Generating Synthetic Magnetic Resonance Images for Deep Learning Applications with Limited Labeled Data

Caio dos S. Felipe[1], Dieine E. B. Schiavon[1], Thatiane A. P. Alva[1], Carla D. L. Becker[1]

[1]Universidade Federal de Ciências da Saúde de Porto Alegre
Porto Alegre, RS, Brasil

{caio.felipe,dieineb,thatiane,carladiniz}@ufcspa.edu.br

Abstract. *This study addresses medical imaging data scarcity in AI development by using Progressive Growing of GANs (PGGAN) and transfer learning (TL) to generate synthetic MRIs of frontotemporal dementia variants. Our pre-trained foundation model achieved high anatomical fidelity (FID: 21.6008), while TL significantly improved image quality compared to models trained from scratch, reducing FID scores from ~115 to ~40. Although clinical validation is still needed, this framework offers a promising approach to expand limited medical datasets while maintaining anatomical accuracy for rare disease analysis.*

1. Introduction

Artificial intelligence (AI) integration in healthcare has experienced unprecedented growth, driving extensive research into the development and implementation of AI models in clinical settings. As these technologies advance, researchers are creating explainable systems that support healthcare professionals in complex medical decision-making while leveraging improved computational capabilities for medical diagnosis, treatment planning, and patient monitoring [Vonder Haar et al. 2023]. At the heart of these technological advances lies a fundamental requirement: high-quality medical data.

Data serves as the cornerstone of machine learning systems, with model performance directly reflecting the quality and quantity of training data. The importance of data quality manifests in multiple dimensions: accuracy of labels, representativeness of diverse patient populations, and adherence to standardized medical protocols. These factors become especially critical in healthcare applications, where model decisions can directly impact patient outcomes and safety [Ehsani-Moghaddam et al. 2021].

However, the healthcare domain presents unique challenges in data acquisition. The substantial costs associated with medical examinations, the complexity of expert annotation, and the regulatory frameworks governing medical data access create significant barriers to assembling comprehensive datasets. Furthermore, the inherent imbalance in medical data distribution, where certain conditions occur rarely compared to others, compounds the challenge of building representative datasets [Ranjbar et al. 2023].

The motivation of this study stems directly from addressing these fundamental data limitations in medical AI development. These constraints particularly affect Deep Learning approaches such as Convolutional Neural Networks (CNNs), broadly used to develop AI-assisted diagnostic tools. CNNs require substantial amounts of diverse training examples to learn robust feature representations and avoid overfitting. While traditional data augmentation techniques partially address the challenges of limited sample sizes, their effectiveness remains constrained by the inherent limitations of synthetic modifications.

In response to these challenges, Generative Adversarial Networks (GANs) have emerged as a promising solution for expanding medical datasets while preserving their clinical relevance. GANs offer the potential to synthesize realistic medical images that capture the complex distributions of anatomical structures and pathological conditions, moving beyond simple geometric transformations to generate entirely new, yet clinically plausible, examples [Celard et al. 2023].

However, training effective GANs for medical image synthesis presents its own challenges, particularly when working with limited data. Traditional GAN architectures require substantial amounts of training data to learn the complex distributions of medical images, making their application to rare conditions or limited datasets problematic.

The objective of this work is to propose a novel approach to leveraging pre-trained GAN models for synthetic MRI generation in limited-data scenarios. Specifically, we employ Progressive Growing of GANs (PGGAN), an architecture designed to generate high-resolution images by gradually increasing the resolution during training [Karras et al. 2018]. Using a PGGAN model initially trained on a comprehensive dataset as a foundation, we develop a methodology for fine-tuning the model on specific classes with limited examples. This transfer learning approach enables the preservation of general MRI image characteristics while adapting to the specific features of target conditions, even with minimal training data, representing a key contribution to the field and potentially opening new avenues for rare disease analysis.

