Differential privacy framework for generating synthetic fMRI data with generative adversarial networks

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Data Analytics
December 2024
Hiba Daafane

UNIVERSITY OF TURKU
Department of Computing

HIBA DAAFANE: Differential privacy framework for generating synthetic fMRI data with generative adversarial networks

Master of Science (Tech) Thesis, 61 p.
Data Analytics
December 2024

Neuroimaging, particularly functional magnetic resonance imaging (fMRI), is one of the most significant tools of medical and cognitive research, as access to high quality neuroimaging datasets provides unparalleled insights into brain activity and functions. Its combination with artificial intelligence applications has increased its scope even further, bringing into view advanced diagnostic tools and predictive models for neurological and psychological disorders. However, the sensitive nature of such data, coupled with strict privacy regulations, significantly limits its accessibility and hinders collaborative research efforts.

Synthetic data has emerged as a valuable tool for generating artificial datasets that replicate the statistical properties of sensitive datasets with a reduced risk of privacy breaches. This thesis takes this as a starting point and builds upon the work done by Zheng et al., which utilized Generative Adversarial Networks (GANs) to generate synthetic task-conditioned fMRI images, and integrated Differential Privacy (DP) into the technical approach as a means to introduce a quantifiable measure of privacy all while preserving utility for downstream machine learning tasks. In this work, DP was integrated into the ICW-fMRI-GAN, using the Opacus privacy engine, and the research investigated three core challenges: the impact of DP on synthetic data sample quality evaluated through the Inception Score (IS), its effect on model performance and classification accuracy in predicting cognitive tasks from the images, and the degree of privacy protection ensured.

The experiments conducted in this work involve two medical institutions of varying sizes and resources, where a DP-wise access protocol is proposed as a potential solution for effective data sharing and research collaboration. The results demonstrated that the use of a combination of real and DP synthetic data achieves a competitive level of predictive accuracy while offering a fair amount of privacy guarantees. The work also underscores the need for future research to refine DP mechanisms for high dimensional data, such as brain images, and to develop synthetic datasets that are capable of maintaining sufficient utility while preserving patients privacy.

Keywords: Neuroimaging, fMRI, Synthetic Data Generation, Differential privacy, Generative Adversarial Networks, Data Collaboration

Contents

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
ASD	Autism Spectrum Disorder
BD	Bipolar Disorder
BOLD	Blood Oxygen Level Dependent
cGANs	Conditional Generative Adversarial Networks
CIFAR	Canadian Institute for Advanced Research
CNNs	Convolutional Neural Networks
DNN	Deep Neural Network
DMN	Default Mode Network
DP	Differential Privacy
EMD	Earth Mover's Distance
FiBI	Finnish Biomedical Imaging Node
fMRI	functional magnetic resonance imaging
GANs	Generative Adversarial Networks

GDPR	General Data Protection Regulation
GP	Gradient Penalty
ICW-GAN	Improved Conditional Wasserstein Generative Adversarial Network
IS	Inception Score
JS	Jensen-Shannon
KL	Kullback-Leibler
MDD	Major Depressive Disorder
ML	Machine Learning
NIST	National Institute of Standards and Technology
PD	Parkinson's Disease
RNNs	Recurrent Neural Networks
SDG	Synthetic Data Generation
SVM	Support Vector Machine
VAEs	Variational Autoencoders
WGAN	Wasserstein Generative Adversarial Network

List of Figures

List of Tables

1 Introduction

1.1 Research background and significance

In the era of explosive data growth, Machine learning, and data analysis techniques have gained significant attention for their ability to derive meaningful insights from large datasets, and across different fields. As a result of this exponential growth of data, various forms of large-scale datasets are being made available, including sensor data, social media feeds, transactional logs, and increasingly healthcare data. Moreover, the integration of Artificial Intelligence and Machine learning algorithms with medical data has become a valuable resource for advancing health care and contributing to advances in diagnostics, treatment plans, early onset of diseases and disorders, as well as an overall better understanding of human health and an improvement in the efficiency and accuracy of healthcare delivery.

A large subcategory, however of this type of data that has gained a lot of attention is neuroimaging data, particularly fMRI, or functional Magnetic Resonance Imaging, which is a non-invasive imaging technique that measures brain activity through the detection of changes in oxygenation and blood flow, that is to say, when neurons in a given area of the brain become active, they consume more oxygen, triggering an increase in oxygen delivery initiated by the nearby blood vessels. These signals are referred to as Blood Oxygen Level Dependent (BOLD), which are captured by fMRI and are used to create detailed spatial maps of brain activity and provide

insights into how different areas of the brain interact during specific tasks or in resting states [1] [2]. Neuroimaging data, such as that from fMRI, is extremely useful in the diagnosis and treatment of neurological disorders, brain mapping, and cognitive neuroscience research.

1.1.1 Neuroimaging and the role of fMRI in diagnostic approaches

The human brain is undoubtedly the most complex machine mankind has ever encountered, and its mechanisms and functionalities still challenge our understanding and remain some of the greatest scientific mysteries. As a result, neuroimaging data has become increasingly important in both clinical and research settings, as researchers are constantly aiming to get a better understanding of this 'black box' we call the brain, and constant efforts are being made to comprehend the underlying mechanisms behind several neurological diseases. For example, research surrounding the diagnostic criteria for Alzheimer's disease [3] was able to demonstrate that neuroimaging techniques, specifically structural imaging, are able to identify unique biomarkers that are crucial for early diagnosis. Similar studies [4] showed that structural imaging alongside other markers can help identify patterns of abnormal brain activity or changes in brain networks that may signal the onset of the disease prior to the appearance of cognitive symptoms and dementia become apparent. Additionally, studies such as the work conducted by (Tessitore et al., 2012) [5] on Parkinson's Disease proved that fMRI can help detect subtle changes in brain activity that aren't usually visible through traditional brain volume measurements. They showed that while standard imaging methods didn't show any differences for example in gray matter size between individuals with Parkinson's Disease (PD) and those without the condition ("healthy controls"), fMRI was able to reveal decreased connectivity in the Default Mode Network (DMN) in unimpaired patients, which demonstrated

that fMRI is able to identify early markers of the disease that are linked to cognitive functions even before any cognitive symptoms appeared. Research on psychiatric disorders such as schizophrenia and psychosis [6] has also benefited from fMRI techniques, as researchers were able to get a better understanding regarding the neural basis of such disorders, and by conducting experiments and comparing results between healthy individuals and those diagnosed with schizophrenia, they found that disrupted functional connectivity in the brain's "default mode network" has been linked with the severity of positive symptoms. Moreover, the list of neuroscience, neurology, and psychiatry research that has benefited from the application of fMRI includes several other disorders, and it extends to developmental disorders as well, such as autism, where certain investigations using fMRI scans [7] have shown that children with Autism Spectrum Disorder (ASD) exhibit unusual patterns of connectivity in certain regions of the brain like the insula and the right superior temporal gyrus, which are not found in neurotypical children. Overall, these insights not only aid in early diagnosis but also guide targeted interventions and therapies, making fMRI a powerful tool for personalized medicine. Therefore, in order to achieve these goals and efficiently improve diagnostics and treatment approaches, the integration of neuroimaging data is of great importance. Finland, in particular, has made significant strides in this field, an example of such work are the contributions made by the Finnish Biomedical Imaging Node (FiBI), as part of the pan-European Euro-BioImaging consortium, particularly through their involvement in advanced imaging technologies and infrastructures, which opens the door for cutting-edge brain research and international collaboration, placing Finland at the forefront of neuroimaging technology development [8].

1.1.2 The role of AI and its applications to neuroimaging

Artificial Intelligence (AI) has given neuroimaging a whole new dimension, making the analyses of complex brain data more accurate, faster, and to some extent more interpretable. Specifically, machine learning and deep learning techniques have been transformative in making sense of the high-dimensional fMRI data. Traditional analytical approaches in neuroimaging research are typically based on a predefined hypothesis concerning areas of interest in the brain with the objective of finding biomarkers that support diagnostic decisions [9]. This however, comes with a high load of feature engineering. On the other hand, AI models, specifically deep learning algorithms, are capable of autonomously uncovering patterns within data, allowing subtle changes to be identified as potentially predictive of specific conditions.

As a result, a growing number of neuroimaging studies are adopting deep learning frameworks for optimized biomarker identification, as well as improved diagnostic accuracy. An example of this is a case study [10] done on Major Depressive Disorder (MDD) patients, where deep learning was proven to be able to successfully classify with competitively high accuracy, different brain states in the patients and distinguish healthy subjects from affected ones, through learning the complex patterns of functional connectivity, the same study also showed that similar approaches provide additional information regarding the brain regions that are involved in mood disorders such as MDD and Bipolar Disorder (BD).

On top of that, AI's contributions to clinical applications in neuroimaging are not just limited to classification and prediction, they also extend to the enhancement of image quality and the extraction of meaningful features from noisy data. AI models can be trained to denoise fMRI data, improving signal quality and reducing the need for lengthy scan times, which in turn makes the imaging process less burdensome for patients [11]. Moreover, this technology can facilitate the analysis of large-scale datasets by automating labor-intensive processes like image segmentation and

reconstruction, ultimately accelerating research timelines and reducing human error.

1.1.3 Privacy concerns and possible solutions

Nevertheless, this type of data is particularly sensitive, as it has the potential to expose confidential information regarding an individual's neurological activity, which in turn can reveal information regarding the individual's cognitive states, emotional responses, mental health conditions, or even susceptibility to certain neurological disorders [12] [13]. Thus the handling of sensitive patient data through these predictive models and the sharing of such information has given rise to several ethical and legal concerns, particularly in the context of privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union [14]. Moreover, considering the complexity of the models needed to accurately represent fMRI images and the fact that high-dimensional imaging data of this sort more often than not comes with a limited sample size, the need for more data is even stronger, as larger datasets help to capture the variability and intricacies present in neural activity. The challenge of balancing the demand for data-driven insights with the responsibility to protect patient privacy has led to a growing interest in methods such as Differential Privacy (DP) [15] [16] and generative models for secure data sharing, and the development of powerful AI models applied to healthcare that enables great advancements in diagnostics, and treatment personalization, all while minimizing the risk of exposing sensitive patient information.

1.2 Objectives of the study

Limited availability of real-world medical data often hinders research progress, especially in domains where strict privacy regulations restrict data access. This scarcity can also lead to issues such as class imbalance, where certain medical conditions or

demographic groups are underrepresented, ultimately impacting the robustness and generalizability of predictive models [17] [18]. To put this challenge into perspective, the experimental setting in this work models a realistic data-sharing scenario involving two health institutions (Hospital A and B) with neuroimaging databases, where Hospital B has limited data availability and seeks to use DP synthetic data generated from the more extensive data set of Hospital A.

This thesis aims to propose a solution to integrate differential privacy into an Improved Conditional Wasserstein Generative Adversarial Network (ICW-GAN) [19] that is specifically tailored to generate synthetic fMRI data for medical applications. By building on the work of (Zhuang et al., 2019) [20] and focusing on a DP version of the originally proposed ICW-GAN, this study addresses critical privacy concerns inherent in sharing medical data, with a particular emphasis on ensuring compliance with the GDPR.

The Generative model is trained on real fMRI data from the NeuroVault collection 1952 [21], with the adopted technical approach for incorporating differential privacy being through the offered functionalities of the privacy engine from the Pytorch Opacus library [22]. This approach ensures a straightforward implementation of a model that is capable of generating synthetic "task-dependent functional brain images" with quantifiable levels of privacy, allowing secure data sharing, and minimizing the risk of revealing identifiable patient information.

The evaluation of the DP ICW-GAN involves assessing the level of privacy achieved in the generated data, examining the utility of the synthetic data for downstream classification tasks, and analyzing the quality of generated samples as measured by the Inception Score (IS) [23]. Moreover, the downstream classification task had the purpose of assessing whether the inclusion of synthetic data (both with and without DP) affected predictive accuracy. This involved training classifiers under three different scenarios: using real data only, a mixture of real and synthetic data,

and a combination of real and DP synthetic data. This setup allows us to compare the level of utility preserved by different data combinations, assess how DP synthetic data influences the classification accuracy, and to some extent help us better understand the trade-offs between maintaining privacy through DP synthetic data and achieving strong predictive performance.

Notably, no hyperparameter tuning or optimization was applied beyond the adjustments for privacy, and as is the case with the original work, this method could still be readily adapted to other datasets or similar data-sharing contexts. In conclusion, this research examines the balance between privacy protection and utility in synthetic fMRI data, offering a potential framework for privacy-preserving data sharing in medical and neuroscience applications. The main research questions guiding this study can be listed as follows:

1. To what extent can differential privacy be integrated into generating fMRI data without significantly impacting the downstream classification utility?

2. How does the inclusion of synthetic data, both with and without DP, influence predictive accuracy when compared to training solely on real data? Does it introduce an improvement in accuracy?

