be combined with GANs to generate high-quality reliable image data. Chapter 4 proposes a Collaborative distillation approach for the image classi cation task. Unlike the previous supervised approaches, this approach overcame more public data requirements or reference data tuning steps. Finally, in Chapter 5, I conclude this thesis.

Chapter 2

Background and Related Works

2.1 Background

Let us review a few de nitions: Generative adversarial networks (GANs), differential privacy (DP), Renyi DP, Gaussian noise, Membership Inference Attack in machine learning, data and knowledge distillation.

2.1.1 GAN

Generative adversarial networks (GANs) [Bea+19; Fri+19; TKP19; ZJW18; Li+20] 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 generated data will be back-propagated 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. (2.1) 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 classi cation task's improvement forcefully lead to high-quality image generation incrementally:

$$\min_{G} \max_{D} E_{x \quad p_{data}(x)}[\log D(x)] + E_{z \quad p_z(z)}[1 \quad \log D(G(z))] \tag{2.1}$$

2.1.2 Di erential privacy

For all data sets P and P', if they di er on at most one training example, any randomized algorithm K (for a set S of outcome where any S Range(K)) gives $"$-di erential privacy (DP) [DR14]. In practice, we add term as a failure probability to Eq. (2.2), which ensures $(";)$-privacy:

$$Pr[K(P) \ 2 \ S] \quad e^{"} \quad Pr[K(P^0) \ 2 \ S] + \tag{2.2}$$

Here, DP algorithm considers epsilon $"$, which indicates the upper bound of privacy loss. Particularly epsilon $"$ 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 $"$ value wisely to maintain the utility-privacy trade-o .

is used to relax the notion. is the estimated probability of breaching the constraints of di erential privacy [DR14]. Here, we used di erential 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 di er when the model has to predict training data or test data.

2.1.3 Renyi di erential privacy

Di erential 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 Di erential Privacy (RDP) [MTZ19] 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:

$$D \quad (P \mathrm{jj} P^{0}) = \frac{1}{1} \log \quad E_{p^{0}(x)} \quad \frac{P(x)}{P^{0}(x)} \quad {}_{1}^{!} \quad " \tag{2.3}$$

This equation calculates Renyi divergence of order 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 2.1.3.

To make our generator deferentially private, we have to ensure that each example may not have any signi cant 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^0 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 N noise which is parameterized by C. The noise is added to modify the distribution in 0 with standard deviation following Eq. (2.4). 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 e cient in our method:

$$M(D) \text{'} \quad f(D) + N(0; \quad {}^2C^2I) \tag{2.4}$$

2.1.4 Membership inference attack in machine learning

Suppose we train a model F in a supervised setting where $(x; y)$ is the corresponding image and true label of the data, and we want to predict $y \triangleq F(x)$. If an adversary model g can access $(x; y)$ and prediction vector $F(x)$ with prediction con dence, then it can exploit such a set of features to infer whether a given labeled image $(x; y)$ belongs to the F model's training set D. The accuracy of model g will be referred to as `privacy accuracy'. In a black box setting, an adversary might train a binary attack model g, assuming they have access to some of the train data of a model F, with the prediction vector $F(x)$. In a white-box manner, an adversary may exploit di erent model features like loss of model F on $(x; y)$ and exploit gradient information $_F(x; y)$. If D^A and D^{A^0} are members and nonmembers of the portion of

the disjoint data accessible to the adversary, attack model g optimizes $_g$ to maximize the following gain I in Eq. (2.5):

$$I^{g}_{D^A;D^{A'}}(g) = \sum_{(x;y)\in D^A} \frac{\log(g(x;y;F(x))}{|D^A|} + \sum_{(x;y)\in D^{A'}} \frac{\log(1 - g(x;y;F(x)))}{|D^{A'}|} \tag{2.5}$$

2.1.5 Dataset distillation

Dataset distillation [Wan+18] tries to keep the model $x$ed and to develop a smaller dataset that can preserve similar test performance on a reserved test set compared to the dataset trained on the original dataset. Like Eq. (2.6), this is a process that helps to squeeze the larger dataset D to a smaller sized dataset S, and the ability to retain the same performance $L(F_{\frac{F}{S}}(;))$ on distilled data S as real data D's training performance $L(F_{\frac{F}{D}}(;))$. Here, L is regular loss (e.g., cross entropy):

$$E_{x \sim P_D}[L(F_{\frac{F}{D}}(x);y)] \approx E_{s \sim P_S}[L(F_{\frac{F}{S}}(s);y)] \tag{2.6}$$

Thus, it maintains a $x$ed model F parameterized by $$ and scales down the amount of data to improve training e$$ectiveness by reducing training complexity; Distilled data S can better capture the feature dynamics of actual data D.

