POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Generating synthetic MR images using state-of-the-art deep learning approaches

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Sebastian Ballesteros Ramirez

Advisor: Prof. Francesca Ieva

Co-advisor: Luca Caldera

Academic year: 2024-2025

1. Introduction

Medical imaging has become an essential tool for diagnosis and biomedical research. The rise of deep learning, a branch of artificial intelligence, has enabled models that improve diagnostic accuracy across various medical imaging tasks, including disease detection and classification. However, training deep learning models requires large volumes of annotated data, which presents challenges in domains like medicine, where data collection and labeling are both costly and constrained by privacy regulations. In this context, data collection and annotation face several obstacles: the acquisition of high-quality imaging data is expensive, expert annotation is costly and time-consuming, and patient privacy regulations impose limitations on data sharing. Together, these challenges lead to small, heterogeneous datasets that limit the effectiveness of deep learning models. A further challenge in medical imaging datasets is class imbalance, where certain classes, such as healthy patients, significantly outnumber others. Class imbalance has proven problematic when training neural networks since it reduces classification performance.

To address these limitations, researchers have turned to data augmentation techniques, which artificially expand the training set by applying transformations or by generating synthetic data. In particular, deep learning–based generative models have been employed to synthesize realistic data, addressing the scarcity of annotated medical images and improving diagnostic model performance and robustness. Generative AI can also be used to mitigate class imbalance by oversampling the class with less representation.

Therefore, this thesis aims to develop a model capable of generating synthetic MR images by investigating two potential approaches:

1. Unconditional generation, where generated images are produced without conditioning on any specific label, is based only on its understanding of the internal distribution of the data.
2. Conditional generation, where images are conditioned on class labels e.g., the scanner type, the medical condition, etc., which influences the appearance of the scans.

By exploring these two approaches, the goal is to find the best model capable of generating synthetic MR images conditioned on a label, which, in this case, will be the scanner type. This provides a guideline for future researchers trying to augment their data in a biomedical setting. The models explored in this research include: Generative Adversarial Networks (GANs) [2], Conditional GANs [5], Variational Autoencoders (VAEs) [4], Conditional VAEs [1], Gaussian Mixture VAEs, CycleGANs. These models were chosen because they represent

the core families of generative frameworks currently applied to medical image synthesis: adversarial approaches, which succeed in producing sharp and realistic images; autoencoding approaches, which provide interpretable latent representations given a large input space; and hybrid approaches, which combine the strengths of both paradigms. By evaluating these models side by side, this work aims to determine their suitability for conditional MRI generation. Based on this comparative analysis, the thesis proposes an improved conditional VAE architecture with a log-cosh loss function, designed to enhance generation quality while maintaining strong conditioning on scanner type.

2. Literature Review

Generative models have emerged as a fundamental area of research in machine learning, driving advances in fields such as computer vision, natural language processing, and medical imaging. Generative Adversarial Networks consist of two models, called the generator and the discriminator, competing against each other. The generator learns to synthesize realistic images, whereas the discriminator learns to detect synthetic images from real images. In this way, both models improve their capabilities and eventually, if learning is successful, the generator will be able to generate realistic images not too different from the original ones. GANs set the ground for future research in generative models. Nevertheless, just a year before, another important generative model had been introduced by Diederik P. Kingma and Max Welling, the Variational Autoencoder (VAEs). VAEs differ fundamentally from GANs in their architecture and training approach. They are composed of two main parts, the encoder and the decoder. The encoder learns to map the input into a reduced latent dimension, and the decoder learns to map this latent dimension to a reconstructed version of the original input. Conditional extensions of GANs and VAEs enable the generation of images guided by specific class labels or features.

Since the inception of GANs, researchers have applied these techniques to the biomedical sector. Notably, by generating, enhancing, and synthesizing biomedical images such as digital mammography, MR images, ultrasound, etc. For example, a GAN-based approach was employed to generate resized sagittal MRI scans using the BRATS 2016 dataset, achieving satisfactory results for 128×128 images. The evaluation was conducted through a Visual Turing Test, in which a physician assessed the realism of the generated images. Subsequently, Zafar et al. proposed a dual-stream contrastive latent projection GAN (DSCLPGAN) for MR image data augmentation, reporting competitive performance based on quantitative metrics such as the Fréchet Inception Distance (FID). Variational Autoencoders (VAEs) and their conditional variants (cVAEs) have also been explored for MR image generation. For instance, Zhang et al. applied a conditional VAE to correct spatial inhomogeneity in chemical exchange saturation transfer MRI, training the model with the target spatial inhomogeneity as a conditional variable. VAEs offer interpretable latent representations and relatively stable training compared with GANs. Nonetheless, the generated images often suffer from blurriness due to the compression of input data into a latent space. Recent research has explored diffusion models and hybrid architectures for conditional MR image generation. Datasets include IXI and OASIS for healthy and dementia subjects, BRATS2021 for glioblastoma, and ISBI for multiple sclerosis. These studies reported competitive results based on FID metrics.

