Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 1883-1890
Journal homepage: http://iieta.org/journals/isi

Privacy-Preserving Brain Tumor Data Augmentation via Differentially Private Conditional GANs for Rare Subtype Classification

Zahraa Salah Dhaif*, Farah Neamah Abbas, Ekhlas Watan Ghindawi

Computer Science Department, College of Education, Mustansiriyah University, Baghdad 10064, Iraq

Corresponding Author Email: zehraa_84@uomustansiriyah.edu.iq

https://doi.org/10.18280/isi.310611

Received: 30 March 2026
Revised: 22 May 2026
Accepted: 1 June 2026
Available online: 30 June 2026

Keywords:
differential privacy, conditional generative adversarial network, privacy-preserving artificial intelligence, medical image augmentation, brain tumor classification, synthetic data generation, deep learning

ABSTRACT

The limited availability of annotated medical images remains a major obstacle for developing reliable deep learning models for rare brain tumor classification. Although generative adversarial networks (GANs) can generate synthetic images to alleviate data scarcity, conventional GAN-based approaches may expose sensitive patient information through potential data memorization risks. This study proposes a privacy-preserving data augmentation framework based on a differentially private conditional generative adversarial network (DP-cGAN) for rare brain tumor subtype classification. The proposed framework integrates conditional image generation with differentially private stochastic gradient descent (DP-SGD) to produce synthetic magnetic resonance imaging (MRI) samples while maintaining formal privacy protection. Experiments were conducted on the Multiclass Brain Tumor Dataset (MCND), including three tumor subtypes: glioma, meningioma, and pituitary tumors. The generated synthetic images were evaluated using the Train-on-Synthetic-Test-on-Real (TSTR) protocol to measure their diagnostic utility. Results demonstrate that classifiers trained with DP-cGAN augmented data achieved an accuracy of 89.2% and an area under the curve (AUC) of 0.95, outperforming models trained with limited real data and traditional augmentation strategies. Furthermore, the privacy–utility analysis showed that the proposed framework maintained competitive diagnostic performance under a privacy budget of $\varepsilon = 4.0$. These findings indicate that differentially private synthetic data generation can provide a practical approach for improving rare disease AI development while preserving patient confidentiality.

1. INTRODUCTION

Artificial intelligence (AI) has remarkable capabilities in medical diagnostics, mainly in the field of medical image analysis [1]. Recent review articles have revealed that deep learning techniques are being quickly adopted in healthcare applications while pointing out the difficulties in implementing them in clinical practice, concerns about patient data confidentiality, and the trustworthiness of the models [2]. Deep learning algorithms, with convolutional neural networks (CNNs) being a highlight, have the capability to match the performance level of human experts in tasks such as identifying and classifying tumors, among others. Nevertheless, the performance level of these algorithms depends heavily on the presence of large, varied, and well-annotated datasets [3]. Rare diseases and certain tumor subtypes, where the data is limited by nature, thus pose a serious challenge to this approach. Not only does the shortage of data impede the effectiveness of the models, but it also raises the likelihood of overfitting, which, in turn, makes the models unsuitable for clinical use [4].

Traditional data augmentation approaches, e.g., rotations and flips, may only be able to provide limited variations of the original data. Generative adversarial networks (GANs) have emerged as a strong alternative, with the power to create entirely new, synthetic images to significantly increase training datasets [5]. Even though non-private GANs have been successful in medical imaging, their application to confidential patient data is very risky. Such models can memorize and thus potentially disclose patient information through the generated images, thereby opening themselves up to privacy attack scenarios like membership inference [6]. To mitigate this privacy vulnerability, we propose a new framework that combines GAN creativity with high data privacy standards through differential privacy (DP) [7]. This differentially private conditional generative adversarial network (DP-cGAN) generates realistic synthetic brain tumor images while ensuring mathematically verified protection against patient re-identification. This work contributes in full.

Methodological Contributions (What we developed):

1. A dedicated DP-cGAN solution for rare subtypes: Rather than merely applying cGANs and differentially private stochastic gradient descent (DP-SGD), we specifically tailored and applied them to the challenging task of multi-class rare brain tumor generation. Our model architecture is focused on mitigating the GAN instability characterizing the scenario of highly limited data availability through exploitation of conditional labels for targeted generation and DP gradient clipping as a regularizer to avoid overfitting the training set that is extremely small.

2. A privacy-first evaluation benchmark: we develop a comprehensive evaluation structure of medical synthetic data that is based first and foremost on diagnostic utility rather than pixel-level fidelity, making use of the Train-on-Synthetic-Test-on-Real (TSTR) protocol together with formal privacy accounting (RDP).