1.1. Related Work

Recent studies have shown fine-tuning pretrained GANs outperforms training from scratch with limited data [Grigoryev et al. 2022]. Their analysis reveals that pretrained checkpoints enhance model coverage rather than sample fidelity, demonstrating that ImageNet-pretrained GANs can effectively initialize diverse target tasks despite potential initial quality limitations.

Mo et al. (2020) introduced FreezeD, selectively freezing lower discriminator layers during fine-tuning, showing significant improvements across multiple datasets using StyleGAN and SNGAN-projection architectures. This modification mitigates overfitting while optimizing performance under computational and data constraints.

Karras et al. (2020) developed adaptive discriminator augmentation, enhancing training stability in limited-data scenarios without architectural or loss function changes. Their approach achieved state-of-the-art performance with 1,000-5,000 images, improving CIFAR-10 FID scores from 5.59 to 2.42.

Despite extensive research on various GAN architectures for transfer learning in data-constrained environments, PGGAN's potential as a pre-training foundation remains underexplored. Our study addresses this gap by developing a methodology for leveraging pretrained PGGAN networks in class-specific image generation with limited data.

2. Materials and Methods

2.1. Experimental Overview

Our study employed a two-phase approach with an initial training of a foundation PGGAN model on the complete dataset, followed by the development of class-specific models through TL. This methodology was designed to evaluate the effectiveness of TL in generating high-quality synthetic MRI images for specific frontotemporal dementia (FTD) variants. In this study, the Fréchet Inception Distance (FID) [Heusel et al. 2017] is adopted as the primary quantitative evaluation metric due to its strong correlation with qualitative analyses and sensitivity to variations in image quality throughout the training and testing phases [Borji 2022].

2.2. Dataset Composition and Preparation

The study utilized data from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI/NIFD). The initiative is coordinated by UC San Francisco's Memory and Aging Center and supported by NIH Grant R01 AG032306, with data distributed by USC's Laboratory for Neuro Imaging. The complete dataset comprised 4,982 T1-weighted brain MRI images. The distribution of images across categories included 2,316 control images (CON), 624 images of non-fluent progressive aphasia (PNFA), 450 images of other FTD variants (OTH), 914 images of behavioral variant (BV), and 678 images of semantic dementia (SV).

The pre-processing pipeline consisted of three main stages. First, we performed image selection by extracting layers 70-75 from full brain volumes and converting DICOM files to PNG format for computational efficiency. Second, we implemented standardization procedures, including resolution normalization to 256×256 pixels, grayscale channel isolation, intensity normalization to the [0,1] range, and statistical standardization to achieve zero mean and unit standard deviation. Finally, we established progressive resolution stages, implementing a resolution hierarchy from 8×8 to 256×256 pixels.

2.3. Foundation model

The foundation model architecture progressively processes resolutions from 4×4 to 256×256. The generator upsamples from latent noise using residual blocks with pixel normalization, incorporating self-attention at 32×32 and 64×64 resolutions. The discriminator mirrors this structure with batch normalization, leaky ReLU activations, and minibatch discrimination to prevent mode collapse.