2.1.6 Knowledge distillation

Hinton et al. [HVD15] improved the concept of Bucilu$$a et al.'s model compression [BCN06] process to transfer the knowledge of an ensemble of cumbersome teacher models into a single compact student model. While teaching the student network, soft labels generated by teacher networks are used for supervision. This ensures the student network can distinguish between di$$erent class values, which helps preserve

similarity information between di erent classes. They used temperature to smooth the output logits' $a_F; a_T$ value distribution to provide higher generalization capability. Some approaches also use intermediate values [Che+20a] than probability. Such generalization helps promotes regularization, which can help mitigate membership privacy information leaked via extremely con dent logit values. If y is the ground truth label and $F(x)$, $T(x)$ are the softened label of student and teacher network, then KL divergence-based loss L_{KL} in Eq. (2.7) is used to compute knowledge distillation loss L_{KD} :

$$L_{KD} = {}^{2}L_{KL}(F(x); T(x)); F(x) = \text{softmax}\left(\frac{a_F}{\ }\right); T(x) = \text{softmax}\left(\frac{a_T}{\ }\right) \tag{2.7}$$

2.2 Related works

2.2.1 Privacy preserving learning

Deep learning is gaining popularity in predictive tasks. But such models are data-hungry, and they use di erent 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 [AC19] proved that it is possible to enforce privacy in deep neural networks with a limited privacy budget. They introduced a di erential private variation of common stochastic gradient descent with the moment accountant technique [Aba+16], which helped to keep track of privacy using each of the moments other than mean, and variance and picks the tightest bounds. They clipped the gradient and added noise, limiting the information learned from any given an example. Clipping bound C is a hyperparameter that needs to be tuned, which is a complex process that can cause bias.

Pepernot et al. [Pap+17] 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 di erentially private, To overcome the drawback of PATE, they proposed a new G-PATE mechanism [Pap+18] where they

used Gaussian distribution instead of Laplacian distribution using Renyi di erential 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 [Xie+18] 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. [Kun+21] proposed DT-GAN for generating tabular synthetic data with privacy analysis by di erential privacy against membership and attribute inference attacks. Tantipongpipat et al. [Tan+19] utilized di erential privacy, and it ensures a private synthetic data generation process that can generate both data and labels. Another approach utilized conditional GAN [TKP19], which provides partial privacy. We were inspired by [COF20] 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 [Aug+20; McM+18] works under the trusted server assumption. Still, it is di cult 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.

2.2.2 Generative models in the medical eld

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 identi er 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 [Goo+14] has already performed signi cantly well in data generation tasks in di erent domains; author Skandarani et al. [SJL21] 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 [RMC15; MO14; Gro16] to heavier style GANs [KLA19] on cine-MRI, liver CT scan, and retina images. The study indicated that good-performing models could develop realistic data with higher Frechet Inception Distance (FID) scores and satisfactory performance with U-net [RFB15]trained on generated data for segmentation. Bermudez et al. [Ber+18] used GAN to synthesize high-quality 2d axial slices of MRI in an unsupervised manner also supported by im-

age de-noising, which proved the power of deep learning in synthetic data generation. Dai et al. [Dai+20] developed a uni ed framework for generating synthetic images for multi-modal MRI. Motion in the images causes quality degradation because of image blurring or artifacts. Johnson et al. [JD19] proposed a GAN model that can predict quality brain images from corrupted data. Lei et al. [Lei+19] 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 [Kaz+20] framework using progressive, growing generative model was able to generate high-quality synthetic data compared to GAN, DCGAN [RMC15], and LAPGAN [Den+15]. Chuquicusma et al. [Chu+18] 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 di erent generative methods to generate synthetic medical data of di erent type. [Bao+19; Wal+18; Gua+18; OOS17]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 di erential privacy guarantee. Tor et al. [TFR22] 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 di erential privacy|to ensure and assess the privacy con dence 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. [Cho+17] 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 arti cial data generation, mode collapse, is a common problem that this article successfully addressed using mini-batch 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 di erential privacy that can still generate high delity image data (high-dimensional) despite a high noise multiplier, and our arti cial data can ensure higher accuracy using simple Resnet18 model [He+16] 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.