3. What is the quality of the generated fMRI samples, as evaluated by the IS, especially when DP is applied?

1.3 Thesis contents

In the remaining parts of this thesis, the content is presented as follows: chapter 2, "Literature review", presents the theory behind the main concepts used through this work, this includes a detailed discussion on Synthetic Data Generation (SDG), the privacy concerns that comes with it, as well as an exploration of the theoretical

framework behind DP. Chapter 3, "Materials and methods", builds on these the theoretical concepts presented in chapter 2 by detailing the practical methodologies and frameworks used in this thesis, it introduces Generative Adversarial Networks (GANs), along with other related subtypes, before going into more detail regarding the modifications and improvements proposed to introduce DP into our implementation. Chapter 4, "Modeling Process and Experimental setup", sets the scene and describes the data used for this work along with its properties, it describes in detail the approach used for the data generation and the introduction of DP, the model training protocols, and it also describes in more detail the hypothetical scenario that constitutes the basis of our downstream classification task. Chapter 5, "Results" presents the outcomes of the work, along with the evaluation process. Finally, the last chapter, "Conclusion" gives a general summary of the work, along with discussions regarding future work.

2 Literature review

Before delving into the practical aspects of this work, this chapter lays the groundwork by introducing foundational concepts central to this thesis. We will explore the privacy challenges inherent in medical data sharing, followed by a discussion on synthetic data generation techniques, highlighting their advantages and acknowledging the associated limitations. In addition, we will explore the concept of Differential Privacy as a potential solution to safeguard sensitive information during data sharing and model training processes.

2.1 Privacy challenges in medical data sharing (GDPR)

The exponential growth of volume in medical data has undoubtedly opened up new opportunities and interest for secondary purposes such as scientific research, statistics, and novelty in different applications [24]. One can even say that this data is believed to be worth its weight in gold when it comes to the development of models that aid health professionals in making more informed, evidence-based decisions, thereby ensuring the ability to improve diagnosis and prognosis, impacting better patient outcomes [25]. However, while healthcare data carries enormous transformative potential, sharing it for these purposes poses a great challenge in maintaining an ideal balance between the protection of individual privacy and facilitating research for potential benefit. This ethical and practical dilemma is often referred to in the literature as the "privacy-utility trade-off" because efforts toward safeguarding sensi-

tive personal information conflict directly with maximizing utility from data-driven insights [26] [27].

As a means to address these privacy concerns, strict regulations were set in place, such as the General Data Protection Regulation in the European Union [14]. The said regulation details a comprehensive set of guidelines and recommendations on how sensitive personal data should be handled, shared, and collected. While the GDPR is of strong importance when it comes to the protection of individual rights, compliance with its principles brings another layer of complexity into the work and efforts that benefit from shared medical data, forcing organizations into navigating a maze of legal requirements, informed consent protocols, data anonymization techniques, and secure transfer protocols to reduce privacy risks effectively. Beyond the borders of the European Union, however, the GDPR also shapes global policies, in a sense that its standards impact international practices and collaborations in data sharing, further complicating the effort to balance privacy and innovation. Therefore, finding sustainable solutions to protect patient privacy all while promoting scientific advancement remains a central challenge in the era of data-driven healthcare.

2.2 Synthetic data generation

Researchers have proposed Synthetic Data Generation (SDG) methods using statistical models as a potential solution to help minimize the risks associated with handling and disclosing sensitive data, as well as a valuable data augmentation technique, particularly in the context of medical applications [28] [29]. Formally, this technique involves creating artificial datasets that closely resemble real data in its statistical properties, without containing any identifiable information about individuals [30]. The goal here is to ensure that the generated synthetic data maintains high quality and statistical accuracy while minimizing the risk of re-identification.

In this sense, SDG methods claim to be inherently private, simply because the data is artificially generated and as a result not directly linked to any original sensitive data records [31]. This claim however has repeatedly been shown not to be correct, and we discuss this in greater detail in section 2.2.3.

Putting the related privacy concerns aside, SDG is still considered an innovative approach that paves the way for advancing healthcare research, by providing a valuable resource for realistic datasets that form the basis for analysis and predictive model development. However, it is important to acknowledge that generating high quality, statistically sound synthetic data requires expertise and robust methodologies, and ensuring the generalizability and validity of insights derived from such data compared to real world data remains an ongoing area of research.

2.2.1 Methods overview

Synthetic data generation has gained a lot of attention in research in past few years and has been applied across various fields. For that matter, the technique has been applied to a wide range of data formats, from tabular data, images, and videos, up to speech generation. Since our work focuses on the generation of synthetic imaging data, it is valuable to outline some foundational methods for synthetic data generation and highlight the specific approaches that are tailored for image data.

The earliest efforts in the development of generative models were very heavily dominated by work on image data, with an attempt to generate realistic images by learning the underlying distribution of the real images. Generative Adversarial Networks techniques together with other GAN variations have been widely applied for generating high quality synthetic images. Additionally, other deep learning-based methods that utilize a variety of neural network architectures include Variational Autoencoders (VAEs) [32] and Diffusion Models [33], both of which are recognized for their ability to generate high resolution and highly realistic images for medical

applications as well as other fields. These models were especially useful in image data applications thanks to their ability to excel at capturing intricate details in images, such as spatial relationships between pixels, textures, and patterns.

Tabular data has also been of great interest, especially since this data type is the most frequently used and readily available kind of data there is. However, work on generating synthetic tabular data has been relatively more modest. This is most likely due to the nature of tabular data itself, which mainly includes structured datasets composed of rows and columns. The nature of this data presents certain challenges, one of which, and probably the most significant being the difficulty of handling mixed data. Tabular data oftentimes contains both numerical and categorical features, requiring distinct analytical techniques to address each feature type separately and accurately. Ensemble methods, Bayesian networks, and statistical modeling methods were among the primary techniques used in tabular data generation prior to the introduction of deep learning. These methods made use of statistical or probabilistic models to produce synthetic data, drawing on the patterns and relationships between variables found in the original dataset [34]. Later on, efforts using deep learning and GAN-based models were also developed for tabular data generation, the results however have yet to demonstrate the same level of success seen in image data generation.

In the specific context of neuroimaging, SDG techniques allow us to generate artificial data that replicates complex brain activity patterns observed for example in fMRI scans [35]. This is particularly useful for data augmentation, especially when real data of such nature is limited, difficult to collect, or subject to privacy regulations. Furthermore, in certain medical research applications, some conditions or classes may be underrepresented, rare, or simply difficult to collect data on, which can translate into issues of class imbalance in the datasets. In such cases, synthetic imaging data could be one way to address these issues and help researchers create

more balanced datasets, hence enhancing model performance and the validity of analytical results.

2.2.2 Advantages and benefits of Synthetic Data Generation

One of the primary benefits that comes from the use of synthetic data is that it enables analyses and model development without requiring the use of real data, thus reducing the direct exposure of sensitive information. However, as we've already established synthetic data on its own does not guarantee privacy, additional methods are often necessary to ensure protection against re-identification risks. For example, a systematic review [36] emphasizes that achieving privacy in synthetic data is a multi-faceted process that often requires a combination of techniques, this includes processes such as Encryption [37], k-anonymity [38], and Differential Privacy which will be the focus of this work and will be discussed in further detail in coming sections.

Beyond the privacy considerations, SDG techniques offer a flexible way of generating datasets tailored to specific research needs while allowing control over the characteristics and distributions of the data. This has great value in overcoming problems inherent in real-life data, which is often noisy and unbalanced. An interesting example of such benefits is demonstrated in the work done by researchers from MIT, and Boston University [39], where they were able to generate a synthetic data set representing human actions, and showed that pretrained machine learning models on this type of artificial datasets offered improved performance compared to real datasets. The study further underlined the flexibility of SDG methods in generating high volumes of "perfectly annotated" data that is perfectly designed for a given use case [40] through careful adjustment of simulation parameters (e.g., lighting, poses).

This can directly mitigate the biases in analyses and model development, by

generating data that is representative of the target population without inheriting biases that naturally occur in real-world data. For instance, in the case of rare-event imbalance in applications like financial fraud analysis, where oftentimes datasets are heavily unbalanced due to the rarity of fraudulent events, the developed models are more susceptible to bias, using SGD can provide reliable analysis for the modeling of such events by generating samples that compensate for underrepresented classes [41] [42].

Furthermore, research by (Frid-Adar et al., 2018) [43] explores how synthetic data generation, particularly using GANs, can be very useful in medical imaging applications by augmenting existing datasets and improving CNN performance for medical image classification. This approach demonstrates that SDG methods not only contribute to expanding data quantity but also to improving model quality and research outcomes. For instance, the study observed increased sensitivity and specificity in models trained with synthetic data augmentation, outperforming classic data augmentation techniques in achieving accurate classifications and ultimately helping create fairer and more reliable models.

2.2.3 Challenges and limitations of Synthetic Data Generation

As we have established already, synthetic data offers a valuable resource for driving responsible innovation, particularly by addressing biases, overcoming limitations of scarce or incomplete datasets, and enhancing opportunities to build more accurate and trustworthy models. Nonetheless, several challenges arise in the use of synthetic data in practice. While synthetic data has been promoted as a promising approach to privacy preservation, several studies revealed that it is not "inherently private", and synthetic datasets can still expose or "leak" sensitive information [44] [45]. Generative models, for example, particularly those trained on real world datasets, have

been demonstrated to be highly susceptible to inference attacks [46] that may expose underlying information about individuals, especially in domains where datasets often include unique, identifiable patterns. In this context, inference attacks [47] refer to adversary attacks that use the output of an algorithm's output, along with known information or complete data records about an individual to identify whether this said individual was included in the dataset used by the algorithm. Even though this may not seem immediately concerning, consider the risk of discovering that a certain public figure, for whom various personal details are known, was part of a dataset on a sensitive issue, such as a specific disease or financial fraud investigation. Such examples have led to concerns that even supposedly anonymized synthetic data could be re-identified by attackers who are able to reverse-engineer the synthetic patterns to approximate the original data, and so carefully implemented privacy mechanisms are of great importance to ensure that the generated data is both useful and meets privacy guarantees.

Another limitation is that synthetic data may not fully represent the complexity of real world scenarios. The variations and unanticipated events common in real life data (particularly rare or extreme cases, otherwise referred to as outliers) are often difficult for generative models to accurately replicate, at least not in a private manner [44] [48], as it is often said, "real life is usually stranger than fiction." As a result, models that are trained solely on synthetic data may not perform as well when applied to real world contexts where these unusual patterns or rare occurrences emerge. An example of this is explained by Stadler et al., where they described a case with a low-probability event like the presence of a multi-billionaire in wealth data. In such a scenario, a synthetic data generator might either fail to replicate the statistics of such rare outliers accurately or reveal potentially sensitive information about the individual in the process. And so, although synthetic data aims to reduce biases inherent in real data, as we have discussed when talking

about its advantages, it can sometimes also introduce new biases if the underlying generative models reflect biased assumptions or fail to capture diverse real-world distributions. Inaccurate synthetic data may reproduce these biases in downstream models, ultimately negatively affecting outcomes.

In short, synthetic data generation presents great opportunities to support and drive efforts in various fields of research by providing accessible and flexible datasets for analysis and model training. However, synthetic data can never be considered a "replacement for real data" [44]. Especially when privacy guarantees are applied, as some degree of distortion is almost always introduced. It is for that reason that synthetic data can best serve as a complementary tool rather than the main focus, and it remains essential that any final tools intended for real world application are rigorously evaluated and, if needed, fine-tuned using real data to guarantee their effectiveness and dependability in practical settings.

2.3 Differential Privacy as a potential solution

As a way to address the shortcomings of standard Synthetic Data Generation, especially relating to privacy preservation, the principle of Differential Privacy (DP) was introduced. A widely recognized definition of Differential Privacy was formally presented by the Cynthia Dwork et al [49], to describe the "Promise of Differential Privacy":

"Differential privacy describes a promise, made by a data holder, or curator, to a data subject: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.' "

At its core, the framework of Differential Privacy ensures that the inclusion or exclusion of a single data point does not significantly affect the outcome of any analysis, and in our use case the outcome of the synthetically generated data. In

Dataset 1

User ID	Age	Location
abc123	34	NY
xyz456	28	CA
qwe789	45	TX
frj022	23	WA
def456	37	NV

Dataset 2

opt-out scenario: one user's data is removed

User ID	Age	Location
abc123	34	NY
xyz456	28	CA
qwe789	45	TX
def456	37	NV

$$D_1 = \{x_1, x_2, x_3, \ldots\} \qquad D_2 = \{x_1, x_2, x_3, \ldots\}$$

Figure 2.1: Adjacent datasets and the opt-out-scenario in DP

other words, it guarantees that a single individual's information is not leaked by not including them in the analysis, otherwise known as the "opt-out-scenario"[50].

To put this into perspective, let's consider two "adjacent datasets" D_1 and D_2. The only thing that sets these two datasets apart is the exclusion of one observation (Figure 2.1). This means that by producing any analysis on D_2, it is impossible to get any identifiable information about the excluded data record. Differential privacy takes advantage of this concept by introducing a controlled amount of randomness or noise into the data (in this case D_1), which would allow to hide the contribution of the data observation in question, all while producing analytical results that strongly converge towards the "opt-out-scenario" results. This introduces a mathematically quantifiable measurement of privacy protection into the analysis.

We will delve further into the mathematical foundations of Differential Privacy in the following chapter.

3 Materials and methods

Generative models are a class of machine learning algorithms designed to learn and approximate complex, high-dimensional probability distributions from data samples, capture intricate patterns and dependencies, and as a result enable the estimation of the likelihood of data observations and generate new samples that closely resemble the underlying distribution of the training data [51]. This makes them a powerful tool that is widely used in applications where synthetic data can serve as a substitute or complement to real data, such as privacy-preserving data sharing, data augmentation, and image synthesis.