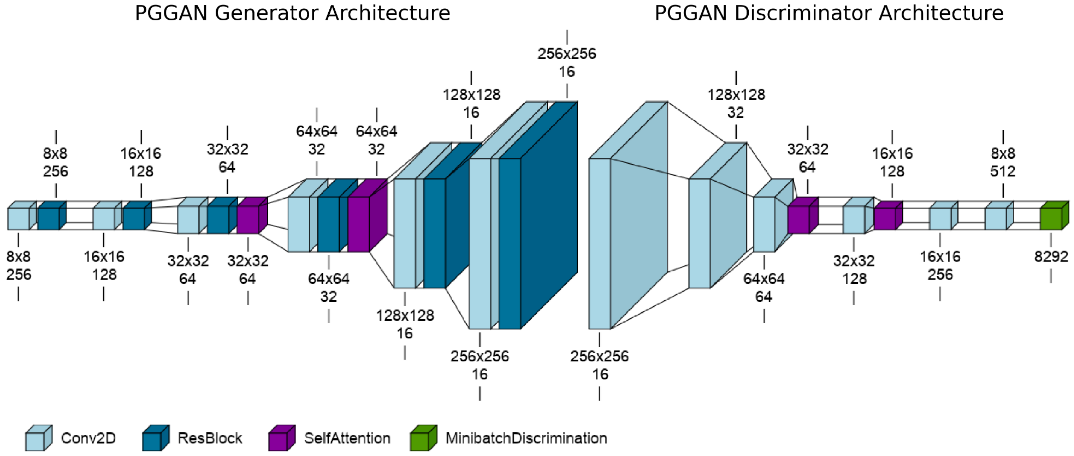

Figure 1. Visual representation of the foundation model architecture.

Training progresses through resolution stages with calibrated learning rates starting at 0.0002 for lower resolutions and decreasing to 0.00015 at higher resolutions. We implemented Wasserstein loss with gradient penalty (λ=10), exponential moving average (decay=0.999), and resolution-dependent sample allocation from 35,000 to 210,000 with decreasing batch sizes from 64 to 2. Smooth transitions use alpha-based fade-in mechanisms, while mixed-precision Adam optimization (β_1=0.5) with gradient clipping ensures stability.

2.4. Class-specific models

The architecture of the class-specific models remained identical to the foundation model, with differences limited to the training parameters. The generator and discriminator weights from the foundation model were initialized in all class-specific models. While no layers were frozen in the generator, the discriminator's convolutional and batch normalization layers corresponding to the lowest resolutions were frozen to preserve its ability to detect fundamental features in the generated images.

For the class-specific models, learning rates started at 0.0002 for lower resolutions and were gradually reduced to 0.00015 for 256×256, with a discriminator learning rate factor reaching 0.75 at the highest resolution. A gradient penalty coefficient of 10 was applied.

3. Results

The foundation model demonstrated the ability to generate high-quality images with high fidelity to anatomical structures and fine-grained details, achieving an FID score of 21.60. The FID calculation utilized the entire dataset of 4,960 real images alongside 4,960 generated images, ensuring a comprehensive assessment of the model's performance. Figure 2 illustrates the comparative performance of our models across different training approaches.

