



OPEN UNIVERSITY OF CATALONIA (UOC) MASTER'S DEGREE IN DATA SCIENCE

## MASTER'S THESIS

AREA: MACHINE LEARNING AND COMPUTER VISION IN HEALTHCARE  
AND MEDICAL APPLICATIONS

# Synthetic Data Generation for MRI Brain Tissue Segmentation

An Evaluation of Model Robustness and Generalization

---

Author: Anna Vidiella de Gonzalo

Tutor: Eloy Martínez de las Heras

Professor: Laia Subirats Maté

---

Barcelona, December 28, 2025



# Credits and Copyright

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International](#) (CC BY-NC-ND 4.0).



The official code repository of this Master's Thesis is licensed under [MIT license](#).



# FINAL PROJECT RECORD

Title of the project:	Synthetic Data Generation for MRI Brain Tissue Segmentation: An Evaluation of Model Robustness and Generalization
Author's name:	Anna Vidiella de Gonzalo
Collaborating teacher's name:	Eloy Martínez de las Heras
PRA's name:	Laia Subirats Maté
Delivery date (mm/yyyy):	01/2026
Degree or program:	Master's Degree in Data Science
Final Project area:	Area 3
Language of the project:	English
Keywords	Brain MRI, deep learning, synthetic data, image segmentation, model generalization



# **Dedication**

Al Carles, el professor que em va fer creure que podia assolir totes les fites acadèmiques que em proposés.



# Acknowledgements

I would like to express my sincere gratitude to everyone who supported me throughout my master's studies and during the development of this master's thesis.

First and foremost, I would like to thank Yesua, for their unconditional support throughout this journey. Their encouragement and continuous support were crucial to the successful completion of this thesis and master's degree.

I would also like to express my sincere gratitude to my supervisor, Dr. Eloy Martínez de las Heras, for his guidance and valuable feedback throughout the development of this project. His expertise was essential to the completion of this project, which has been a great opportunity for my academic growth.

I am also grateful to the NITRC team for granting access to the OASIS-3 dataset, which made the development of this work possible.

Finally, I would like to thank my family and friends for their encouragement and support throughout my studies.



# Abstract

Deep learning models have been very helpful in the medical field, especially for image analysis tasks such as diagnosis and tissue segmentation. However, their performance requires large, diverse training datasets to achieve good generalization. In the case of brain MRI, the availability of real data can be limited due to cost, patient privacy, or ethical concerns. In this scenario, synthetic data has emerged as an alternative for model training. This master's thesis proposes the generation of synthetic brain MRI images to evaluate the performance of a deep learning model in tissue segmentation tasks. The images were generated using a generative model that synthesizes MRI scans by sampling from a Gaussian Mixture Model, together with the corresponding segmentation labels for gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Domain randomization was implemented to introduce variability in contrast and anatomical shapes, with the aim of improving the model's generalization. Subsequently, an nnU-Net was trained exclusively on the generated synthetic images, and its performance was evaluated on real brain MRI scans. Although the full training pipeline could not be completed due to time constraints, the results show that the model achieves an acceptable performance, implying that synthetic data can be a viable strategy for model training when real brain MRI images are unavailable.

**Keywords:** Brain MRI, deep learning, synthetic data, image segmentation, model generalization.



# Resum

Els models d'aprenentatge profund han esdevingut eines de molta utilitat per al camp de la recerca mèdica, especialment per a les tasques d'anàlisi d'imatge com el diagnòstic i la segmentació de teixits. Tanmateix, el seu bon rendiment requereix conjunts de dades grans i diversos per aconseguir una bona generalització. En el cas de la ressonància magnètica cerebral, la disponibilitat de dades reals pot ser limitada degut al cost, la privacitat dels pacients o qüestions ètiques. En aquest escenari, les dades sintètiques han emergit com una alternativa a les dades reals per a l'entrenament dels models. Aquest treball de final de màster proposa la generació d'imatges sintètiques de ressonància magnètica cerebral per avaluar el rendiment d'un model de aprenentatge profund en tasques de segmentació de teixits. Les imatges s'han creat fent servir un model generatiu que sintetitza ressonàncies magnètiques a partir d'un model de Barreja Gaussiana, amb les màscares de segmentació corresponents per a la substància grisa (SG), substància blanca (SB) i líquid cerebroespinal (LC). S'ha implementat l'aleatorització de domini per introduir variabilitat en contrast i formes anatòmiques, amb l'objectiu de millorar la generalització del model. Una nnU-net s'ha entrenat exclusivament amb les imatges sintètiques generades, i posteriorment s'ha avaluat la capacitat de generalització del model sobre imatges de ressonància magnètica reals. Tot i no haver-se completat el procés d'entrenament en la seva totalitat a causa de restriccions temporals, els resultats mostren que el model assoleix un rendiment acceptable, suggerint que l'ús de dades sintètiques és una estratègia viable a l'hora d'entrenar models quan no hi ha dades reals de RM disponibles.

**Paraules clau:** RM cerebral, aprenentatge profund, dades sintètiques, segmentació d'imatges, generalització de models.



# Contents

<b>Abstract</b>	<b>ix</b>
<b>Table of Contents</b>	<b>xiii</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1 Context and Motivation . . . . .	3
2 Goals . . . . .	4
3 Sustainability, Diversity, and Social Challenges . . . . .	5
3.1 Sustainability . . . . .	5
3.2 Diversity, Gender and Human Rights . . . . .	6
3.3 Ethical Behavior and Social Responsibility . . . . .	6
4 Approach and Methodology . . . . .	7
5 Project Planning . . . . .	8
5.1 Available Resources . . . . .	8
5.2 Schedule . . . . .	9
<b>2 State of the Art</b>	<b>13</b>
1 Introduction to Medical Image Analysis . . . . .	13
2 Deep Learning Applications for MRI Segmentation . . . . .	14
2.1 Machine Learning Algorithms . . . . .	14
2.2 Convolutional Neural Networks and CNN-based models . . . . .	14
2.3 Other Architectures and Applications . . . . .	15
3 Synthetic Data for Model Training . . . . .	16
4 Synthetic Data Generation Methods . . . . .	17
4.1 Basic Image Transformation . . . . .	17

4.2	Computer Aided Design and Computer Generated Imagery . . . . .	17
4.3	Generative Adversarial Networks . . . . .	17
4.4	Variational Autoencoders . . . . .	18
4.5	Generative Diffusion Models . . . . .	18
5	Model Development with Synthetic Data . . . . .	18
<b>3</b>	<b>Methods and Resources</b>	<b>21</b>
1	Dataset . . . . .	21
1.1	Training Data . . . . .	22
1.2	Test Data . . . . .	23
2	Data Pre-processing . . . . .	23
2.1	Training Data . . . . .	23
2.2	Test Data . . . . .	24
3	Image Quality . . . . .	26
4	Modeling and Training . . . . .	27
<b>4</b>	<b>Results</b>	<b>29</b>
1	Quality of Synthetic MRI Scans . . . . .	29
2	Model Evaluation . . . . .	31
<b>5</b>	<b>Conclusions and Future Work</b>	<b>35</b>
<b>6</b>	<b>Glossary</b>	<b>37</b>
	<b>Bibliography</b>	<b>38</b>
<b>7</b>	<b>Appendices</b>	<b>45</b>

## List of Figures

1.1	Process diagram showing the different stages of CRISP-DM (Shearer 2000) . . . . .	9
1.2	Gantt chart for the project. . . . .	11
2.1	T1-weighted MRI slices of the brain in axial, sagittal, and coronal planes. Image extracted from OASIS-3 dataset (LaMontagne et al. 2019). . . . .	13
2.2	Synthetic brain MRI scans generated with <code>lab2im</code> (Billot et al. 2020). . . . .	20
3.1	Slices from an instance extracted from training set: a synthetic brain MRI and its corresponding segmentation mask, generated with <code>lab2im</code> (Billot et al. 2020). . . . .	22
3.2	Results of image pre-processing on real MRI, synthetic MRI and synthetic segmentation masks. . . . .	25
3.3	Image depicting the training metrics through this process, including loss and pseudo dice metrics per fold, throughout 1000 epoch. . . . .	28
4.1	Intensity densities for both real and synthetic MRIs. . . . .	30
4.2	Distribution of SSIM and PSNR per contrast. . . . .	31
4.3	Confusion matrices for each ROI (GM, WM and CSF). . . . .	32
4.4	Example of segmentation performed by nnU-Net on a T2-weighted MRI. From left to right, the real MRI, the ground-truth segmentation and the model segmentation. . . . .	33



# List of Tables

4.1	Overall voxel intensity statistics for real and synthetic MRI datasets.	29
4.2	SSIM and PSNR statistics across contrasts T1w and T2w.	30
4.3	Mean Dice score obtained for each fold.	31
4.4	nnU-Net segmentation metrics per class and foreground mean.	32



# Chapter 1

## Introduction

### 1 Context and Motivation

In the recent years, artificial intelligence (AI) has become a popular tool in many areas, but specially in the medical field. According to [Jiang et al. \(2017\)](#), the use of deep learning in the field of medical research nearly doubled in 2016, with imaging analysis being the most predominant area of application. Machine learning models rely on data, and the quantity and quality of it directly affects the success of ML-based applications ([Thambawita et al. 2022](#)). However, many real imaging datasets lack the diversity and scope to train robust models for computer vision tasks ([Gopinath, Hoopes, Alexander, Arnold, Balbastre, Billot, Casamitjana, Cheng, Chua, Edlow et al. 2024](#)). This is due to two reasons: economic cost and patient privacy.

Regarding the monetary cost associated to obtaining or processing medical images, the use and maintenance of MRI equipment is highly expensive. The overall expense includes not only the purchase of the scanner, but also the infrastructure, workforce salaries, maintenance, contrast agents, and the energy required to operate it ([Jalloul et al. 2023](#)). As a result, economic constraints often limit MRI availability to clinically prioritized applications, making it less accessible for research purposes. Moreover, even when medical images are obtainable, acquiring high-quality annotations is time-consuming and requires expert knowledge, further increasing the associated costs of obtaining labeled data for model training purposes ([Gopinath et al. 2024](#)).

As for the second reason, medical data may not be available due to privacy or ethical concerns. The first reason is because patients should give consent to allowing their data to be used for research purposes, which constraints the available data. The second reason is because medical data, even if given consensually, represents sensitive personal information, and as such, must be treated and protected according to legal and ethical standards ([Conduah et al. 2025](#)), which may prove a difficult task for some organizations. Furthermore, privacy concerns and

regulations can also vary across countries and regions, making it difficult to obtain data from certain areas, and to ensure diversity in medical datasets.

In this scenario, synthetic data becomes a valuable resource as a substitute for real data in image analysis tasks. First, it does not face the same privacy issues as real data ([Gopinath et al. 2024](#), [Paproki et al. 2024](#)). Second, the development of deep learning models dedicated to segmentation, such as SynthSeg ([Billot, Magdamo, Cheng, Arnold, Das & Iglesias 2023](#)), drastically reduces the costs associated to annotation ([Valabregue et al. 2024](#)). Lastly, even if the generation of synthetic images may present an elevated computational cost, it is not as expensive as real MRI, making synthetic data more accessible economically.

Taking this into consideration, the present thesis aims to address and evaluate the use of synthetic data to train models for brain MRI segmentation. The choice of this topic was heavily influenced by my academic background. As a researcher in the behavioral sciences, I recognize the high value of neuroimaging in many fields, including neuropsychology, clinical psychology, cognitive psychology, among others. However, as previously stated, access to high-quality MRI images is often limited.

Additionally, some research groups may lack access to advanced computational resources, such as dedicated GPUs or high-performance computing infrastructure, a limitation that also applies to certain clinical settings. This thesis also aims to address the lack of powerful computational infrastructure by proposing a methodology that remains feasible in resource-constrained environments.

In summary, through this master's thesis, I aim to contribute to behavioral sciences and medical research by applying deep learning techniques to address challenges I have previously encountered in my career, particularly the limited availability of real brain MRI data and the lack of access to advanced computational infrastructure.