2.2.3 Membership privacy attack and defenses

Usually, there are two types of membership inference attacks : Black box and white box attacks. Nasr et al. [Sho+17] came up with the concept first. They trained shadow models on synthetic data (selected from distribution based on confidence score) and mathematically showed and quantified membership privacy leakage given black-box access only. Long et al.'s [Lon+18] findings demonstrated that models with good generalization capabilities can still be vulnerable to inference attacks, as they can be indirectly targeted by accessing associated data. Nasr et al. [NSH19] extended this attack by introducing a white box attack that exploits gradient information and showed how such gradient difference between member and non-member makes deep learning algorithms vulnerable in both central and federated settings. Atiqur et al. [Rah+18] also showed that guaranteed differential privacy could be susceptible if the trade-off expects acceptable utility. [SM20] improved the performance of MIA by using a class-dependant threshold based on prediction entropy and proposed a new risk score metric. Label-only attacks [Cho+21; LZ21] reduced dependency on prediction confidence and proved that the confidence masking defense is not very efficient. Certain studies [Ben+20] have indicated that to prevent attacks, it is necessary to address common enemies such as overfitting and generalization gap. Chen et al. [Che+20b] showed how attack models could leak the training data information of different generative models in different settings.

Some of the defense strategies are discussed here. For example, Nasr et al. [NSH18] maximized the classification performance of the model and, at the same time, minimize the most potent adversary's membership inference attack's gain based

on prediction con dence. But, adversarial regularization comes with a signi cant toll on utility. Other regularizations include L2 regularization, early stopping, Weight Decay, Weight Normalization, DropOut and label smoothing [KHD20] are also used to ensure privacy. Such regularizers deteriorate the utility due to a sub-optimal trade-o . Another approach is MemGuard which alters the output of the resultant model by adding noise to confuse the attack model, but it does not work for white-box attacks. DP-SGD [Aba+16] uses di erential privacy-based optimization and adds noise to the gradient of micro-batches. Similarly, PATE [Pap+18; Pap+16], PATE with GN-MAX used an ensemble of teachers on di erent subsets of the data and started using knowledge distillation for privacy. DP-based GAN [Ho+21; Fai+22] tried to generate synthetic data with additional noise. Song et al. [SM20] developed a new privacy risk score and recognized both model sensitivity and generalization error pose a mutual threat. Caruana et al. [CLG00] showed how early stopping can help reduce over tting problems and mitigate privacy risk.

DMP [SH21] used knowledge transfer initially, where the student model is trained on public data with private teacher-generated soft labels to achieve a superior utility trade-o . But, they require available reference data with desirable properties, like low entropy, to produce better results, and synthetic data-based DMP achieved moderate performance. Managing public data with such properties may be challenging in medical domains. KCD, SELENA [Cho+22; Tan+22] used data splitting and repeated private data usage. However, KCD's limitation is that it might not work if there is duplication in train data or class imbalance and outlier.

2.2.4 Dataset condensation and distillation

This is an emerging topic, and many works have been done in this domain, focusing on neural architecture search and training e ciency. The central concept behind this approach is maintaining the model while simultaneously creating a condensed dataset that allows the model to achieve comparable performance on the reduced data, as it does when trained on the complete dataset. The initial approach was introduced by Wang et al. [Wan+18] where they used an approach similar to meta-learning where a randomly initialized dataset is optimized in a few steps of gradient descent to come up with a smaller dataset that can ensure a similar performance of the model on the real data. To simplify the nested meta-training loop, Zhao et al. [ZMB21] concentrated on aligning the model's gradients based on generated and real data, which aims to ensure that the model follows the same path towards a solution. They [ZB21] also utilized di erentiable Siamese augmentation where the actual images and synthetic images are transformed using the same augmentation before matching gradient, which helps to improve the performance signi cantly due to shared transformation. Some tried to use soft labels via label distillation [BYH20] and also considered text data [SS19]. Nguyen et al. [NCL20] used Kernel Ridge Regression and label solve technique via closed form solution, increasing e ciency. Lee et al. [Lee+22] emphasized class-wise di erences in the loss and improved optimization stability using bi-level warm-up. Distribution Matching in embedded space [ZB22] helped to ensure faster computation, and Dong et al. [DZL22] demonstrated how data condensation is connected with privacy theory and training e ectiveness. They provided theoretical justi cations for the relationship between di erential privacy and data condensation. However, their

assumption regarding adversary access to the compressed data becomes vulnerable in worst case, which occurs quite often practically. So, we combine model's knowledge transfer with privacy-oriented data distillation, restricting adversary access to distilled data to ensure better empirical privacy. Moreover, Cazenavette et al.'s [Caz+22] parameter-based imitation learning improved distillation e ciency. But, their work does not optimize data utilizing privacy, and their goal and deliverable are compressed data, not a private model. So, we embrace parameter-matching loss during privacy-oriented data distillation and sanitize the model using knowledge transfer to ensure compact private model delivery.