Across GANs, VAEs, and diffusion-based methods, each framework exhibits distinct strengths and limitations in the context of MR image synthesis. GANs are capable of generating visually sharp and realistic images but often require careful architecture design and may suffer from training instability. VAEs provide a principled framework for modeling MR images with conditional inputs and capturing latent structure. Their main limitation is the lower sharpness of generated images compared with GANs, which can reduce their practical utility in certain medical applications. Overall, prior work demonstrates significant progress in generating MR images, including pathology-conditioned synthesis; however, there remains a lack of exploration in scanner-conditioned MR image generation, motivating the present study to investigate models capable of producing MR images conditioned on scanner type for improved data harmonization and augmentation.

3. Evaluation metrics

The Fréchet Inception Distance (FID) [3] is a widely used metric for evaluating synthetic images in the context of Generative AI. Both the set of real images and the set of generated images are embedded into a feature space. The distributions of these feature embeddings are then approximated as multivariate Gaussian distributions. FID calculates the distance between the real and generated feature distributions by comparing their means and covariances. In other words, it quantifies how similar the overall statistics of the generated images are to those of the real images. In the context of synthetic MRI generation, FID provides a quantitative measure of how closely the distribution of generated im-

ages matches that of real MR images, capturing both global structural patterns and statistical consistency. Lower FID values indicate a closer match between the generated and real image distributions. While FID is widely adopted, it primarily captures global statistics and may not fully reflect fine-grained local anatomical details, which are critical in medical imaging. Nonetheless, it remains a standard benchmark for evaluating generative models in this domain. Interpreting FID scores requires considering the application domain. In general computer vision tasks, values around 100 are considered good quality, and values above 150–200 are often associated with low-quality results. In contrast, medical imaging studies often operate with different requirements: research has shown that FID values around 60–70 start reflecting realistic synthetic images.

4. Dataset

The Information eXtraction from Images dataset (IXI) and the Southwest University Adult Lifespan Dataset (SALD) were the two main datasets used to train the generative models. IXI is a well-known dataset collected at three different hospitals by Imperial College London. It consists of 581 MR images from healthy subjects. The collection details of these MR images are critical since each hospital used different scanner types. In particular, two scanners (*Intera* and *Gyroscan Intera*) from manufacturer *Philips*, and one *Unspecified* scanner model (GE manufacturer). In addition, the Southwest University dataset consists of 494 healthy subjects acquired from a single scanner, *Magnetom TrioTim*, from the *SIEMENS* manufacturer.

In total, there are $1,075$ MR images coming from a set of four different scanners that will be used as our condition labels: *Gyroscan Intera*, *Intera*, *Magnetom TrioTim*, and an *Unspecified* scanner. The detailed information is shown in table 1.

Dataset	Scanner Type	Image Count
IXI	Gyroscan Intera	322
	Intera	185
	Unspecified	74
SALD	Magnetom TrioTim	494
		1075

Table 1: Scanner details for the IXI and SALD datasets

4.1. Pre-processing

Each scan is stored as a 3D volume with dimensions $182 \times 218 \times 182$ voxels. Since the central slices contain the most relevant anatomical information, such as brain structures and tissue composition, whereas peripheral slices mainly capture boundary regions with limited diagnostic value, only the central portion of each scan was used. This choice also helps reduce noise and variability introduced by less informative peripheral slices. Hence, for each 3D volume, the 10 central axial (transverse) slices were extracted, with each slice being a 2D image of size 182×218 pixels. Therefore, the dataset consists of a total of $10,750$ images of size 218×182. Examples of the final images, categorized by scanner type, can be seen in Figure 1 and Figure 2. The similarities between the MR images are immediately noticed (e.g., shape, noise levels). However, by looking more closely, differences in contrast, intensities, and sharpness can be perceived. Notably, images recorded with *Gyroscan Intera* show a higher contrast, while the images recorded with *Intera* are slightly brighter across slices. Lastly, the images coming from the *Unspecified* scanner exhibited lower sharpness, likely due to acquisition differences.

The final dataset consisting of $10,750$ images was randomly shuffled prior to splitting. A standard data split was applied, separating 90% of the images for training and the remaining 10% for testing. This resulted in a training set consisting of $9,675$ images and a test set containing $1,750$ images.