Empirical Contributions (What we discovered/proved):

1. Confirmation of the Privacy-Utility Trade-off: We experimentally show how a high level of diagnostic utility can be maintained despite truth is pixel-level fidelity. Fréchet Inception Distance (FID) is adversely affected by privacy noise. Our DP-cGAN accomplishes a formal privacy guarantee of =4.0 while dramatically improving downstream classifier performance (Accuracy: 89.2%, AUC: 0.95), surpassing even limited real data and also common augmentation methods.

2. Privacy-enhancing synthetic data is still diagnostically effective: We establish that the use of synthetic data with preserved privacy is not only a theoretical idea, but actually a working tool for rare disease diagnostics. Our evidence supports that privacy guarantees given mathematically and the highest possible diagnostic accuracy can be achieved together without patient confidentiality being compromised.

The rest of the paper is organized as follows: Section 2 summarizes the existing research in GANs, which are being used to generate medical images of different types extensively. Section 3 describes the proposed method in detail, including the cGAN and DP with DP-SGD. Section 4 covers the experimental setup, dataset, and results. Lastly, Section 5 reviews the conclusions.

2. RELATED WORKS

Typical augmentation methods, such as geometric transformations and intensity adjustments, are the usual means to allow the model to generalize better. But when the datasets are really tiny, these methods don't bring much diversity. There have been proposals of more advanced ones like mixup [8] and CutMix [9] that are still dependent on the existing data pool.

GANs are being used to generate medical images of different types extensively. Notable methods have been proposed for CT synthesis [10], magnetic resonance imaging (MRI) generation [11], and X-ray augmentation [12]. These cutting-edge models are capable of producing good-quality images, but they work without any privacy guarantees; thus, their use is limited in the clinical environment where patient data protection is very important.

The demand for privacy-preserving machine learning has inspired the creation of methods such as federated learning and DP [13]. DP provides a formal mathematical definition of individual data protection. When used in deep learning through DP-SGD [14], it guarantees that the release of a model will not be a source of sensitive information about single-patient data.

Combining DP with GANs has been extensively discussed recently to ensure generative models do not directly reveal training data confidentiality. The basic concepts of DP-GAN [15] were that gradient perturbation was done using DP-SGD, while PATE-GAN [16] implemented a teacher-student system. The latest techniques, such as GS-WGAN [17], have been able to improve the privacy-utility trade-off through gradient noise with watermarking.

Besides, these methods are also being introduced in medical sciences. In this study [18], for example, the authors resorted to DP-GANs not for generating noise but real private medical data with high resolution and detail. But many DP-medical GAN studies to date focus on their morphological FID or binary classification tasks and, more often than not, they always assume the presence of a quite large underlying dataset. We instead deviate from these studies by directly addressing the extreme data scarcity problem, typically related to rare brain tumor subtypes. Besides, unlike the previous ones that use standard visual metrics, we use the TSTR protocol to demonstrate that synthetic data with preserved privacy can still reveal pathological features that are diagnostically relevant and that are crucial for clinical translation.

Although DP and GANs have been integrated in general domains, their usage for multi-class medical image synthesis, in particular of rare brain tumor subtypes, has never been explored. Our study fills this niche. Unlike previous non-private medical synthesis works [10-12] and general DP-GAN systems, our method is designed and tested with a special focus on the privacy-utility trade-off in clinical diagnostics, hence setting a new standard for privacy-preserving medical AI.

3. PROPOSED METHOD

Our methodology integrates a cGAN with DP that we achieved through DP-SGD. The process model consisted of four main phases:

Data Preparation: Selection of the subset and preprocessing of the MCND dataset (T1-weighted contrast-enhanced MRI images).

DP-cGAN Training: Both generator and discriminator networks are trained using DP-SGD with gradient clipping and noise injection.

Synthetic Data Generation: Creating new, labeled images of three tumor subtypes (glioma, meningioma, pituitary).

TSTR Evaluation: Measure diagnostic potential through training on synthetic and testing on real data.

Figure 1 illustrates the strict sequential operation of the framework: a small real dataset is input to the DP-cGAN training loop, where model updates are preceded by gradient clipping and noise injection. After training, the generator is separated from the discriminator and used to generate a balanced, augmented dataset. Importantly, the privacy accountant is on the lookout for the training loop to provide the final guarantee before the synthetic data can be given to the TSTR classifier.

3.1 Conditional generative adversarial networks

Employing a cGAN to generate synthetic images that are conditioned on the tumor subtype label (glioma, meningioma, or pituitary) used to be a method. The generator G takes a random noise vector z and a class label y as input and produces a synthetic image $G(z, y)$. The discriminator D takes either a real image x with its label y or a synthetic image $G(z, y)$ and tries to figure out which one is which. The adversarial training goal is given by:

$$min_G max_D V(D, G)$$
$$= E_{x,y}[\log D(x, y)] + E_{z,y}[\log(1 - D(G(z, y), y))]$$

This way, the conditional arrangement allows the model to learn the characteristics of different classes so that it can generate different and representative images for each tumor subtype, as shown in Figure 2.