## 2 Goals

The goals of this project are divided into primary and secondary objectives. The primary goals represent the general objectives of this master's thesis, while the secondary ones consist of the specific tasks and steps required to achieve the main goals.

Primary goals:

- To assess the potential of synthetic data as a substitute for real brain MRI images in training deep learning models when real data are unavailable. This will be done through evaluation of model generalization on real brain MRI images.

Secondary goals:

- Generate a set of synthetic brain MRI images. This set will contain MRI scans and their corresponding masks segmenting gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF).
- Build and train a nnU-Net with the generated data, then evaluate the model performance.

## 3 Sustainability, Diversity, and Social Challenges

### 3.1 Sustainability

Artificial intelligence has proven a powerful tool in many fields, especially in medical research and healthcare. However, one of the biggest concerns is the requirements for computational resources, energy demands, and the environmental impact associated with model training and deployment.

The present project involves three major computational tasks: synthetic image generation, data pre-processing and model training. They will be carried out using standard computing resources and infrastructure, such as a personal laptop and a personal computer (see section 5.1 for further information on the software and hardware resources). Thus, the development of this project does not require energy use or resource consumption beyond the typical research or academic activity.

Furthermore, the present project aims to assess the potential of synthetic data for segmentation model training purposes, aiming to explore a solution to resource-intensive real MRI, thus promoting innovation and optimization of resources in medical research.

Considering the previous statements, this project aims to address the following Sustainable Development Goals (SDG):

- SDG 7 - Affordable and Clean Energy: Decrease energy consumption in model development and synthetic image generation.
- SDG 9 – Industry, innovation, and infrastructure: Promote innovation and optimize resource utilization in medical research.
- SDG 12 - Responsible Consumption and Production: Optimize the use of computational resources and minimize the sustainability impact during model development.
- SDG 13 - Climate Action: Decrease energy and carbon footprint by optimizing resource use and energy consumption.

### 3.2 Diversity, Gender and Human Rights

The result of this project is primarily technical, as it focuses on the development and validation of technical tools rather than on social aspects. There may not be any direct positive or negative impact in terms of gender or diversity; good practices, nonetheless, will be followed to ensure the use of gender-neutral language, in accordance with the guidelines outlined in the *Gender Toolkit* of the Universitat Oberta de Catalunya ([Diez 2021](#)).

Moreover, this project does not directly involve human participants. All the datasets containing real data used are publicly available and do not include any identifiable information. Therefore, it does not have an impact on human rights, laws, or regulations, and is compliant with data protection standards and ethical research practices.

Finally, regarding accessibility, disability, and ergonomics, this project is considered to have no direct impact on those aspects, since it is conducted within an academic research environment. Given the purely technical scope of this project, it will not entail accessibility-related factors. However, best practices will be applied to all plots and figures to ensure accessibility for individuals with color blindness.

In summary, regarding diversity, this project aims to address the following Sustainable Development Goals:

- SDG 5 – Gender Equality: Commitment to use inclusive language throughout the development of the project and in any document related to it.
- SDG 16 – Peace, Justice, and Strong Institutions: By adhering to ethical standards and promoting transparency, accountability, and responsible practices in medical research, and by following good practices to make the manuscript accessible.

### 3.3 Ethical Behavior and Social Responsibility

While the result of this project is primarily technical, it may have a positive social impact regarding innovation related to medical research and data accessibility.

As its main objective is to evaluate the use of synthetic data as a substitute for real MRI for training models, a benefit could be found for agents such as research groups or clinicians who may not have access to neuroimaging resources or real MRI data. Additionally, synthetic data can be openly shared with fewer limitations than real data ([Gopinath et al. 2024](#)), becoming a useful resource as a substitute for real images in specific scenarios.

As stated in previous sections, real MRI data were used as ground truth for synthetic image generation and model evaluation in order to assess model generalization. These data were obtained from the NITRC repository, with access granted upon compliance with applicable data

protection and ethical guidelines. In addition, no personal, sensitive, or identifying information was used on this project, thereby protecting the privacy of individuals adhering to ethical standards in medical research. Good practices suggested by the Universitat de Catalunya's Research Ethics Committee were followed to ensure data protection throughout the whole project, covering the whole data lifecycle, ensuring correct management, protection, and deletion of the real data used to perform this work once the project is completed.

To ensure ethical behavior and social responsibility, this project aims to address the following Sustainable Development Goals:

- SDG 10 – Reduced Inequalities: Working on developing alternatives to real MRI data, which can be useful for research groups or clinical environments with low computational or MRI-related resources.
- SDG 16 – Peace, Justice, and Strong Institutions: By ensuring proper data management and patient privacy protection, this project promotes transparency, accountability, and responsible practices in medical research.

## 4 Approach and Methodology

The present project begins with a literature review on the state of the art regarding image synthesis, segmentation tools and the use of synthetic data for brain tissue segmentation, followed by a definition of a problem, which is the lack of real MRI data in certain contexts. A proposed approach to address this issue is the generation of synthetic data followed by model training and evaluation of model generalization on real data.

Considering the tasks described for the development of this project, a structured approach is required to ensure that it can be carried out successfully. Given its academic nature, it is necessary that the chosen methodology ensures reproducibility and transparency, while also allowing some degree of flexibility, not only to assume certain overlap between phases, but to also allow the possibility of going back and forth between them to refine aspects if adjustments are needed. For these reasons, the best-suited option is the CRISP-DM methodology, which combines a systematic structure while allowing to iterate throughout the phases as needed.

The CRoss-Industry Standard Process for Data Mining is a data mining model that defines a series of good practices to be applied in data mining projects ([Shearer 2000](#)). Despite the specificity of this methodology, it has been designed to be generic enough to be applied to any industry ([Gouvea 2021](#)). CRISP-DM organizes the processes into six distinct stages (see Figure 1.1). Considering the scope and objectives of this project, and the chosen methodology, the planning of the project would be the following:

- **Business Understanding.** This phase is centered on determining the objectives of the project, and developing a project plan. In the case of this thesis, this stage includes the revision of the available literature on the matter on hand. Additionally, it also includes the preparation of the state of the art review.
- **Data Understanding.** In this phase, the data mining process takes place. For this project, this includes the generation of a set of synthetic data, as well as requesting access to OASIS-3 dataset for real brain MRI images. Once the data mining process is completed, the exploration and verification of the data takes place.
- **Data Processing.** During this phase, images are prepared to be analyzed for quality assessment and for modeling, performing tasks such as skull-stripping, intensities normalization, etc.).
- **Modeling.** After data processing, the model (nnU-Net) is implemented. Once ready, the model is trained with the synthetic data.
- **Model Evaluation.** Once the model is trained, the model generalization is evaluated on real data.
- **Deployment.** This is the final phase of the project. This includes any work related to the model deployment, and the elaboration of the final report, which is the thesis manuscript. In the case of the present project, the thesis manuscript was written throughout all stages to comply with the different milestones set to follow the Continuous Assessment Test planning. However, this phase is crucial to write the final chapters and make additional modifications, so the manuscript serves as a final report of the project.

## 5 Project Planning

### 5.1 Available Resources

The software and hardware resources used to carry out this project are:

- Laptop with Intel(R) Core(TM) i7-10870H processor @ 2.20GHz, NVIDIA GeForce RTX 1650 VRAM 4GB, 16.0 GB RAM, and Windows 11.
- Personal Computer with AMD Ryzen 5 5600 6-core Processor @ 3.50GHz, NVIDIA GeForce RTX 3060 VRAM 12GB, and Windows 11.
- TeXstudio version 4.8.9 to write the thesis in LaTeX.

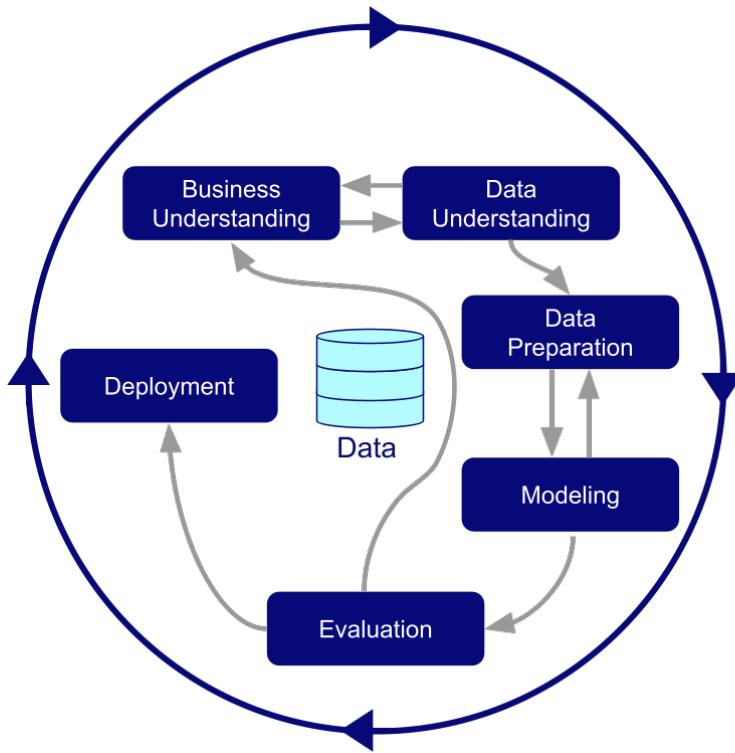


Figure 1.1: Process diagram showing the different stages of CRISP-DM ([Shearer 2000](#))

- Conda 24.11.3 as environment manager.
- PyCharm Community Edition 2024.3 as IDE to elaborate python files and programs.
- Jupyter Notebook version 7.4.5 with IPython kernel v.9.7.0 to execute specific coding tasks and analyses.
- Python versions 3.7.1, 3.10.19 and 3.12.12 in different conda environments created to avoid incompatibilities with tools and libraries.

Additionally, there are other resources that have been used to develop this project:

- Neuro Imaging Tools & Resources Collaboratory (NITRC), a web-based resource that offers resources and information related to neuroinformatics software and data, such as dataset OASIS-3.

## 5.2 Schedule

The project has been divided into six different stages according to the CRISP-DM methodology (see section 4), and its development has been scheduled according to the dates set for the

Continuous Assessment Tests. Taking all these factors into consideration, the project has been planned as follows:

- **Module 1: Thesis proposal.** This module includes the definition and planning of the master's thesis.
- **Module 2: State of the Art.** This module includes a review of the literature regarding the topic of the project, and includes the tasks under the business understanding phase, which are the literature research and the state of the art review.
- **Module 3: Implementation.** In this module, the tasks specific to the project are carried out. This includes the following stages and tasks:
  - Data Understanding, compromising all the tasks related to the data mining process, that is, the generation of synthetic data, the research for real data and the data overview.
  - Data Processing, in which all the necessary transformations are applied on the data so it can be correctly processed by the model.
  - Modeling, in which the nnU-Net is implemented and trained with the synthetic data.
  - Evaluation, in which the model evaluation is carried out.
- **Module 4: Thesis manuscript.** This module is focused on finishing the thesis manuscript, which is redacted throughout the different project stages. During this module, the model deployment is carried out, and the final chapters of the thesis are written concurrently.
- **Module 5: Audiovisual presentation.** This module includes the preparation of the oral presentation of the thesis, which should take place once the manuscript has been written.

The described schedule has been represented in a Gantt diagram in Figure [1.2](#).