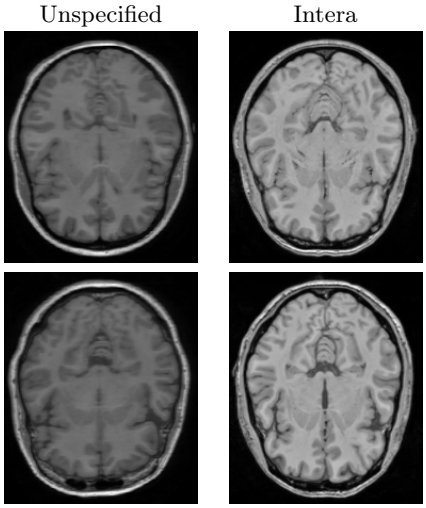

Figure 1: Selected MR images from the *Unspecified* and the *Intera* scanner type.

5. Proposed Model

Given prior discussions, the conditional VAE emerged as the most suitable model, considering the training time, the quantitative and qualitative results, and the overall simplicity of the model. Nevertheless, the VAE's loss function was adjusted, as a standard conditional VAE continues to produce slightly blurred images. This modification consisted of replacing the standard Mean Squared Error (MSE) loss with the log-cosh loss, which behaves

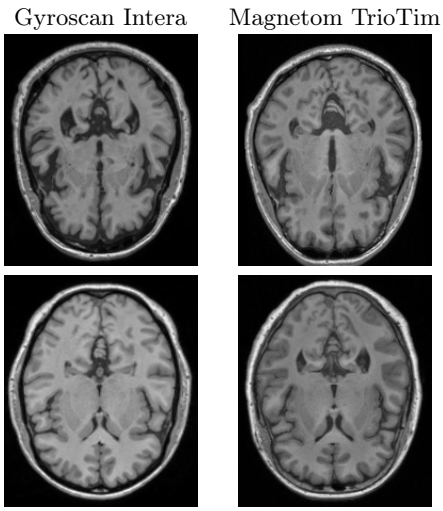

Figure 2: Selected MR images from the *Gyroscan Intera* and the *Magnetom TrioTim* scanner type.

similarly to MSE near zero but is less sensitive to outliers, encouraging sharper and more stable reconstructions without over-penalizing larger errors.

5.1. Architecture

The architecture is divided into three main components: the encoder, the reparameterization module (which implements the reparameterization trick discussed in section 2), and the decoder. The encoder takes the concatenation of the images (218 × 182) and their labels (MRI scanner type). In other words, the representation of the 218 × 182 images is flattened into a vector concatenated with the one-hot encoded scanner labels. The decoder reconstructs images from the latent vector z, sampled using the reparameterization trick, concatenated with the one-hot label vector. This vector is passed through a mirror architecture of the encoder, which consists of linear and ReLU layers and ends with a sigmoid activation to ensure output values are within the [0,1] range, representing pixels.

The design of the proposed conditional VAE follows a fully connected architecture rather than a convolutional one. This choice was motivated not only by empirical results obtained during model development but also by computational constraints, particularly the limited availability of GPU resources. Fully connected layers kept the model lightweight and computationally efficient, allowing training within the hardware limits. Convolutional variants were implemented and evaluated under comparable training protocols. However, these models demanded substantially more memory and longer training times, which were challenging given the restricted GPU capacity, and did not yield improved reconstructions in the experiments. A potential drawback of the fully connected design is its reduced ability to capture local spatial patterns compared to convolutional architectures, which are typically better at modeling spatial hierarchies in image data. Moreover, the latent dimensionality was set to 256. Since the original image size was kept, the latent dimension was a compromise between expressiveness and regularization. A smaller latent space risks underfitting and losing important information, while a much larger latent space can lead to overfitting and unstable training. This value is consistent with related works in the literature, where latent sizes of 100 and 200 are commonly used for medical imaging tasks.

A log-cosh loss is used instead of mean squared error. It has been proven that the log-cosh loss improves the reconstruction mechanism without jeopardizing the latent space optimization, balancing reconstruction accuracy with the quality of generated samples. The images generated by the standard VAE exhibit blurry features and fuzzy edges and shapes. In contrast, the images generated using the log-cosh VAE are noticeably sharper, with clearer features and improved contrast. This enhanced sharpness suggests that the log-cosh loss encourages more detailed and sharper results, and it offers visually improved reconstructions that are more similar to the original input images, regardless of the scanner.