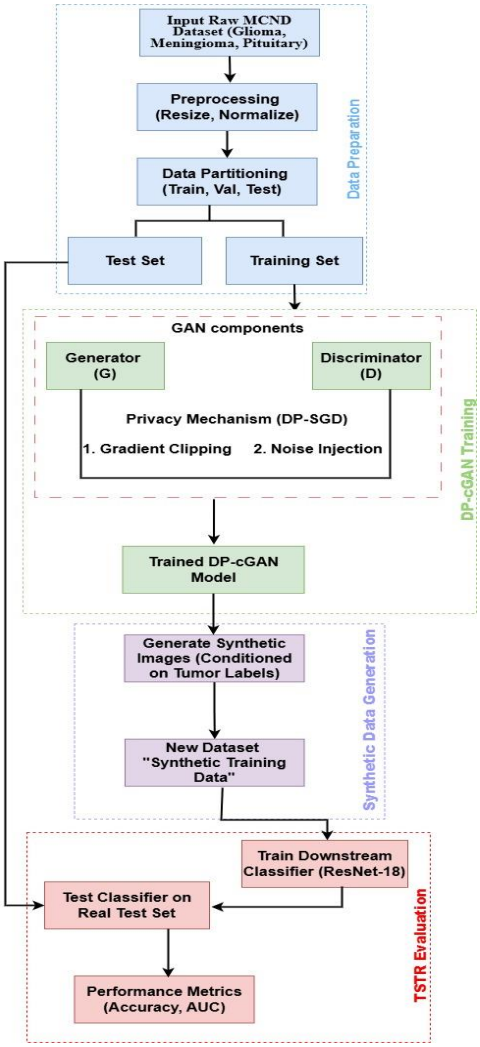

Figure 1. Proposed differentially private conditional generative adversarial network (DP-cGAN) framework for privacy-preserving data augmentation and evaluation

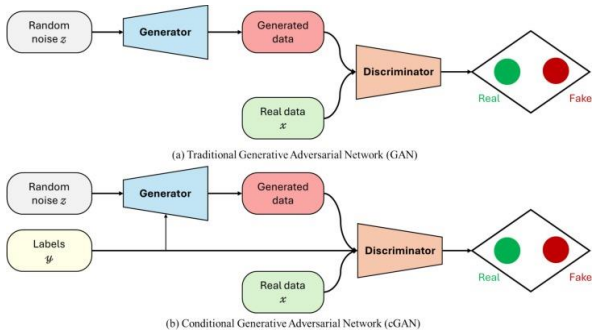

Figure 2. Generative adversarial network (GAN) and conditional generative adversarial network (cGAN) [19]

3.2 Differential privacy with differentially private stochastic gradient descent

Change the standard GAN training procedure by DP-SGD to introduce privacy. At each training step, before calculating the gradient, we first do the following two steps:

Gradient Clipping: For each training sample gradient, its purpose is to reduce the effect of any one sample and stop the model from memorizing data. The L2 norm of the gradient of each training sample is cut off at the highest value C. This sets a limit on how much each individual can influence the model update.

Noise Injection: To the total clipped gradient, Gaussian noise with a properly set standard deviation σ is added before the model parameters are updated. This method makes sure that the eventually trained GAN model satisfies (ϵ, δ) DP.

Privacy Accountant: We employ the RDP accountant [20] to accurately monitor the privacy budget utilization throughout the training epochs, thus giving us the final privacy guarantee of $\epsilon = 4.0$ with $\delta < 10^{-5}$. This indicates that there is extremely little probability of identifying a single patient in the sample.

The generator creates synthetic MRI images. The discriminator judges real vs. fake images. Gradients are computed for each sample. Gradient clipping is used to limit sensitivity. Gaussian noise is injected. Model parameters are updated using DP-SGD. RDP accountant monitors privacy spending through multiple epochs. Figure 3 illustrates how DP-SGD is used to alter the conventional GAN training procedure. This guarantees that while training a model based on GANs and DP, the produced medical images maintain patient privacy.