2.3 Summary

In this chapter, we reviewed a few de nitions|namely, generative adversarial networks (GANs), di erential privacy (DP), Renyi DP, Gaussian noise, membership inference attack in machine learning, data and knowledge distillation. We also discussed related works on privacy-preserving learning, generative models in the medical eld, membership privacy attack and defenses, as well as dataset condensation and distillation.

As a preview, in Chapter 3, we develop a private data generator to utilize a noise-free discriminator, which helps to produce high-quality private radiology data whereas previous methods could not guarantee high-resolution radiology data with such a tight privacy budget. In Chapter 4, we come up with a private classi er model without directly accessing any private data directly, which ensures better privacy. Again, it can ensure a smaller model with high performance which requires half the

parameters compared to related works.

Chapter 3

Privacy-Preserving Synthetic Image Generation

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 ndings. 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-identi cation 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 signi cant because a less accurate diagnosis may lead to the wrong course of treatment and/or loss of life. In this chapter, we develop a di erentially private approach that can generate high-quality and

high dimensional synthetic medical image data with guaranteed di erential privacy. It can be used to create su cient quality data to train a deep model. Moreover, we use W-GAN for bounded gradient guarantee, which eliminates the need for an extensive clipping hyperparameter search. We also add noise selectively to the generator to maintain the privacy-utility trade-o . 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 di erent hospitals can contain their private images in the local server and use a central server to generate synthetic radiology images without storing patient data.

3.1 Method

We designed a privacy-preserving method to generate synthetic data. In our case, we have utilized Wasserstein GAN for a speci c purpose. Some of the previous approaches [Cho+17; Gua+18; Xu+19; Wal+18] 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 di erent 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 di erential 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 classi er models to generalize well.

3.1.1 Renyi di erential privacy implementation

In previous "-DP approaches, the model creates some problems due to noise accumulation using strong composition [DR14]. 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 $(\ ,\ )$-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 $= 1$. $(\ ;\ )$-RDP ensures $(\ + \frac{\log(1/\ )}{1})$-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^0 as two distributions, and $Pr(M(D^0))$ is the probability of D after applying the generative mechanism M. is a parameter of that equation. Here, different epoch's generation task is considered as different mechanisms:

$$D\ (M(D)\,||\,M(D^0(x))) = \frac{1}{1}\log E_{x\ \ M(D)}\ \frac{Pr(M(D))}{Pr(M(D^0))}^{\ 1} \qquad (3.1)$$

3.1.2 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 generator iteration. If x is the input data, it tries to minimize the following loss in Eq. (3.2). So, the loss function in Eq. (3.2) consists of

$_D$ and $_G$ parameters for discriminator D and generator G and g^t from Eq. (3.4) is the loss for generator and discriminator:

$$\min_G \max_D E_{x \quad p_{data}(x)}[\log D(x)] + E_{z \quad p_z(z)}[1 \quad \log D(G(z))] \tag{3.2}$$

$$\min_G \max_D E_{x \quad p_{data}(x)}[D(x)] \quad E_{z \quad p_z(z)}[D(G(z))] \tag{3.3}$$

So, we decided to use Wasserstein loss following Eq. (3.3) 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 ght the vanishing gradient problem, and continuous feedback helps to keep the learning with feedback consistent:

$$g^t = r \quad L(\ _G;\ _D) \tag{3.4}$$

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 $jj gjj_2$ then we do the clipping by following $g = g(\max(1; \frac{jj gjj_2}{C})$. This process helps to ensure that $jj gjj_2 \quad C$ 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 $jj gjj_2 \quad 1$. 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. (3.5). We will calculate the loss as the distance between the real image x from the P

distribution, and the fake imagey 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 , 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 . We deduct one from the gradient of discriminator's norm r_D in Eq. (3.6), 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:

$$L_D = E_{x \sim P}[D(x)] + E_{\hat{x} \sim Q}[D(\hat{x})] + E[(\|r_D(\ x + (1 \quad)x\| \quad 1)^2] \tag{3.5}$$

$$L_G = E_{z \sim p_z(z)}[D(G(z))] \tag{3.6}$$

3.1.3 Privacy preserving training with santization

At this moment, by sanitization, we indicated re ning 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 di erential privacy so that one individual sample cannot impact the overall dataset. Previous approaches applied sanitization on both the discriminator and generator. Still, following some recent works [COF20], we decided to add noise to the gradient of the generator G in