Overall, these architectural decisions reflect a balance between computational feasibility, stability, and alignment with prior research. One plausible explanation for the observed superiority of the fully connected configuration is that scanner-conditioned generation in this dataset relied predominantly on global intensity and contrast characteristics (e.g., scanner-specific contrast), which are more directly modeled by dense connections spanning the whole image. In contrast, convolutional filters emphasize local texture and edges and, given the available dataset size and the need to preserve full-resolution images, tended to overfit local noise or required aggressive downsampling. Consequently, the fully connected design provided a more favorable trade-off between reconstruction fidelity and computational cost.

5.2. Training

The final model was trained for 100 epochs with a batch size of 8 (due to memory limitations). An Adam optimizer [27] with a learning rate of 0.0003 was used. The training and test loss curves of the proposed conditional VAE model trained with the log-cosh loss function over 100 epochs are presented in 3. Both losses are significantly improved as the

training progresses. The training loss seems to reach a plateau around epoch 90, whereas the test loss exhibits more spikes and variability. In addition, the gap between training and test losses remains relatively small, suggesting good generalization and limited overfitting. These results show that the proposed model successfully learned to minimize the loss function during training.

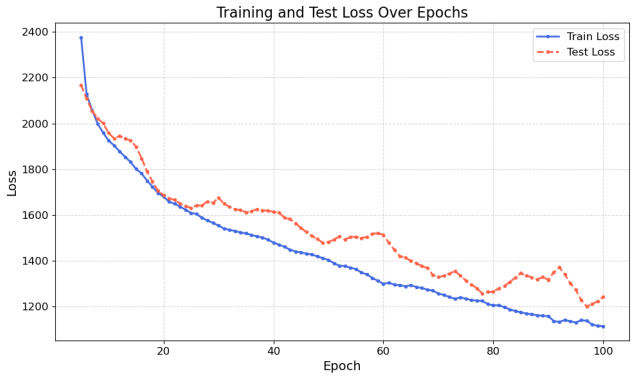

Figure 3: The train and test losses of the proposed conditional VAE model with a log-cosh loss over 100 epochs.

6. Results

The FID results, organized by scanner type, are presented in Table 2 alongside the reported scores two models of a similar work (DCGAN and WGAN-GP). To enhance the reliability of our evaluation, each FID score was computed 20 times, with the mean and standard deviation reported.

Scanner Type	$FID_{ImageNet}$ $(\mu \pm \sigma)$
Unspecified	87.19 ± 0.92
Intera	117.32 ± 1.31
Gyroscan Intera	135.66 ± 0.85
Magnetom TrioTim	127.37 ± 0.75
Model	$FID_{ImageNet}$
DCGAN	224.43
WGAN-GP	300.52

Table 2: FID scores of the generated images by scanner type compared against other models.

Among the scanner-specific results, images generated for the *Unspecified* scanner type achieved the best FID score (87.17 ± 0.92), indicating the lowest distance between the real images' distribution. For a better visualization, the FID scores are also presented in a boxplot format in Figure 4. This plot highlights the distribution and variability of FID scores across different scanner types. As shown, the *Unspecified* scanner type consistently results in the lowest and most stable FID score, while the *Gyroscan Intera* exhibits the highest scores. The boxplot makes it easier to compare not only the average performance but also the consistency of image quality across scanners.

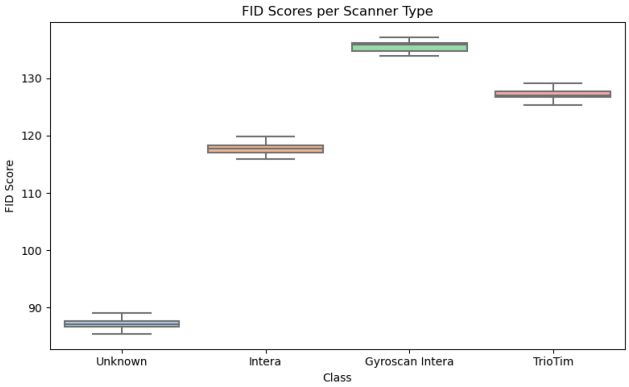

Figure 4: A boxplot depicting the FID scores of the generated images by scanner type.

Selected MR images directly generated from the latent space are presented in Figure 5 and 6, which are also categorized by scanner type for easier visualization.