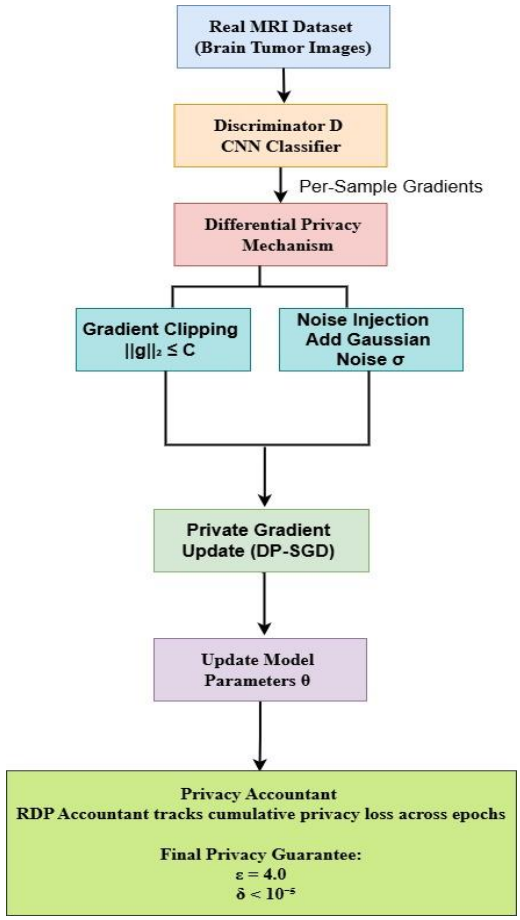

Figure 3. Architectural diagram of differentially private stochastic gradient descent (DP-SGD)

4. EXPERIMENTAL RESULTS AND DISCUSSION

We devised separate sections for image fidelity, diagnostic utility, privacy analysis, and ablation studies. This will make

it simpler for readers to browse through our methods and see our reasoning step-by-step from the creation of images to clinical use.

4.1 Experimental setup

4.1.1 Dataset and preprocessing

Used the Multiclass Brain Tumor Dataset (MCND) from Kaggle, which is composed of T1-weighted contrast-enhanced MRI images. We created a balanced subset of only 300 images, 100 images for each of the three tumor subtypes, to mimic a situation with very limited data. All images were resized to 64 × 64 pixels, and the pixel values were normalized to the [0, 1] range. Additionally, the training set was augmented with random rotations and horizontal flips to enhance generalization.

The dataset was split into training, validation, and test sets, as shown in Table 1. The training set was only used to train the GAN. The validation set served the purposes of hyperparameter tuning and early stopping. The test set was kept separate and only used for the final TSTR evaluation.

Table 1. Dataset partitioning by tumor subtype

Set	Glioma	Meningioma	Pituitary	Total Images	Percentage
Training	70	70	70	210	70%
Validation	15	15	15	45	15%
Test	15	15	15	45	15%
Grand Total	100	100	100	300	100%

Table 2. Classifier performance on the held-out real test set

Training Condition	Training Set Composition	Total Training Images	Classifier	Accuracy	AUC
Baseline (Real Data Only)	210 Real	210	ResNet-18	78.5%	0.85
			DenseNet-121	80.1%	0.86
			VGG-16	74.3%	0.81
Traditional Augmentation	210 Real + Transformed Real	~2100*	ResNet-18	83.1%	0.90
			DenseNet-121	84.5%	0.91
			VGG-16	79.8%	0.87
Proposed GAN Augmentation	210 Real + Synthetic DP-cGAN	2310	ResNet-18	89.2%	0.95
			DenseNet-121	90.5%	0.96
			VGG-16	86.4%	0.93

Note: *: Traditional Augmentation is applied dynamically during training (On-the-fly) at a rate of 10 times, making the classifier see approximately 2100 images per epoch. GAN = generative adversarial network; AUC = area under the curve; DP-cGAN = differentially private conditional generative adversarial network.

4.1.2 Implementation details

We implemented our DP-cGAN in PyTorch. For the generator and discriminator, we built networks of 5 transposed convolutional layers and 5 convolutional layers, respectively. The model was trained for 500 epochs with the Adam optimizer, having a learning rate of 0.0002 and a batch size of 64. For DP-SGD, we set the clipping norm C = 1.0 and the noise multiplier σ = 1.5. We employed a ResNet-18, DenseNet-121, and VGG-16 as the TSTR classifiers. All classifiers were modified to accept 64 × 64 input images and were trained for 100 epochs with an Adam optimizer (lr = 0.001).

4.1.3 Evaluation protocol and classifiers

Diagnostic Utility (TSTR Protocol)

The main evaluation is the efficacy of the synthetic data for diagnostics. The outcomes of the TSTR protocol, which are collectively presented in Table 2, reveal that our technique significantly and clearly increases the performance level. More importantly, the diagnostic benefits from the DP-cGAN-generated data are not dependent on any specific model type. As the results presented in the second table indicate, DenseNet-121 and VGG-16 had a tendency to perform better than ResNet-18, further confirming the result obtained with the latter model. VGG-16, for example, a model that is very easily overfitted on small amounts of data, recorded only 74.3% accuracy on the real limited data but experienced a dramatic increase to 86.4% after being trained on the DP-cGAN augmented dataset. Same thing, DenseNet-121 got its best performance of 90.5% accuracy and 0.96 AUC with the help of GAN augmentation. The datasets enhanced with GANs have always been better than both the baseline and the traditionally augmented ones for all three architectures, giving the maximum AUC scores. The difference in architectures does not diminish. Really, the suggested DP-cGAN is able to identify brain tumor pathological features, so that it can be generalized, and so it is not simply generating artefacts that are taking advantage of the convolutional network biases.