3.1 Generative Adversarial Networks

Introduced by Goodfellow et al., (2014) [52], Generative Adversarial Networks (GANs) represent a class of generative models that operate through an adversarial process between two distinct models: a generator, denoted as G, and a discriminator, denoted as D. These models are trained simultaneously, with the generator G aiming to capture an estimated distribution, denoted as p_{model}, from a training set consisting of samples drawn from a distribution p_{data}. On the other hand, the discriminator D tries to estimate the probability of whether a given data point comes from the training set or is generated by G. The effectiveness of GANs stems from their training procedure, which involves optimizing both the discriminator and generator simultaneously through a competitive "two-player minmax game" [53].

More formally, in order to learn the distribution p_{model}, the generator G constructs generated outputs $G(\mathbf{z})$, taking in a noise vector $\mathbf{z}$ as input from a known distribution. Subsequently, the discriminator D operates as a binary classifier, outputting the likelihood that the input data ($\mathbf{x}$) is real or synthetically generated ($G(\mathbf{z})$).

These models involve various architectures for both the generator and the discriminator, which can include neural networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and autoencoders. As a result, the discriminator and the generator update their weights according to two loss functions, denoted as $Loss_D$ and $Loss_G$, facilitating the update of network parameters (Figure 3.1). During training, the generator adjusts its parameters only based on backpropagation signals originating from the generated outputs $G(\mathbf{z})$. However, the discriminator receives more information, incorporating both fake and real outputs to update its weights [54]. The training objective of GANs can be formulated with the value function $V(G, D)$ as a two-player minmax game where the discriminator and generator optimize their respective objectives :

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \tag{3.1}$$

This framework underscores the dynamic interplay between the generator and discriminator, ultimately driving the convergence towards an equilibrium where the generated samples closely resemble real data distributions.

3.1.1 Conditional Generative Adversarial Networks (cGANs)

While 'vanilla' GANs have shown remarkable success in training generative models with applications in various domains, oftentimes, they lack the ability to control the

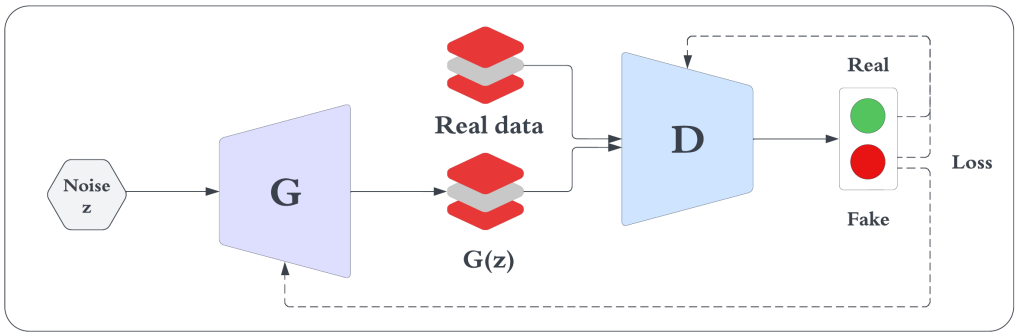


Figure 3.1: GAN architecture

generation process based on specific conditions, which makes them "unconditional". This has led to the introduction of Conditional Generative Adversarial Networks (cGANs), which was an extension of the original GAN framework. The concept was proposed by Mirza and Osindero (2014) [19], and it addresses this limitation by incorporating additional information to guide the data generation process and potentially yield better results.

This additional information is represented in a conditioning variable y that is added to both the generator and the discriminator through auxiliary input layers. y here can be simple labeled data, or for more complex conditioning, y can also represent other relevant attributes or modalities, (Figure 3.2) illustrates how the conditioning inputs change the structure of the network.

This extension is also translated into the updated minmax game objective function shown in Equation 3.2 :

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x}|\mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z}|\mathbf{y})))] \quad (3.2)$$

By conditioning the data generation on specific inputs, cGANs enable more controlled and meaningful generation of synthetic data. This is especially useful in applications where generating data with certain characteristics is important, such

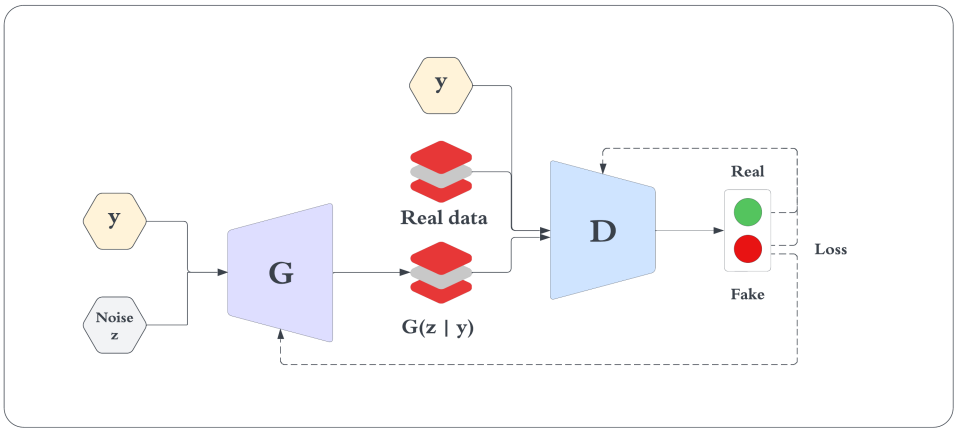


Figure 3.2: cGAN architecture

as in medical imaging or data augmentation for training machine learning models.

3.1.2 Wasserstein GAN (WGAN)

Wasserstein GAN is one of the newest introductions into the world of generative models, which came as another alternative to improve the 'vanilla' GAN training, as they have been proven by Arjovsky et al., (2017) [55] to face problems with training stability and mode collapse. For reference, **mode collapse** refers to a scenario where the generator fails to produce an output that represents the real data distribution and instead generates a limited variety of outputs, and **training instability** which can include oscillating losses and the problem of vanishing gradients, where the gradients used to update the model parameters become very small and approach zero, eventually slowing down or even stopping the training process.

The key innovation of Wasserstein GAN, introduced by Arjovsky et al., [56], is the reformulation of the loss function previously used to train GANs. Wasserstein Generative Adversarial Network (WGAN) makes use of the Earth Mover's (EM) distance, otherwise known as the Wasserstein-1 distance, ergo the name. EM represents a measure of distance between the real data distribution and the distribution of the generated data. In less formal terms, this would mean that the discriminator

(or the critic following the WGAN terminology), instead of only deciding whether or not the output is real, it quantifies the amount and distance that the "mass" should be moved to transform the fake output into something indistinguishable from the real one, resulting in more meaningful gradients for the generator, and leading to more stable convergence.

To further explain the advantage of using the Wasserstein-1 distance, we will consider the analogy of moving mass, let's suppose moving a pile of boxes. The Earth Mover's distance measures the minimum cost of transforming the collection of boxes from one location to the other, where the cost here is defined as the amount of "boxes" that need to be moved (or the weight of the pile) times the distance the mass needs to be moved. We denote $\gamma(\mathbf{x}, \mathbf{y})$ as a coupling of a joint probability distribution that describes how much mass should be moved from a point $\mathbf{x}$ in the source distribution p, to a point $\mathbf{y}$ in the target distribution q. The goal eventually is to find the optimal transport plan γ that reduces the total transportation "cost" across all points in the distributions.

Mathematically, the Earth Mover's Distance $W(p, q)$ between two distributions p and q is the "cost" of this optimal plan, and as defined in [56] can be formulated as:

$$W(p, q) = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[\|\mathbf{x} - \mathbf{y}\|]$$

where $\Pi(p, q)$ represents the set of all possible joint distributions for which the marginal distributions are p and q.

In the context of generative models, Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence [57] [58] are commonly used measures to quantify the difference between the real data distribution p_{data} and the generated data distribution p_{model}. However, in cases where these distributions do not overlap (the generated output has a distribution far away from that of the real data), these measures can

fall short and be quite problematic for GANs, as they can lead to vanishing gradients making the discriminator too powerful, and ultimately hindering the training process. This is because the KL and JS divergence are based on the logarithms of probabilities, and can become highly unstable when distributions do not overlap.

The Wasserstein-1 Distance on the other hand is characterized by a linear loss, which ensures that small changes in the distribution result in proportional changes in the distance. Thus, the gradients derived from the Wasserstein-1 Distance are more stable and informative for the generator, leading to more reliable and efficient updates during the training.

With that, the WGAN framework minimizes the Wasserstein-1 distance to replace the traditional GAN's loss function, and based on the Kantorovich-Rubinstein duality the Wasserstein loss for the discriminator (or critic) and the generator can be represented separately in Equations 3.3 and 3.4

$$L_D = -\mathbb{E}_{\mathbf{x}\sim p_{data}}[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z}))] \tag{3.3}$$

$$L_G = -\mathbb{E}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z}))] \tag{3.4}$$

Where $\mathbf{z}$ is the input noise sampled from the noise distribution p_z, $\mathbf{x}$ is the real data and $G(\mathbf{z})$ is the data generated by G.

In this context, the critic D must belong to the set of 1-Lipschitz functions [56], which means that for any two points x and y in the input space, the critic's output difference is bounded by the distance between x and y, scaled by a maximum constant of $L = 1$. Mathematically, this is expressed as:

$$|D(x) - D(y)| \leq L\|x - y\|$$

Gradient penalty

To address these limitations, Gulrajani et al., [59] introduced the Gradient Penalty (GP) term, which is an effective regulation method that ensures the Lipschitz-1 continuity condition, by making sure that the gradient norm remains close to 1 and penalizing deviations. In other words, and as shown in Eq 3.5 the GP is defined as the squared deviation of the gradient norm from 1, calculated along samples interpolated between real and generated data. This penalty term is added to the loss function as:

$$L_D = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[D(G(\mathbf{z}))] + \underbrace{\lambda \mathbb{E}_{\tilde{\mathbf{x}}\sim p_{\tilde{\mathbf{x}}}}\left[(\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1)^2\right]}_{\text{gradient penalty}} \tag{3.5}$$

$$\tilde{\mathbf{x}} = \alpha G(\mathbf{z}) + (1-\alpha)\mathbf{x}, \quad \alpha \sim \text{Uniform}(0,1) \tag{3.6}$$

Where $\tilde{\mathbf{x}}$ represents the samples interpolated between real and generated data, λ is the regularization coefficient controlling the strength of the gradient penalty, α is the interpolation coefficient, and $p_{\tilde{\mathbf{x}}}$ denotes the distribution of interpolated points.

3.2 (ϵ, δ)-Differential Privacy

Formally, following Cynthia Dwork's work [49], a randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for any adjacent datasets D_1 and D_2 varying on at most one data observation, and for any subset of possible outputs $S \subseteq \text{Range}(\mathcal{A})$,

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(D_2) \in S] + \delta$$

Where $\mathcal{A}$ is the algorithm that represents the actions performed on the data, in our case, the GAN model responsible for generating the synthetic data, ϵ is the

parameter that serves as an upper bound to control the privacy loss, and δ represents the probability of violating the privacy guarantee. We will further discuss these parameters in more detail throughout this thesis.

It is important to note that this definition does not "create differential privacy" in and of itself; rather it represents a measure of the level of privacy protection. In other words, it quantifies "how much privacy is afforded" by a given algorithm [60].

Epsilon (ϵ) parameter

The parameter ϵ, otherwise known as the **"privacy budget"**, is crucial in the differential privacy framework as it controls the amount of **privacy loss**. Specifically, ϵ defines the degree to which the outputs of a given algorithm are indistinguishable when applied to two datasets differing by a single observation or individual (adjacent datasets), which subsequently quantifies the amount of privacy loss when there is a "differential" change in the data. This variation in the outputs, or lack thereof, is how the formal definition of privacy is determined in DP.

By definition, a smaller ϵ value indicates stronger privacy guarantees, thus more similarity between the outputs, this means that by looking solely at the outputs, an adversary would find it difficult to infer a certain individual's presence in the dataset. Nonetheless, naturally, this often comes at the cost of reduced utility of the synthetic data.

Defining an 'acceptable' range for ϵ can rather be a tricky task, as it is highly dependant on the specific application as well as the type of data being protected. In 2019, Dwork et al., emphasized that "there is no clear consensus on how to choose ϵ, nor agreement on how to approach this and other key implementation decisions. Given the importance of these details, there is a need for shared learning amongst the differential privacy community" [61]. More recently, the National Institute of Standards and Technology (NIST) has also acknowledged the difficulty of setting ϵ

values and has suggested that an ϵ in the "low single digits" (generally between 0 and 5) provides strong privacy for most applications. They also note that, in specific cases, values up to 20 or higher can offer a reasonable privacy balance, especially where utility is prioritized over strict privacy constraints [62].

It is also worth mentioning that given a desired value of ϵ, which defines the total privacy budget, each time the algorithm performs operations or calculations that include an observation relating to an individual, a part of their privacy budget is consumed.

Delta (δ) parameter

The parameter δ was introduced to the original definition of DP [15] in hopes of introducing a relaxation of the strict constraint that imposes how much the output of an algorithm can change when a single individual's data is modified. The presence of a non-zero δ allows for a small probability that the differential privacy guarantee might be violated, as long as the chosen value for δ is ideally smaller than the inverse of the dataset size [49]. This addition is especially useful to accommodate particular scenarios where the algorithm's output could accidentally leak sensitive information about individuals.