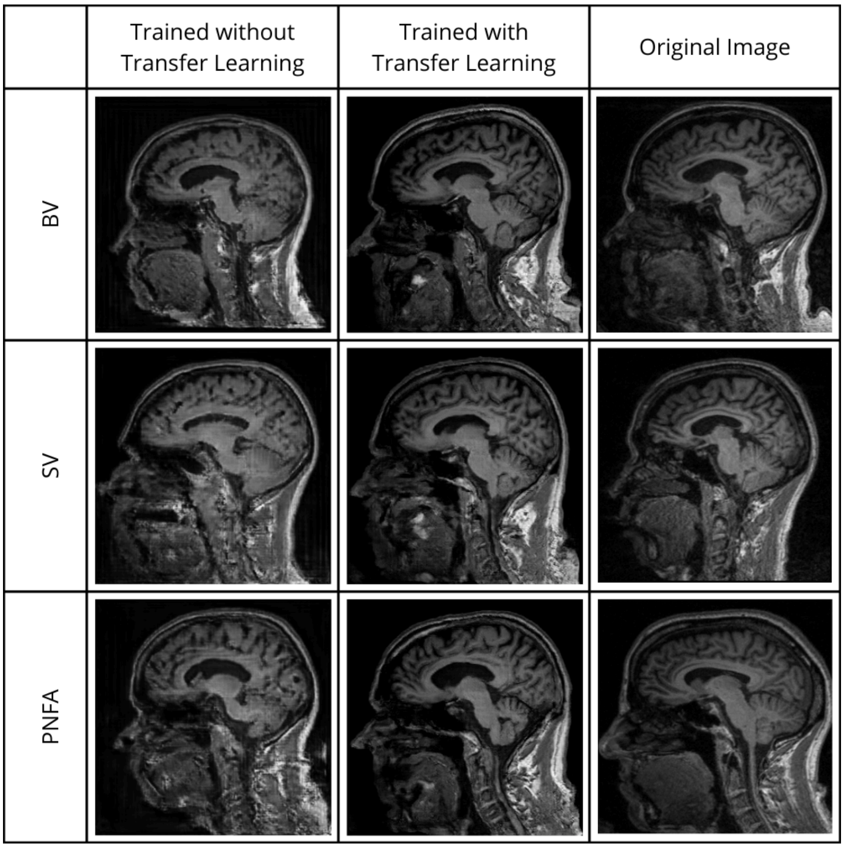

Figure 2. Comparison of generated images from models trained with and without the use of TL, displayed alongside the original images for clear visual comparison.

Table 1 highlights the FID results for both training approaches, demonstrating the clear advantages of TL in terms of both training efficiency and model performance. The models trained using the foundation model's weights consistently achieved significantly lower FID scores, indicating higher fidelity and better alignment with the real anatomical structures. This difference underscores the value of TL in accelerating training while maintaining or improving the fidelity of the generated images.

Table 1. FID comparison of class-specific models trained with and without the TL approach.

	FID without TL approach	FID with TL approach
BV	115.93	38.14
SV	113.31	42.18
PNFA	118.85	45.03

4. Conclusion

Our research demonstrates the significant potential of TL in medical image synthesis, particularly in scenarios with limited data availability. The success of our methodology across different FTD variants suggests broad applicability in specialized medical imaging contexts.

Several important directions for future research emerge from our findings. First, rigorous clinical validation by expert radiologists and neurologists is essential to ensure the diagnostic relevance and accuracy of the generated images in representing specific pathological features. Second, the foundation model could be further enhanced by incorporating additional anatomical constraints or domain-specific knowledge to improve consistency.

The implications of this work extend beyond FTD imaging to various medical conditions. Our approach could accelerate the development of AI-based diagnostic tools while reducing the computational resources required for training effective models. Future work should focus on establishing standardized protocols for synthetic medical image validation and exploring their integration into clinical AI training pipelines.

References

Vonder Haar, L., Elvira, T. and Ochoa, O. (2023) “An analysis of explainability methods for convolutional neural networks”, https://doi.org/10.1016/j.engappai.2022.105606.

Ehsani-Moghaddam, B., Martin, K. and Queenan, J.A. (2021) “Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data”, https://doi.org/10.1177/1833358319887743, January–May.

Ranjbar, A. and Ravn, J. (2023) “Data Quality in Healthcare for the Purpose of Artificial Intelligence: A Case Study on ECG Digitalization”, https://doi.org/10.3233/SHTI230534, June.

Celard, P. et al. (2023) “A survey on deep learning applied to medical images: from simple artificial neural networks to generative models”.

Karras, T., Aila, T., Laine, S. and Lehtinen, J. (2018) “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, https://arxiv.org/abs/1710.10196.

Grigoryev, T., Voynov, A. and Babenko, A. (2022) “When, Why, and Which Pretrained GANs Are Useful?”, https://arxiv.org/abs/2202.08937.

Mo, S., Cho, M. and Shin, J. (2020) “Freeze discriminator: A simple baseline for fine-tuning gans”, https://arxiv.org/abs/2002.10964.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2020) “Analyzing and improving the image quality of stylegan”.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S. (2017) “Gans trained by a two time-scale update rule converge to a local nash equilibrium”.

Borji, A. (2022) “Pros and cons of GAN evaluation measures: New developments”.

Frontotemporal Lobar Degeneration Neuroimaging Initiative (2010) “FTLDNI Database”.