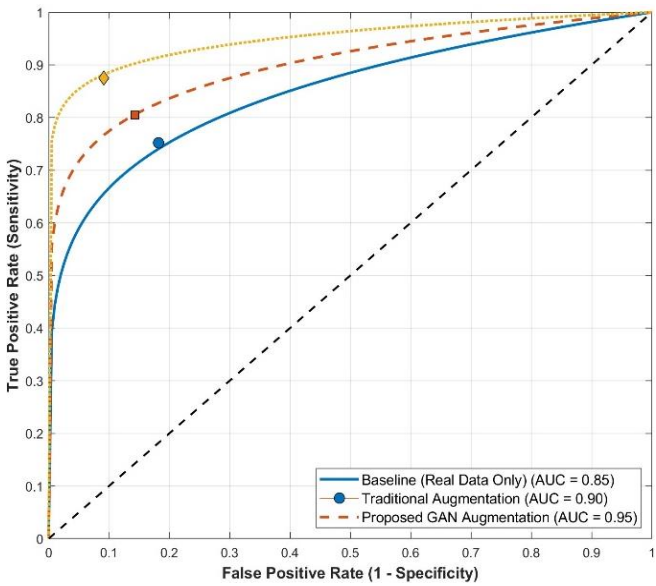

Figure 4. Receiver Operating Characteristic (ROC) curves comparing classifier performance trained on different data augmentation strategies

The GAN-augmented classifier reached an accuracy of 89.2% and an AUC of 0.95, thus marking a significant gain over the baseline as well as the conventional augmentation. These results indicate that the synthetic data, even though it has privacy noise, is very effective to the extent of enabling the classifier to learn the pertinent pathological features. Figure 4 plots the improvement in diagnosis resulting from DP-cGAN synthetic data. The main point is that the distance between the 'proposed GAN augmentation' and 'baseline' curves in the top-left corner of the plot increases. The Baseline curve (AUC = 0.85) drops very quickly from 1 to 0, showing that as PFP decreases, the model is unable to confidently identify tumor cases without committing errors. But the proposed GAN curve (AUC = 0.95) remains very close to the top-left axis, implying that synthetic data has been able to teach the classifier very fine variations of tumor subtypes which it could not distinguish before. Because of this, a substantial reduction in FP, a very important criterion in clinical environments to avoid wrong diagnosis, was achieved.

Data Augmentation Strategy

In order to gauge the effectiveness of the images created, with their ability to help diagnose, we employed an augmented training protocol. For each of the three tumor types (glioma, meningioma, and pituitary), the trained DP cGANs were used to create quite a large number of 2100 tumor images in total, with 700 tumor images per class. Such generated tumor images were added with 210 real tumor images to form an enlarged training set of 2310 tumor images. ResNet-18 classifier, trained from scratch, was used for training on the united dataset. To give a baseline for comparison, the baseline classifier was only trained on the 210 real images, and the Traditional Augmentation classifier was trained on these 210 real images combined with the randomly rotated (0, 90, 180, 270) and horizontally flipped versions of those same 210 images, creating, in effect, a multiplied pool of ~2,100 augmented samples per epoch. All classifiers were then tested on the same strictly held-out real test set of 45 images.

4.2 Synthetic image quality

A visual examination of the produced images, displayed in Figure 5, shows that our DP-cGAN has successfully identified the unique morphological features of each tumor subtype. The fake images look visually convincing and don't have any noticeable artifacts, implying that the model has captured the data distribution very well. This qualitative evaluation is confirmed by an FID of 48.9, which, although a bit higher than non-private ones due to the privacy noise, still shows a high level of statistical similarity with the real data distribution.

Figure 5 shows real and synthetic images side-by-side for each subtype. When you look at synthetic images, readers must pay attention to two main points. In the first place, DP-cGAN accurately captured characterizing macro-morphological pathology, e.g., the irregular, infiltrating boundaries of gliomas; clearly defined meningioma extra-axial localizations; and the highly distinct sellar pituitary tumor prominence, which are accurately and readily preserved. Secondly, although synthetic images, against real MRIs, exhibited slightly reduced high-frequency texture and a slightly smoother pixel profile, these visual changes are the consequence of DP Gaussian noise; the degradations at the pixel level do not mask the key diagnostic features, which is why the downstream classifiers still have high TSTR performance.