The originally proposed definition [16] represented the case where the value of δ is set to zero or is negligible. This is referred to as (ϵ,0)-differential privacy, or simply ϵ-differential privacy. This form, while it brings more simplicity, also imposes a stricter privacy guarantee, as it does not allow for any probability of privacy breaches beyond the bound set by ϵ.

3.3 The adopted model in this thesis

The model adopted in this thesis is a modified version of the 3D Conditional Wasserstein GAN introduced in the original work by Zhuang et al., [20]. This adaptation

incorporates both conditional generation and the Wasserstein GAN framework, making it capable of generating synthetic data conditioned on specific labels while also ensuring a more stable training through the use of the Wasserstein distance. To address privacy concerns, our approach introduces an improved version of the model that enforces DP during model training with the help of the Opacus Privacy Engine [22]. We have to note, however, that a key adjustment that was made when incorporating Opacus was the removal of the gradient penalty term, which is, as previously discussed, typically used in standard WGANs, and instead set an upper bound on the gradient norm. The reason behind this is that the Privacy Engine API modifies the training procedure to preserve privacy by limiting the sensitivity of gradients, and thus eliminating the need for an explicit penalty term, as the differential privacy mechanism inherently controls the gradient magnitudes, and introduces some regularization.

4 Modeling Process and Experimental setup

4.1 Data overview

4.1.1 Description of the (fMRI) dataset

NeuroVault [21] is a widely recognized web-based repository dedicated to the storage and sharing of unthresholded statistical maps, parcellations, and other results obtained from neuroimaging studies. The platform was built to promote open science and reproducibility in neuroimaging research. With its user-friendly interface and robust metadata, the site allows researchers to explore a large bank of imaging data on the human brain, as well as share or reanalyze these datasets, thus supporting a broad range of neuroimaging projects, and serving as a critical resource for the neuroimaging community.

The dataset used in this thesis is sourced from NeuroVault's largest comprehensive compilation of fMRI statistic maps; collection 1952[1], also known as BrainPedia. The collection is a result of several neuroimaging research efforts, namely the Human Connectome Project, Neurospin research center, and OpenfMRI. The fMRI images in the collection were acquired from a diverse group of participants undertaking various task-based protocols. These tasks are designed to examine different aspects of

[1]The collection is available on NeuroVault at: https://identifiers.org/neurovault.collection:1952.

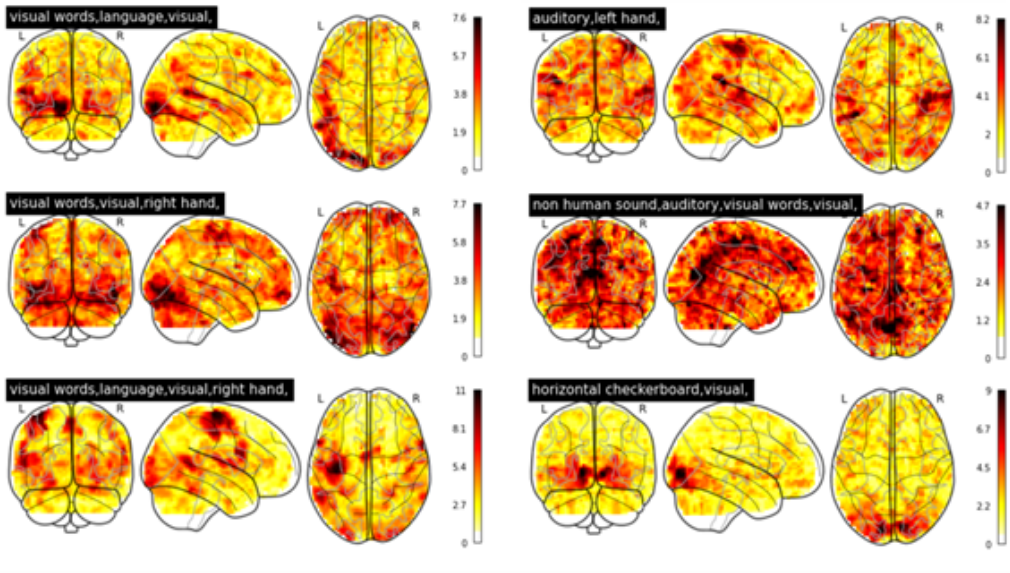


Figure 4.1: Real data samples from collection 1952

cognitive functions and brain states, hence providing a wide range of neural activity patterns for possible analysis and modeling. The dataset has also been preprocessed to ensure consistency and reliability across the different studies.

Collection 1952 represents a large dataset of 6,573 images, each with dimensions of 53x63x46. The dataset is categorized into 45 distinct classes, which are multiclass labels representing various cognitive tasks and correspond to specific experimental conditions. The classes are derived from 19 meta-labels (e.g. ‘visual,’ ‘language,’ ’faces,’ ’auditory,’). For example, a class like "visual, right hand, faces" refers to the brain activity captured during tasks such as viewing a face and responding to a 0 or 2-back working memory task versus fixation, whereas "human sound, auditory, language, right hand," represents tasks like listening to a strongly compressed (40%) auditory sentence. These class labels are a representation of the cognitive processes involved in the tasks, like visual recognition, or auditory function, etc.

4.1.2 Preprocessing

The preprocessing pipeline for the data collection followed the procedures and methods outlined in the original work [20], with additional steps to ensure the images were suitable as input for our models. The preprocessing steps are detailed as follows.

Downsampling of images

To manage the high dimensionality of the brain images and reduce computational load and memory requirements, the images were resampled. This process involved rescaling each image to a smaller size, determined by the specified downsampling factor, in our case of 0.25, meaning that each dimension of the original image was reduced to 25% of its original size, thereby reducing the total number of voxels while preserving the most relevant spatial information in the statistical maps.

For reference, in the context of fMRI images, voxels are the three-dimensional equivalent of a pixel, they are typically 3D units of brain volume, with each voxel representing a small cube of brain tissue.

Each image also has an affine transformation matrix, which is a mathematical representation that maps the 3D space of the image to a physical space (like the actual brain in a scanner). After resampling the images (changing the resolution or size), we also needed to adjust this matrix so that the new, smaller images still accurately represent the same physical space in the brain.

Lastly, since the images are being resized, some new voxels might not directly correspond to the original ones. For that reason, continuous interpolation was used during the resampling process to estimate the intensity values of the new voxels, and subsequently help in smoothly transitioning these intensity values, so the image doesn't look "blocky" or lose important details.

Normalization

Normalization is a critical preprocessing step that was performed to standardize the intensity values of the images, facilitating the comparison and integration of data across different samples. This step involved a **min-max normalization** of the voxel intensity values to a [0, 1] range. Ensuring that all statistical maps had comparable intensity scales, and reducing the impact of variability in the brain images, that might be introduced by varying brightness or contrast between scans.

One-Hot encoding of labels

The dataset's categorical labels, which represent different cognitive tasks, were transformed into a one-hot encoded format. In this format, each label is represented as a binary vector, with each dimension corresponding to a specific class, and the presence of a class is indicated by a '1' in the corresponding position, and all other positions are set to '0'. This encoding was necessary in order to allow for the representation of multiple classes without implying any ordinal relationship between them.

Computation of mask

Typically, an fMRI volume includes both brain regions and surrounding areas that do not contribute meaningful information. Since the analysis is usually focused on brain voxels, applying a mask is essential to isolate these relevant regions. This mask identifies the brain voxels of interest by preserving their original values while setting all voxels outside the mask to a value of zero [63], effectively excluding them from the analysis. In this work, the nilearn.masking library was used to implement this masking process.

4.2 Experimental setting

To illustrate the practical implications of our research, we examine a scenario involving two hospitals as illustrated in Figure 4.2. Hospital A, in this setting, is a large institution with access to a significant amount of data, derived from various sources such as clinical trials and research studies. However, due to strict privacy concerns and regulations, Hospital A is unable to share its extensive dataset directly. On the other hand, Hospital B is a much smaller establishment with very limited resources and facing data scarcity. Despite having some existing datasets, which consist of fMRI imaging data available from past studies, Hospital B struggles to improve its analytical capabilities due to the lack of diverse and extensive data.

To address this challenge, Hospital A employs a DP generative model to synthesize private data. By doing so, Hospital A ensures that the privacy of its dataset is preserved while enabling data sharing with Hospital B, allowing it to augment its limited dataset. In this scenario, this combined dataset will then be used for a classification task to evaluate whether the addition of private synthetic data improves predictive accuracy and analytical outcomes. Furthermore, to simulate a real-world scenario more closely, we assume that the available data in this experiment covers only a limited number of cognitive tasks. This setup will eventually help us to highlight the importance of using differentially private generative models to bridge the data gap between institutions of varying sizes and resources.

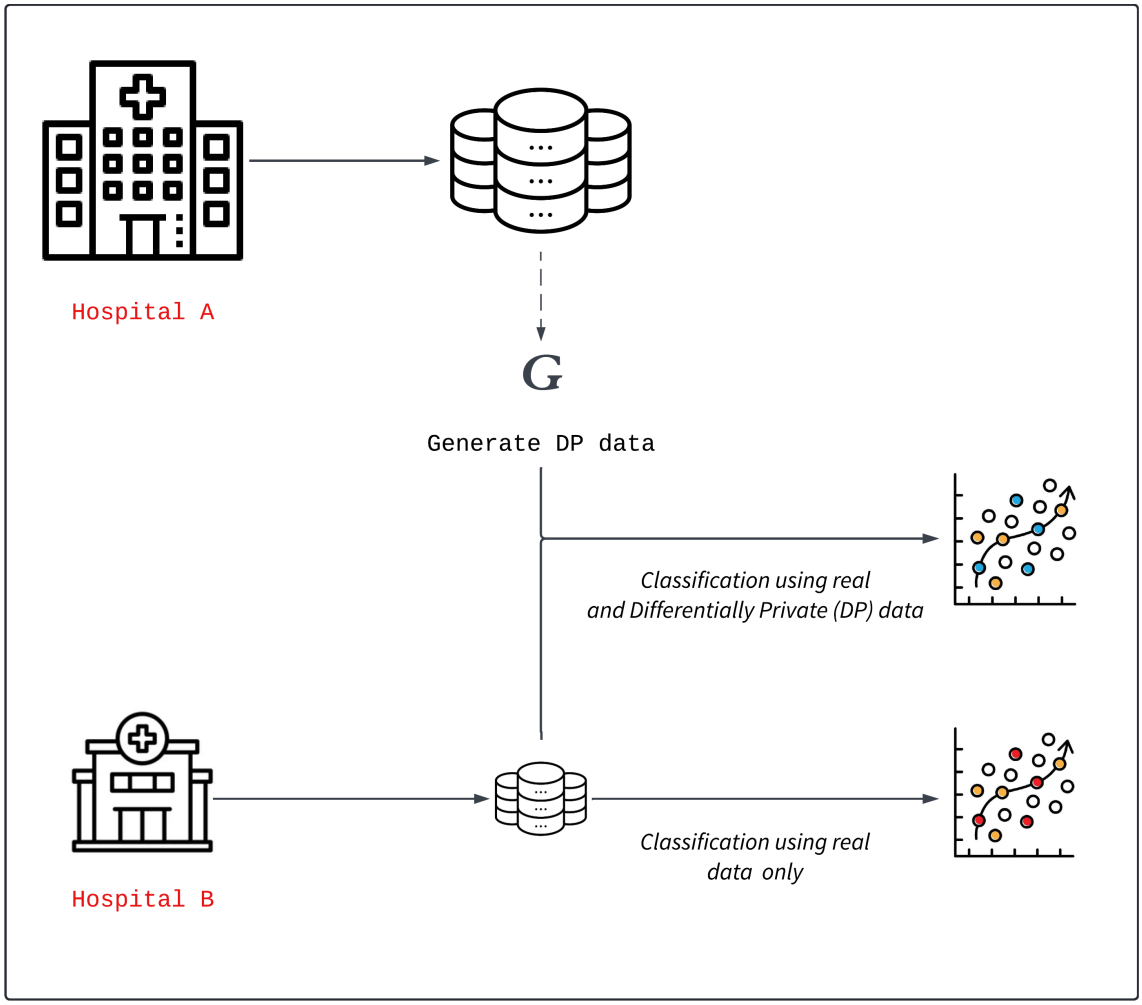


Figure 4.2: Flow chart of the proposed scenario

4.3 Modeling

4.3.1 Model training and synthetic data generation

The training process for our generator and critic networks, as we already know at this point, is adversarial, meaning that the two components are trained simultaneously in a back-and-forth manner to produce realistic and label-conditioned synthetic fMRI data. Below, we describe the key aspects of this process, including the initialization, loss computation, optimization, and privacy integration.

Initialization and hyper-parameters

Before initiating model training, the dataset was first filtered to include only the top ten most frequent classes, ensuring alignment with the described experimental setting, and making sure that the model only had access to a number of well-represented classes. The processed data was then divided into two subsets to simulate the scenarios for Hospital A and Hospital B, maintaining a 2:1 ratio to reflect the differing data availability at each institution. A data loader was then configured to handle the Hospital A subset, with a batch size of 50, enabling efficient sampling and training.