Figure 3.1: Data sanitization work ow

Eq. (3.10) only. We will not clip and add noise to the discriminator D in Eq. (3.9) because we are going to release the generator for data generation. $gr_G^{(t)}$ is the gradient of the generator G we apply gradient clipping and noise-adding mechanism $M_{;C}(gr_G^{(t)})$ to get that the modi ed gradient gr_G^t so that each example cannot have a huge impact on the dataset as in Eq. (3.7) $M_{;C}(gr^{(t)})$ adds noise from Gaussian distribution with variance . We will not provide the discriminator to the client, and discriminator gradient $gr_D^{(t)}$ 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.

$$gr^t = M_{;C}(gr^{(t)}) \tag{3.7}$$

$$^{(t+1)} = {}^{(t)} \quad :gr^{(t)} \tag{3.8}$$

$$^{(t+1)}_{D} = {}^{(t)}{}_{D} \quad : gr^{(t)}_{D};$$

$$f \text{ Discriminator} \quad : gr^{(t)}_{D} := gr^{(t)}_{D} g \tag{3.9}$$

$$^{(t+1)}_{G} = {}^{(t)}{}_{G} \quad : gr^{(t)}_{G};$$

$$f \text{ Generator} : gr^{(t)}_{G} := M_{;C}(gr^{(t)}_{G}) g \tag{3.10}$$

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 di erence 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- delity 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-o that can ensure both image quality and privacy, which is later proved by the satisfactory performance mentioned in our result section,

According to Figure 3.1, there are two parts to the generator's gradient. One part is local, which 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 a ected 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 the 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 xed and then update the discriminator and keep the generator xed. According to the gure, 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 Figure 3.1. 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 Figure 3.1 red arrow indicates the sensitive gradient and the green arrow indicates sanitized gradient. The red X D indicates the real X-ray image data, which is sensitive so the gradient coming back from discriminator D's loss is indicated with a red arrow. The green arrow going out to generator G from Mechanism M is a 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.

3.1.4 Federated approach

We also ensured a Federated learning approach where there will be N_D a 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. (3.11) 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- delity synthetic data. We followed a pre-trained starting approach where the discriminators will be previously pre-trained in di erent 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, similar to the centralized approach. The iterative process increases the privacy budget, so 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 speci c 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:

$$D_{i=1:::N}^{(t+1)} = {}^{(t)D} \quad : g_D^{(t)}; f \text{ Discriminator} \quad : \hat{g}_D^{(t)} := g_D^{(t)} g \tag{3.11}$$

3.2 Experiment for medical data

For medical purposes, we considered Kaggle Chest X-ray Images (Pneumonia) [Ker+18] and also used MNIST dataset for qualitative and quantitative comparison purposes. Because most of the previous privacy-based data generation models used MNIST for study purposes to compare generated image-based data. This is the rst 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 datasets is reliability. So to ensure reliability and to defend against mode collapse we used W-GAN, which is famous for its high- delity data synthesis performance. In each iteration, we generated di erent Bernoulli or Gaussian noise depending on user's choice to preserve the diversity of the dataset. Observed from Table 3.1, our approach's performance in terms of CNN is much closer to real data. In the 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

Table 3.1: 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

according to Table 3.1. As observed from the gure, 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 the case of X-ray images, it also gained really good results with MLP: 74.4% accuracy. We used a highly regularized CNN model to train using our synthetic data and such regularized model also will help to make it free from membership inference attacks. Using such a simple ResNet18 model for X-ray images, it gained an accuracy of 76.245% using CNN on the synthetic image, which is close to 74.78% accuracy for real image (according to Table 3.1). In the case of MLP, it also gained 74.484% accuracy based on arti cial radiology data, which is really amazing and it is closer to the model's accuracy of 76% using real images. In our case, we used synthetic data in the training set and real data in the 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 3.1, the row for G-PATE is missing for the x-ray image because G-PATE is not applicable to our dataset type.