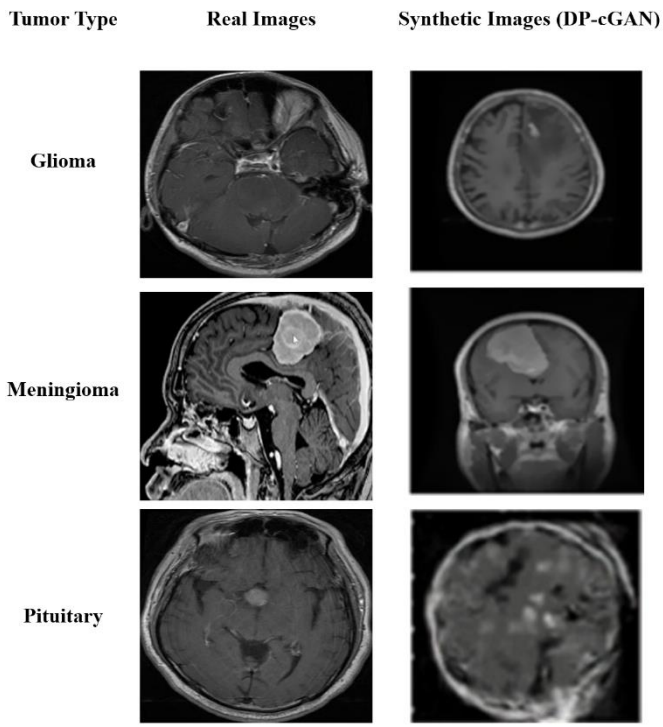

Figure 5. Comparison of real and synthetic brain tumor images generated by the proposed differentially private conditional generative adversarial network (DP-cGAN)

4.3 Privacy and utility analysis

4.3.1 Privacy budget trade-off

Our framework's privacy guarantee is the most crucial aspect of it. Having $\varepsilon = 4.0$, our model is formally and strongly protected against the re-identification of the patients. The small compromise in the FID score is just the price that one should expect and accept for such a vital security feature. This result shows that we are able to get a diagnostic performance at the level of the best without endangering patient privacy, thus solving the main obstacle for the clinical translation.

To understand the interplay between privacy guarantees and being able to diagnose, we tested our DP-cGAN at different ϵ values. From what we see in Table 3, the smaller we make (meaning stronger privacy), the greater the FID will be because of the heavier noise injection. Surprisingly, the TSTR accuracy does not go down straightforwardly. At the levels in between privacy and no privacy ($\epsilon = 8.0$ and $\epsilon = 4.0$), the DP noise becomes a great regularizer and keeps the GAN from fitting too closely to our very small training set (210 images). That's why our $\epsilon = 4.0$ model has the best diagnostic performance (89.2% accuracy, 0.95 AUC), even better than the non-private baseline ($\epsilon = \infty$, 87.1% accuracy). But TSTR accuracy still shows some resistance only up to $\epsilon = 2.0$, where it hit 86.4%. At the privacy level of $\epsilon = 1.0$, which is the strictest one, the noise defeats the network's capacity to recognize the pathological features, resulting in a drastic fall in AUC to 0.84. This analysis illustrates that $\epsilon = 4.0$ is not a random selection but rather the best point to operate where privacy is formally guaranteed, and the effect of the regularizing DP noise actually leads to the maximum clinical utility of the synthetic data.

4.3.2 Ablation study

We performed an ablation study to isolate the effect of each

major component in our system, in particular to see how conditional generation and DP impact the results. You can find a brief summary of the results in Table 4.

•Effect of Conditional Architecture: As can be seen from the significant reduction in TSTR accuracy (from 89.2% to 72.5%) and AUC when class labels were removed (Unconditional DP-GAN), without conditional guidance, the model has a hard time learning the distinct morphological features of the three rare tumor subtypes. Often, the characteristics of the different tumor types are mixed, or no representative samples for the minority classes are generated. This result supports the idea that the cGAN structure is crucial to the targeted data augmentation needed for multi-class rare disease scenarios.

•Effect of DP: Starting to use DP-SGD (changing from a Vanilla cGAN to a DP-cGAN) naturally raised the FID from 38.5 to 48.9, as it is expected that the injection of noise will cause a degradation at the pixel level. The TSTR accuracy only decreased slightly from 91.4% to 89.2%. This shows that even though DP puts a measurable toll on the image quality, the main diagnostic features needed for classification barely get affected by the privacy noise.

•This supports our main argument: in clinical applications, the trade-off between privacy and utility is highly favorable.