Differential Privacy in training

Differential privacy was systematically integrated into the training process of the ICW-fMRI-GAN using Opacus [22], a library that facilitates the training of PyTorch models under differential privacy constraints. The Privacy engine is the primary tool from Opacus used for introducing privacy by modifying the gradients during training, adding calibrated noise and applying gradient clipping. More specifically, per-sample gradients were clipped to a predefined maximum norm of 1.0 to limit sensitivity, and ensure that no single training example exerts too much influence on the model. The privacy parameters also included a noise multiplier of 1.4, which defines the amount of noise added by calculating the ratio between the Gaussian noise's standard deviation and the L_2-sensitivity of the model's computations [22]. Additionally we set a target value for the parameter δ to 1×10^{-5}, representing the upper bound on the probability of the privacy guarantee being violated. Making use of Opacus' privacy accounting mechanism, we were able to monitor and get an estimate of the cumulative privacy loss over the course of training, which resulted in a final ϵ value of 15.

Model training process

At each training step the critic attempts to distinguish between real images (from the dataset of Hospital A) and fake images (produced by the generator), while the generator in turn, attempts to fool the critic by producing increasingly realistic fMRI images. The training was conducted over 200,000 steps, and in each iteration, the critic was updated multiple times before a single update to the generator to ensure the critic remained well-optimized and provided meaningful gradients to guide the generator's learning process.

The critic's primary objective was to give higher scores to real samples and lower scores to generated ones, with the loss calculated as the difference in scores between these two groups. The generator, in turn, aimed to maximize the scores assigned by the critic to its generated outputs, effectively learning to produce samples indistinguishable from the real data. To achieve this, noise vectors drawn from a standard normal distribution were fed into the generator, and the resulting outputs were evaluated by the critic to compute the generator loss.

To set up the basis for our comparison, two training setups were employed: one where differential privacy was enabled, and another without privacy constraints. In the DP training, gradient clipping and noise addition ensured compliance with privacy guarantees, while in the non-DP training, these mechanisms were excluded and the gradient penalty was maintained as part of the loss computation.

Synthetic Data Generation

After training the models, the synthetic data generation process utilized the pretrained generator. Two generators, one trained with differential privacy and the other without, were loaded with their saved state and were used to generate separate synthetic datasets. In order to generate the synthetic samples, noise vectors of length 128 were sampled from a standard normal distribution and fed into the gener-

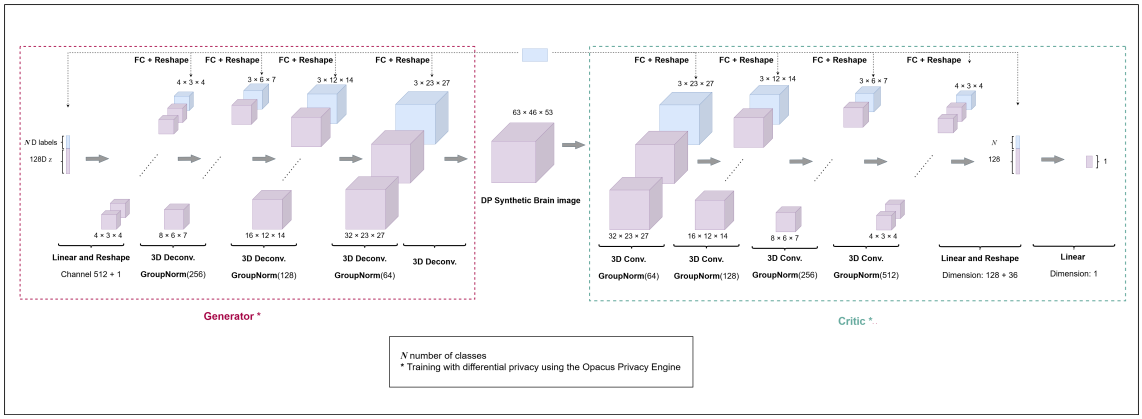


Figure 4.3: Model architecture with modifications for differential privacy. Group Normalization (GroupNorm) replaces Batch Normalization (BatchNorm) to support privacy-preserving training, and the Opacus Privacy Engine is used to ensure differential privacy.

ator alongside conditioning labels. During generation, these conditioning labels were selected in a cyclic manner and played a critical role in ensuring that the synthetic samples represented the classes from the training dataset accurately and uniformly, as each conditioning label corresponded to one of the unique classes present in the real dataset. We also note that the decision to generate the labels uniformly was made following the same approach used in the original work [20], however, since the generator was trained with differential privacy, the use of these conditioning labels should not directly reveal sensitive information about the individual samples, and it also hides, to some extent, the overall distribution of our data tags. Finally, the outputted synthetic brain images were subsequently upsampled to match the original data dimensions using a reverse scaling process consistent with the preprocessor applied during training.

4.3.2 Model architecture

The architecture of the proposed model in this work is illustrated in Fig 4.3, and it is almost identical to the architecture of the model in the original work [20] with some minor changes in order to support the addition of differential privacy. The generator

is designed to produce synthetic fMRI images conditioned on multi-tag labels, this is achieved by incorporating a multi-dimensional binary label vector, which encodes the presence or absence of the various tags. This label vector is concatenated with the random noise vector (128D **z** drawn from a multivariate Gaussian), forming the input to the generator. The generator then processes this input through a series of fully connected layers, each followed by normalization and ReLU activations, ultimately creating label-conditioned synthetic images. As the data passes through successive deconvolution layers, the image size is progressively upscaled, and the label information is reintroduced at each stage through additional label processing layers. Tanh activation is used at the final layer to produce the synthetic images which are then used as input for the critic, which mirrors the generator's architecture, using convolutional layers instead of deconvolution, with the major difference being the use of linear activations in the final layer.

To incorporate differential privacy in this model, small but crucial adjustments were made:

1. **Replacing Batch Normalization with Group Normalization:** Batch Normalization (BatchNorm) is often used in deep learning, particularly in many GAN architectures, to make training more stable and faster. BatchNorm achieves this by normalizing the outputs of a layer using batch-level statistics (mean and variance). In a differentially private setting, however, this can introduce issues, especially with small batch sizes, as it may compromise privacy. We recall that in the context of DP, we aim to ensure that the inclusion or exclusion of any single data point in a dataset does not significantly affect the results. BatchNorm in this sense might violate this principle because it normalizes a sample's value based on the other samples in the same batch. This means that the output for a given sample "depends on who else is in the batch" [64]. In theory, to address this issue and make BatchNorm DP-

friendly, we would need to add a privacy-preserving mechanism at this level as well, which complicates the implementation significantly. On the other hand, Opacus suggests using alternative normalization methods such as LayerNorm, InstanceNorm, or their generalization GroupNorm, which do not rely on batch-level statistics. Instead, they normalize based on the sample itself or small groups of its features, making them inherently "privacy-safe".

In our work, we replaced BatchNorm with Group Normalization (GroupNorm) throughout the network, as it also avoids dependency on batch size, which can pose privacy risks by introducing variability in the statistics across training samples. By normalizing features within each sample independently of others in the batch, GroupNorm ensures stable training and reduces the potential for privacy leakage.

2. **Privacy Engine:** As discussed in earlier sections, the Opacus privacy engine was integrated into the implementation to ensure differential privacy. This integration primarily modifies the gradient computation process by clipping and adding noise to the gradients during backpropagation. As a result, the privacy engine does not directly alter the data flow or affect the shape of the network. Therefore, its inclusion did not introduce significant changes to the model architecture.

4.3.3 Monitoring training progress

During training, at each visualization interval, which is set to every 1000 steps, the model saves visualizations comparing a real image to one generated by the generator, with both images corresponding to the same tag. For each interval, we display the real image and the generated image side by side, along with computed correlation scores for each image, which will be further explained in the following section. An example of such an output is displayed in Fig 4.4.

The visualization process begins with the dataloader, which fetches batches of data while ensuring privacy compliance through the privacy engine. For each batch, the generator uses random noise as input, along with the real labels to produce synthetic images conditioned on these labels. Since the noise is randomly generated and not derived from the real data, this process does not introduce immediate privacy concerns. Opacus adds further privacy guarantees by clipping and adding noise to the gradients during training, this guarantees that no individual data point has an excessive influence on the model.

It is also important to emphasize that the model learns primarily through the loss functions rather than the visualization process itself. These visualizations are purely a tool for monitoring the training progress and providing qualitative insights into the generator's performance. At each visualization step, the comparison focuses on the first image from both the real and synthetic batches. Since the generator is conditioned on the same labels as the real batch, the corresponding indices align, allowing a side-by-side comparison of two images associated with the same tag.

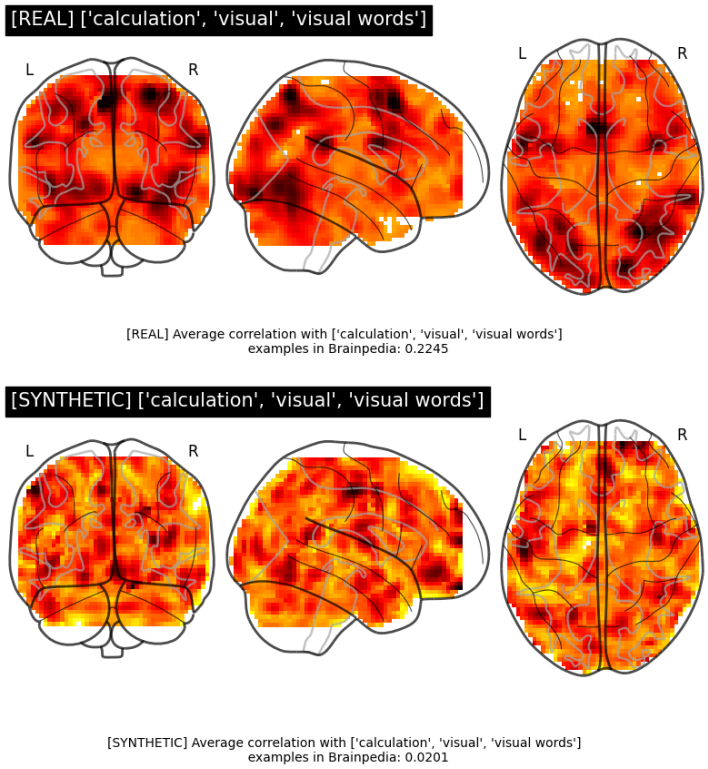


Figure 4.4: Example of training output

Correlation as a similarity measure for fMRI imaging

Comparing the results of generated images to real images with the naked eye can be challenging, especially when working with statistical maps. While blurry or pixelated images can be a clear indicator that the model is not performing very well and our results are largely noise, more subtle differences between images are much harder to detect visually and make it difficult to accurately assess the quality of the generated images.

To address this, we use correlation as a more reliable metric. In the context of neuroimaging, correlation is a widely used score to quantify similarity in brain activity patterns across different images, whether they are from multiple individuals, or the same individual, such as in cases where images are taken at different times.

As discussed in previous sections, brain images such as those from fMRI scans are made up of several small areas known as voxels, with each voxel representing the brain's activity at that point in space. When we calculate the correlation between two brain images, we are essentially comparing the activity levels in corresponding voxels across both images.

In this work this was implemented using the Neurosynth Image Decoder [65], where the values portrayed in the decoding table represent the Pearson correlation values. One thing to note however is that the current implementation of this decoder does not necessarily support the standard interpretation of correlation scores, where higher similarity in voxel activity results in a higher correlation. In other words, the values generated do not have a straightforward explanation, for example, a correlation score of 0.6 or 0.12 doesn't directly translate to "high" or "low" similarity in the way that we would interpret it in other contexts, which is why we will use the values as a comparison measure rather than focus on the absolute correlation values.

How the process works

At specific intervals during training, we visualize a real image and a generated image for the same task or label (referred to as a tag). For each visualization step, we compute the mean correlation of the real image with other images in the training set that share the same tag, we then calculate the same mean correlation for the generated image as well.

Our approach is to monitor the change in correlation over time. This means that as we evaluate the model at different training steps, we are able to track how the difference in correlation between the original and synthesized images evolves. If this difference tends to decrease consistently as the training progresses, it means that the synthesized images are becoming more similar to the original ones, which

in turn means that the model is learning effectively. Hence, by using correlation as a comparison metric, we are in a position where we can quantitatively examine the model's effectiveness or the quality of the generated images, without relying solely on visual inspection.

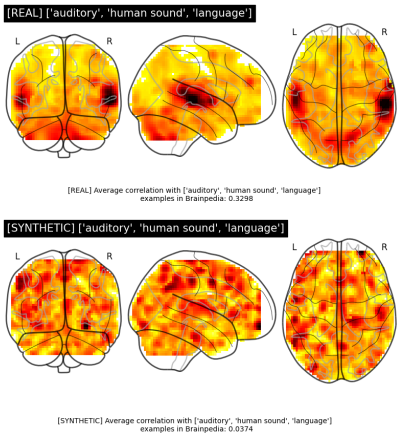
Fig 4.5 illustrates the progression of the model's performance over different training steps, namely at 1000, 50000, 100000, 150000, and 200000 iterations. Meanwhile, Table 4.1 shows the comparison of the correlation differences between these different training steps. Additionally, Table 4.2 provides a similar comparison for non-DP images. As expected, while both sets of findings exhibit a decrease in correlation difference over time, the differences are smaller in the case of non-DP images, indicating better image quality when privacy constraints are not applied. This aligns with the trade-off between utility and privacy, where some level of utility is sacrificed in order to ensure privacy. Nevertheless, these results can initially guide us to prove that our model is capable of generating increasingly accurate images that are closer to real ones. However, this will not be the sole piece of evidence taken into account, and additional metrics for evaluating the quality of the generated images will be further discussed in the results chapter for more comprehensive analysis.