To analyze the impact of privacy parameters like noise multiplier we performed some experiments with varying noise level. In Figure 3.2, we showed our model's data

Figure 3.2: Normal patients (rst 2 columns) & pneumonia patients (last 2 columns) with noise multiplier 0.07 (1st, 3rd column) vs. noise multiplier 1.02 (2nd, 4th columns)

quality concerning the noise multiplier. The rst two columns indicate the standard patient images where the rst 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

Figure 3.3: Generated normal patients data with noise multiplier 0.07

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 Figure 3.3, which is developed with value of 10 and noise multiplier 0.07. We also displayed the generated pneumonia images in

Figure 3.4: Generated pneumonia patients data with noise multiplier 0.07

Figure 3.4. Observed from those two grid views, our model can di erentiate between normal and pneumonia patients based on semantic structure. We used a sampling rate of $\frac{1}{1000}$ and we considered a number of iterations to be 2000. According to the experiment in the worst case, our highest privacy budget for 24000 data and 2000 epochs is $3.194 \quad 10^{4}$.

Table 3.2: Privacy communication

Method	epsilon	delta	CT bytes
FedAVGGan	$9.99 \quad 10^{6}$	$1 \quad 10^{-5}$	$3.94 \quad 10^{7}$
Ours	$7.89 \quad 10^{2}$	$1 \quad 10^{-5}$	$1.95 \quad 10^{5}$

Previous approaches did not use high-resolution images for high stake domains, like { the medical radiology image classi cation task, so we had some trouble comparing with baselines. We used the scaled-down images and modi ed the di GAN [TFR22] 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 a noise multiplier of 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 Figure 3.5.

3.2.1 Federated learning based experiment

In the federated approach, it ensured a much more e cient communication cost than previous approaches. Communication cost indicates how many bytes it consumes to perform one generator step by transferring the gradient to the server. It takes fewer 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 3.2 shows that Fed AVG GAN's total" value was $9.99 \quad 10^{6}$ with CT bytes $3.94 \quad 10^{7}$. In contrast, in our approach," value was $7.89 \quad 10^{2}$ and CT bytes $1.95 \quad 10^{5}$. It has much higher gains in gradient communication in terms of CT bytes. Fed-AVG GAN cannot perform well with a noise multiplier value that is

Figure 3.5: Modi ed di GAN (top) vs. ours (bottom)

more than 0.1 whereas we have used 1.02 for the noise multiplier value and still our approach is able to generate quality data.

3.3 Summary

This chapter of the thesis presented a novel GAN structure that ensures di erential privacy while generating high- delity and reliable arti cial X-ray images. This architecture facilitates the generation of radiology data using both central and distributed processes. It' main feature is to introduce noise only to the generator part that uses real data while maintaining the integrity of the discriminator to produce high-quality images.

Chapter 4

Privacy-Preserving Learning via Data and Knowledge Distillation

In the current era of deep learning, computer vision, and image analysis have become ubiquitous across various sectors, ranging from government agencies and large corporations to small end devices, due to their ability to simplify people's lives. However, the widespread use of sensitive image data and the high memorization capacity of deep learning present signi cant privacy risks. Now, a simple Google search can yield numerous images of a person, and the knowledge that a speci c patient's record was utilized for training a model associated with a disease may reveal the patient's ailment, potentially leading to advanced attacks in the future. Furthermore, these unprotected models may also su er from poor generalization due to this over tting to train data. Previous state-of-the-art methods like di erential privacy (DP) and regularizer-based defenses compromised functionality, i.e., task accuracy, to preserve privacy. Such a trade-o raises concerns about the practicability of such defenses.

Other existing knowledge-transfer-based methods either reuse private data or require more public data, which could compromise privacy and may not be viable in certain domains. To address these challenges, where privacy is of utmost importance and utility cannot be compromised, we propose|in this chapter|a novel collaborative distillation approach that transfers the private model's knowledge based on a minimal amount of distilled synthetic data, leading to a compact private model in an end-to-end fashion. Empirically, our proposed method guarantees superior performance compared to most advanced models currently in use, increasing utility by almost 8%, 34%, and 6% for CIFAR-10, CIFAR-100, and MNIST, respectively. The utility resembles non-private counterparts almost closely while maintaining a respectable level of membership privacy leakage of 50-53.5%, despite employing a smaller model with 50% fewer parameters.