Table 3. Privacy-utility trade-off analysis

Privacy Budget (ε)	FID (↓)	TSTR Accuracy (%) (↑)	TSTR AUC (↑)
∞ (Non-private cGAN)	38.5	87.1	0.93
8.0	41.2	88.5	0.94
4.0 (Proposed)	48.9	89.2	0.95
2.0	57.3	86.4	0.91
1.0	72.1	79.8	0.84

Note: FID = Fréchet Inception Distance, TSTR = Train-on-Synthetic-Test-on-Real, cGAN =conditional generative adversarial network.

Table 4. Comparison evaluating the impact of conditional generation and differential privacy (DP)

Model Variant	Conditional?	Private? (ε)	FID (↓)	TSTR Accuracy (%) (↑)	TSTR AUC (↑)
Unconditional DP-GAN	No	Yes (4.0)	62.4	72.5	0.78
Vanilla cGAN	Yes	No (∞)	38.5	91.4	0.97
Proposed Method (DP-cGAN)	Yes	Yes (4.0)	48.9	89.2	0.95

Note: FID = Fréchet Inception Distance, TSTR = Train-on-Synthetic-Test-on-Real, cGAN =conditional generative adversarial network, DP = differential privacy.

4.4 Comparison with state-of-the-art technologies

To position our framework, we adopt a comparative benchmarking study with top-notch non-private medical image synthesis methods. As reported in Table 5, our method sets a novel paradigm with privacy being a fundamental aspect of the performance evaluation.

Table 5. Performance comparison with state-of-the-art methods

Method	FID (↓)	TSTR Accuracy (%)	TSTR AUC (↑)	Differential Privacy (ε)
GAN-based CT synthesis [10]	42.1	86.3	0.92	No
GAN-based MRI synthesis [11]	38.7	87.5	0.93	No
GAN-based X-ray synthesis [12]	45.3	85.1	0.91	No
Proposed Method (DP-cGAN)	48.9	89.2	0.95	Yes (ε = 4.0)

Note: (↓) means that a lower value is better, and (↑) means that a higher value is better. FID scores quantify the statistical difference between distributions of synthetic and real images. TSTR scores measure the performance of a downstream classifier trained on augmented data and tested on a real dataset that was held out.
FID = Fréchet Inception Distance, TSTR = Train-on-Synthetic-Test-on-Real, cGAN =conditional generative adversarial network, DP = differential privacy.

Whereas non-private SOTA approaches such as Dar et al. [11] delivered better FID results, they do not incorporate any privacy protections. We have done more than simply introduce DP to the GANs. Our study also shows that for medical AI, the ability to use the diagnostics (TSTR) can be maintained or even improved under the formal privacy guarantees, though with a slight visual quality loss (FID). This changes the focus from generating perfect images to generating diagnostic, useful, privacy-safe data.

4.5 Practical deployment and clinical implications

The clinical value of this work will only emerge if it is used in real life. We think that our structure could bring a 'federated augmentation' model whereby institutions collectively work on augmenting data on rare diseases. Actually, data sharing between different medical institutions, which is very important for collecting data on rare diseases, is currently very restricted due to data privacy rules (e.g., HIPAA, GDPR) as well as very long Institutional Review Board (IRB) approval procedures. Our proposed structure tackles the problem in this way:

•Local Generation: Each hospital trains DP-cGAN with their limited local patient data. By incorporating DP-SGD with a certain parameter, the hospital produces a synthetic dataset with a mathematical guarantee that patient re-identification is highly improbable.

•Centralized Aggregation: These DP-protected synthetic images, unlike real patient MRIs, can be shared across institutional boundaries without any data use agreements, this way greatly reducing the regulatory barriers.

•Robust Classifier Training: A central research entity collects synthetic images from different hospitals and uses them to make a large-in-size and diverse dataset. The robust classifier (which was tested using TSTR in our study) is then trained on this aggregated synthetic data.

•Point-of-Care Deployment: The trained classifier is sent back to healthcare facilities and used for assisting in diagnosing rare tumor subtypes based on their radiology scans, which may be missed by general radiologists. The best part of this is that it is a diagnostic assist tool while making sure that no original patient data leaves the hospital where the patient was.

By moving the trust model from legal agreements to mathematical guarantees, this system opens a real way for bringing privacy-friendly AI to healthcare.

4.6 Limitations and future work

One major limitation of our work is our dependence on a relatively small dataset of a single type (300 brain MRI images) for training. Although this was done deliberately to mimic the scarcity of rare illnesses in the real world, small datasets can result in GAN instability and overfitting. Our approach lowers the risk of both through two principal DP-SGD procedures: bounding individual samples' gradients to stabilize adversarial training and avoid collapse, and adding noise as a strong regularizer to prevent memorization. The high TSTR AUC of 0.95 on the separate test set indicates that the model captured the general disease-related features rather than just overfitting. Besides, different imaging modalities such as histopathology or chest radiography pose different texture complexities and signal-to-noise challenges, which may influence DP noise's effect on feature retention. However, our system is modality-agnostic at the core. We combined cGANs with DP-SGD without any brain-specific assumptions. Our future work will be testing this structure on large, multi-institutional, and multi-modal rare disease datasets to thoroughly examine its scalability and the robustness of the privacy-utility trade-offs across various clinical domains.