Table 4.1: Comparison of correlation scores across training steps - DP data

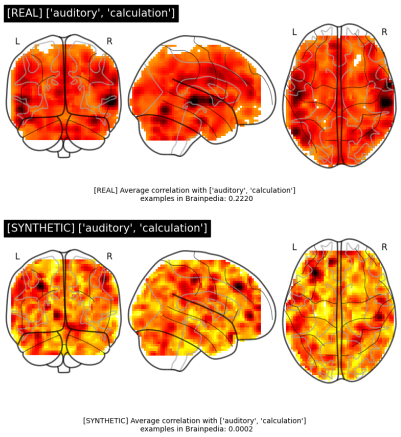
Training Step	**Correlation Real**	**Correlation Generated**	**Correlation Difference**
1000	0.3298	0.0374	0.2924
50000	0.222	0.0002	0.2218
100000	0.2245	0.0201	0.2044
150000	0.1711	0.0505	0.1206
200000	0.0729	0.0284	0.0445

Table 4.2: Comparison of correlation scores across training steps - non-DP data

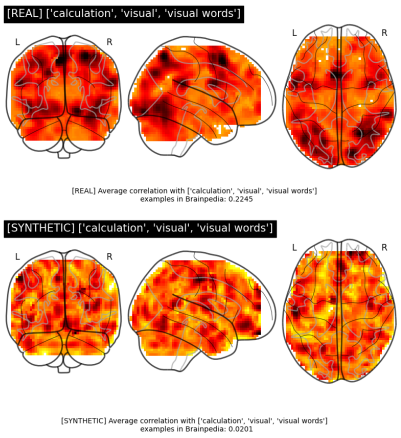
Training Step	**Correlation Real**	**Correlation Generated**	**Correlation Difference**
1000	0.3579	0.2407	0.1172
50000	0.4166	0.3263	0.0903
100000	0.1630	0.0955	0.0675
150000	0.1706	0.1586	0.0120
200000	0.1016	0.1100	0.0084



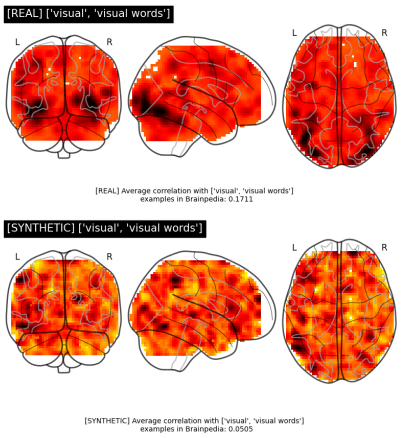
(a) Step 1000



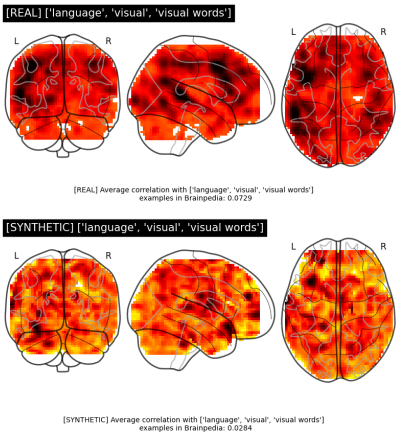
(b) Step 50000



(c) Step 100000



(d) Step 150000



(e) Step 200000

Figure 4.5: Comparison of real and generated images over different training steps

4.4 Description of the downstream classification tasks

The downstream classification task involved multi-classification prediction of cognitive tasks based on the fMRI data. This experiment also closely followed the same classification methodology used in the original paper [20], with one key adaptation to align with the the scenario described in the experimental setting: instead of using all the 45 available classes, only 10 were selected.

In order to ensure the comparability of the results, the same two models used in the work of Zhuang et al., [20] were employed for performance evaluation: a Support Vector Machine (SVM) and a custom Deep Neural Network (DNN) classifier. These models were a natural choice for their proven ability to handle high-dimensional fMRI data, making them the standard machine learning approach in this domain [66].

The **SVM** [67] classifier was applied to fMRI data after dimensionality reduction of the brain volumes through masking, as explained in section 4.1.2 where non-informative voxels were discarded, and valid voxels were flattened into a 1-dimensional vector. This approach is a common practice when working with fMRI data, as it simplifies the data and allows the SVM to focus on the most relevant features.

The **Deep Neural Network** [68] classifier followed a 3D convolutional architecture, similar to the discriminator used in the generative model. It included convolutional layers with Leaky ReLU activations, adapted to classify brain volumes without incorporating label information into intermediate layers.

4.4.1 Train-Test split and data setup

For the classification experiments, the real fMRI data (from NeuroVault Collection 1952) was first filtered to include only the 10 classes of interest, resulting in a dataset of 3,000 images. Of these, 2,000 images were used to train our generator,

which simulates the role of Hospital A (see Fig 4.2 for reference).

The remaining 1,200 images on the other hand represented the data from Hospital B, which was further divided into training and testing sets, ensuring that test data was always kept separate and was not used in neither the classification training nor the GAN training process. This approach differs from the procedure followed in [20], where the same dataset was used for both GAN training and the classification task.

Two experimental setups were used to assess the impact of synthetic data on classification performance:

Real-Data-Only classification: In this baseline setup, the models were trained solely on real fMRI data from the Hospital B training set. The goal was to establish a performance baseline for each model when no synthetic data was introduced into the training process.

Real + Synthetic (DP and non-DP) Data classification: In the second setup, models were trained on a combination of real data and synthetic data, both with and without differential privacy, generated by the ICW-fMRI-GAN. The goal here was to determine whether augmenting the training set with synthetic data, either private or non-private could improve the model's ability to generalize to unseen real test data. Additionally, this setup sought to evaluate the impact of differentially private versus non-private synthetic data on utility and classification performance.

For both setups, the same test data (the remaining images from Hospital B) was used to evaluate the classification performance, ensuring that the results reflected the models' ability to generalize to real world data. The training data was split in a 75/25 ratio, with 75% used for training and 25% reserved for testing. The inclusion of synthetic data increased the overall size of the training set, potentially providing the models with a more diverse set of examples to learn from.

4.4.2 Tag distribution

In order to get a closer look into how the different tags in the data were affected by this data augmentation, Fig 4.6 presents bar plots showing the distribution of tags across the real training data, real test data, and the mixed datasets that include synthetic data, both with and without differential privacy. From these plots, we can observe how the introduction of synthetic data (especially differentially private synthetic data) impacts the balance and frequency of specific tags. The synthetic data generally preserves the distribution trends of the real data, and this, in theory shouldn't mean that we are leaking information about our data, since we are using the synthetic data as an augmentation to the real data. Nonetheless, certain tags exhibit slight shifts in frequency due to the added privacy mechanism, but the effect is not significantly different from that of the non-DP synthetic data. One thing we do notice however, is that the added synthetic data reduces the disparity between the frequencies of different tags in comparison to the real data, which might be one of the reasons we noticed an improvement in accuracy (refer to chapter 5), since the data augmentation seems to compensate in a way for the lack of representation in certain tags.

4.5 Evaluation metrics

In this section we will shortly present the different metrics that were used for evaluating the performance of our model, as well as the quality of the generated images.

4.5.1 Downstream classification utility evaluation

The performance of the SVM and NN models was evaluated using accuracy as the primary metric, calculated as the percentage of correctly predicted labels over the total number of predictions.

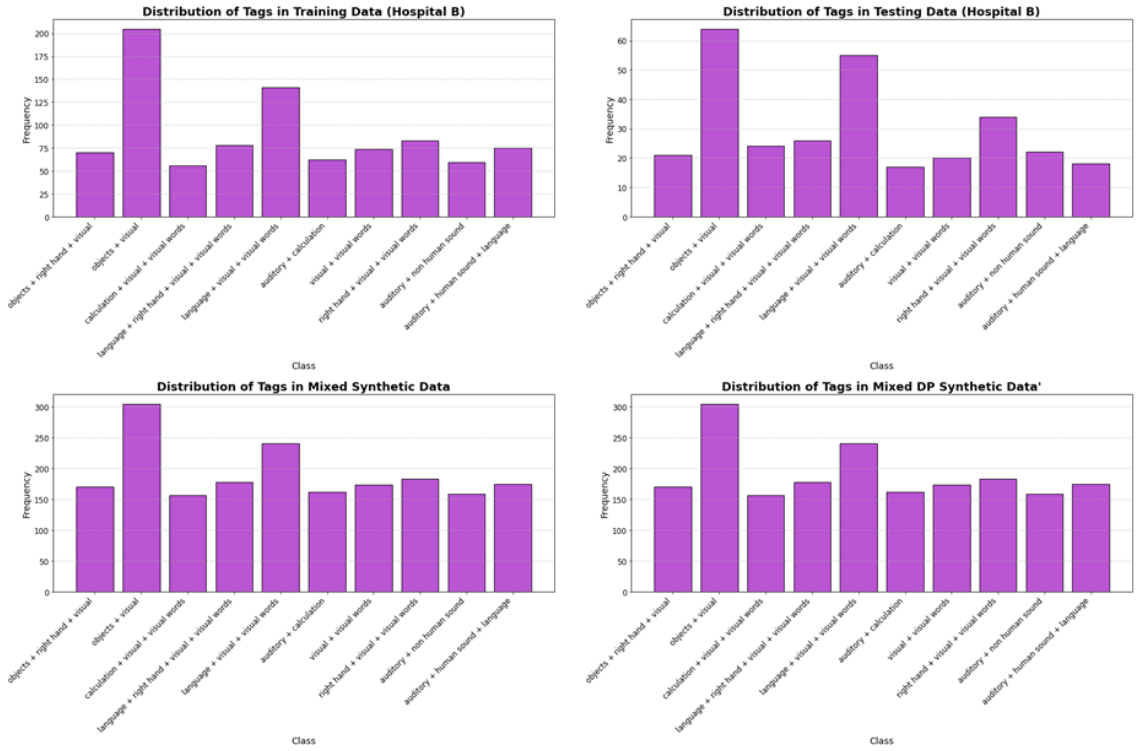


Figure 4.6: Tag distributions in the different datasets

In addition, given the multi-class nature of the task, and in order to further investigate the performance of the classifiers, we decided to also include additional evaluation metrics aside from overall accuracy. More specifically we wanted to get a closer look at how our model is performing for each class separately, and if maybe the model could be performing considerably better in specific classes, to do so we extracted per-class accuracies and constructed confusion matrices for each model.

The per-class accuracy, as the name suggests represents how the classifier performs on individual classes, and it offers further insight into strengths and weaknesses in scenarios involving multi-class classification. This becomes particularly crucial when handling imbalanced data, which is exactly what we are dealing with.

On the other hand, a confusion matrix is a tabular summary of the classification results which represents a comparaison between actual true labels and predicted labels; hence, we are able to identify any pattern of misclassifications. In the case of multiclass classification, each row of the matrix represents instances of the true

class, while each column represents instances of the predicted class. The diagonal elements show the number of correct predictions for each class (i.e., true positives for that class), and the off-diagonal elements indicate misclassifications between classes. Unlike binary classification, the confusion matrix in multiclass settings reveals how the model confuses specific classes with one another, helping to assess its performance on individual classes and detect patterns of misclassification.

With the help of these metrics, we can better understand which classes the model confuses, helping to highlight any overlap between the features learned by the classifiers and interpret their generalization capability when applied to both real and synthetic data, particularly in the context of differential privacy where the data distribution may be slightly altered.

4.5.2 Generated image quality evaluation

Apart from utility, another important aspect to consider when it comes to the performance evaluation of our generative model is the sample quality of the generated images. While it can be challenging to quantify this aspect using a single metric, especially one that allows for comparable results across different model architectures and varying datasets [69], a number of methods have been proposed to assess image quality, such as the Inception Score.

The **Inception Score**, [70] often just referred to as IS, is one of the most popular metrics for generative model image quality assessment, especially in the context of Generative Adversarial Networks (GANs). The main motive behind the development of this metric was to replace the "subjective evaluation by humans". It quantitatively evaluates both the **diversity** of the synthesized images (e.g., each image is a different cognitive task rather than repetitive) as well as their **resemblance** to real ones (e.g., the synthesized images show characteristics similar to actual fMRI scans); thus, giving an indication of how well the generative model has

captured the underlying data distribution.

How the Inception Score is calculated

The Inception Score (IS) calculation originally involves the use of a pre-trained image classifier from Google, commonly known as the Inception network [71]. While the specific architecture of the Inception model is optimized for general image classification tasks, these details are not directly relevant to our application, and the important thing to note here is that this network is responsible for computing class probability distributions for each given image, which is what matters for the IS computation, as these distributions make up the basis for the score calculation on the generated images. In our case, we replace the Inception network with our custom deep neural network classifier (discussed earlier in Section 5.1) to better suit the domain-specific nature of fMRI data. This adaptation ensures that the classifier reflects the intricacies of brain imaging data rather than relying on a general purpose image classifier.