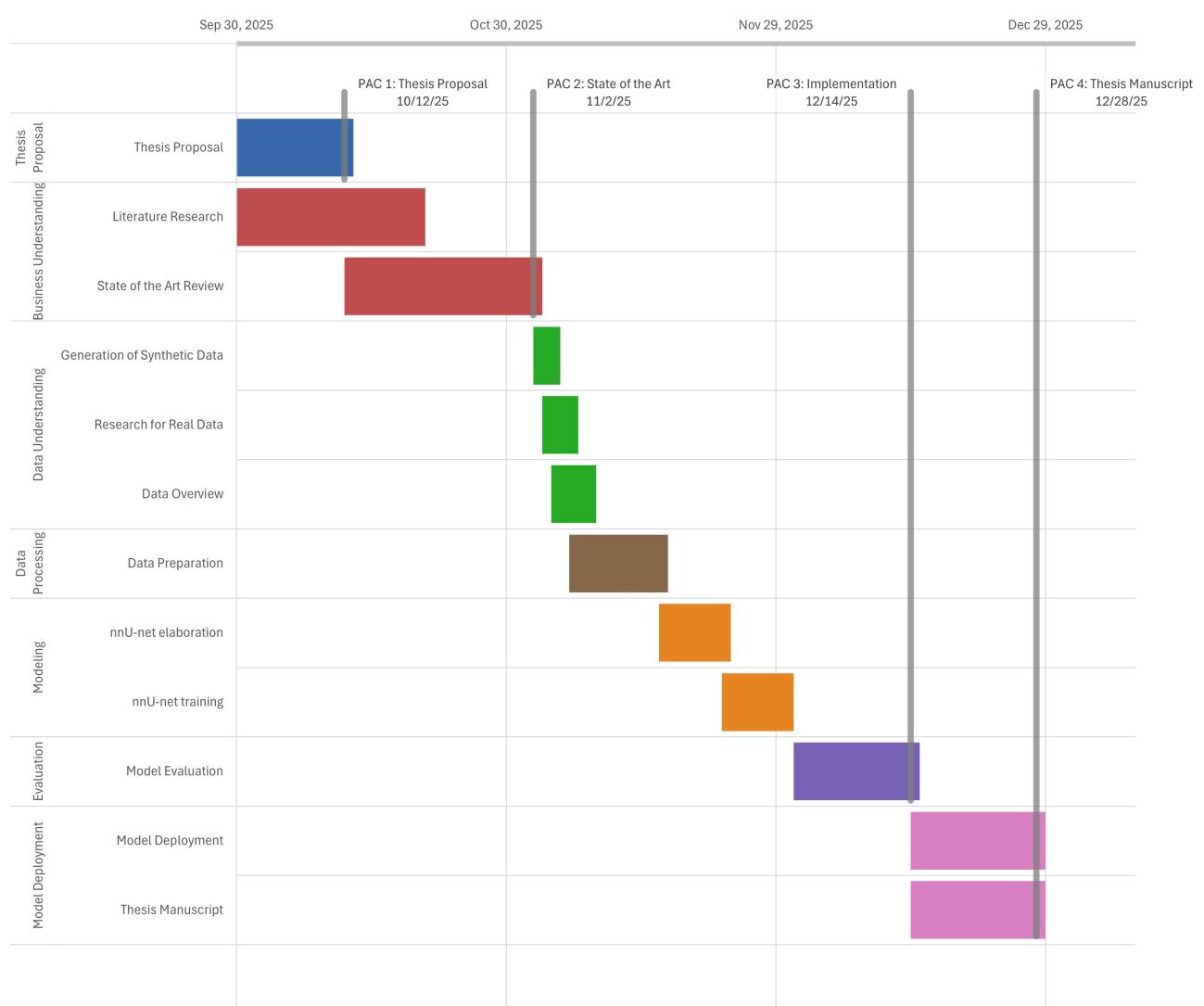


Figure 1.2: Gantt chart for the project.



# Chapter 2

## State of the Art

### 1 Introduction to Medical Image Analysis

Medical imaging plays a critical role in modern healthcare, allowing clinicians to visualize the internal structures and functions of the body non-invasively. In recent decades, medical imaging methods, including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), mammography, ultrasound, and X-ray, have played a crucial role in early detection, diagnosis, and management of various diseases ([Shen et al. 2017](#)).

For brain imaging, MRI remains the preferred modality for structural analysis, given the lack of ionizing radiation and its superior contrast resolution for soft tissues ([Chau et al. 2025](#)). In clinical and research contexts, MRI is used extensively in tasks such as detection of anatomical and cellular structures, tissue segmentation, and computer-aided disease prognosis and diagnosis ([Shen et al. 2017](#)). As this project has as its main topic brain image segmentation, this review will be focused on that specific task.

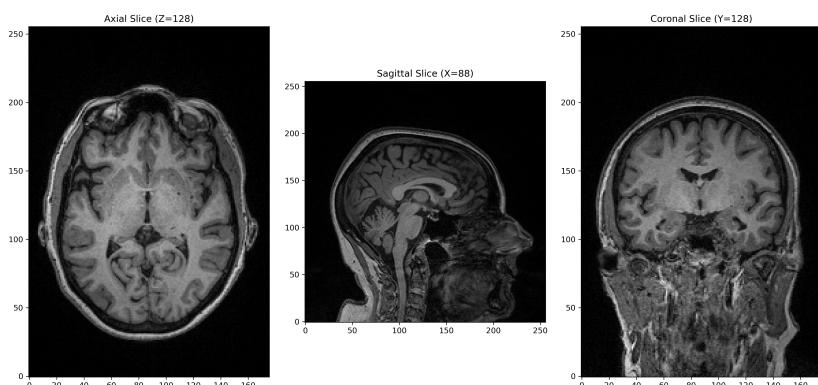


Figure 2.1: T1-weighted MRI slices of the brain in axial, sagittal, and coronal planes. Image extracted from OASIS-3 dataset ([LaMontagne et al. 2019](#)).

Medical image segmentation is the process of dividing an input medical image into meaningful areas to identify and isolate distinct segments considered regions of interest (ROIs), such as tissues, organs, or lesions. On brain imaging, segmentation has diverse applications such as delineating brain tumor boundaries in MRI images ([Soppari et al. 2024](#)); the identification of lesions and abnormalities associated with medical conditions ([Chau et al. 2025](#)); and the segmentation of GM, WM and CSF ([Chau et al. 2025](#), [Shen et al. 2017](#)). The latter is relevant for detecting atrophy and identifying morphological changes associated with neurodegenerative conditions such as Alzheimer’s disease, Parkinson’s disease, multiple sclerosis or psychiatric disorders like schizophrenia ([Shen et al. 2017](#)).

Given the wide variations in pathology and the potential fatigue of human experts, researchers, and doctors, the development of tools that can perform automated or semi-supervised analysis of these images has become a valuable resource in both clinical practice and medical research. The next section provides a general overview of the different techniques used on computer vision for brain MRI segmentation.

## 2 Deep Learning Applications for MRI Segmentation

According to the systematic review performed by [Chau et al. \(2025\)](#), ML models, convolutional neural network-based models (CNN) and FreeSurfer have been the AI-related tools more used for brain segmentation in the last decade. Other methods, although less popular, include Statistical Parametric Mapping (SPM), Voxel-Based Morphometry (VBM), and hybrid approaches.

### 2.1 Machine Learning Algorithms

While many machine learning algorithms can be applied to MRI Segmentation tasks, Support Vector Machine (SVM) and RandomForest, both classification algorithms, are the most frequently used, mainly to classify brain structures or predict outcomes based on segmented data. These methods are, however, not specifically designed for image segmentation, and thus they usually need to be combined with additional methods to perform a segmentation analysis such as Voxel-Based Morphometry ([Chau et al. 2025](#)).

### 2.2 Convolutional Neural Networks and CNN-based models

CNNs have proven a powerful tool for computer vision, especially for medical imaging analysis, thanks to their ability to directly learn hierarchical representations from raw pixel data ([Thakur et al. 2024](#)). In recent years, CNNs have demonstrated superior performance than ML

algorithms in brain MRI segmentation (Chau et al. 2025). Hence why it is one of the most popular methods.

The most prominent architecture of this kind in biomedical image segmentation is the U-Net, introduced by Ronneberger et al. (2015), a method designed for that specific purpose. Its symmetric encoder-decoder architecture includes a contracting path that captures contextual information and an expanding path that enables precise localization of the ROIs (Ronneberger et al. 2015).

Due to the greater suitability for segmentation tasks exhibited by U-Net, over the past decade different architectures have been proposed to improve its performance, efficiency, and adaptability across different data types, and hardware limitations. Some well-known examples include 3D U-Net (Çiçek et al. 2016), Attention U-Net (Oktay et al. 2018), Inception U-Net (Szegedy et al. 2015), Residual U-Net (He et al. 2015), Dense U-Net (Huang et al. 2018), nnU-Net (Isensee et al. 2018) and UNet++ (Zhou et al. 2018). More recently, novel architectures based on U-Net have been introduced, such as SSA U-Net (Jiang et al. 2024), DPF-Unet (Li et al. 2025), FFLUnet (Kundu et al. 2025), and NNDM (Makanaboyina 2025).

## 2.3 Other Architectures and Applications

Although not based on deep learning or widely adopted, there are alternative tools to approach brain MRI segmentation.

### 2.3.1 FreeSurfer

FreeSurfer is a software platform for neuroimaging analysis widely used as a tool for brain MRI segmentation, and has proven to be very powerful in the specific task of cortical and subcortical structures segmentation (Fischl 2012). Although it is not a standalone AI model or architecture, it incorporates AI-based methodologies, and given that it is in the top three most used segmentation methods in the last decade (Chau et al. 2025), it is worth mentioning as a state-of-the-art segmentation tool.

### 2.3.2 Statistical Parametric Mapping and Voxel-based Morphometry

Statistical Parametric Mapping (SPM) is a software package designed for the analysis of brain imaging data sequences developed by Friston (2003). Its integrated segmentation algorithm uses a Gaussian Mixture Model (GMM) that assigns each voxel a probability of belonging to a specific tissue based on its intensity and spatial location (Chau et al. 2025).

VBM is an application that performs a voxel-wise comparison of the local concentration of gray matter between two groups of subjects, finding patterns of differences between individuals

and a reference dataset ([Ashburner & Friston 2000](#)).

### 2.3.3 Hybrid Approaches

Hybrid approaches are those defined by the combination of different methods responding to different research needs. For instance, [Pérez-Millan et al. \(2023\)](#) performed gray matter volumes and cortical thickness measures using FreeSurfer, applied principal component analysis (PCA) for dimensionality reduction, and then used a SVM to classify patients with Alzheimer's disease, fronto-temporal dementia and healthy controls. Even with accessible and powerful deep learning methods such as U-Net, hybrid approaches still obtain valuable results and enhance segmentation accuracy in clinical cohorts ([Chau et al. 2025](#)).

## 3 Synthetic Data for Model Training

As discussed previously, access to medical data can be restricted due to privacy concerns, and obtaining or processing MRI images can be resource-intensive, making it challenging to obtain sufficient data for research purposes. Additionally, real world data can present class imbalance, considering that in many medical images presenting a medical condition are scarce in comparison to healthy samples ([Qu et al. 2020](#)).

An alternative approach to these limitations is the generation of synthetic data, which can be shared with less limitations ([Gopinath et al. 2024](#)), and used for data augmentation purposes ([Thambawita et al. 2022](#), [Paproki et al. 2024](#)), reducing class imbalance and dataset bias ([Sun et al. 2017](#)), and improving model generalization through the creation of different image variations.

One limitation often attributed to artificial data is its potential dissimilarity to real data, which can have an adverse effect on model generalization. The domain gap existent between real and synthetic data (sim2real) remains an open challenge when training models with synthetic data, as it often leads to overfitting to their data distribution and poor performance on real data ([Paproki et al. 2024](#)). However, there are methods that mean to tackle this issue, such as domain adaptation, transfer learning, multicenter collaboration ([Thakur et al. 2024](#)), or domain randomization.

An especially successful approach is the latter, domain randomization, introduced by [Tobin et al. \(2017\)](#). This method is based on the hypothesis that if the variability in the simulated data is great enough, models will generalize to the real world data with no need of additional training. In MRI analysis, domain randomization has been successfully applied in tasks such as structural segmentation, skull stripping, registration, feature extraction, image-to-image translation and super-resolution reconstruction ([Hoffmann 2025](#)).

In tandem with techniques aimed to address the domain gap, synthetic data represents a promising approach to overcome the limitations related to the lack of real MRI data. The following section presents the methods used for synthetic data generation.

## 4 Synthetic Data Generation Methods

Since 2015, a variety of generative models have been described for creating synthetic medical images ([Koetzier et al. 2024](#)). Amongst of the most popular methods are generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models. While GAN-based architectures are yielding the best results in the medical domain ([Thambawita et al. 2022](#)), other interesting approaches exist.