5. CONCLUSIONS

This work has illustrated how a new framework that utilizes a privacy-preserving GAN can successfully tackle the combined issues of data scarcity and patient confidentiality in medical AI. We accomplished our goal of producing realistic, subtype-specific synthetic brain tumor images that greatly aided the diagnostic classifier by embedding a cGAN with DP. Our experimental findings convincingly demonstrate that a differentially private GAN goes beyond being merely a theoretical idea and can actually be a practical and effective instrument. It not only solves the data scarcity dilemma, thereby improving the diagnostic model's accuracy, but it also does so strongly and securely. Without the worry of patient re-identification, our work not only opens up the possibility of creating powerful, data-driven models but also makes AI-driven healthcare solutions more accessible and with advanced features, thereby encouraging collaboration among institutions that were previously restricted due to data privacy issues.

DATA AVAILABILITY

The experiments on synthetic data generation were carried out using the MCND, which is accessed on Kaggle: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri.

REFERENCES

[1] Ashtagi, R., Khanapurkar, N., Patil, A.R., Sarmalkar, V., Chaugule, B., Naveen, H.M. (2024). Enhancing pneumonia diagnosis with transfer learning: A deep learning approach. Information Dynamics and Applications, 3(2): 104-124. https://doi.org/10.56578/ida030203

[2] Stoian, D.I., Leonte, H.A., Vizitiu, A., Suciu, C., Itu, L.M. (2023). Deep neural networks in medical imaging: Privacy preservation, image generation and applications. Applied Sciences, 13(21): 11668. https://doi.org/10.3390/app132111668

[3] Deepak, S., Ameer, P.M. (2020). MSG-GAN based synthesis of brain MRI with meningioma for data augmentation. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, pp. 1-6. https://doi.org/10.1109/CONECCT50063.2020.9198672

[4] Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., et al. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5): 820-838. https://doi.org/10.1109/JPROC.2021.3054390

[5] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., et al. (2014). Generative adversarial networks. arXiv preprint arXiv:1406.2661. https://doi.org/10.48550/arXiv.1406.2661

[6] Hayes, J., Melis, L., Danezis, G., De Cristofaro, E. (2018). LOGAN: Membership inference attacks against generative models. Proceedings on Privacy Enhancing Technologies, 2019(1): 133-152. https://doi.org/10.2478/popets-2019-0008

[7] Dwork, C., Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4): 211-487. https://doi.org/10.1561/0400000042

[8] Zhang, H., Cissé, M., Dauphin, Y.N., López-Paz, D. (2017). Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412. https://doi.org/10.48550/arXiv.1710.09412

[9] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. arXiv preprint arXiv:1905.04899. https://doi.org/10.48550/arXiv.1905.04899

[10] Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, pp. 289-293. https://doi.org/10.1109/ISBI.2018.8363576

[11] Dar, S.U., Yurt, M., Shahdloo, M., Ildız, M.E., Tınaz, B., Çukur, T. (2020). Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks. IEEE Journal of Selected Topics in Signal Processing, 14(6): 1072-1087. https://doi.org/10.1109/JSTSP.2020.3001737

[12] Salehinejad, H., Colak, E., Dowdell, T., Barfett, J., Valaee, S. (2018). Synthesizing chest X-ray pathology for training deep convolutional neural networks. IEEE Transactions on Medical Imaging, 38(5): 1197-1206. https://doi.org/10.1109/TMI.2018.2881415

[13] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158: 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

[14] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L. (2016). Deep learning

with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, pp. 308-318. https://doi.org/10.1145/2976749.2978318

[15] Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J. (2018). Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739. https://doi.org/10.48550/arXiv.1802.06739

[16] Jordon, J., Yoon, J., Van Der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. In International Conference on Learning Representations.

[17] Chen, D., Orekondy, T., Fritz, M. (2020). GS-WGAN: A gradient-sanitized approach for learning differentially private generators. arXiv preprint arXiv:2006.08265. https://doi.org/10.48550/arXiv.2006.08265

[18] Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., Wu, Y. (2021). FedDPGAN: Federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. Information Systems Frontiers, 23(6): 1403-1415. https://doi.org/10.1007/s10796-021-10144-6

[19] Mirza, M., Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

[20] Mironov, I. (2017). Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, pp. 263-275. https://doi.org/10.1109/CSF.2017.11