Before we get into the details of the score calculation, we will briefly explain the kind of outputs or probability distributions we expect from our "Inception Network." For each individual image, we expect to see conditional label distributions $p(y|x)$ (a probability distribution that shows the likelihood of the image belonging to each possible label) where y is the set of labels and x is the image. Depending on the images passed through the model, we generally expect to see either uniform distributions of the labels, indicating that the image can equally likely belong to any of the available labels, or ideally a narrow distribution with a distinct peak, indicating that the image is more confidently classified into a single category, meaning a low entropy for the conditional label distribution $p(y|x)$. The original authors also discussed the concept of 'marginal distributions', which combine the label probability distributions across all synthetic images to give us an idea of the overall 'variety'

present in the dataset, meaning a high entropy for the marginal distribution $p(y)$. In other words, "We want each image to each be distinct and to collectively have variety" [23].

The score itself is computed as the KL-divergence between the conditional label distribution and the marginal distribution, averaged over all the generated images, and since this metric is, in essence, a measure of similarity/ difference between distributions, it is especially useful for assessing how distinct the conditional label distributions are from the marginal ones, as the greater the divergence (i.e., lower entropy compared to higher entropy), the better the quality, and naturally, the higher the score. The lowest possible value for IS is zero, and in theory, the highest possible value is infinity.

$$\text{IS} = \exp\left(\mathbb{E}_x \left[D_{KL}(p(y \mid x) \| p(y))\right]\right) \tag{4.1}$$

5 Results

The following chapter presents the outcomes of the experiments conducted to evaluate the performance of the DP ICW-fMRI-GAN model.

5.1 Performance evaluation: Classification utility

The results of the classification experiments are presented in Table 5.1. As can be seen, both the SVM and neural network classifiers experienced improvements in accuracy when real data from Hospital B was combined with the DP synthetic data from Hospital A. For the SVM classifier, the accuracy increased from 77.08% when using only real data to 78.07% when using a combination of real and synthetic (non-DP) data. Similarly, for the neural network classifier, the accuracy improved from 80.15% with real data only to 83.46% when mixing real and synthetic data. These results highlight the benefit of including synthetic data in improving classification performance.

When adding differentially private synthetic data, both models still demonstrated an improvement over the real data alone. The SVM classifier's accuracy increased to 77.41%, and the neural network classifier achieved 80.50%, both of which are higher than their respective real-data-only baselines. However, as expected, the results with non-DP synthetic data performed better than those with DP synthetic data, reinforcing the trade-off between privacy and utility. Although DP synthetic data boosts accuracy, the results with non-DP data were superior,

confirming that some level of utility is sacrificed to achieve privacy.

Table 5.1: Classification test accuracies for SVM and Neural Networks

Model	**Test Accuracy (%)**
SVM Classifier	
Real Data Only	77.08
Real + Synthetic (non-DP)	78.07
Real + DP Synthetic	77.41
Neural Network Classifier	
Real Data Only	80.15
Real + Synthetic (non-DP)	83.46
Real + DP Synthetic	80.50

When comparing the results across different experiments, it is clear that the addition of differential privacy resulted in a trade-off between data quality and classifier performance. While the quality of synthetic fMRI data was reduced under privacy constraints, the classifier's performance remained competitive, particularly with the use of mixed real and synthetic data, highlighting the utility of differentially private synthetic data in augmenting real datasets.

Per-class performance

Class	**SVM Classifier (Real Data Only)**	**Mixed Synthetic SVM**	**Mixed DP SVM**
0	84%	84%	84%
1	71%	71%	71%
2	83%	83%	80%
3	80%	82%	80%
4	66%	66%	66%
5	100%	100%	100%
6	90%	80%	95%
7	55%	64%	55%
8	88%	92%	88%
9	79%	83%	83%

Table 5.2: Per-Class test accuracies for SVM classifiers (Real Data, Mixed Synthetic, and Mixed DP)

The results presented in Table 5.2 compare the per-class accuracies for the three different SVM classifiers: one trained on real data only, one trained on mixed real and synthetic data, and one trained on mixed real and DP synthetic data. In addition, for easier readability Table 5.3 summarises the full class labels descriptions. Overall, all three classifiers show similar performance across most classes. For Class 0, Class 1, Class 4, and Class 5, all three classifiers yield the exact same accuracy, indicating consistent performance regardless of the training data.

Notably, for Class 6, the classifier trained on mixed DP synthetic data outperforms the others with a 95% accuracy, compared to 90% for the real data classifier and 80% for the mixed synthetic classifier. This suggests that the addition of DP synthetic data may benefit certain classes more than others by providing more diverse training samples. On the other hand, the classifiers perform quite poorly for Class 7, with a 55% for both the mixed DP classifier and the real data classifier, however, the mixed synthetic classifier performs considerably better with an accuracy of 64%.

In addition, these results does confirm the fact that the models can achieve varying performances in different classes, in this scenario Class 5 showed by far the highest level of correct predictions and Class 7 showed the lowest. From the tag distributions in Fig 4.6, we can also observe that these two classes do not display any particularly distinctive distributions (neither is the most represented nor the least represented). Therefore, we can say that the classifiers' performance is not directly linked to the class distributions, which suggests that other factors, such as the inherent difficulty of the classification task for specific tags, might be the main factor influencing the results.

Class	**Description**
0	auditory, calculation
1	objects, right hand, visual
2	right hand, visual, visual words
3	language, visual, visual words
4	objects, visual
5	auditory, non human sound
6	auditory, human sound, language
7	calculation, visual, visual words
8	language, right hand, visual, visual words
9	visual, visual words

Table 5.3: Class Definitions and Descriptions

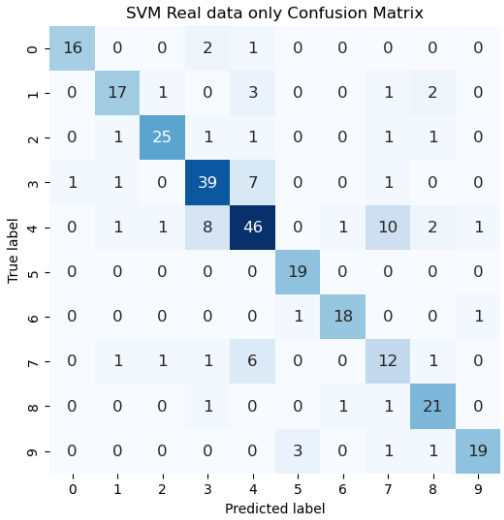
These results are also supported by Fig 5.1 which presents the three confusion matrices for the three classifiers. These provide further insights into the models performance by highlighting where exactly do the misclassifications occur across the different classes. At a first glance, it looks like yet again the models perform very similarly and most values fall in the diagonals of the matrices which means that more often than not the models predictions are accurate.

The specific missclassification trends are consistent through all classifiers, for example we observe misclassifications primarily in Class 3, Class 4 and Classs 7 in all three confusion matrices. We also can notice that these three classes are quite often confused for each other, if we take the Mixed Synthetic classifier as an example, in Class 4, 13 samples were misclassified as Class 7, and in the Mixed DP classifier, 9 samples from Class 4 were misclassified as Class 3, this might suggest that these three classes may share common features, leading to confusion. Regardless of that, all three classifiers yield a high number of correct predictions in Classes 3 and 4 (ie,

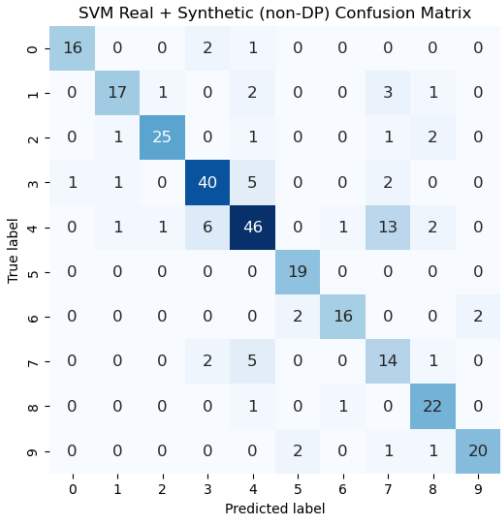
39 and 46 for first classifier), which explains why the per class accuracies are still relatively high for these two classes.

The overall distribution of predictions also suggests that all classifiers collectively perform well for Class 5 (19 correct predictions and 0 misclassifications) and Class 6 (16 to 19 correct predictions), matching the high accuracies reported in Table 5.2, this most likely means that these classes are well-represented and more separable in the feature space. An initial concern we had looking at the 100% accuracy level for Class 5 was that the classifiers might have had an easier time predicting this class due to the availability of more data, but now we can safely conclude that this was not the case.

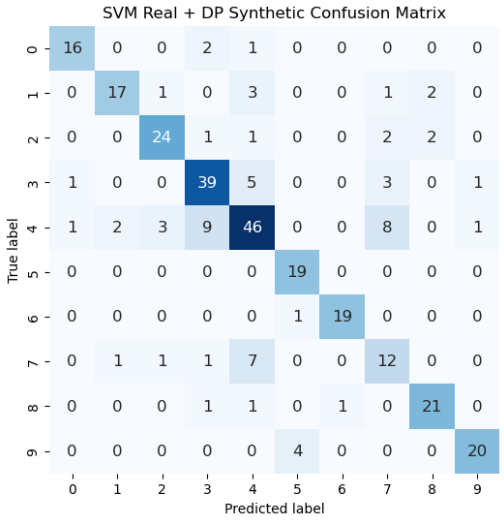
Looking at individual trends, we can see that the three classifiers achieve a total number of misclassifications of 69, 66 and 68, respectively. This reinforces the idea that the introduction of synthetic data introduces a level of improvement in accuracy even though quite small, and while adding DP synthetic data improves performance in certain classes in comparaison to using real data alone, it may lead to overfitting or misclassification in others. Additionally, this slightly improved performance in these classes across all three classifiers could indicate that the generative model was successful in preserving the key characteristics of these classes during synthetic data generation.



(a) Real Data Only



(b) Mixed Synthetic



(c) Mixed DP

Figure 5.1: Confusion Matrices for SVM Classifiers

5.2 Performance evaluation: Generated image quality

In our work, we implemented the Inception Score calculation to evaluate both the synthetic data and the differentially private (DP) synthetic data generated by our

DP ICW-fMRI-GAN model. The process followed was the same as described in the previous section with the addition of splitting the data into 10 subsets to compute the score across multiple batches for stability. We also refer to the work of (Salimans et al., 2016) on the CIFAR-10 dataset for comparability of our scores as it represents a common IS benchmark. The results are presented in Table 5.4

Table 5.4: Comparison of Inception Scores (IS) for CIFAR-10, Synthetic Data, and DP Synthetic Data

Dataset	**Score ± std**
CIFAR-10	11.24 ± 0.12
Synthetic Data	9.7240009 ± 0.1405271
DP Synthetic Data	9.7240037 ± 0.1405263

The Inception Score values indicate that the synthetic and DP synthetic fMRI images produced by the model are of high quality and diversity. The fact that the DP synthetic data scores almost identically to the non-DP synthetic data suggests that the introduction of differential privacy did not degrade the quality of the generated images, meaning that our implementation of privacy did not adversely affect the utility of the data for downstream tasks. This finding aligns with the results we observed in classification accuracies, which were very similar across both datasets. Additionally, we notice that the IS scores of our model are very comparable to the benchmark values from a dataset like CIFAR-10. This is a significant finding, as it demonstrates the capability of our DP ICW-fMRI-GAN to produce realistic and diverse fMRI images while ensuring privacy, which is a crucial factor in medical data generation.

6 Conclusion

Getting access to neuroimaging data, such as functional magnetic resonance imaging (fMRI) data is a persistent challenge in the field of medical research, as the acquisition process of such images, along with the effective screening of patients, is often expensive, time-intensive, and logistically complex, which limits the availability of high-quality datasets that can be used for analysis and innovation. Collaborative data sharing between institutions has emerged as a promising solution that bridges the gaps of lack of data in many fields, however, in the context of medical research, this process, particularly between hospitals, introduces significant privacy concerns due to the sensitive nature of medical data and the strict regulatory frameworks that are in place to to ensure the privacy of patients. To address these challenges, this thesis explored the integration of Differential Privacy into synthetic data generation applied specifically to fMRI data, building on the efforts of Zhuang et al., [72]. To our knowledge, this work represents the first attempt to combine these concepts to generate differentially private synthetic fMRI images, and as a result, the technical choices made throughout this research were guided by the dual objective of ensuring rigorous privacy guarantees while preserving the utility of the synthetic data for downstream tasks.

The research questions presented in section 1.2 were aimed at addressing the core challenges of this study, specifically examining the impact of integrating DP on (1) downstream classification utility, (2) the performance when using real data

alone compared to the addition of synthetic data, and (3) whether the generated images maintained a satisfactory sample quality. To address this, we evaluated the performance of classifiers trained on real data by incorporating additional training examples generated as either synthetic data or DP synthetic data to assess their impact. Our results demonstrated that the inclusion of synthetic data (both DP and non DP) did indeed lead to a small but noticeable improvement in predictive accuracy, with non DP synthetic data generally resulting in slightly higher accuracy compared to DP synthetic data due to the utility loss introduced as a result of the noise added to ensure Differential Privacy. In addition, the overall quality of the generated samples, as evaluated by the IS score, was shown to be quite competitive with benchmark values. Overall, we proved that it is possible to achieve comparable results using both non-DP and DP data, but naturally, neither of these can match the performance of real data, but the DP synthetic data provides a privacy preserving alternative that still enables meaningful downstream analysis and supports data sharing initiatives.