### 4.1 Basic Image Transformation

Image transformation is not exactly a form of synthetic data generation, falling within the category of semi-synthetic data generation methods. It involves creating new images from real images using operations like geometric transformations (such as rotations or cropping), color space transformations, or kernel filtering ([Paproki et al. 2024](#)).

### 4.2 Computer Aided Design and Computer Generated Imagery

These include all applications that create 3D surface models of artificial objects, usually with software tools such as Unity3D, Blender, or the Unreal graphic engine, to simulate specific situations, and then render imaging data ([Paproki et al. 2024](#)). Computer aided design (CAD) and computer generated imagery (CGI) have proven to be useful tools, as the created images can be tailored to specific applications. The image generation itself is, however, labor intensive. Some popular CGI based datasets include the surreal dataset ([Varol et al. 2017](#)) and the human 3.6M dataset ([Ionescu et al. 2014](#)).

### 4.3 Generative Adversarial Networks

GANs are among the most widely used architectures for synthetic image generation in the medical field, as they can produce highly realistic representations of image data. Their architecture consists of two different networks: a discriminator, which determines if an image is real or fake, and a generator, which attempts to create realistic images to fool the discriminator. The dynamic established between generator and discriminator allows to progressively produce more refined images with each iteration. Given the potential of this architecture, many variations of

the GAN architecture have been developed to surpass the limitations of vanilla GANs and allow for the synthesis of high-resolution images (Paproki et al. 2024), such as Deep-convolutional GAN (Radford et al. 2016) Laplacian GAN (Denton et al. 2015), pix2pix (Isola et al. 2018) and Cyclic GAN (Zhu et al. 2017).

#### 4.4 Variational Autoencoders

VAEs are a type of artificial neural network introduced by Kingma & Welling (2022). Their architecture is based on an encoder and a decoder, where the encoder maps input data to compress it into a lower dimensional space, and the decoder reconstructs the data to regenerate a plausible version of the original input. Based on this framework, Larsen et al. (2016) introduced a variation of VAEs in tandem with GANs named VAE-GANs, that combine both the generative capabilities of a VAE with the adversarial training of GANs.

#### 4.5 Generative Diffusion Models

Generative diffusion models are a family of generative models that apply a reverse diffusion process to synthesize images from random noise (Yeğin & Amasyali 2024). During the training phase, real images are gradually corrupted with Gaussian noise, and the model learns to reverse this process to recover detailed and realistic data representations (Koetzier et al. 2024). Although image generation is time-intensive, the high quality of the resulting images makes diffusion models a promising tool in the medical imaging field.

### 5 Model Development with Synthetic Data

Many high-performing tools related to MRI processing and analysis have been developed by training deep learning models on synthetic data (Gopinath et al. 2024).

- **SynthSR** (Iglesias et al. 2023). A CNN for super-resolution that turns a MRI scan or CT scan of any orientation, resolution and contrast into 1 mm isotropic Magnetization Prepared Rapid Gradient Echo (MP-RAGE). The method applied by the authors was based on applying random deformations to segmentation masks, which were later used to generate synthetic scans. This strategy exposes the model to a wide range of acquisition conditions during training, resulting in strong generalization across MRI sequences and high robustness to overfitting.
- **SynthStrip** (Hoopes et al. 2022): A deep learning tool for brain extraction (skull stripping) from MRI scans. It was entirely trained on synthetic data, enabling robust gen-

eralization across a wide range of real MRI scans, including both adult and newborn anatomies. SynthStrip has demonstrated accurate and robust brain extraction, outperforming baseline methods such as BET.

- **SynthMorph** ([Hoffmann et al. 2021](#)): A deep learning-based framework for MRI image registration trained on synthetic images and label maps with arbitrary contrasts and shapes, enabling accurate and robust registration across real MRI scans. It achieves performance that surpasses state of the art tools both within and across imaging contrasts.
- **EasyReg** ([Iglesias 2023](#)): A CNN for symmetric multi-modality registration of brain MRI that is also contrast-agnostic because of the training with synthetic data. This tool has demonstrated the same accuracy as classical registration methods, but is also much more accurate across modalities and resolutions thanks to domain randomization.
- **SynthSeg** ([Billot, Greve, Puonti, Thielscher, Van Leemput, Fischl, Dalca, Iglesias et al. 2023](#)): A CNN for segmentation trained on synthetic data, adopting a domain randomization strategy that randomizes image contrast and resolution to obtain a model capable of segmenting real scans from a wide range of target domain without requiring fine-tuning, while achieving great robustness to a wide range of morphological variability.

The current project is based on the methodology followed for the development of SynthSeg ([Billot, Greve, Puonti, Thielscher, Van Leemput, Fischl, Dalca, Iglesias et al. 2023](#)). The main objective is to train a model on synthetic MRI scans and corresponding segmentation masks while applying domain randomization. Through the inclusion of spatial deformations, different contrasts, orientations, and noise levels, the model learns not to rely on these features when estimating the output, becoming agnostic to them. To carry out these tasks, this work uses Gaussian Mixture Models for synthetic image generation and a nnU-Net to implement a brain MRI segmentation model.

Images are generated with `lab2im`, a library for generating images by sampling a GMM conditioned on label maps ([Billot et al. 2020](#)). For each image, the GMM parameters are sampled from prior distributions of predefined hyperparameters, creating differences for appearance, deformation, noise, and bias field, thus generating a dataset diverse enough that the trained segmentation model becomes contrast-agnostic ([Billot et al. 2020](#)).

Additionally, data augmentation is also performed including spatial deformation of the label maps, random cropping, random blurring, and intensity augmentation steps such as bias field corruption, and gamma augmentation ([Billot et al. 2020](#)).

An example of synthetic MRI and segmentation masks generated with `lab2im` can be found in Figure 2.2

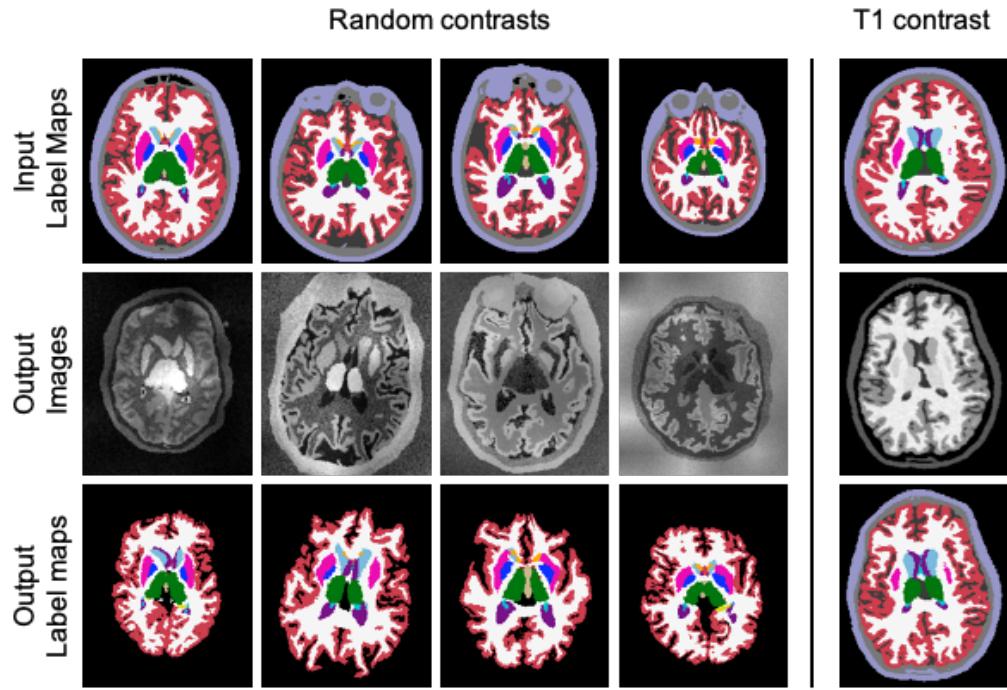


Figure 2.2: Synthetic brain MRI scans generated with `lab2im` ([Billot et al. 2020](#)).

The implemented model is a nnU-Nnet ('no new net'), a U-Net architecture introduced by [Isensee et al. \(2018\)](#) with a strong performance on segmentation tasks. nnU-Net is a self-configuring deep learning-based segmentation method that automatically configures itself, including preprocessing, network architecture, training and post-processing.

Thanks to its automated configuration, nnU-Net is fast and simple to implement, and requires very little computational resources beyond standard model training ([Isensee et al. 2018](#)). Additionally, nnU-Net is dataset-agnostic, and automatically adapts to a new dataset by systematically addressing the configuration of the segmentation pipelines ([Isensee et al. 2018](#)). Finally, it is worth noting that nnU-Net is considered 'data efficient', as its encoding design is based on a wide range of datasets that serves as inductive bias for application to datasets with scarce training data ([Isensee et al. 2018](#)). In the following sections, the implementation of the project using these methods is be detailed.

# Chapter 3

## Methods and Resources

### 1 Dataset

The dataset used for this project is The Open Access Series of Imaging Studies 3 (OASIS-3), by [LaMontagne et al. \(2019\)](#). OASIS-3 is a multimodal collection of data from 1378 participants collected across several projects through the Charles F. and Joanne Knight Alzheimer Disease Research Center during the course of 30 years.

Along with PET scans, this dataset includes a total of 2842 magnetic resonance sessions (including different contrast modalities), most of them accompanied by volumetric segmentation files produced with FreeSurfer, of 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42 to 95 years ([LaMontagne et al. 2019](#)).

All neuroimaging scans were conducted by the Knight Alzheimer Research Imaging Program at Washington University in St. Louis. Specifically, MRI was collected on 3 different Siemens Medical Solutions USA, Inc. scans: Vision 1.5T, TIM Trio 3T, and BioGraph mMR PET-MR 3T. High-resolution structural sequences include T1-weighted (T1w), T2-weighted (T2w), FLAIR, and Turbo Spin Echo (TSE), susceptibility-weighted imaging (SWI), diffusion-weighted imaging (DWI), arterial spin labeling (ASL), resting-state BOLD, field maps, and time of flight.

T1w images collected on 1.5T scanners were processed with FreeSurfer v5.0 or v5.1, while T1w images from 3T scanners were processed with FreeSurfer v5.3. After segmentation, a trained specialist reviewed the images to ensure accuracy and quality control, deciding whether they were valid or should be rerun through the XNAT pipeline.

For the current project, 150 subjects were randomly selected and split into two independent subsets: a training set and a test set. As the OASIS-3 dataset is designed for longitudinal analyses and allows multiple sessions per subject, only a single session per subject was selected to avoid similarities between scans in the selected sample. Considering this restriction, random selection was applied to ensure representativeness and avoid selection bias. Additionally, as

this project does not specifically focus on finding pathological markers or structural differences among groups, information about cohorts of any kind was not included in the analysis and has not affected in any way the sample selection.

## 1.1 Training Data

For the training data, a total of 50 different subjects were randomly selected from the OASIS-3 sample, including MRI scans in different contrast modalities ( $T1w=160$ ,  $T2w=92$ ,  $FLAIR=10$ ,  $TSE=90$ ), along with their corresponding segmentation masks.

Since the aim of this project was to train a model exclusively on synthetic data, the segmentation masks of the selected sub-sample were used as the ground truth to generate a set of 500 synthetic images for model training. For each subject, a total of 10 synthetic MRI and segmentation masks were generated using `lab2im` ([Billot et al. 2020](#)), which randomizes image contrast, noise, and deformations, and applies data augmentation on the synthetic images generated. Consequently, although multiple images are generated from the same ground truth mask, each synthetic image is unique. The code used for generating these images is publicly available on [this project's Github Repository](#). See an example of some slices from a synthetic MRI, with its synthetic segmentation mask created with this library, in Figure 3.1.