4.1 Method

Trained machine learning models demonstrate di erent behavior for data from other distributions. If the data is from a train set, they show a higher con dence score but o er lower con dence scores if it does not belong to the train set. The central concept of Membership privacy is to produce similar con dence scores and gradient values for members and non-members. To promote that indistinguishability, we can use the distilled data instead of the private training data while training the target student model so it can mitigate the spike in prediction's con dence value in the presence of seen private train data. Three types of networks are used in two phases: expert, teacher, and student. Experts and student models are used for data

distillation, and experts have the same architecture as the student model. Here, the teacher and reinitialized student are used for model distillation. Firstly, we use private data, a trained expert set of models trained on real data, and a student model trained on generated data to optimize the distilled data to develop a smaller amount of representative synthetic data that helps preserve the general pattern of the original dataset. It can generate similar performance while using real private data. Secondly, we train a large teacher model with private data and use it to generate soft labels of the distilled synthetic dataset. This data is used to train the target student model. As this distilled dataset size is smaller and balanced and the features are highly discriminative, this student training becomes more computationally e cient and fair regarding class balance. We also want our student model to be (almost half of the teacher model in terms of parameters) lightweight so that it does not over t the train data; Instead, their limited capacity and soft label help promote generalizability. So, this collaboration of data distillation and knowledge transfer ensures that the resultant model behaves similarly for members and non-members due to sanitized knowledge of the private data. The high-level process is illustrated in detail in Figure 4.1.

The steps are as follows: (i) Generate a small number of synthetic data S with which the student model $\hat{F}$ can be trained to have similar performance and parameters to a set of expert networks of same architecture when trained on the full private data. (ii) Train unprotected large capacity teacher model T on private data $(x; y)$. (iii) Use the trained teacher model T to compute soft labels $T(s)$ of distilled data S with higher temperatures to smooth out logit z_i's distribution. (iv) Then,

Figure 4.1: Transfer private teacher knowledge to a smaller student using distilled data

we train the reinitialized student model F using KL-divergence-based loss $\mathcal{L}_{KL}$ on distilled data and soft label $T(s)$ to predict $\hat{y}$ as $F(s)_{\frac{F}{S}}$ to transfer and sanitize the knowledge of the private model. We will talk about distilled data generation in Section 4.1.1 and how we transfer knowledge of private model via distilled synthetic data in Section 4.1.2.

4.1.1 Generating distilled data

We generate a smaller number of synthetic distilled data where the performance of the student model trained on such synthetic data will be similar to that of an ideal model trained on full private data. This data will be used later (in Section 4.1.2) for knowledge transfer to the student network. Initially, we train multiple expert networks that have similar architecture to our target student network on whole private

data using the regular cross-entropy loss like Eq. (4.1). We save the best expert network parameters for each time step interval during training and store these parameter snapshots as , indicating the upper bound of this k-class classi cation task on real data where c; p_c; y_c are class variable, predictions and true labels:

$$\text{Loss}_{\text{experts}} = \sum_{c=1}^{k} y_c \log(p_c) \tag{4.1}$$

In that case, we utilized a set of models to generate expert snapshots, which increases the generalizability of the distilled data set by capturing all the discriminative features possible. The best among saved snapshots will guide the optimization of distilled data, where the student model trained on generated distilled data will be used to test the performance of distilled data. This student trained on synthetic data will be encouraged to have similar parameters to ideal experts when trained on private data.

We consider these saved parameters $_t$ with time t as snapshots of the exemplary model behavior that the student model F should follow while optimizing generated distilled data S. From Eq. (4.2), we see that $_t$ indicates the expert time sequence parameters or trajectories that are learned using real private data. After saving the parameters of experts, we fetch a random expert parameter $_t$ from a random time step/epoch t from the saved list of expert parameters to initialize the student network F with that speci c model weight named as $\hat{}^F_t$ for time step t. After the student model is initialized at a randomly sampled time step t, it is trained for a de ned no of iterations N on distilled data where the synthetic data S is updated/optimized based on the parameter matching loss L_{traj} shown in Eq. (4.2). This process is illustrated in Figure 4.2(a):

$$L_{\text{traj}} = \frac{\| \hat{}^F_{t+N} \quad {}_{t+M} \|_2^2}{\| {}_t \quad {}_{t+M} \|_2^2} \tag{4.2}$$