These findings have broader implications for medical research, as the generation of DP synthetic data paves the way for enabling secure data sharing across institutions, and fostering collaborative innovation with a reduced risk of exposing sensitive patient information. As of date, a number of efforts and platforms such as MDClone's Synthetic Data Engine [73] exemplify ongoing efforts to accelerate research by transforming electronic health records into synthetic datasets that preserve statistical integrity. However, while MDClone ensures privacy through heuristic methods that prevent re-identification, it does not explicitly implement formal frameworks like Differential Privacy, or at least doesn't explicitly claim to do so. Incorporating DP into such platforms in the same way that was achieved in this work could possibly enhance their ability to meet the strict privacy standards and provide measurable assurances of data protection.

In addition, we also highlight the key trade-off between utility and privacy, as the utility loss introduced by DP mechanisms remains a limitation, particularly in high dimensional medical imaging data. The 'hybrid synthetic datasets' used in this work which combine real and synthetic records in a sense was proven to be an efficient workaround to introduce privacy while still retaining valuable levels of utility. This suggests, however, that future work should explore the development of DP mechanisms that can generate synthetic data with sufficient utility even when used as stand alone, without relying on real data. Achieving this would reduce the dependence on hybrid datasets and help make synthetic data more viable for such applications.

Furthermore, while the authors of the original work [72] encouraged extending the ICW-fMRI-GAN framework to other datasets, the scope of our work was more narrowly focused on the integration of privacy mechanisms and the evaluation of their impact within the specific context of fMRI data. This suggests that testing this architecture on a wider range of datasets can offer a broader perspective on the generalizability of the framework and could further validate its applicability across diverse datasets, along with possibly yielding better results in the application of differential privacy, potentially improving both sample quality and the strength of privacy guarantees.

From a technical standpoint, while we do acknowledge that GANs have demonstrated competitive advantages for synthetic data generation, diffusion models are increasingly emerging as the new gold standard in generative modeling due to their stability and capability to generate highly diverse outputs [74] [75]. For that reason, we believe it is worth investigating how similar architectures or emerging generative models might enhance the quality of synthetic data, and potentially lead to better models for capturing the complexity and nuances inherent in brain imaging data.

In conclusion, this thesis demonstrates the feasibility and usefulness of incorpo-

rating DP synthetic data into medical research and data-sharing initiatives. While several challenges are still a source of concern, the findings underscore the potential that privacy-preserving approaches offer as a means to mitigate and bridge data gaps and facilitate secure collaboration. It serves as a foundation for further innovation in privacy-aware synthetic data generation, paving the way for robust, secure, and impactful applications in healthcare technologies.

References

[1] G. H. Glover, "Overview of functional magnetic resonance imaging", *Neurosurgery Clinics*, vol. 22, no. 2, pp. 133–139, 2011.

[2] P. M. Matthews and P. Jezzard, "Functional magnetic resonance imaging", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 1, pp. 6–12, 2004.

[3] B. Dubois, H. H. Feldman, C. Jacova, *et al.*, "Research criteria for the diagnosis of alzheimer's disease: Revising the nincds–adrda criteria", *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[4] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in alzheimer disease", *Nature reviews neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[5] A. Tessitore, F. Esposito, C. Vitale, *et al.*, "Default-mode network connectivity in cognitively unimpaired patients with parkinson disease", *Neurology*, vol. 79, no. 23, pp. 2226–2232, 2012.

[6] A. G. Garrity, G. D. Pearlson, K. McKiernan, D. Lloyd, K. A. Kiehl, and V. D. Calhoun, "Aberrant "default mode" functional connectivity in schizophrenia", *American journal of psychiatry*, vol. 164, no. 3, pp. 450–457, 2007.

[7] A. Di Martino, C. Kelly, R. Grzadzinski, *et al.*, "Aberrant striatal functional connectivity in children with autism", *Biological psychiatry*, vol. 69, no. 9, pp. 847–856, 2011.

[8] Aalto University. "Finnish biomedical imaging node accepted to the euro-bioimaging research infrastructure". Accessed: 11-11-2024. (Dec. 2020), [Online]. Available: `https://www.aalto.fi/en/news/finnish-biomedical-imaging-node-accepted-to-the-euro-bioimaging-research-infrastructure`.

[9] W. Yin, L. Li, and F.-X. Wu, "Deep learning for brain disorder diagnosis based on fmri images", *Neurocomputing*, vol. 469, pp. 332–345, 2022.

[10] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction", *CNS neuroscience & therapeutics*, vol. 24, no. 11, pp. 1037–1052, 2018.

[11] Y. Lui, P. Chang, G. Zaharchuk, *et al.*, "Artificial intelligence in neuroradiology: Current status and future directions", *American Journal of Neuroradiology*, vol. 41, no. 8, E52–E59, 2020.

[12] A. S. Jwa and R. A. Poldrack, "Addressing privacy risk in neuroscience data: From data protection to harm prevention", *Journal of Law and the Biosciences*, vol. 9, no. 2, lsac025, 2022.

[13] R. A. Poldrack, "Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding", *Neuron*, vol. 72, no. 5, pp. 692–697, 2011.

[14] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. [Online]. Available: `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN`.

[15] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation", in *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, Springer, 2006, pp. 486–503.

[16] C. Dwork, "Differential privacy", in *International colloquium on automata, languages, and programming*, Springer, 2006, pp. 1–12.

[17] M. Rezaei, T. Uemura, J. Näppi, H. Yoshida, C. Lippert, and C. Meinel, "Generative synthetic adversarial network for internal bias correction and handling class imbalance problem in medical image diagnosis", in *Medical Imaging 2020: Computer-Aided Diagnosis*, SPIE, vol. 11314, 2020, pp. 82–89.

[18] R. J. Ellis, R. M. Sander, and A. Limon, *Twelve key challenges in medical machine learning and solutions*, 2022.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets", *arXiv preprint arXiv:1411.1784*, 2014.

[20] P. Zhuang, A. G. Schwing, and O. Koyejo, "Fmri data augmentation via synthesis", in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, IEEE, 2019, pp. 1783–1787.

[21] K. J. Gorgolewski, G. Varoquaux, G. Rivera, *et al.*, "Neurovault. org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain", *Frontiers in neuroinformatics*, vol. 9, p. 8, 2015.

[22] A. Yousefpour, I. Shilov, A. Sablayrolles, *et al.*, "Opacus: User-friendly differential privacy library in PyTorch", *arXiv preprint arXiv:2109.12298*, 2021.

[23] D. Mack, *A simple explanation of the inception score*, Accessed: 17-10-2024, 2019. [Online]. Available: `https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a`.

[24] Finnish Ministry of Social Affairs and Health, *Act 552/2019 on the secondary use of health and social data, 2019*, Accessed: 25-03-2024. [Online]. Available: `https://stm.fi/en/secondary-use-of-health-and-social-data`.

[25] J. W. de Kok, M. Á. A. de la Hoz, Y. de Jong, *et al.*, "A guide to sharing open healthcare data under the general data protection regulation", *Scientific data*, vol. 10, no. 1, p. 404, 2023.

[26] P. Movahedi, V. Nieminen, I. M. Perez, T. Pahikkala, and A. Airola, "Evaluating classifiers trained on differentially private synthetic health data", in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2023, pp. 748–753.

[27] P. Movahedi, V. Nieminen, I. M. Perez, *et al.*, "Benchmarking evaluation protocols for classifiers trained on differentially private synthetic data", *IEEE Access*, 2024.

[28] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare", *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, 2021.

[29] M. Hernadez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions", *Methods of information in medicine*, vol. 62, no. S 01, e19–e38, 2023.

[30] A. Torfi, "Privacy-preserving synthetic medical data generation with deep learning", Ph.D. dissertation, Virginia Tech, 2020.

[31] J. Abowd, "Session 12b-recent advances in confidentiality protection–synthetic data", 2007.

[32] Y. Yao, X. Wang, Y. Ma, *et al.*, "Conditional variational autoencoder with balanced pre-training for generative adversarial networks", in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2022, pp. 1–10.

[33] Z. Han, Y. Wang, L. Zhou, *et al.*, "Contrastive diffusion model with auxiliary guidance for coarse-to-fine pet reconstruction", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 239–249.

[34] V. C. Pezoulas, D. I. Zaridis, E. Mylona, *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods", *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024, ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2024.07.005.

[35] K. I. Vaden Jr, M. Gebregziabher, M. A. Eckert, D. D. Consortium, *et al.*, "Fully synthetic neuroimaging data for replication and exploration", *NeuroImage*, vol. 223, p. 117 284, 2020.

[36] P. A. Osorio-Marulanda, G. Epelde, M. Hernandez, I. Isasa, N. M. Reyes, and A. B. Iraola, "Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review", *IEEE Access*, 2024.

[37] S. A. Khowaja, K. Dahri, M. A. Jarwar, and I. H. Lee, "Spike learning based privacy preservation of internet of medical things in metaverse", *IEEE Journal of Biomedical and Health Informatics*, 2023.

[38] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression", 1998.

[39] Y.-w. Kim, S. Mishra, S. Jin, *et al.*, "How transferable are video representations based on synthetic data?", *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 710–35 723, 2022.

[40] K. Emam, *Accelerating AI with synthetic data*. O'Reilly Media, Incorporated, 2020.

[41] M. Ye-Bin, N. Hyeon-Woo, W. Choi, N. Kim, S. Kwak, and T.-H. Oh, "Exploiting synthetic data for data imbalance problems: Baselines from a data perspective", *arXiv preprint arXiv:2308.00994*, 2023.

[42] Synthesized, *Solving data imbalance with synthetic data*, Accessed: 30-10-2024, 2021. [Online]. Available: `https://www.synthesized.io/post/data-imbalance`.

[43] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification", *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[44] J. Jordon, L. Szpruch, F. Houssiau, *et al.*, "Synthetic data–what, why and how?", *arXiv preprint arXiv:2205.03257*, 2022.

[45] I. M. Perez, P. Movahedi, V. Nieminen, A. Airola, and T. Pahikkala, "Does differentially private synthetic data lead to synthetic discoveries?", *arXiv preprint arXiv:2403.13612*, 2024.

[46] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models", *arXiv preprint arXiv:1705.07663*, 2017.

[47] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models", in *2017 IEEE symposium on security and privacy (SP)*, IEEE, 2017, pp. 3–18.

[48] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data-a privacy mirage", *arXiv preprint arXiv:2011.07018*, 2020.

[49] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy", *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[50] A. Kopp, "Microsoft smartnoise differential privacy machine learning case studies", *Microsoft Azure White Papers*, vol. 14, 2021.

[51] L. Ruthotto and E. Haber, "An introduction to deep generative modeling", *GAMM-Mitteilungen*, vol. 44, no. 2, e202100008, 2021.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets", *Advances in neural information processing systems*, vol. 27, 2014.

[53] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks", *arXiv preprint arXiv:1701.00160*, 2016.

[54] P. Salehi, A. Chalechale, and M. Taghizadeh, "Generative adversarial networks (gans): An overview of theoretical model, evaluation metrics, and recent developments", *arXiv preprint arXiv:2005.13178*, 2020.

[55] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks", *arXiv preprint arXiv:1701.04862*, 2017.

[56] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan", 2017.

[57] S. Kullback and R. A. Leibler, "On information and sufficiency", *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[58] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[59] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans", *Advances in neural information processing systems*, vol. 30, 2017.

[60] Shaistha Fathima, *Differential Privacy Definition*. [Online]. Available: https://medium.com/@shaistha24/differential-privacy-definition-bbd638106242.

[61] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!", *Journal of Privacy and Confidentiality*, vol. 9, no. 2, 2019.

[62] Joseph Near, David Darais, *Differential privacy: Future work & open challenges*, Cybersecurity Insights a NIST blog, Accessed: 07-11-2024, 2022. [Online]. Available: https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges.

[63] A. Jahn, *Andy's Brain Book*. 2022, Accessed: 28-08-2024. DOI: 10.5281/zenodo.5879293. [Online]. Available: https://andysbrainbook.readthedocs.io/en/latest/index.html.

[64] Opacus, *Guide to module validators and fixers*. [Online]. Available: https://opacus.ai/tutorials/guide_to_module_validator.

[65] Tal Yarkoni, *Neurosynth image decoder*. [Online]. Available: https://neurosynth.org/decode/.

[66] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns", *Neuroimage*, vol. 43, no. 1, pp. 44–58, 2008.

[67] C. Cortes, "Support-vector networks", *Machine Learning*, 1995.

[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, vol. 25, 2012.

[69] C. Wang, Z. Chen, K. Shang, and H. Wu, "Label-removed generative adversarial networks incorporating with k-means", *Neurocomputing*, vol. 361, pp. 126–136, 2019.

[70] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans", *Advances in neural information processing systems*, vol. 29, 2016.

[71] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[72] P. Zhuang, B. Chapman, R. Li, and S. Koyejo, "Synthetic power analyses: Empirical evaluation and application to cognitive neuroimaging", in *2019 53rd asilomar conference on signals, systems, and computers*, IEEE, 2019, pp. 1192–1196.

[73] MDClone, *Synthetic data for healthcare innovation*, Accessed: 02-12-2024. [Online]. Available: `https://www.mdclone.com/`.

[74] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis", *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[75] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, *et al.*, "Brain imaging generation with latent diffusion models", in *MICCAI Workshop on Deep Generative Models*, Springer, 2022, pp. 117–126.