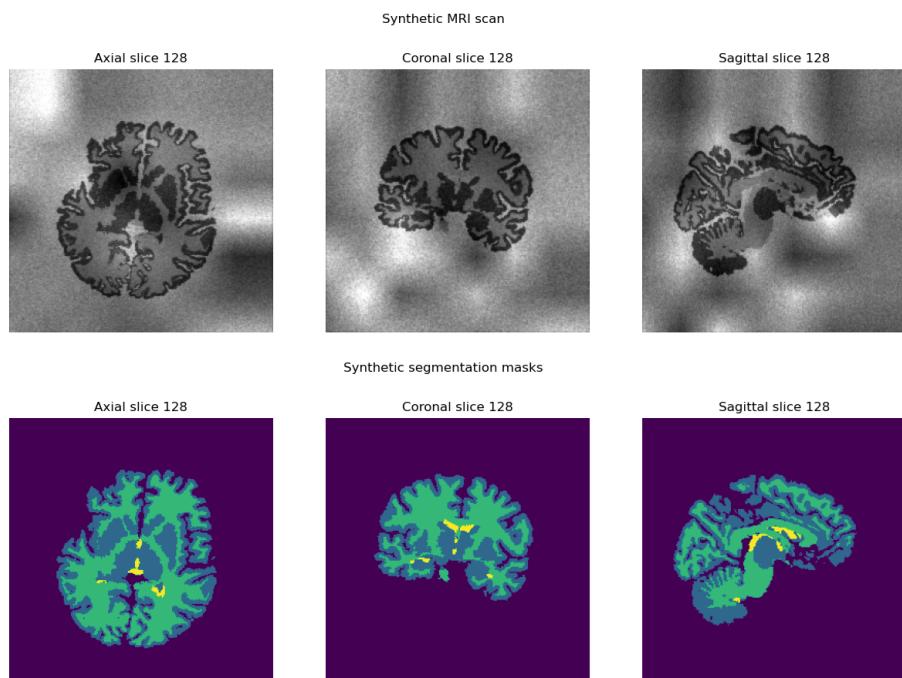


Figure 3.1: Slices from an instance extracted from training set: a synthetic brain MRI and its corresponding segmentation mask, generated with `lab2im` ([Billot et al. 2020](#)).

## 1.2 Test Data

For the test set, 100 subjects that were not used in the synthetic image generation phase were randomly selected from the dataset, corresponding to approximately 20% of the total training data. From these subjects, also containing MRI scans in different contrast modalities, a balanced selection was performed to test model generalization on T1w, and T2w volumes, considering the availability of each modality (T1w=53, T2w=47). TSE and FLAIR were not included in the training set because of the differences in volume size and voxel sizes on the time axis.

# 2 Data Pre-processing

In this section, the different tasks performed to pre-process data before modeling is detailed.

## 2.1 Training Data

Before the generation of the synthetic data, it was necessary to pre-process the selection of 50 subjects to be used as ground truth. Both synthetic MRI volumes and segmentation masks were pre-processed for different purposes that are detailed in the following sections.

### 2.1.1 Synthetic MRI generation

To generate synthetic MRI volumes and synthetic segmentation masks, it was necessary to pre-process the segmentation files provided by the original dataset.

The segmentation of the MRI scans by the OASIS-3 project was performed with FreeSurfer (v.5.0, v.5.1, or v.5.3, depending on the scan), hence producing an automated segmentation for different anatomical structures. Library `lab2im` creates an output with the same anatomical labels as its input. Then, considering that this project only focuses on identifying GM, WM, and CSF, segmentation masks produced by FreeSurfer were pre-processed to segment only these areas, with labels re-coded appropriately.

To do so, a `dictionary` was created based on the code-ROI correspondence extracted from `FreeSurferColorLUT.txt`, grouping different cortical and subcortical areas under the GM or WM label and liquid spaces as CSF. The labels were then recoded as follows: 0 = background, 1 = GM, 2 = WM, 3 = CSF.

After pre-processing the segmentation masks, they can be used as ground truth to generate a set of 500 synthetic MRI scans and segmentation masks using the library `lab2im`. The code can be also found in this project's [GitHub repository](#).

### 2.1.2 Synthetic Data Evaluation

To evaluate the quality of the synthetic data, the traditional MRI preparation steps detailed in [Kondrateva et al. \(2024\)](#) were followed, following some considerations by [Gourdeau et al. \(2022\)](#), and adding some additional steps to accommodate the needs of the dataset.

- Check voxel spacing. For a reliable evaluation, images with voxel spacing different from  $1 \times 1 \times 1 \text{ mm}^3$  were excluded from the quality assessment. This has led to dropping both TSE and FLAIR volumes, which presented different voxel spacing, for this analysis.
- Reorient segmentation mask. FreeSurfer uses LIA (Lateral Inferior Anterior) orientation for all of its volumetric data files, so the synthetic MRI and masks maintain it. Those were reoriented to RAS (Right Anterior Superior) to match the real MRI images and the ground truth segmentation masks.
- Create a brain mask. Synthetic MRI volumes have a noisy background, so a brain mask is generated from the segmentation masks, merging the labels for GM, WM, and CSF into a whole brain mask so the brain tissue can be extracted from the background in the synthetic MRI.
- Reorient MRI. As with the segmentation mask, the MRI volume should be reoriented from LIA to RAS to match the real MRI.
- Remove background. From the brain mask created from the segmentation mask, extract the brain tissue and remove the background, recoding it with label 0.
- Clip and normalize intensities inside the brain. Extreme values are clipped at the 1st and 99th percentiles, and intensities are normalized to the [0,1] range.

The code used to perform all these tasks to generate the synthetic training set is available in a [dedicated section](#) on this project's GitHub Repository.

## 2.2 Test Data

Because of the features of the nnU-Net, which performs pre-processing of the volumes on its own before the training phase, most of the pre-processing steps defined in [Kondrateva et al. \(2024\)](#) were not needed. However, there are a few tasks that should be performed before using them for testing.

- Skull stripping. Real MRIs include bone tissue, which is not useful to the current project and can induce noise in the model ([Kondrateva et al. 2024](#)). To avoid this, the brain tissue is extracted from the skull using library HD-BET ([Isensee et al. 2019](#)).

- Resize image to have same axes. The real images have varying dimensions along the three axes, which differ from those of the synthetic images. Considering that their volume does not condition the size of the brain masks, the chosen method to resize the real MRI is zero-padding, and it is meant to reconstruct background spaces to fit size  $256 \times 256 \times 256$ , avoiding any brain mask modification.
- Extract brain mask. This step is needed because even if real MRIs do not have a similar background to that of synthetic MRIs created with `lab2im`, the brain tissue should be isolated from the background to normalize the intensities accurately.
- Clip and normalize intensities inside the brain. Extreme values are clipped at the 1st and 99th percentiles, and intensities are normalized to the [0,1] range.

Both the code for [skull-stripping](#) and [MRI pre-processing](#) can be found on the GitHub repository. See an example with slices of pre-processed real MRI, synthetic MRI and segmentation labels on Figure 3.2.

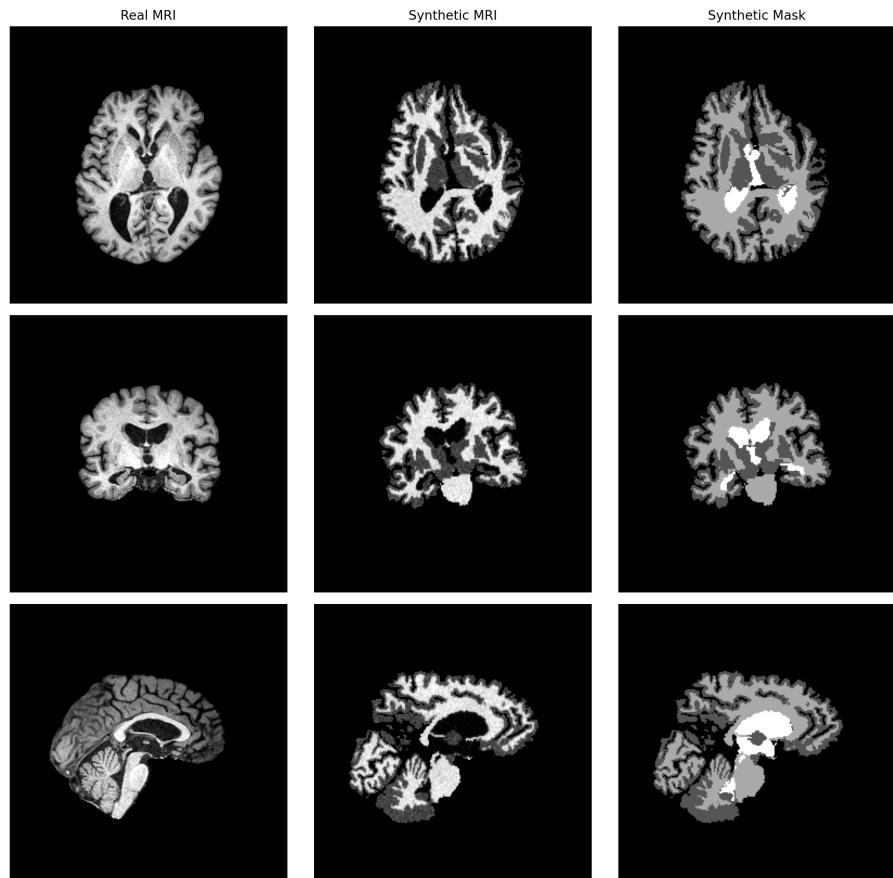


Figure 3.2: Results of image pre-processing on real MRI, synthetic MRI and synthetic segmentation masks.

### 3 Image Quality

To evaluate the fidelity of the synthetic MRI images generated for this project, several quantitative analyses were performed comparing the synthetic scans to their corresponding real images. In the first place, the distribution of voxel intensities within the brain was visualized using histograms for both synthetic and real images, providing an overall view of intensity ranges and potential differences between real and synthetic images.

The Structural Similarity Index Measure is a popular image quality assessment metric that combines the evaluation of the luminance, contrast, and structure. This metric intends to reproduce certain features of the human visual system following Weber’s law, and according to some authors, it correlates better than MAE with human assessment of image quality ([Gourdeau et al. 2022](#)). Thus, this metric is useful to evaluate the quality of the synthetic MRI in terms of brain anatomy, but it also allows to evaluate how close synthetic images are to real MRI sequences in terms of contrast and luminance, and it also can provide an approximation of how realistic they can look to the human eye.

The Peak Signal-to-Noise Ratio was originally developed to measure the reconstruction quality of a lossy compressed image compared to the uncompressed reference image ([Dohmen et al. 2025](#)), but it is frequently used to evaluate image similarity. PSNR performs a voxel-wise comparison of intensity differences, but unlike SSIM, it is insensitive to structural information. Nevertheless, it can serve as an approximation of how closely the intensity distribution of a synthetic image matches that of a real one. This makes this metric useful not only to assess how similar a synthetic image is to a specific MRI sequence but also for quantifying how different the synthetic brain is from its ground truth, reflecting variability in the generated data.

Considering that from each real image belonging to a single subject, 10 synthetic images were generated, evaluation metrics have been performed pairwise, comparing each synthetic MRI to a ground truth. For SSIM, the metric was calculated slice by slice and averaged across slices to obtain a single 3D SSIM value. Because subjects may have scans in multiple sequences, every scan was included as potential ground truth. For each synthetic image, both SSIM and PSNR were computed against all real volumes of that subject. The highest SSIM was selected as the best match, identifying the real image most similar to the synthetic one, including its MRI sequence. This approach allows quantifying not only similarity in intensity and structure but also the similarity of the synthetic image, which is not in any specific MRI variant, with real sequences.

## 4 Modeling and Training

The nnU-Net is a self-configuring, deep learning-based segmentation method that automatically configures itself, including preprocessing, network architecture, training, and post-processing. For this reason, no manual pre-processing was needed for the training data. A specific function of the nnU-Net performs the pre-processing considering the characteristics of the sample. In the case of the present training data, voxel spacing was modified to  $1.26677 \times 1.26677 \times 1.26677$  mm<sup>3</sup> with voxel intensities standardized with Z-score normalization.