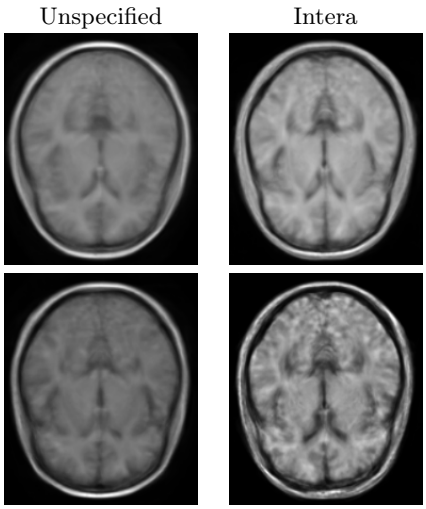

Figure 5: Selected synthetic MR images generated by the proposed conditional variational autoencoder from the *Unspecified* and *Intera* scanner type.

The model captures key anatomical features consistently across different scanner types, with realistic contrast and similar features. Images generated for the *Unspecified* scanner type show the same characteristics as the original images (e.g., unclear and blurry). In contrast, samples from the *Intera* and *Gyroscan Intera* exhibit more variability in texture and intensity, with *Intera* showing brighter images, as expected. Finally, the *Magnetom TrioTim* images demonstrate anatomical alignment with the original images shown in 4.1. Overall, while the image quality differs slightly across scanner types, the results in-

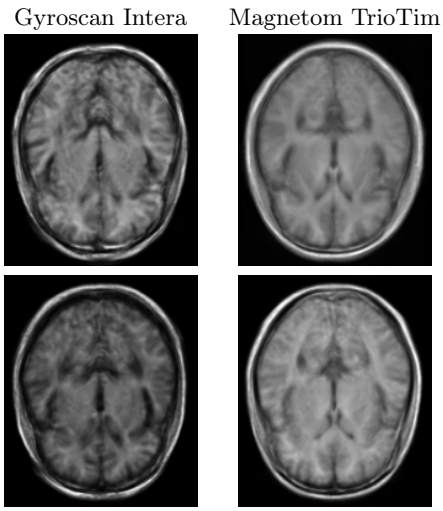

Figure 6: Selected synthetic MR images generated by the proposed conditional variational autoencoder from *Gyroscan Intera* and *Magnetom TrioTim*

dicate that the model effectively conditions on scanners and generates images that mimic the features of real MR images.

7. Conclusions

This thesis explored the state-of-the-art of deep conditioned generative models for the synthetic generation of brain MR images. Multiple generative models were contemplated as possible candidates. Notably, the conditional GAN, the CycleGAN, the conditional VAE, and the Gaussian Mixture Variational Autoencoder (GMVAE). After carefully analyzing the preliminary results and overall efficacy of the models, the final proposed model was a conditional VAE architecture trained using a log-cosh reconstruction loss to address the blurry results typically observed in VAEs.

Quantitative results, assessed using Fréchet Inception Distance (FID) with Inception-V3 pre-trained on ImageNet, demonstrate that the final model is capable of generating realistic images . Specifically, our conditional VAE achieved scanner-conditioned FID scores ranging from 87.19 to 135.66, positioning itself as a competitive candidate for conditional generation. These scores suggest an improvement in generating high-quality synthetic images that follow the distribution and the feature of specific scanner types. Additionally, visual assessments further confirm the quality of our outputs: images generated directly from noise exhibit clear anatomical structures and reflect scanner-specific characteristics. Nevertheless, some limitations were encountered along the process. Generated and reconstructed images still suffer from a degree of blurriness, and the model's ability to preserve very detailed anatomical features varies across scanner domains. Another limitation relates to evaluation: while FID provided a general measure of image quality, it relies on ImageNet-trained features that are not specific to the medical domain. As such, the current assessment may not fully capture clinical realism or subtle anatomical fidelity, highlighting the need for domain-specific evaluation metrics.

For future steps, several promising directions can follow this research. One evident extension is to explore more advanced generative frameworks such as diffusion models, which have recently demonstrated successful high-resolution image synthesis tasks, including medical imaging. Another important avenue for future work is the integration of direct validation from clinical experts to ensure that synthetic images meet practical diagnostic standards. This method sets the ground for conditional image generation, where conditioning is not limited to scanner type but can be extended to other relevant attributes, such as pathology, acquisition protocol, etc.

In conclusion, this work has demonstrated the feasibility and effectiveness of using conditional VAEs for scanner-specific synthetic MRI generation. While challenges remain, the strong quantitative and qualitative performance enables further exploration for conditional generation of medical imaging. Ultimately, bridging the gap between visual realism, clinical accuracy, and medical applicability will be key to transforming conditional generative models into reliable tools for healthcare.

References

[1] Mohamed Debbagh. Learning structured output representations from attributes using deep conditional generative models, 2023.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024.

[4] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.