After pre-processing, nnU-Net produces a file with recommendations on the best architecture and hyperparameters for the model. Considering the characteristics of the sample and the computational resources available, the recommended pipeline is training a model `3d_lowres` and, after that, a `3d_cascade_fullres`. This pipeline consists of, firstly, training a 3D U-Net that operates on low-resolution images on a 5-fold cross validation scheme and then, secondly, a high-resolution 3D U-Net refines the predictions of the former. However, because of time and computational constraints, only `3d_lowres` training has been performed on 5 folds with a split of 400 training and 100 validation cases for each one.

Each fold of the `3d_lowres` architecture uses a plain 3D U-Net following the standard encoder-decoder structure, consisting of six encoder-decoder stages. The number of feature channels per stage is [32, 64, 128, 256, 320, 320], with early stages using fewer channels and deeper stages using more to capture complex spatial patterns. All convolutional layers use  $3 \times 3 \times 3$  kernels. Strides are set to [1, 1, 1] in the first stage and [2, 2, 2] in the following stages. Each stage contains two consecutive convolutional layers, mirrored in the decoder. All convolutions are followed by a LeakyReLU activation, and no dropout is used.

Folds are trained for 1,000 epochs each, which is the default value set by nnU-Net. The initial learning rate is 0.01 for learning network weights, and its value decays throughout the training following the formula  $(1 - \frac{\text{epoch}}{\text{epoch}_{\max}})^{0.9}$ . The loss function is computed with the sum of cross-entropy and pseudo-dice loss.

The training metrics reported by nnU-Net are shown in Figure 3.3, and show how for each fold, validation losses decreased rapidly in the initial 100 epochs until reaching a plateau, indicating stable convergence. The pseudo-Dice metric increases correspondingly, reaching a high value early in training stages for all folds, and remaining stable throughout the epoch with pseudo-Dice scoring over 0.9.

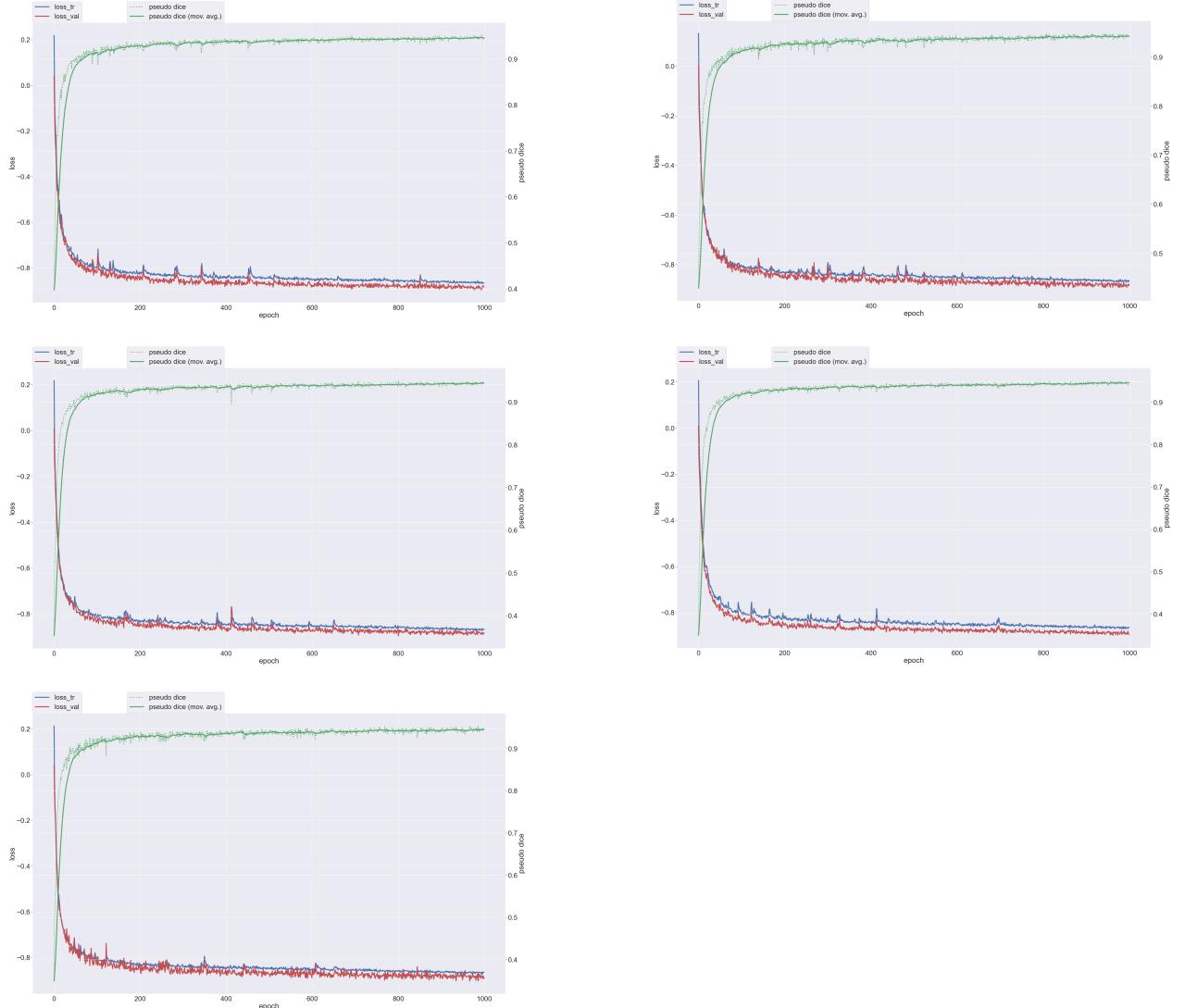


Figure 3.3: Image depicting the training metrics through this process, including loss and pseudo dice metrics per fold, throughout 1000 epoch.

# Chapter 4

## Results

In this chapter, the results of the image generation, as well as those of the model training are reported.

### 1 Quality of Synthetic MRI Scans

To assess the quality of the generated synthetic MRI scans, in the first place a qualitative analysis of image intensities is performed (see Table 4.1).

Table 4.1: Overall voxel intensity statistics for real and synthetic MRI datasets.

MRI Type	Mean	Std	Min	Max	Median
Real	0.473	0.249	0.002	1.0000	0.448
Synthetic	0.465	0.282	0.004	1.000	0.456

Overall, results show that the intensity distribution is similar between real and synthetic MRI. However, when plotting a histogram to evaluate the distribution of those intensities, notable differences can be observed (see Figure 4.1).

Real MRI concentrate most of the intensities around values 0.2 to 0.4, stabilizing for the subsequent intensities and showing few low or high-intensity voxels at the extremes, around the range 0.0–0.1 and 0.9–1.0. In contrast, synthetic images show a pronounced peak at low intensities between 0.0 and 0.2, and another increase around 0.6–0.9. Therefore, synthetic images exhibit a much more varied intensity distribution than real images, since real MRIs present specific intensities according to the image contrast of the scan. These differences reflect the particular characteristics of real MRIs and their sequences (T1w, T2w or FLAIR), but also that of synthetic MRIs generated with `lab2im`, which do not replicate intensities of any specific real sequence. Additionally, synthetic images show deformations and random noise, which contributes to intensity variability when compared to real images.

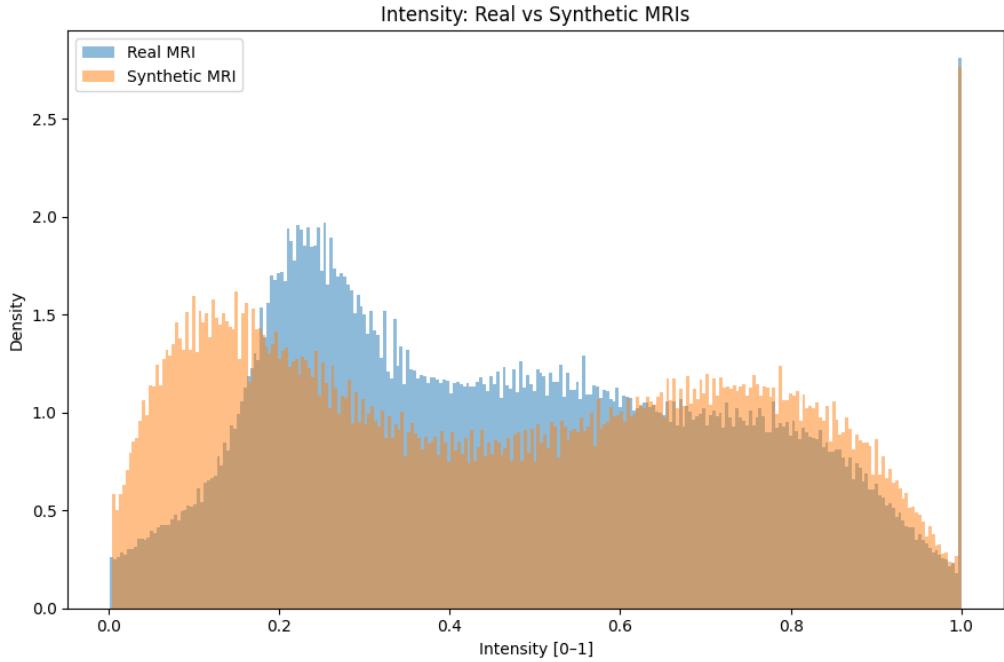


Figure 4.1: Intensity densities for both real and synthetic MRIs.

To perform a quantitative analysis of the quality of the synthetic images, SSIM and PSNR were computed across all scans. As described in the previous section, each synthetic image was matched to a real scan from the same subject, using the highest SSIM value to determine the best correspondence. Since synthetic MRIs do not represent any specific sequence, this highest match serves as an approximation for identifying the most similar real contrast. The results are summarized in Table 4.2.

Table 4.2: SSIM and PSNR statistics across contrasts T1w and T2w.

Contrast	SSIM				PSNR (dB)			
	Min	Max	Mean	Std	Min	Max	Mean	Std
T1w	0.7799	0.8771	0.8251	0.0174	15.66	21.35	18.24	0.94
T2w	0.7643	0.8514	0.8108	0.0175	16.05	22.86	18.89	1.49

As shown in Figure 4.2, which presents the same results, a slightly higher number of synthetic images correspond to T1w real MRIs compared to T2w volumes. Additionally, T1w matches reach higher SSIM values. This observation suggests that the synthetic MRIs generated using `lab2im` may be somewhat biased toward resembling T1w images than T2w.

However, when examining the PSNR distribution, the results differ slightly. Many synthetic images tend to achieve higher PSNR values for T2w scans, indicating greater similarity to real images. At the same time, the T2w images exhibit a wider PSNR distribution suggesting that, while some synthetic T2w images closely resemble real scans, there is also considerable

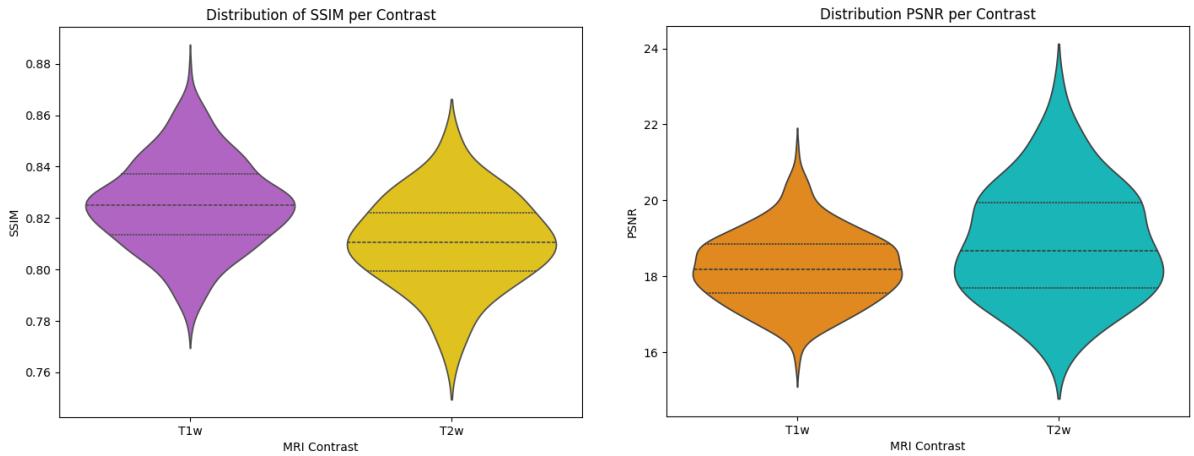


Figure 4.2: Distribution of SSIM and PSNR per contrast.

variability in their quality, possibly due to structural differences that PSNR cannot capture. In contrast, T1w images display a narrower distribution, indicating that `lab2im`-generated synthetic MRIs produce more consistent T1w outputs.

In conclusion, synthetic MRIs matching T1w scans are structurally consistent and visually similar to T1w real MRIs, as reflected by their slightly higher SSIM values, but also exhibit subtle intensity differences, leading to lower PSNR values. On the other hand, synthetic images that match T2w scans achieve higher PSNR values, indicating closer pixel-wise similarity to real T2w scans, but also show more structural variability, explained by slightly lower SSIM values when compared to synthetic T1w matches.

## 2 Model Evaluation

After training the nnU-Net model on the dataset in 5 folds, its performance was evaluated using multiple quantitative metrics.

Overall, all folds achieved very similar performance, with mean Dice scores of approximately 0.93 (see Table 4.3), indicating good stability and consistency of the training process. To further assess the model's performance, inference and evaluation were performed using an ensemble of the five folds, yielding slightly improved segmentation results than the single folds alone.

Table 4.3: Mean Dice score obtained for each fold.

Fold	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Mean Dice	0.939	0.933	0.932	0.936	0.934

To assess the model performance and robustness, metrics reported by nnU-Net were used (see Table 4.4).

First, the model results provide with Dice Similarity Coefficient (DSC), which is a statistic used to evaluate the similarity of two samples (in the specific case of segmentation, between the FreeSurfer and the nnU-Net segmentation). It is equivalent to the F1-score. Second, the Intersection over Union (IoU) metric is used to evaluate how well a model's prediction overlaps with the ground truth. It is a more strict metric than DSC, so lower values are expected. Additionally, the recall was also computed from the confusion matrix components provided by the nnU-Net output, with the aim to assess the proportion of actual correct segmentation that was properly identified by the model.

Table 4.4: nnU-Net segmentation metrics per class and foreground mean.

Class	Dice	IoU	Recall
Foreground Mean	0.762	0.664	0.799
Grey Matter (GM = 1)	0.663	0.519	0.719
White Matter (WM = 2)	0.812	0.694	0.889
Cerebrospinal Fluid (CSF = 3)	0.812	0.693	0.846

For the foreground mean, model shows moderately good results (DSC=0.762, recall=0.799), but as the model training pipeline was not fully completed, there is still room to improve. The Intersection Over Union metric (IoU=0.664) reflects that.

Regarding the segmentation of specific ROIs, the metrics show good results to identify WM (DSC=0.812, recall=0.889, IoU=0.694) and CSF (DSC=0.812, recall=0.846, IoU=0.693), being GM the tissue with moderate-to-low results (DSC=0.663, recall=0.719, IoU=0.519).

The confusion matrix for each ROI (see Figure 4.3) shows that, for GM, there is a high percentage of false negatives, that is, 28% of the voxels that are actual GM are not identified as such. Although the overall results for WM and CSF are better, they show a similar tendency: areas that are actual WM and CSF are not labeled correctly.

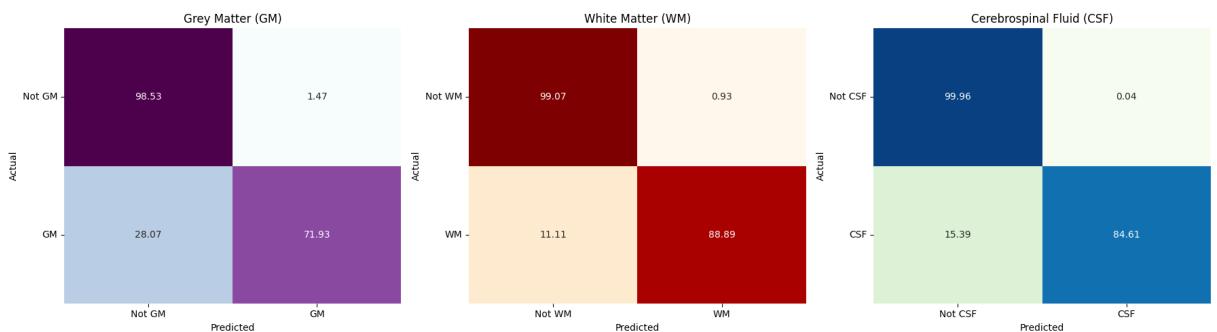


Figure 4.3: Confusion matrices for each ROI (GM, WM and CSF).

Finally, a qualitative example of the segmentation results is shown in Figure 4.4.

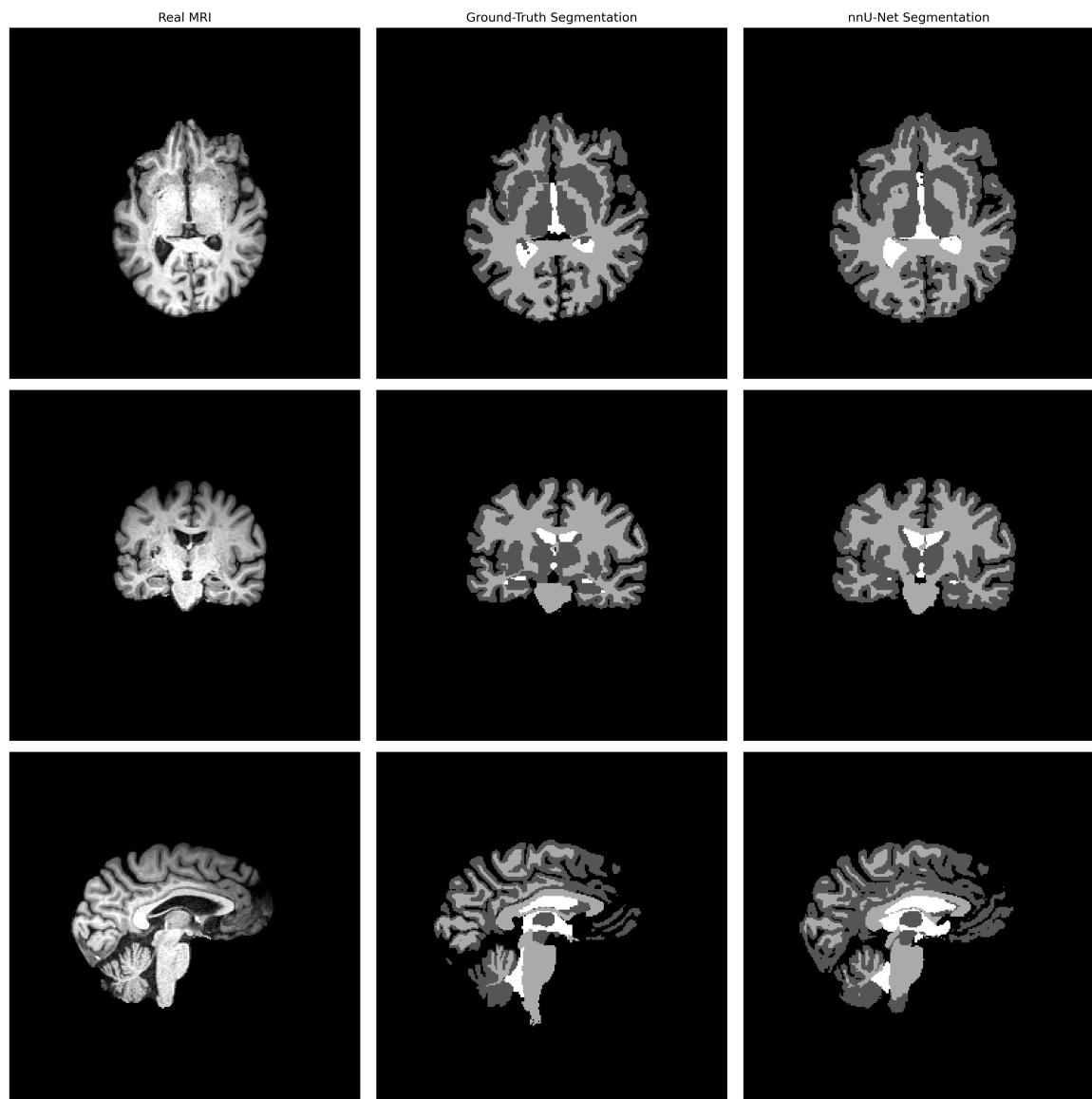


Figure 4.4: Example of segmentation performed by nnU-Net on a T2-weighted MRI. From left to right, the real MRI, the ground-truth segmentation and the model segmentation.



# Chapter 5

## Conclusions and Future Work

This work has explored the potential use of synthetic data for model training to address the problem of data scarcity, costly imaging and patient privacy restrictions. To tackle these issues, a methodology similar to that of SynthSeg ([Billot, Greve, Puonti, Thielscher, Van Leemput, Fischl, Dalca, Iglesias et al. 2023](#)) has been adopted to implement a segmentation model trained on synthetic data with domain randomization, making the model agnostic to MRI sequences.

Specifically, this work implements this method with the use of library `lab2im` by [Billot et al. \(2020\)](#), which generates synthetic MRI scans and segmentation masks from a ground-truth label maps, applying a variety of modifications to successfully implement domain randomization. These images were then used to train a nnU-Net ([Isensee et al. 2021](#)), a self-configuring segmentation framework that does not require prior image pre-processing or hyperparameter optimization. This makes it an easy-to-use tool accessible to medical specialists without expertise in deep learning architecture design.

The main limitations encountered during the development of this project were primarily related to time constraints and hardware infrastructure. First, as this work was conducted as part of a master's thesis developed over less than four months, the available time was limited. Many processes involved in data manipulation, such as pre-processing, applying skull-stripping algorithms like HD-BET ([Isensee et al. 2019](#)), and training the nnU-Net, are time-intensive. Consequently, only the low-resolution model version in 5 folds could be trained, preventing completion of the full nnU-Net pipeline. Second, restrictions in hardware resources further constrained the project. While platforms such as Kaggle or Google Colab provide GPUs suitable for model training, library incompatibilities and Python environment requirements made it impossible to rely on these resources. Instead, training was performed on a personal computer with limited computational capacity, which increased training times. These limitations also led to using the low-resolution variant of the nnU-Net configuration, which was not the most optimal option for this task.

Despite the evident limitations, when performing an ensemble of the 5 folds, the model achieved moderately robust results. During training, it reached high pseudo-Dice scores (above 0.9) and demonstrated the ability to generalize to real data, producing good segmentation results for white matter and cerebrospinal fluid. For gray matter, the model encountered more difficulties accurately identifying the tissue, though a similar trend was observed in the other regions of interest. Overall, these results are promising and could likely be improved by completing the full training pipeline proposed by the nnU-Net framework.

Additionally, a further refinement of the pipeline could have been performed to incorporate imaging modalities on the model test phase, but due to the lack of time, it was not possible. As TSE and FLAIR could not included in this project due to differences in voxel spacing, a direct comparison between the model segmentations and the ground-truth annotations could not be performed. Consequently, while the model appears to generalize well to Tw and T2w images, it remains unclear whether these results would extend to other contrasts such as FLAIR.

For future work, it would be of interest to train the model on larger datasets and to expand the anatomical variability of the synthetic training data in order to improve generalization to pathological cases. Incorporating clinical cohorts and pathological hallmarks such as tumors or reclassifying labels such as white-matter hyperintensities or non-white-matter hypointensities (which were classified as white matter and grey matter, respectively, in this project) would increase the clinical relevance of this work.

While based on the SynthSeg framework ([Billot, Greve, Puonti, Thielscher, Van Leemput, Fischl, Dalca, Iglesias et al. 2023](#)), this project proposes an alternative pipeline designed for common working environments, requiring modest computational resources common energy consumption that does not go beyond typical academic activities. As it has been successfully applied in a personal computing environment, this pipeline can also be accessible to low-resource clinical settings and research groups that do not dispose of great, up-to-date hardware infrastructures, democratizing the access to deep learning tools for medical image analysis.

# Chapter 6

## Glossary

- **AI:** Artificial Intelligence
- **CNN:** Convolutional Neural Network
- **CSF:** Cerebrospinal Fluid
- **CT:** Computed Tomography
- **DSC:** Dice Similarity Coefficient
- **GM:** Gray Matter
- **IoU:** Intersection Over Union
- **LIA:** Lateral Inferior Anterior
- **ML:** Machine Learning
- **MRI:** Magnetic Resonance Imaging
- **PET:** Positron Emission Tomography
- **PSNR:** Peak Signal-to-Noise Ratio
- **RAS:** Right Anterior Superior
- **ROI:** Region of Interest
- **SDG:** Sustainable Development Goals
- **SPM:** Statistical Parametric Mapping
- **SSIM:** Structural Similarity Index Measure

- **SVM:** Support Vector Machine
- **T1w:** T1-weighted
- **T2w:** T2-weighted
- **VBM:** Voxel-based Morphometry
- **WM:** White Matter

# Bibliography

Ashburner, J. & Friston, K. J. (2000), ‘Voxel-based morphometry—the methods’, *Neuroimage* **11**(6), 805–821.

Billot, B., Greve, D., Leemput, K. V., Fischl, B., Iglesias, J. E. & Dalca, A. V. (2020), ‘A Learning Strategy for Contrast-agnostic MRI Segmentation’.

**URL:** <https://arxiv.org/abs/2003.01995>

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., Iglesias, J. E. et al. (2023), ‘SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining’, *Medical image analysis* **86**, 102789.

Billot, B., Magdamo, C., Cheng, Y., Arnold, S. E., Das, S. & Iglesias, J. E. (2023), ‘Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets’, *Proceedings of the National Academy of Sciences* **120**(9), e2216399120.

Chau, M., Vu, H., Debnath, T. & Rahman, M. (2025), ‘A scoping review of automatic and semi-automatic mri segmentation in human brain imaging’, *Radiography* **31**(2), 102878.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016), 3d u-net: learning dense volumetric segmentation from sparse annotation, in ‘International conference on medical image computing and computer-assisted intervention’, Springer, pp. 424–432.

Conduah, A. K., Ofoe, S. & Siaw-Marfo, D. (2025), ‘Data privacy in healthcare: Global challenges and solutions’, *Digital Health* **11**, 20552076251343959.

Denton, E., Chintala, S., Szlam, A. & Fergus, R. (2015), ‘Deep generative image models using a laplacian pyramid of adversarial networks’.

**URL:** <https://arxiv.org/abs/1506.05751>

Diez, A. (2021), ‘Toolkit de gènere’. Accessed: 2025-10-25.

**URL:** <https://genere-toolkit.recursos.uoc.edu/uoc-us-no-sexista-de-la-llengua/>

- Dohmen, M., Klemens, M. A., Baltruschat, I. M., Truong, T. & Lenga, M. (2025), ‘Similarity and quality metrics for mr image-to-image translation’, *Scientific Reports* **15**(1), 3853.
- Fischl, B. (2012), ‘Freesurfer’, *Neuroimage* **62**(2), 774–781.
- Friston, K. J. (2003), Statistical parametric mapping, in ‘Neuroscience databases: a practical guide’, Springer, pp. 237–250.
- Gopinath, K., Hoopes, A., Alexander, D. C., Arnold, S. E., Balbastre, Y., Billot, B., Casamitjana, A., Cheng, Y., Chua, R. Y. Z., Edlow, B. L. et al. (2024), ‘Synthetic data in generalizable, learning-based neuroimaging’, *Imaging Neuroscience* **2**, 1–22.
- Gourdeau, D., Duchesne, S. & Archambault, L. (2022), ‘On the proper use of structural similarity for the robust evaluation of medical image synthesis models’, *Medical Physics* **49**(4), 2462–2474.
- Gouvea, A. (2021), ‘The crisp-dm methodology’. Accessed: 2025-10-5.  
**URL:** <https://almirgouvea.github.io/The-Crisp-DM-Methodology/chapters/intro.html>
- He, K., Zhang, X., Ren, S. & Sun, J. (2015), ‘Deep residual learning for image recognition’.  
**URL:** <https://arxiv.org/abs/1512.03385>
- Hoffmann, M. (2025), ‘Domain-randomized deep learning for neuroimage analysis’, *arXiv preprint arXiv:2507.13458*.
- Hoffmann, M., Billot, B., Greve, D. N., Iglesias, J. E., Fischl, B. & Dalca, A. V. (2021), ‘Synthmorph: learning contrast-invariant registration without acquired images’, *IEEE transactions on medical imaging* **41**(3), 543–558.
- Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. (2022), ‘Synthstrip: skull-stripping for any brain image’, *NeuroImage* **260**, 119474.
- Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. (2018), ‘Densely connected convolutional networks’.  
**URL:** <https://arxiv.org/abs/1608.06993>
- Iglesias, J. E. (2023), ‘A ready-to-use machine learning tool for symmetric multi-modality registration of brain mri’, *Scientific Reports* **13**(1), 6657.
- Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P. & Fischl, B. (2023), ‘Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry’, *Science advances* **9**(5), eadd3607.

- Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. (2014), ‘Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(7), 1325–1339.
- Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. (2021), ‘nnu-net: a self-configuring method for deep learning-based biomedical image segmentation’, *Nature methods* **18**(2), 203–211.
- Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. et al. (2018), ‘nnU-net: Self-adapting framework for U-net-based medical image segmentation’, *arXiv preprint arXiv:1809.10486*.
- Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W. et al. (2019), ‘Automated brain extraction of multisequence mri using artificial neural networks’, *Human brain mapping* **40**(17), 4952–4964.
- Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2018), ‘Image-to-image translation with conditional adversarial networks’.
- URL:** <https://arxiv.org/abs/1611.07004>
- Jalloul, M., Miranda-Schaeubinger, M., Noor, A. M., Stein, J. M., Amiruddin, R., Derbew, H. M., Mango, V. L., Akinola, A., Hart, K., Weygand, J. et al. (2023), ‘Mri scarcity in low-and middle-income countries’, *NMR in Biomedicine* **36**(12), e5022.
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. & Wang, Y. (2017), ‘Artificial intelligence in healthcare: past, present and future’, *Stroke and vascular neurology* **2**(4).
- Jiang, S., Chen, X. & Yi, C. (2024), ‘Ssa-unet: Whole brain segmentation by u-net with squeeze-and-excitation block and self-attention block from the 2.5 d slice image’, *IET Image Processing* **18**(6), 1598–1612.
- Kingma, D. P. & Welling, M. (2022), ‘Auto-encoding variational bayes’.
- URL:** <https://arxiv.org/abs/1312.6114>
- Koetzier, L. R., Wu, J., Mastrodicosa, D., Lutz, A., Chung, M., Koszek, W. A., Pratap, J., Chaudhari, A. S., Rajpurkar, P., Lungren, M. P. et al. (2024), ‘Generating synthetic data for medical imaging’, *Radiology* **312**(3), e232471.
- Kondrateva, E., Druzhinina, P., Dalechina, A., Zolotova, S., Golanov, A., Shirokikh, B., Belyaev, M. & Kurumkov, A. (2024), ‘Negligible effect of brain mri data preprocessing for tumor segmentation’, *Biomedical Signal Processing and Control* **96**, 106599.

- Kundu, S., Dutta, S., Mukhopadhyay, J. & Chakravorty, N. (2025), ‘Fflunet: Feature fused lightweight unet for brain tumor segmentation’, *Computers in Biology and Medicine* **194**, 110460.
- LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G. et al. (2019), ‘Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease’, *medrxiv* pp. 2019–12.
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H. & Winther, O. (2016), Autoencoding beyond pixels using a learned similarity metric, in ‘International conference on machine learning’, PMLR, pp. 1558–1566.
- Li, X., Lan, Z., Sun, Y., Sun, Y., Guo, Y., Wang, Y. & Yuan, A. (2025), ‘Dpf-unet: a cnn-swin transformer fusion network for 3d brain tumor segmentation in mri images’, *The Journal of Supercomputing* **81**(7), 832.
- Makanaboyina, S. (2025), ‘Nndm: Nn\_unet diffusion model for brain tumor segmentation’, *arXiv preprint arXiv:2510.09681* .
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B. et al. (2018), ‘Attention u-net: Learning where to look for the pancreas’, *arXiv preprint arXiv:1804.03999* .
- Paproki, A., Salvado, O. & Fookes, C. (2024), ‘Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias’, *ACM Computing Surveys* **56**(11), 1–37.
- Pérez-Millan, A., Contador, J., Juncà-Parella, J., Bosch, B., Borrell, L., Tort-Merino, A., Falgàs, N., Borrego-Écija, S., Bargallo, N., Rami, L. et al. (2023), ‘Classifying alzheimer’s disease and frontotemporal dementia using machine learning with cross-sectional and longitudinal magnetic resonance imaging data’, *Human brain mapping* **44**(6), 2234–2244.
- Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J. & Tyrrell, P. N. (2020), ‘Assessing and mitigating the effects of class imbalance in machine learning with application to x-ray imaging’, *International journal of computer assisted radiology and surgery* **15**(12), 2041–2048.
- Radford, A., Metz, L. & Chintala, S. (2016), ‘Unsupervised representation learning with deep convolutional generative adversarial networks’.
- URL:** <https://arxiv.org/abs/1511.06434>

- Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* ‘International Conference on Medical image computing and computer-assisted intervention’, Springer, pp. 234–241.
- Shearer, C. (2000), ‘The CRISP-DM model: the new blueprint for data mining’, *Journal of Data Warehousing* **5**(4), 13–22.
- Shen, D., Wu, G. & Suk, H.-I. (2017), ‘Deep learning in medical image analysis’, *Annual review of biomedical engineering* **19**(1), 221–248.
- Soppari, K., Putnala, S., Borala, S. S. & Gondhi, S. K. (2024), ‘A survey on brain mri segmentation’, *World Journal of Advanced Research and Reviews* **21**(03), 1702–1710.
- Sun, C., Shrivastava, A., Singh, S. & Gupta, A. (2017), Revisiting unreasonable effectiveness of data in deep learning era, *in* ‘Proceedings of the IEEE international conference on computer vision’, pp. 843–852.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015), Going deeper with convolutions, *in* ‘Proceedings of the IEEE conference on computer vision and pattern recognition’, pp. 1–9.
- Thakur, G. K., Thakur, A., Kulkarni, S., Khan, N. & Khan, S. (2024), ‘Deep learning approaches for medical image analysis and diagnosis’, *Cureus* **16**(5).
- Thambawita, V., Salehi, P., Sheshkal, S. A., Hicks, S. A., Hammer, H. L., Parasa, S., Lange, T. d., Halvorsen, P. & Riegler, M. A. (2022), ‘Singan-seg: Synthetic training data generation for medical image segmentation’, *PloS one* **17**(5), e0267976.
- Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. & Abbeel, P. (2017), Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, *in* ‘2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)’, IEEE, pp. 23–30.
- Valabregue, R., Girka, F., Pron, A., Rousseau, F. & Auzias, G. (2024), ‘Comprehensive analysis of synthetic learning applied to neonatal brain MRI segmentation’, *Human Brain Mapping* **45**(6), e26674.
- Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I. & Schmid, C. (2017), Learning from synthetic humans, *in* ‘CVPR’.
- Yeğin, M. N. & Amasyalı, M. F. (2024), ‘Generative diffusion models: A survey of current theoretical developments’, *Neurocomputing* **608**, 128373.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. (2018), ‘Unet++: A nested u-net architecture for medical image segmentation’.

**URL:** <https://arxiv.org/abs/1807.10165>

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017), Unpaired image-to-image translation using cycle-consistent adversarial networks, *in* ‘Proceedings of the IEEE international conference on computer vision’, pp. 2223–2232.

# **Chapter 7**

## **Appendices**

GitHub repository for this project: <https://github.com/avidiella/synthetic-mri-segmentation>